A Simulation Study of Computer-Adaptive Testing for Measuring Treatment-Related Change in Confrontation Naming

by

Pauline Bayotas

Communication Sciences and Disorder, School of Health and Rehabilitation Sciences, University of Pittsburgh, 2022

Submitted to the Graduate Faculty of the

University Honors College in partial fulfillment

of the requirements for the degree of

Bachelor of Philosophy

University of Pittsburgh

2022

Committee Membership Page
UNIVERSITY OF PITTSBURGH

UNIVERSITY HONORS COLLEGE

This thesis was presented

by

Pauline Bayotas

It was defended on

March 30, 2022

and approved by

Gerasimos Fergadiotis, Ph.D., CCC-SLP

Michael Dickey, Ph.D., CCC-SLP

Sarah Wallace, Ph.D., CCC-SLP

William Hula, Ph.D., CCC-SLP

Abstract

A Simulation Study of Computer-Adaptive Testing for Measuring Treatment-Related Change in Confrontation Naming

Pauline Bayotas, BA

University of Pittsburgh, 2022

Abstract

Computer adaptive testing (CAT) is an approach that can be used to shorten assessments without sacrificing their psychometric properties. Recent studies (Fergadiotis, Kellough, & Hula, 2015; Hula, Kellough, & Fergadiotis, 2015; Fergadiotis, Hula, Swiderski, Lei, and Kellough, 2019; Hula, Fergadiotis, Swiderski, Silkes, & Kellough, 2020) produced a CAT with an item bank consisting of the Philadelphia Naming Test (PNT; Roach et al., 1996). The main advantage of CAT is to maximize the precision of the test, requiring fewer testing items while having the same or better level of accuracy as traditional brief naming assessments. However, before a CAT can be used to measure change in anomia severity, it is important to understand how the algorithm interacts with commonly used aphasia interventions and whether it is as responsive to treatment-related change as standard static assessments. This simulation study investigated the sensitivity of a computer adaptive version of the Philadelphia Naming Test (PNT-CAT) to treatment-related change in three different treatment conditions: item-general, item-specific and partially item-specific. For each condition, we simulated responses using a one-parameter logistic item response theory model and computed pre- to post-treatment change scores for the PNT-CAT and the full PNT. For the item-general condition, both tests performed similarly well. However, the PNT-CAT overestimated the effects of item-specific and the partially item-specific treatment relative to the full test. These results provide useful information about the conditions in which CAT can be validly

used to measure treatment outcomes. Present results suggest that when treatment affects underlying naming ability the PNT-CAT30 is appropriately responsive to treatment and provides an efficient alternative to the administration to the full PNT. On the other hand, when treatment effects are item-specific, the PNT-CAT30 may overestimate or underestimate treatment effects depending on the baseline ability level and the number of treated items included in post-treatment CAT.

Table of Contents

Preface ..... **ix**

1.0 Introduction **1**

2.0 Methods **11**

3.0 Results **13**

3.1 Research Question 1: Item-General Condition **13**

3.2 Research Question 2: Item-Specific Condition **13**

3.3 Research Question 3: Partially Item-Specific Condition **14**

4.0 Discussion **22**

Conclusion **29**

List of Tables

Table 1. Summary of t-tests comparing simulated full PNT and PNT-CAT30 change scores in item-general and item-specific treatment conditions. 15

List of Figures

Figure 1. Histogram of the number of treated items administered in the post-treatment PNT-CAT30 in condition 2, item-specific treatment. 16

Figure 2. Scatterplot of PNT-CAT30 change scores over the number of treated items administered in the post-treatment PNT-CAT30. 17

Figure 3 Histogram of the number of treated items administered in the post-treatment PNT-CAT30 in condition 3, partially item-specific treatment. 18

Figure 4 Scatterplot of PNT-CAT30 change scores over the number of treated items administered in the post-treatment PNT-CAT30. 19

Figure 5 . Scatterplot of the number of treated items in pre-treatment PNT-CAT30 over true naming ability level. 20

Figure 6 Scatterplot of the number of treated items in post-treatment PNT-CAT30 over true naming ability level. 21

Preface

My sincerest gratitude to my thesis advisor and lab mentor, Dr. Hula, for his support and guidance on this thesis and my growth overall. I could not have accomplished this without his advice, expertise, and encouragement. I would also like to thank the defense committee for their input and participating throughout the process. Finally, without the consistent emotional support and encouragement from my family, my friends and colleagues, the completion of this thesis would not have been possible. I could not be more appreciative of all they have done for me.

1.0 Introduction

Aphasia is a language impairment commonly caused by stroke in the left hemisphere of the brain. This disorder is associated with a variety of impairments that relate to communication, such as speaking, reading, writing, and understanding language. Anomia, the cardinal deficit of aphasia, is the inability to access and retrieve words (Goodglass & Wingfield, 1997). Anomia typically manifests as a failure to produce the intended name of a person, object, or action. These failures can take the form of semantically-related words, phonologically-related words or nonwords, unrelated nonwords, descriptions, or a complete lack of response, among other forms. As a result, individuals with anomia are negatively affected by their reduced ability to convey what they want accurately and efficiently. In turn, this can be debilitating for people with aphasia in activities of daily living (Goodglass, 1993).

One of the most commonly used methods to assess and diagnose the severity of word-finding impairments is confrontation picture naming. Picture naming tests are used to assess anomia in people with aphasia (PWA) and correlate highly with overall aphasia severity (Walker & Schwartz, 2012). Picture naming tests are an effective choice for assessing anomia because they can be used to quantify a person's overall ability to access and retrieve words (Fergadiotis, Hula, Swiderski, Lei, & Kellough, 2019) as well as provide consistent assessment results (Goodglass, 1993).

Currently, there are several different confrontation picture naming tests in common use. The most widely used naming test in the United States is most likely the Boston Naming Test (BNT; Kaplan et al., 2001), and most widely used aphasia batteries, such as the Western Aphasia Battery (WAB-R; Lippincott, Williams, & Wilkins, 2007) and the Comprehensive Aphasia Test

(CAT; Swinburn, Porter, & Howard, 2004) have confrontation naming subtests. One other test that is widely used, especially in clinical research contexts, is the Philadelphia Naming Test (PNT; Roach et al., 1996). The PNT is a 175-item test developed to assess naming ability in people with aphasia (PWA). The items of the PNT are one to four syllables in length, vary in lexical frequency, and age of acquisition (Roach et., al 1996). It has strong psychometric properties including high test-retest reliability, low correlation with premorbid educational level, and high correlation with aphasia severity (Walker & Schwartz, 2012).

Despite the widespread use of the PNT, its utility for quantifying anomia has limitations. The PNT, like most other currently available tests, was developed under classical test theory, which relies on often unrealistic assumptions (Hula, Fergadiotis, Swiderski, Silkes, & Kellough, 2020). For instance, score precision for naming tests is typically expressed as a single standard error of measurement that is constant regardless of the ability level of the client being tested. This assumption disregards the idea that standard error of measurement varies as it relates to the ability of the test-taker (de Ayala, 2013). In addition, the PNT is too long to give in many clinical settings due to time constraints on clinicians. Furthermore, the length of the test increases testing burden on PWA and may result in fatigue and frustration that can affect an individual's performance, leading to inaccurate results and conclusions. To address this problem, Walker and Schwartz (2012), developed two 30-item short forms of the PNT. Based on their findings, both short forms correlate highly with the original long form PNT (Walker & Schwartz, 2012).

Although the PNT short-forms (Walker & Schwartz, 2012) offer the advantage of shorter administration time, while also maintaining a strong correlation with the long form of the PNT, there are limitations with this method. Because these short forms are static, containing a fixed item set, they are more precise for people with average severity and less precise for those who are at

the extreme and low ends of the spectrum (Fergadiotis et al., 2019). To demonstrate this point, in a hypothetical scenario where a naming test contains only easy items, scores for individuals with severe aphasia would differ meaningfully from one another, permitting useful rank ordering of the individuals. By contrast, scores for individuals with mild aphasia would be uniformly high, making rank ordering of them less useful if not impossible (Fergadiotis et al., 2019). Additionally, the PNT short-forms assume the standard error of measurement is uniform regardless of ability level, which as discussed earlier, is unrealistic in most testing applications.

An alternative approach to shortening the PNT is to employ item response theory (IRT) and computerized adaptive testing (CAT) methods. IRT (Lord, Novick, & Birnbaum, 1968) is a psychometric framework used for the development and analysis of tools for educational, psychological, and related kinds of measurement. A major advantage of IRT is that it can be used to support computer adaptive testing. An IRT-based computer adaptive test (CAT) uses an algorithm that selects and administers only items targeted to maximize statistical information at a certain ability level (Fergadiotis et.al., 2019). The main advantage of CAT is that it maximizes the precision of the test, requiring fewer testing items while having the same or better level of accuracy as traditional naming assessments of similar length. However, IRT and CAT requires more data than what is required in classical test theory approaches in order to estimate item parameters and investigate validity. To address these issues, Fergadiotis, Hula, and colleagues (Fergadiotis, Kellough, & Hula, 2015; Hula, Kellough, & Fergadiotis, 2015; Fergadiotis et. al., 2019; Hula et. al., 2020; Fergadiotis, Casilio, Hula, and Swiderski, 2021) have investigated the applicability of IRT models to the PNT and developed an IRT-based computer adaptive test version of the Philadelphia Naming Test.

As mentioned previously, IRT is a psychometric framework used for the development of educational and clinical assessments. The essential characteristic being measured in IRT is an unobservable or, *latent*, trait, which is inferred from an individual's observed performance on a set of calibrated test items (Baylor et. al., 2011). Unlike classical test theory, which is focused on test-level properties, IRT emphasizes item-level characteristics, such as item difficulty, and models the responses of a test-taker to the individual items.

A commonly used IRT model is the one-parameter logistic model (1-PL), which uses only one parameter, *item difficulty*, to explain the relationship between the item, the individual's latent trait or ability level, and the individual's response to the item. Item difficulty refers to the location of the item on the trait or ability range (Baylor et. al., 2011). If an individual has a latent trait level higher than an item's difficulty, it will increase the likelihood of answering the item correctly. On the other hand, as items become more difficult, participants would need to have a higher level of the latent trait to correctly respond to the item. Item difficulty and person ability are typically scaled such that when they are the same, the probability of a correct response is 50%.

In the 1-PL model, all items are assumed to have the same discrimination. The 2-PL model adds another parameter to account for variability in item discrimination. *Discrimination* refers to how well an item distinguishes among individuals located at different points along the ability continuum (Baylor et., al 2011). In the 2-PL model, items that have higher discrimination values are more closely related to the trait level, give more information in estimating person ability levels, and are more likely to elicit different responses from individuals with different trait levels (Baylor et al., 2011). While the 2-PL model may better fit the data, it also requires larger sample sizes to ensure the parameter estimates are accurate and stable.

The last IRT model that will be discussed is the three-parameter logistic model (3-PL), which adds a parameter that takes into account the possibility of individuals answering items correctly by chance. This third parameter, often called the guessing or 'psuedo-guessing' parameter, models the probability of a person with infinitely low ability getting an item correct (Baylor et. al., 2011). For the current investigation, a 1-PL model will be used. The 2-PL model could be appropriate for naming assessment in aphasia, and this possibility will be considered in the Discussion. By contrast, because the probability of naming a picture correctly purely by chance is negligible, the 3-PL model is not appropriate for this application and will not be discussed further.

Another important feature of IRT models that makes them useful for computer adaptive testing is the concept of statistical *information*. In an IRT model, each item in the test is associated with an item information function, which defines the degree to which the item increases the precision of an individual's ability estimate (Hula et al., 2015). This function reaches its peak at the ability level where it corresponds to the item's difficulty (Hula et al., 2015). Item information is additive, indicating that as more items are distributed close to the individual's ability level, the information for the overall test is maximized to its potential, producing more precise results (Hula et al., 2015). In this way, the item information function allows selection of items targeted to a person's ability level during computer adaptive test administration.

Recent studies (Hula et. al., 2015; Fergadiotis et al., 2015; Fergadiotis et al., 2019; Hula et. al., 2020) produced computer adaptive tests (CATs) with an item bank consisting of the PNT. An item bank is a set of items that measures a common underlying ability and uses a scale that is defined by person ability and item difficulty (Fergadiotis et.al., 2019). The CAT begins with the assumption that the person has an average score estimate. The first item is selected as the one best

targeted to this ability level. The item is presented, a response is collected, and the score estimate is updated based on that response. The next item is then selected as the one that provides the most information at this new ability level. As responses are collected, the score estimate is revised, and previous steps are repeated until the stopping rule is satisfied. The term *stopping rule* refers to a rule to stop the CAT, typically either when the standard error of the trait estimate falls below a threshold or when a predetermined number of items has been administered. Once the criterion is met, the final score estimate and standard error are presented.

Fergadiotis and colleagues (2015) found that the 1-PL model was adequately fit PNT data collected from a large sample of persons with aphasia and provided reliable estimates of the PNT item difficulties. In a second study, Hula and colleagues (2015) conducted a simulation experiment to test whether the CAT could produce the same results as the full PNT. They investigated two CAT versions using different stopping rules: one thirty-item form (PNT-CAT 30) and one variable length form (PNT-CAT-VL). In both cases, they found that the CAT version correlated highly with the full test (0.95) and provided a valid and efficient measurement of anomia severity in aphasia. They concluded that these results have good implications for the use of an IRT based CAT version of the Philadelphia Naming Test.

In order to confirm these simulation-based results, Fergadiotis, Hula, and colleagues (Fergadiotis et. al., 2019; Hula et. al., 2020) conducted two empirical studies. In their 2019 study, they investigated agreement between independent administration of the full PNT and the PNT-CAT-30. They found a correlation of 0.95, in high agreement with prior simulation study. In the second study, they (2020) evaluated agreement between the PNT-CAT-30 and the PNT-CATVL, which excluded items that were administered in the PNT-CAT-30. They found that the two CAT versions correlated highly (0.90) and were stable in the absence of treatment.

Despite this positive evidence that the PNT-CAT versions agree well with the full test and with each other and are stable in the absence of treatment, there is currently no evidence about their responsiveness to treatment. One previous study of the responsiveness of CAT to treatment for low vision (Massof, 2013) suggested that adaptive testing may underestimate the effects of treatment in some conditions. Massof (2013) used simulation methods to study how well a computer adaptive test measured change in response to different kinds of treatment for low vision. He found that when the simulated treatment affected vision generally (e.g., cataract surgery), the CAT performed well and measured treatment effects accurately. However, when the simulated treatment did not affect vision generally, but only improved performance on some items included in the CAT (e.g., magnification glasses), the CAT underestimated the effects of the treatment. The results suggested that if the intervention produces a change in the properties of the items that were selected, CAT will experience difficulty making an accurate estimate of the person ability score. In these cases, the estimated person ability will depend on the responsiveness to the treatment of the particular items presented. It is important to note that the items in Massof's study had more response categories than the PNT, and thus required a more complicated IRT model. This difference in IRT model structure may influence CAT performance and the effects of the items that were administered.

Following Massof's categorization of treatments for low vision as item-general or item-specific, some treatments for aphasia can affect naming ability generally, while others are specific to the items that are directly trained. In general, behavioral treatments for anomia have larger effects on items that are directly treated and practiced than on items that are not treated directly (Wisenburn and Mahoney 2009; Schuchard and Middleton., 2018; Qique, Evans, and Dickey 2019; Kendall, Moldestad, Allen, & Nadeau., 2019). In Wisenburn and Mahoney's (2009) meta-

analysis of word-finding treatments for aphasia, their findings showed evidence of strong gains for directly trained words and minor gains for untrained words regardless of the treatment approach (semantic, phonological, or mixed). These results parallel other more recent studies of naming treatment such as semantic feature analysis (SFA; Qique, Evans, and Dickey 2019) and phonomotor treatment (Kendall, Moldestad, Allen, & Nadeau., 2019). In Quique, Evans, and Dickey's (2019) meta-analysis of SFA results, they found that improvements were larger for treated words than untreated words. Theoretically, because SFA treatment activities target the semantic system, they cause activation and retrieval of similar concepts in the semantic system, leading to improvement for semantically related untreated items in addition to directly treated items. Results were similar for a study comparing SFA and phonomotor treatment (Kendall, Moldestad, Allen, & Nadeau., 2019), which showed that the directly trained words had the largest improvement. Kendall and colleagues also found that treatment effects generalized to untrained words that shared features (semantic features or phonological sequences, respectively), but to a lesser degree than directly trained words. In both studies, there was no significant generalization to untrained words that did not share semantic features or phonological sequences.

While behavioral treatments for anomia show their strongest effects on the treated items, it is also important to consider non-behavioral interventions, such as pharmacological and non-invasive brain stimulation treatments, which can be provided with or without concurrent behavioral treatment. For example, Hong, Zheng, Luo, Yin, Deng, and Hu (2021) conducted a meta-analysis of 14 studies in order to determine whether transcranial magnetic stimulation (TMS), a kind of non-invasive brain stimulation, had a positive effects on severity of impairment, expressive language, and receptive language in persons with aphasia. They found that TMS combined with behavioral treatment had favorable immediate and long-term effects on language

recovery in patients with post-stroke aphasia. Because some studies (e.g., Gravier et al., 2021; Barwood et al., 2013) have shown positive effects of TMS without concurrent behavioral treatment, the question of how well CAT measures the effects of treatments that affect general naming ability is also relevant.

It is also important to examine treatments that are completely item specific and only affect directly trained items. Errorless learning is one example of a treatment approach that can affect directly trained items through repeated exposure of the stimuli and for which there is limited evidence of generalization to untreated stimuli (Fillingham, Hodgson, Sage, & Lambon 2003; Middleton & Schwartz 2012). This treatment approach is motivated by evidence that some individuals who make errors may strengthen incorrect responses (Fillingham, Hodgson, Sage, & Lambon 2003; Middleton & Schwartz 2012;). In naming treatments, errorless learning reduces the occurrence of errors by removing spontaneous naming attempts. Instead, clinicians administer repeated exposure of the training stimuli to patients in order to activate semantic and phonological features between the target objects and their names (Fillingham, Hodgson, Sage, & Lambon 2003). While a minority of studies of error-reducing treatments showed positive generalization, there is no strong evidence that completely errorless techniques generalize well to untreated items (Fillingham, Hodgson, Sage, & Lambon 2003).

The purpose of this study is to investigate how well the PNT-CAT measures change due to treatment by assessing the sensitivity of PNT-CAT in three different simulated treatment conditions. One condition is treatment that affects naming ability generally, without any item-specific effects (e.g., like some applications of rTMS). The second condition is treatment where the effects are specific to particular items, and therefore only affect directly trained items (e.g., errorless learning). The final condition is treatment that is partially item specific, where the effects

are stronger for treated items and weaker, but *present*, for untreated items (e.g., SFA or phonomotor treatments). For each condition, we will test how well the PNT-CAT and the full PNT measure change due to treatment.

In regard to the first condition in which treatment effects are completely item-general, we hypothesize that the PNT-CAT and the full PNT will perform similarly. For the second condition in which treatment effects are completely item-specific, we presume that the PNT-CAT will underestimate the effects of the treatment relative to the full PNT and the simulated effect. This prediction is motivated by Massof's (2013) study in which CAT underestimated the effects of item-specific treatment because it is dependent on the responsiveness of the particular items presented. Our hypothesis for the third condition is similar to our hypothesis for the second condition. In this case, we hypothesize that the PNT-CAT will underestimate treatment effects relative to the full PNT, but to a *lesser* degree than in condition two. This is because the third condition is only partially item specific and will also affect untreated items. As a result, the third condition may perform slightly better than the second condition. The results of these proposed study will provide needed validity evidence about the responsiveness of CAT to treatment related change.

The proposed research examines how well the different conditions of the PNT-CAT compare with the full PNT. In doing this, we hope to determine how well the PNT-CAT measures change due to anomia treatment. By implementing treatments that target item-general, item-specific, and partial item-specific through repeated structured simulations, we can draw stronger conclusions about the determinants of changes in overall naming performance and examine how different variables affect the likelihood of correct naming responses during CAT administration.

2.0 Methods

The three conditions in this study were naming treatments that 1) affect all items equally, 2) affect only the trained items, and 3) primarily affect trained items with smaller effects on untreated items. The R packages *catR* (Magis & Barrada, 2017) and *catIrt* (Nydick, 2013) were used to simulate administration of the full PNT and computer adaptive PNT (PNT-CAT30) before and after two simulated treatment conditions (item-general, item-specific) under a 1-parameter logistic (1-PL) item response theory model.. The simulation parameters were based on studies conducted in Dr. Hula's lab (Fergadiotis et al., 2015; Hula et al., 2015) and use the ability and item parameter estimates reported by Huston (2021). Huston (2021) refit IRT models originally reported by Fergadiotis and colleagues (2015) within a Bayesian framework and produced item parameter estimates based on a larger participant sample.

We drew 1000 simulated pre-treatment naming ability values from a skew-normal distribution with mean 50 and SD 10 (and thus on a T-score scale) based on empirical data (Fergadiotis et al., 2015; Fergadiotis et al., 2019). Using these ability values and the PNT 1-PL model item parameters, we simulated pre-treatment responses for all 175 items of the PNT for each of the 1000 simulated participants (simulees).

For the item-general condition, we simulated post-treatment responses based on naming ability values increased by a constant 0.4 logits, approximately 2.2 T-score units on the current scale, corresponding to an approximate maximum seven percentage point increase depending on baseline score. This effect size was based on Gravier and colleagues' (2021) study of rTMS as a treatment for anomia without concurrent behavioral treatment.

For the item-specific condition, we simulated post-treatment responses by first selecting 20 items to approximate 25% correct at baseline, within the limits of the item bank, for each simulee individually. We then subtracted 12.08 T-score units (~2.2 logits) from the difficulty of these 20 treated items in order to simulate improvement from ~25% at baseline to ~75% at post-treatment, and then simulated a new set of post-treatment responses for all 175 items. These item-specific treatment effect sizes were based on Quique et al. (2018). We used the simulated responses at pre and post-treatment to estimate scores on both the full PNT and PNT-CAT30 using the original item parameters.

For the partially item-specific condition, we combined the item-general condition effect and item-specific condition effect and implemented those conditions to simulate the post treatment responses. We used the same procedures as for condition 1 and condition 2 where we made a selection of 20 items to approximate 25% correct at baseline, subtracted 12.08 T-score units from the difficulty of these treated items to simulate improvement from ~25% at baseline to ~75% at post-treatment, and also added a constant amount to simulees' ability levels (~0.4 logits/2.2 T-score units) at post-treatment. We then used the simulated responses at pre and post-treatment to estimate scores on both the full PNT and PNT-CAT30 using the original item parameters.

Following Fergadiotis et al. (2019), scores for both the full PNT and the PNT-CAT30 were estimated using Bayesian expected a posteriori (EAP) scoring with a normal prior with a mean of 50, a standard deviation of 10, and possible scores ranging from 10 to 90. The CAT items were selected using the maximum of the Fisher information function at the current ability estimate and the CAT terminated after 30 items.

3.0 Data and Results

3.1 Research Question 1: Item-General Condition

Results are presented in Table 1. For condition 1, paired sample t-tests indicated that the change scores for both the full PNT (*M* = 2.27, 95%CI: 2.16, 2.38, *SD* = 1.83, *t*(999) = 39.3, *p* = < 2e-16) and PNT-CAT30 (*M* = 2.11, 95%CI: 1.92, 2.30, *SD* = 3.06, *t*(999) = 21.7, *p* = < 2e-16) were significantly greater than 0 with confidence intervals that included the generating value of 2.2 T-score units. The difference of change score between the full PNT and the PNT-CAT30 (*M* = 0.16, 95%CI: -0.06, 0.39, *SD* = 3.65) showed that the two tests were not significantly different from one another, *t*(999) = 1.41, *p* = 0.16. Therefore, direct comparison of the full PNT and PNT-CAT30 indicated that they performed similarly.

3.2 Research Question 2: Item-Specific Condition

Results are presented in Table 1. For condition 2, paired sample t-tests indicated that the change scores for both the full PNT (*M* = 1.91, 95%CI: 1.79, 2.03, *SD* = 1.96, *t*(999) = 30.9, *p* = < 2e-16) and PNT-CAT30 (*M* = 5.12, 95%CI: 4.83, 5.40, *SD* = 4.54, *t*(999) = 35.6, *p* = < 2e-16) showed that both PNT versions obtained significant positive change scores. However, the difference of change score between the full PNT and the PNT-CAT30 (*M* = -3.20, 95%CI: -3.51, -2.90, *SD* = 4.09) indicated that the two tests were significantly different from one another, *t*(999) = -20.5, *p* = <2e-16, with the CAT obtaining a larger effect size. A histogram of the number of treated items

administered in the post-treatment PNT-CAT30 is shown in Figure 1. A scatterplot of the PNT-CAT30 change scores over the number of treated items administered is shown in Figure 2. There was a strong correlation between the change score and the number of treated items administered (Pearson r(999) = .70, p < .001). A scatterplot of the number of treated items in the pre-treatment PNT-CAT30 over true baseline naming ability level is shown in Figure 5. A scatterplot of the number of treated items in the post-treatment PNT-CAT30 over true baseline naming ability level is shown in Figure 6.

3.3 Research Question 3: Partially Item-Specific Condition

Results are presented in Table 1. For condition 3, results showed that both PNT versions performed similarly to Condition 2 with both the full PNT (*M* = 4.22, 95%CI: 4.10, 4.33 , *SD* = 1.90, *t*(999) = 70.1, *p* = < 2e-16) and the PNT-CAT30 (*M* = 7.41, 95%CI: 7.14, 7.68, *SD* = 4.33, *t*(999) = 54.2, *p* = < 2e-16) obtaining significant positive change scores, but with the CAT obtaining a larger effect size. The difference of change score between the full PNT and the PNT-CAT30 was significant (*M* = -3.19, 95%CI -3.04, -3.35, *SD* = 3.88, *t*(999) = -21.4, *p* = < 2e-16) and similar in size to Condition 2. A histogram of the number of treated items administered in the post-treatment PNT-CAT30 is shown in Figure 3. A scatterplot of the PNT-CAT30 change scores over the number of treated items administered is also shown in Figure 4. Similar to Condition 2, there was a strong correlation between the change score and the number of treated items administered (r(999) = .63, p < .001).

Table 1. Summary of t-tests comparing simulated full PNT and PNT-CAT30 change scores in item-general and item-specific treatment conditions.

Comparison	Mean (95%CI) Change or Difference Score	SD of Change or Difference Score	t-statistic	p-value
Condition 1: Item-General Treatment				
Full PNT	2.27 (2.16, 2.38)	1.83	39.3	<2e-16
PNT-CAT30	2.11 (1.92, 2.30)	3.06	21.7	< 2e-16
Full PNT – PNT-CAT30	0.16 (-0.06, 0.39)	3.65	1.41	0.16
Condition 2: Item-Specific Treatment				
Full PNT	1.91 (1.79, 2.03)	1.96	30.9	< 2e-16
PNT-CAT30	5.12 (4.83, 5.40)	4.54	35.6	< 2e-16
Full PNT – PNT-CAT30	-3.20 (-3.51, -2.90)	4.09	-20.5	< 2e-16
Condition 3: Partial Item-Specific Treatment				
Full PNT	4.22 (4.10, 4.33)	1.90	70.1	< 2e-16
PNT-CAT30	7.41 (7.14, 7.68)	4.33	54.2	< 2e-16
Full PNT – PNT-CAT30	-3.19 (-3.04, -3.35)	3.88	-21.4	< 2e-16

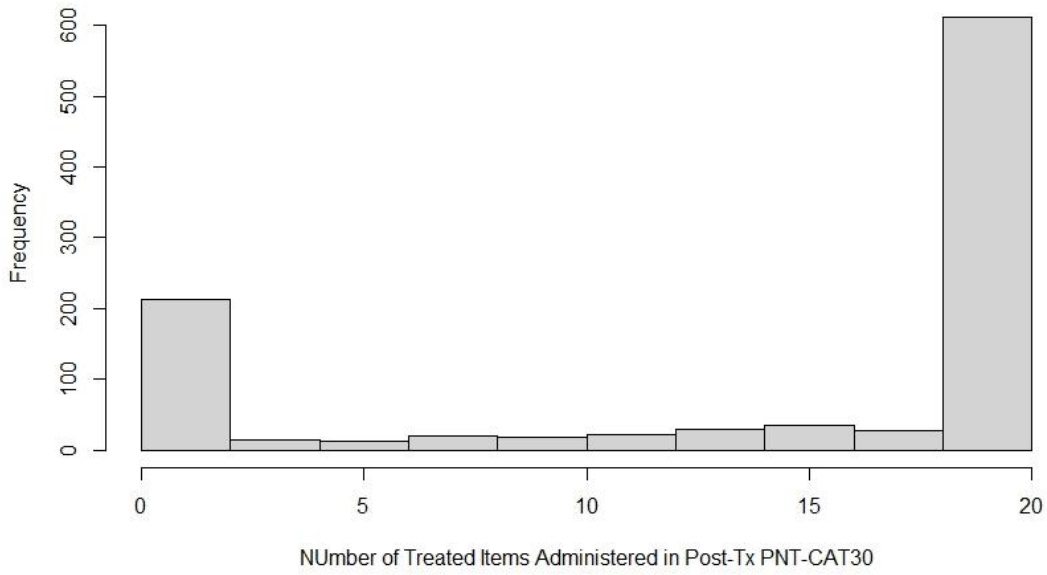

Figure 1. Histogram of the number of treated items administered in the post-treatment PNT-CAT30 in condition 2, item-specific treatment.

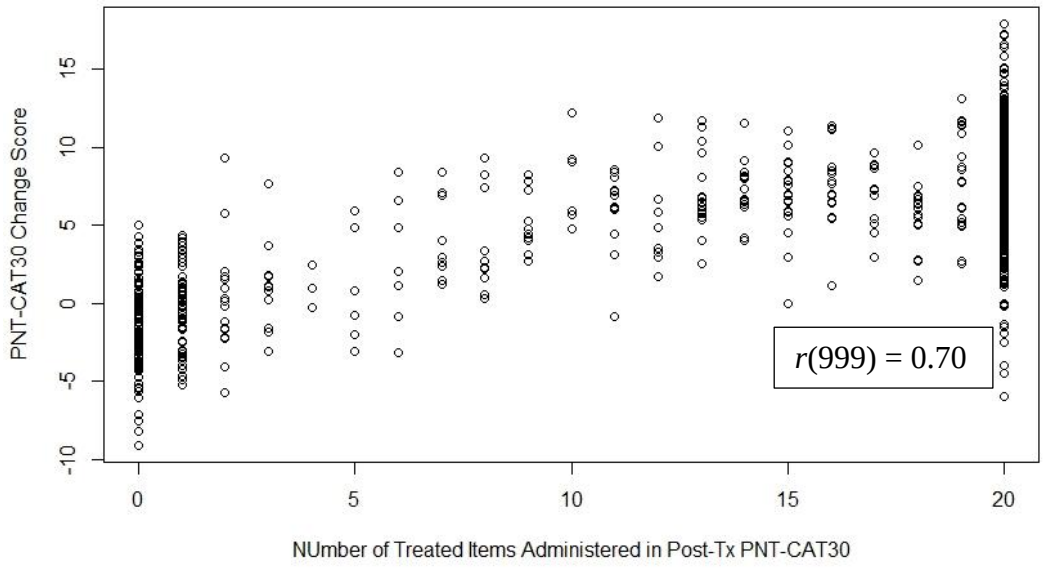

Figure 2. Scatterplot of PNT-CAT30 change scores over the number of treated items administered in the post-treatment PNT-CAT30.

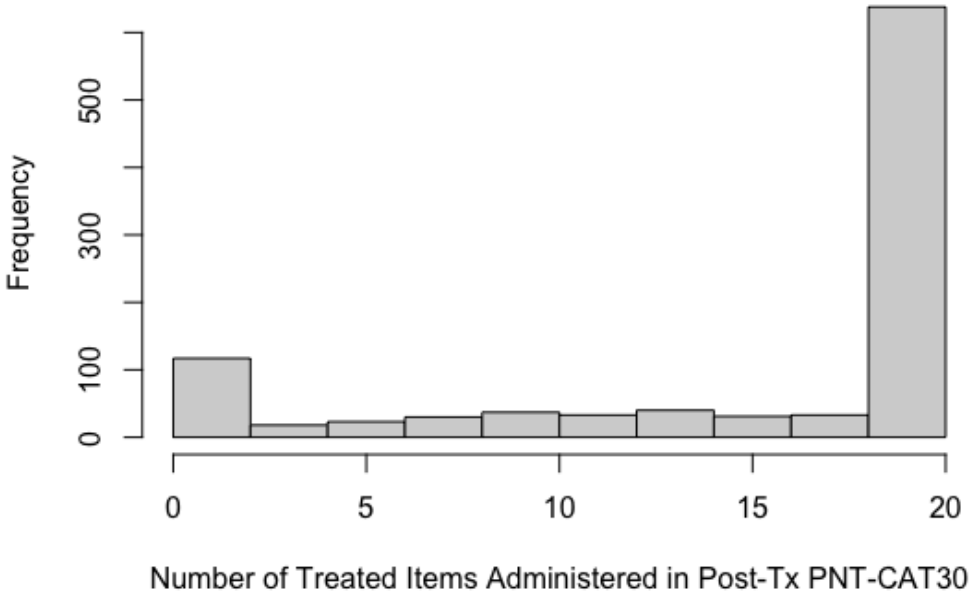

Figure 3 Histogram of the number of treated items administered in the post-treatment PNT-CAT30 in condition 3, partially item-specific treatment.

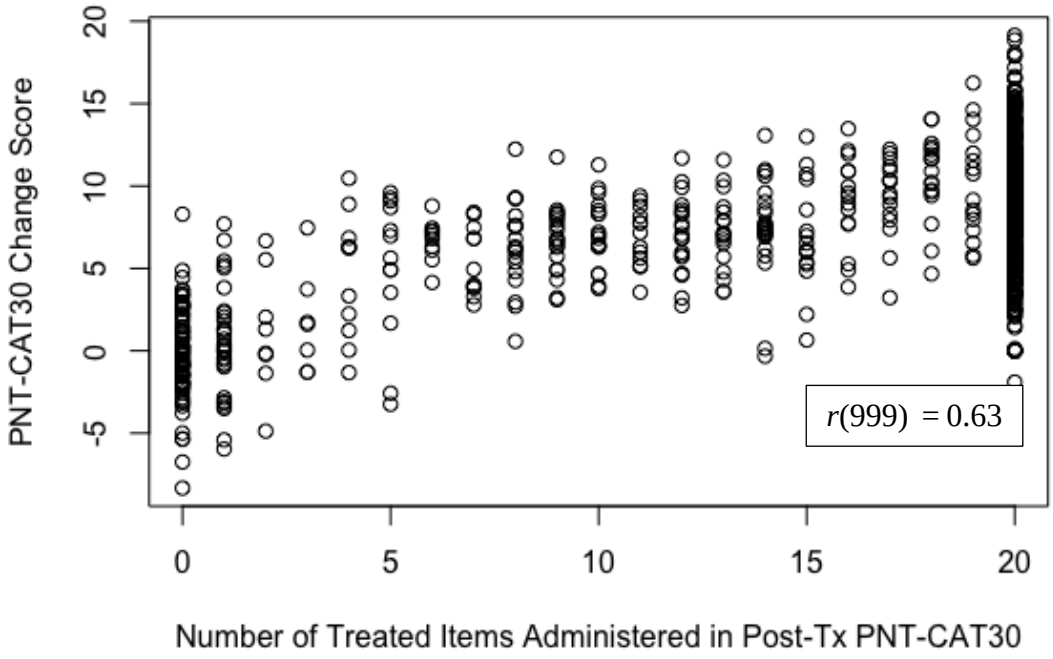

Figure 4 Scatterplot of PNT-CAT30 change scores over the number of treated items administered in the post-treatment PNT-CAT30.

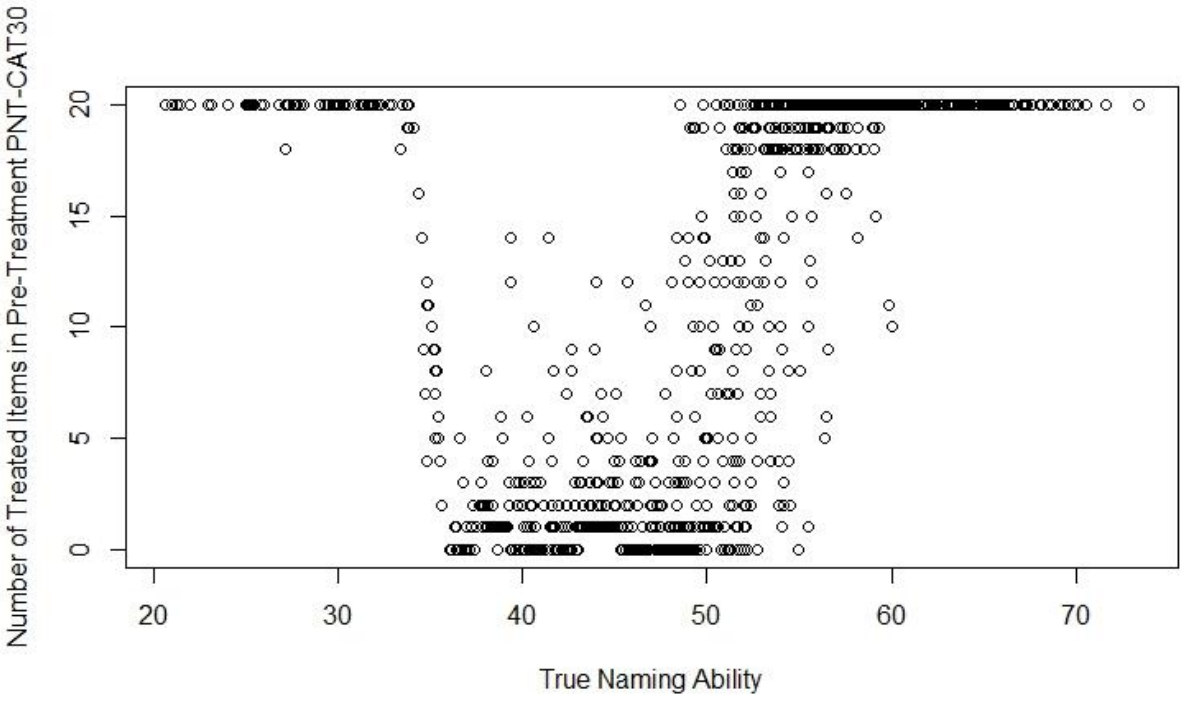

Figure 5 . Scatterplot of the number of treated items in pre-treatment PNT-CAT30 over true naming ability level.

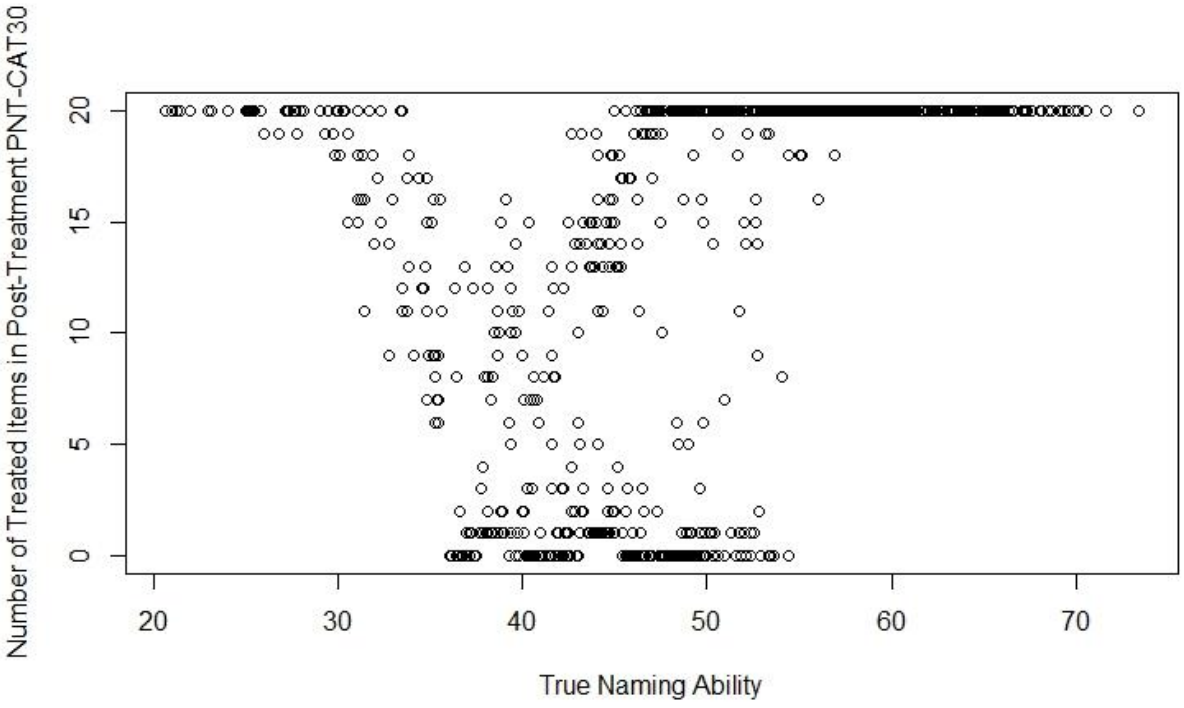

Figure 6 Scatterplot of the number of treated items in post-treatment PNT-CAT30 over true naming ability level.

4.0 Discussion

The overall purpose for this research was to investigate the validity of the PNT-CAT with respect to its responsiveness to treatment related change under three different conditions. The first condition was an item-general treatment, where the treatment affected naming ability generally, without the inclusion of any item-specific effects (e.g., like some applications of rTMS). The second condition was an item-specific treatment where the effects were only specific to a set of items, and therefore only affected the trained items (e.g., errorless learning). The final condition was partially item-specific, with effects that were stronger for directly trained items and weaker, but present, for untrained items (e.g., SFA or phonomotor treatments). For each condition, we compared the change scores estimated by the PNT-CAT and the full PNT to assess the sensitivity of PNT-CAT and determine the conditions under which it would be valid for measuring treatment-related outcomes.

We hypothesized that for the first condition, under which treatment effects were completely item-general, the PNT-CAT and the full PNT would perform similarly. For the second condition, in which treatment effects were completely item-specific, we predicted that the PNT-CAT would underestimate the effects of the treatment relative to the full PNT and the simulated effect. This prediction was motivated by Massof's (2013) findings in which CAT underestimated the effects of the item-specific treatment because of the CAT algorithm's dependence on the responsiveness of the set of items that were presented. Finally, we hypothesized that the third condition would perform similarly to the second condition, in which the PNT-CAT would underestimate the treatment effects relative to the full PNT, but to a lesser degree, because both item-general and item-specific effects were present.

In Condition 1, we found that our results were as predicted. Paired sample t-tests indicated that the change scores for both the full PNT and the PNT CAT were significantly greater than 0 with confidence intervals that included the generating value of 2.2 T-score units. Also, direct comparison of the full PNT and the PNT-CAT30 suggested that both tests performed similarly. These results have implications for the use of the PNT-CAT in a clinical setting. If a treatment is expected to have a general effect on naming ability and none of the items in the PNT are specifically trained, then PNT-CAT30 change scores are valid, if somewhat less precise than full PNT scores.

For Condition 2, results indicated that both PNT tests obtained significant positive change scores. However, a direct comparison of the two showed that our predictions were refuted because the PNT-CAT obtained a significantly larger average effect size. In addition to finding that CAT over-estimated item-specific treatment effects on average, we found a strong relationship between the number of treated items administered in the post-treatment CAT and the change score. As previously stated, the PNT-CAT would be a valid tool for measuring response to treatments with item-general effects. However, if the treatment is item-specific, the assessment should contain only the treated items in order to maximize its responsiveness (Massof 2013).

While Massof's (2013) study suggested that the PNT-CAT would underestimate item-specific treatment effects, we found that the PNT-CAT overestimated them. Two possible reasons that our results differed from Massof's were the differences in the specific IRT model used and the selection of treated items. For this study, we used a simpler IRT model for dichotomous items, which are items that are scored as simply correct versus incorrect. However, in Massof's study, he used a more complicated IRT model for polytomous items that contained 5 categories of response for each item. It is possible that a naming test with more response categories, such as the object

naming subtest of the Comprehensive Aphasia Test (Swinburn, Porter, & Howard, 2004), would have produced different results.

Another potential reason why our findings were different from Massof's were the differences in how items were selected for treatment and how items were selected for inclusion in post-treatment CAT. For the present study, we selected treated items based on participants' ability level specifically to choose treated items that were difficult at pre-treatment. However, we did not control how many of the treated items were selected for the post-treatment CAT as we simply allowed the CAT algorithm to generate the results. By contrast, Massof randomly selected items for treatment and administered the simulated treatments in such a way that 33% of the post-treatment CAT was composed of treated items for each simulee. In our case, allowing the CAT algorithm to freely select items at post-treatment resulted in variable numbers of treated items across all simulees and on average a higher number of treated items at post-treatment CAT, which explained why our CAT was more responsive to the item-specific treatment.

In addition, CAT performance, in terms of both item selection and change score estimation, was dependent on the particular properties of the PNT item-bank relative to the distribution of the simulees ability level. Figures 5 and 6 show that simulees who had extreme high or low naming ability at baseline administered more treated items at pre- and post-treatment, respectively. In turn, this suggests that if items in the PNT are selected to be difficult at baseline within the limits of the item bank and subjected to item-specific treatment, the PNT-CAT30 will overestimate the effects of the treatment relative to the full PNT for the people located at either extreme of the ability continuum. If the item bank had contained more items at the high and low extremes of difficulty, then the results of the present simulation study might have been different.

Given the difference between the present Condition 2 results and Massof's findings, it is important to assess the impact of that difference on how we selected the treated items and how many treated items should be selected at post-treatment as a proportion of the number of items in the post-treatment CAT. It may be useful to conduct an experiment mirroring Massof's framework where we would randomly select 1/3 of the item bank for treatment for each simulee and allow the CAT algorithm to generate the results with the predictions that the results should be similar to Massof's (2013). Thus, further simulations would be needed to investigate this outcome.

The implications of this study are also relevant to all IRT-based tests, regardless if they are adaptive or static. The present results indicate that the responsiveness of any test to an item-specific treatment will be dependent on the number of the treated items that are contained in the test. However, because the 1-PL IRT model assumes that there is an underlying cause (i.e., naming ability) for all the simulees' responses, valid interpretations of the score estimates are different when modeling the effects of item-specific treatment. The PNT-CAT30 performs well in the item-general condition because the intervention affects the simulees' underlying naming ability, which affects how they respond to all of the items in a probabilistic manner. By contrast, the item-specific condition influences only the observed response to a subset of items. As a result, this causes the IRT model to not function properly, resulting in change score estimate that are less valid within the IRT framework. We are explicitly measuring the response to treatment of a specific set of items, negating the ability to generalize inferences about that person's naming ability to other items. Therefore, a change score estimate following an item-specific treatment should be interpreted differently than one obtained following an item-general treatment. In the item-specific case, inferences do not generalize beyond the trained items whereas in the item-general case, in theory, inferences apply to all items.

The results for Condition 3 were similar to the results for Condition 2 in that both tests had change scores that were significantly greater than 0 but the CAT had a larger effect size. It is important to consider that Condition 3 was the additive combination of Conditions 1 and 2. In Condition 1, we increased ability values by 2.2 T-score units whereas for Condition 2, we found a difference of 3.2 T-score units between the CAT and the full test. The results for Condition 3 indicated we accurately recovered the item-general effects similarly to Condition 1 (i.e., the change estimate of PNT-CAT for Condition 3 is 7.4 which is approximately 2.2 higher than Condition 2's change estimate of 5.1) and recovered the same between-test difference that occurred in Condition 2 (i.e., the change estimate difference between the full PNT and the PNT-CAT for Condition 3 is -3.19, while the change estimate difference between the full PNT and the PNT-CAT for Condition 2 is -3.20). As a result, in Condition 3 the PNT-CAT overestimated the treatment effect relative to the full PNT. Based on these results, if a treatment is expected to contain both effects, we would advise users to administer the PNT-CAT but exclude the treated items from the item bank and assess those trained items separately.

Limitations

While computer simulations have advantages, as they are quick, inexpensive, and offer a high degree of experimental control, their external validity may be limited by how much the results may transfer to realistic settings. Previous work on the PNT-CAT (Fergadiotis et al., 2015; Hula et al., 2015; Fergadiotis et al., 2019; Hula et al., 2020) found that the empirical studies produced results similar to those from the simulation studies. Therefore, evidence suggest the that current simulations results are useful. However, because this is a new area of research and there are various treatments that function differently from one another, it is crucial to replicate these results in an

empirical study with real people with aphasia. The present simulation study can provide useful guidance on how to accommodate different treatment conditions, therefore making future empirical studies more efficient and preserving resources. By implementing treatments that target both item-general, item-specific, and partially item-specific through repeated structured simulations, we can draw stronger conclusions about the determiners of the overall naming performance and examine how different variables affect the likelihood of correct naming responses during CAT administration. Overall, the results of these studies will provide an essential framework for guiding clinicians and researchers in appropriate, evidence-based use of computer adaptive testing for measuring treatment-related change in aphasia rehabilitation.

Future Directions

Additional research that can be implemented based on this study would be to conduct a simulation that contains a range of different treatment effect sizes based on previous studies. This potential study could have multiple levels item-general effect sizes, ranging from null to large, as well as a range of item-specific treatment effect sizes that could be operationalized by applying the item-specific treatment to different number of items, ranging from 0 to 20 items. We could then study the agreement between the full PNT CAT and the CAT change scores as a function of underlying change in general in naming ability and of the number of directly treated items and examine whether these results parallel those of our present study.

Another potential direction for future studies is to repeat this simulation with a different IRT model such as the two-parameter logistic (2-PL) model. As suggested earlier in the Discussion, using a specific IRT model may impact the selection of the treated items. The 1PL model only estimates one parameter, item difficulty, for each item and assumes that the discrimination

parameter is constant across all items. The main advantages of this model are its parsimony and the fact that it can be fitted with smaller sample sizes. However, these properties are useful only if the model fits the data adequately. In some situations, the assumptions of the 1PL model do not hold, and more complex models, like the 2PL, would be more appropriate. The 2PL model, which also models binary items, allows them to vary in both their difficulty and discrimination (Baylor et. al., 2011). Because of the additional flexibility of the discrimination parameter, the 2PL model shows that higher discriminating items perform better at differentiating between people of two different ability levels. Given the same test and the same participants, the 2PL model may provide a more accurate and precise ability estimates. In addition, the 2PL model does a better job at fitting and reproducing the data, which can potentially increase the confidence in interpreting the item parameters and person score outputs. Because of this, if the 2PL model becomes an accepted measurement model for the PNT, it would be useful to repeat and verify that the present results hold given that the 2PL model could produce different results.

5.0 Conclusion

In this study, we investigated the validity of PNT-CAT for measuring change in response to three different treatment conditions. The PNT-CAT is valid for measuring change if a treatment is item-general as it performed similarly to the full PNT in this condition. However, the PNT-CAT performed differently in both the item-specific and partially-item specific conditions, with its responsiveness depending on the number of treated items included in the post-treatment CAT. Therefore, if the PNT-CAT encounters a treatment that is item-specific, it is recommended that the assessment should only contain the treated items in order to maximize its responsiveness. Similarly, if the treatment contains both effects, it is advised to separate the treated items from the item bank and assess those items separately. This study demonstrates the usefulness of computerized adaptive testing in clinical aphasiology and provides results that can be used to improve aphasia assessment.

Bibliography

Barwood, C. H., Murdoch, B. E., Riek, S., O'Sullivan, J. D., Wong, A., Lloyd, D., & Coulthard, A. (2013). Long term language recovery subsequent to low frequency rTMS in chronic non-fluent aphasia. *NeuroRehabilitation*, *32*(4), 915-928.

Baylor, C., Hula, W., Donovan, N. J., Doyle, P. J., Kendall, D., & Yorkston, K. (2011). An introduction to item response theory and Rasch models for speech-language pathologists. *American journal of speech-language pathology*, *20*(3), 243–259. https://doi.org/10.1044/1058-0360(2011/10-0079)

de Ayala, R. J. (2013). *Theory and practice of item response theory*. Guilford Publications.

Fergadiotis, G., Kellough, S., & Hula, W.D. (2015). Item Response Theory Modeling of the Philadelphia Naming Test. *Journal of speech, language, and hearing research : JSLHR, 58 3*, 865-877 .

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 : JSLHR*, *62*(6), 1724–1738. https://doi.org/10.1044/2018_JSLHR-L-18-0344

Fergadiotis, G., Casilio, M., Hula, W. D., & Swiderski, A. (2021, June). Computer adaptive testing for the assessment of anomia severity. In Seminars in Speech and Language (Vol. 42, No. 03, pp. 180-191). Thieme Medical Publishers, Inc..

Fillingham, J. K., Hodgson, C., Sage, K., & Lambon Ralph, M. A. (2003). The application of errorless learning to aphasic disorders: A review of theory and practice. *Neuropsychological rehabilitation*, *13*(3), 337–363. https://doi.org/10.1080/09602010343000020

Goodglass, H. (1993). *Understanding aphasia*. San Diego: Academic Press.

Goodglass, H., & Wingfield, A. (1997). Anomia: Neuroanatomical and cognitive correlates. Academic Press.

Hong, Z., Zheng, H., Luo, J., Yin, M., Ai, Y., Deng, B., Feng, W., & Hu, X. (2021). Effects of Low-Frequency Repetitive Transcranial Magnetic Stimulation on Language Recovery in Poststroke Survivors With Aphasia: An Updated Meta-analysis. *Neurorehabilitation and neural repair*, *35*(8), 680–691. https://doi.org/10.1177/15459683211011230

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

Hula, W. D., Fergadiotis, G., Swiderski, A. M., Silkes, J. P., & Kellough, S. (2020). Empirical Evaluation of Computer-Adaptive Alternate Short Forms for the Assessment of Anomia Severity. *Journal of Speech, Language, and Hearing Research*, *63*(1), 163–172. https://doi.org/10.1044/2019_JSLHR-L-19-0213

Huston, H. (2021). *Bayesian Item Response Theory Modeling* (Unpublished doctoral dissertation or master's thesis). Portland State University, OR.

Gravier, M.L., Dickey, M.W., Hula, W.D., Johnson, J.P., Autenreith, A.V., Doyle, P.J., & Forman, S. (May 2021). Excitatory-primed inhibitory rTMS is more effective than inhibitory TMS alone. Presentation to the Clinical Aphasiology Conference, virtual.

Kendall, D. L., Moldestad, M. O., Allen, W., Torrence, J., & Nadeau, S. E. (2019). Phonomotor Versus Semantic Feature Analysis Treatment for Anomia in 58 Persons With Aphasia: A

Randomized Controlled Trial. *Journal of speech, language, and hearing research : JSLHR*, *62*(12), 4464–4482. https://doi.org/10.1044/2019_JSLHR-L-18-0257

Lippincott Williams & Wilkins-Kertesz, A. (2007). Western Aphasia Battery – R. Grune & Stratton

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

Magis, D., & Barrada, J. (2017). Computerized Adaptive Testing with R: Recent Updates of the Package catR. *Journal of Statistical Software*, *76*(Code Snippet 1), 1 - 19. doi:http://dx.doi.org/10.18637/jss.v076.c01

Massof, R. W. (2013). 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

Magis, D., & Barrada, J. R. (2017). Computerized Adaptive Testing with R: Recent Updates of the Package catR. *Journal of Statistical Software, Code Snippets*, *76*(1), 1–19. https://doi.org/10.18637/jss.v076.c01

Nydick, S. W. (2013). catIrt: An R package for simulating IRT-based computerized adaptive tests. R package version 0.4-1.

Quique, Y. M., Evans, W. S., & Dickey, M. W. (2019). Acquisition and Generalization Responses in Aphasia Naming Treatment: A Meta-Analysis of Semantic Feature Analysis Outcomes. *American journal of speech-language pathology*, *28*(1S), 230–246. https://doi.org/10.1044/2018_AJSLP-17-0155

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for

Statistical Computing, Vienna, Austria. 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

Schuchard, J., & Middleton, E. L. (2018). Word repetition and retrieval practice effects in aphasia: Evidence for use-dependent learning in lexical access. *Cognitive Neuropsychology, 35*(5-6), 271–287. https://doi.org/10.1080/02643294.2018.1461615

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

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)

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