7-9-2021

Ancillary Data for Refining Computer Adaptive Algorithms for the Assessment of Anomia

Emily Kathryn Tudorache
Portland State University

Recommended Citation
Tudorache, Emily Kathryn, "Ancillary Data for Refining Computer Adaptive Algorithms for the Assessment of Anomia" (2021). *Dissertations and Theses.* Paper 5730.
https://doi.org/10.15760/etd.7601

Ancillary Data for Refining Computer Adaptive Algorithms for the Assessment of Anomia

by

Emily Kathryn Tudorache

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science
In
Speech and Hearing Sciences

Thesis Committee:
Gerasimos Fergadiotis
William Hula
Maria Kapantzoglou

Portland State University
2021

Abstract

Computer adaptive testing formats, based in item response theory (IRT), are becoming an increasingly popular approach to testing in healthcare because they offer numerous psychometric and practical advantages to assessment when compared to static tests that rely on classical test theory. Fergadiotis and colleagues (2015) have developed computer adaptive versions of the Philadelphia Naming Test (PNT) short-forms, which have demonstrated acceptable precision and standard error of measurement when compared to the static short-forms and original full-length assessment. This study sought to use synthetic data simulations using the catIrt R package (Nydik, 2014) to investigate possible advantages of the use of tailored provisional ability scores at the start of a CAT PNT. Results revealed no significant improvement in the performance of the test when starting at a tailored provisional ability score. These results further guide next steps in developing more precise computer adaptive tests for assessing anomia and additionally demonstrated the advantages of computer simulations in advancing this line of work.

Acknowledgements

My utmost gratitude goes to my advisor, Gerasimos Fergadiotis, for his support and guidance on this thesis. I could not have accomplished this without his advice, expertise, and encouragement. I would also like to thank Will Hula for his advice in developing the method of this project. Finally, without the consistent emotional support and encouragement from my family, particularly my sister, Kaitlyn, and my friends and classmates, the completion of this thesis would not have been possible. I could not be more appreciative of all they have done for me.

Table of Contents

Abstract i
Acknowledgements ii
Introduction 1
Methodology 10
  Synthetic Data Post Hoc Simulation 10
  Statistical Analysis 12
Data and Results 13
  Research Question 1: Differences on Average SEM across Conditions 13
  Research Question 2: Differences on Correlations with “True” Theta across Conditions 14
Discussion 16
  Limitations 19
  Future Directions 20
References 21
Appendix A: catIrt Simulation Code 23

List of Tables

Table 1: Descriptive Statistics for the SEM Based on the Full Item Bank and the Two Simulated Conditions 12
Table 2: Mean, SD’s, and Correlations among “true” and CAT generated thetas 14

List of Figures

Figure 1: A graphical representation of the standard error of measurement associated with the full item bank and the two conditions 13

Introduction

Aphasia is a general term that describes impairments of language following brain damage, typically when the damage occurs in the language dominant hemisphere (Goodglass & Wingfield, 1997). It is characterized by a variety of impairments related to communication, such as difficulty with speaking, understanding language, reading, or writing. Anomia, the inability to access, retrieve, and phonologically encode the name of an object or concept, is the most prevalent deficit associated with aphasia (Goodglass, 1993). Individuals with anomia experience major negative impacts to their ability to communicate what they want efficiently or accurately. This can become incredibly frustrating or even debilitating in day-to-day life (Goodglass, 1993).

One of the most commonly used methods for assessing severity of word-finding difficulty is confrontation picture naming. Picture naming tests are the primary choice for assessing anomia because they provide consistent assessment content (Goodglass, 1993). In connected speech or discourse tasks, individuals can modify their language output for certain words; thus, compensating for their word-retrieval difficulties. With picture naming tasks, all individuals are presented with a specific stimulus that has a specific target and must produce the same word (Goodglass, 1993). Thus, standardized picture naming tests present an avenue to even and constrain the testing demands for more accurate assessment of the underlying deficit.

There are a variety of standardized tests for confrontation picture naming (e.g., Boston Naming Test [BNT; Kaplan et al., 2001], Philadelphia Naming Test [PNT; Roach et al., 1996], naming subtest of the Western Aphasia Battery Revised [WAB-R;

Lippincott, Williams, & Wilkins, 2007], etc.). Of these, the PNT is a commonly used tool in research investigations due to its strong psychometric properties (Walker & Schwartz, 2012). A study by Walker and Schwartz (2012) on this test found excellent test-retest reliability and that the PNT as a measure of anomia severity is a strong predictor of overall aphasia severity as quantified by the Western Aphasia Battery Revised (WAB-R) Aphasia Quotient (Walker & Schwartz, 2012). Additionally, the items on the PNT have high naming agreement, as indicated by 85% or greater of a group of 30 control participants responding to the image with the appropriate target word (Roach et al., 1996).

The PNT is composed of 175 black-and-white line drawings of objects that present a range of occurrence in everyday language. However, the considerable length of the PNT means that, despite its superior reliability and positive psychometric properties, it is often inappropriate for use in many clinical settings where productivity demands impact time available for assessment and patient stamina determines the feasibility of completing the assessment. To address this problem, Walker and Schwartz (2012), developed two 30-item short-forms of the PNT. Both forms have been shown to 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 maintaining a strong correlation with the long-form PNT, there are limitations associated with their use. The PNT short-forms are based in classical test theory, and therefore have certain psychometric weaknesses (Hula et al., 2020). As static tests containing a fixed item set, they are most precise for people with average severity, while being less precise for those at the extreme high or low ends (Hula et al., 2020). Additionally, the PNT short-forms assume the standard error of measurement is uniform regardless of the ability level of the individual taking the test (Walker & Schwartz, 2012), while in reality the standard error of measurement varies with the difficulty of the test as it relates to the ability of the test-taker (de Ayala, 2013). As a result, measures for monitoring change in naming ability pre- and post-treatment may be distorted as the individual's naming ability does or does not change.

To address these limitations, Fergadiotis, Hula and their colleagues have developed an item response theory (IRT) version of the PNT. IRT (Lord & Novick, 1968) is a psychometric framework used often for psychological testing that assumes the characteristic being measured is on an unobservable, or *latent*, continuum that jointly references the difficulty of test items and the ability level of the individuals being tested (Embretson & Yang, 2006). IRT models use information about the difficulty of a test item and the ability level of the test-taker to predict how the individual will respond to an item. A commonly used IRT model is the one-parameter logistic model (1PL) which calculates the probability an individual will respond correctly to an item based on the item's difficulty and the individual's estimated ability level (theta). The model often has ability level scaled to a mean = 0 and standard deviation = 1, and item difficulty is then placed on the same scale (Fergadiotis et al., 2015). When item difficulty and person ability level are equal, the probability a patient will answer the item correctly is 50%. When item difficulty is greater than person ability level, the probability of a correct response is lesser, while the inverse is true for person ability level greater than item difficulty.

Fergadiotis, Kellough & Hula (2015) have investigated the applicability of IRT models to the PNT and its associated short-forms and have found that models based on item difficulty and person ability are quite precise when assessing object naming ability using the PNT short-forms. In their 2015 study, Hula, Fergadiotis, and Kellough developed two computer adaptive versions of the PNT, one thirty-item form (PNT CAT-30) and one variable length (PNT CAT-VL). Both forms were compared to the full-length PNT and the two PNT short forms developed by Walker and Schwartz (2012). Results found that both the PNT CAT-30 and PNT CAT-VL correlated significantly greater with the full-length PNT than the PNT30A form and the PNT CAT-VL correlated with the full-length test more strongly than the PNT30B (Fergadiotis et al., 2015). These results have good implications for the use of an IRT based computer adaptive test (CAT) version of the Philadelphia Naming Test.

An important feature of IRT models in adaptive testing is *information.* In IRT, each item is associated with an item information function which shows the level of information each item contributes for the estimation of ability (Hula, Kellough, & Fergadiotis, 2015). The function reaches its peak precision at the ability level that corresponds to the item’s difficulty (Hula et al., 2015). The square root of the inverse of this function represents the standard error for that item, thus information refers to the extent an item reduces the uncertainty of an ability estimate (Hula et al., 2015). In an IRT-based CAT, the algorithm utilizes each item’s information function so that the computer can discern which items are most informative at a given ability level score (Hula et al., 2015). When item information functions for multiple items are overlaid, we can produce a test information function that models the information for the collective, administered items for an individual test-taker (Hula et al., 2015). As more items closer to the test-taker's "true" ability level score are administered, the information for the overall test is maximized and the result is more precise (Hula et al., 2015).

Computer adaptive testing is a testing format in which a computer algorithm collects responses from the test-taker throughout the testing process and presents items that have been calculated to provide the most information about the test-taker's "true" ability level (Fergadiotis et al., 2015). CAT begins with the individual being assigned a preliminary ability level score, typically the mean (= 0), and then are presented with a first item with a difficulty level that matches the preliminary ability score. The individual responds either correctly or incorrectly, and the computer uses the individual's response to update the provisional ability score, and then selects either an easier or more difficult item that would be maximally informative based on the updated ability level score. The individual is presented with the new item, and the process repeats until a stop rule is met. For fixed-length tests, this is after the target number of items have been administered. For variable-length tests, this is after a predetermined level of precision is met.

By presenting a series of optimally informative items to the test-taker, computer adaptive tests can maintain the precision of measurement of a test while being relatively shorter than a non-adaptive test. Some adaptive tests are administered without the use of a computer through basal and ceiling rules (such as with the BNT), that determine a start point based on prior notions of the individual’s ability (due to age, scores on related external measures, etc.) and a stop point that is determined by their performance on the test so far (e.g. three consecutive questions incorrect) (Mills & Stocking, 1996). This requires the test to be organized so that items increase in difficulty as the test goes on and each individual must be presented with each item in order until the stop rule is reached, a feature not found in the BNT. Items are not presented in order of increasing difficulty, so some participants may not be presented with easier items that they would have named correctly, thus resulting in an incorrectly lowered score.

The majority of IRT based computer adaptive tests begin by selecting an item at the mean ability level and proceeding from there; however, clients’ “true” ability scores are not always equivalent to the mean. Additional information from clinical impressions and other measures can inform a general sense of an individual’s severity of anomia. This general sense of aphasia severity can be used to select a testing start point that may be closer to their actual ability level than the mean. It is possible that estimating a person’s general aphasia severity and starting them at a start point closer to this perceived severity can help optimize the results of the test by either requiring fewer items to be administered to reach the stop rule or resulting in a greater number of optimally informative items to be administered.

The purpose of this study is to determine the possible advantages of developing computer adaptive testing software for confrontational naming tests that allows clinicians to select a starting provisional ability score based on clinical impressions and/or additional measures, such as scores from other standardized aphasia assessments. Testing software that allows clinicians to choose a starting point as an alternative to software that starts testing for all clients at a mean provisional ability score may have implications for improving the efficiency and accuracy of computer adaptive naming tests on an individual basis, since the starting point for the test would be tailored to the unique presentations of aphasic individuals. To address this premise, we ask the following research questions: 1) Is there a difference between mean level of uncertainty for ability estimate scores for computer adaptive naming tests that begin at an average provisional ability score and those that begin at a provisional ability estimate set equal to the "true" ability score? and 2) Is there a difference between the mean concordance of computer adaptive naming test ability estimate scores and "true" ability estimate scores for computer adaptive naming tests that start at an average ability estimate and computer adaptive naming tests that begin at a provisional ability estimate set equal to the "true" ability score?

Regarding the first question, we hypothesize that level of uncertainty will be more precise for the tests that start at the test-taker's "true" ability score compared to tests that start at a default provisional score. This is because in the latter scenario, the test would start at the mean and then present items that would gradually approach the client's "true" ability estimate. In a fixed-length computer adaptive test, this would mean potentially fewer maximally informative items would be presented, especially for individuals who are severely or mildly impaired. With an overall less informative item set, the confidence interval for the final ability estimate score would be wider, and therefore uncertainty about the "true" ability score would be greater. For the prior scenario, it is probable the test would present more items that are closer to the client’s “true” ability estimate and therefore are more informative. The more informative the item set presented, the narrower the confidence interval for the final ability estimate score.

Our hypothesis for the second question is similar to the hypothesis above. We presume ability estimate scores for tests that allow clinicians to select starting points based on prior knowledge of the client will be more precise than ability estimate scores for tests that begin at a predetermined provisional ability score. Similarly, to narrowing confidence intervals through presenting a greater number of maximally informative items for the individual, the test would present items that would gradually approach the individual's “true” ability estimate score, resulting in a more precise estimate within a fixed number of items.

The proposed research questions explore the limiting case, in which we compare current conditions of computer adaptive testing with starting provisional ability score set at the mean to the ideal conditions: computer adaptive testing with a starting provisional ability score equivalent to test-taker’s “true” ability score, conditions that will not realistically occur in a clinical setting since the “true” ability score cannot be known. In doing this, we optimize the outcomes of the test by presenting the test-taker with a maximally informative initial test item. The difference between these conditions will create the theoretical space for improvement and demonstrate the greatest potential advantages for creating CAT software that allows clinicians to select a starting point based on other factors (e.g. clinical impressions, alternative measures). If the improvement between the two conditions is insignificant, then we can know advantages in realistic clinical conditions will also be insignificant and proceed with exploring alternative approaches to augment the estimation process. However, if the space for improvement between current conditions and ideal conditions is significant, it may warrant further investigation of potential advantages in more realistic conditions and guide future research.

Methodology

Synthetic Data Post Hoc Simulation

The *catIrt* package was used in R Studio to simulate the results of the computer adaptive version of the Philadelphia Naming Test under the two conditions. The *catIrt* package (Nydik, 2014) simulates computer adaptive tests based on a vector of ability values, a matrix of item parameters, and item selection and termination criteria. The available code can be found in Appendix A. To begin, generating (i.e., "true") ability scores were created for 1000 simulees by generating random thetas using the *rnrom* function in R. These thetas represent the generating ability parameters of the hypothetical cases that were ran through the simulated CAT PNT under the two conditions. For this simulation, the 1000 cases were generated to adhere to the distribution from the study by Fergadiotis et al. (2015) according to which the mean equaled 0.1 and the standard deviation equaled 1.44

The next step was to generate simulated responses for all 175 items of the PNT for each of the 1000 simulees. To this end, the generating ability parameters (i.e. thetas) and the known PNT item parameters from Fergadiotis et al. (2015) were used. The responses were generated under a 1-parameter logistic model.

After the datasets were generated, a classical post hoc simulation using synthetic data was performed. For the uninformed Condition 1, every CAT simulation was based on the same initial ability estimate across simulees (ability estimate = 0) whereas for Condition 2, the generating thetas were provided as the initial ability estimate. Other than that, *catIRT* was configured identically across conditions.

Specifically, *catIRT* code required the specification of start, middle, and final phase of the simulation. First, the options for specifying the start of the CAT were configured. These settings applied to the first five items. The key parameter in this block of code was “init.theta”. If “init.theta” was a scalar, every simulee had the same starting value (i.e., Condition 1). Otherwise, simulees had different starting values based on the respective element of “init.theta” (Condition 2). Further, no constraints were imposed on the level of difficulty parameters for items selected during the starting portion of the CAT. Additionally, the adaptive algorithm was specified to select the single best item that maximized the unweighted Fisher information precisely at theta. Finally, thetas were estimated based on the expected a posteriori scoring and theta was constrained to fall between -4 and 4.

Next, the middle phase of the simulation was specified. The phase was configured similarly to the start phase. The only difference was that after the first five items, theta estimation was based on maximum likelihood.

Finally, the final phase of the simulation was configured. The key parameter here was "term" which can be configured to terminate the CAT either after a specified number of items ("fixed") or after a certain standard error measurement is achieved ("precision"). For this study, each simulation was terminated after the administration of 30 items.

Statistical Analysis

The first research question asked whether there was a difference between mean level of uncertainty for ability estimate scores for computer adaptive naming tests that begin at an average provisional ability score and those that begin at a provisional ability estimate set equal to the “true” ability score. To answer this question, the average standard errors of measurement across the two conditions were compared using a paired samples *t-test* with an alpha level set equal to .001. The second research question asked whether there was a difference between the mean concordance of computer adaptive naming test ability estimate scores and “true” ability estimate scores for computer adaptive naming tests that start at an average ability estimate and computer adaptive naming tests that begin at a provisional ability estimate set equal to the “true” ability score. To answer the second question, the correlations between the thetas generated under each condition and the generating thetas were compared statistically using Steiger’s approach (Steiger, 1980) as implemented in Lee and Preacher’s (2013) online Java application.

Data and Results

Research Question 1: Differences on Average SEM across Conditions

Descriptive statistics for the standard error of measurement associated with the administration of the full item bank and each of the two conditions can be found in Table 1. A graphical representation of the three empirical SEM's as a function of the generating theta can be seen in Figure 1. As expected, and as can be seen in Figure 1, the SEM of measurement associated with the administration of the full item bank was considerably lower compared to the two CAT conditions.

Table 1

Descriptive Statistics for the SEM Based on the Full Item Bank and the Two Simulated Conditions

	Full Item Bank	Condition 1	Condition 2
Mean	0.187	0.336	0.334
SD	0.145	0.091	0.088
Min	0.131	0.291	0.291
Max	0.081	0.815	0.815

Note. Condition 1 refers to the uninformed CAT simulation and Condition 2 refers to the CAT simulation for which initial values were provided.

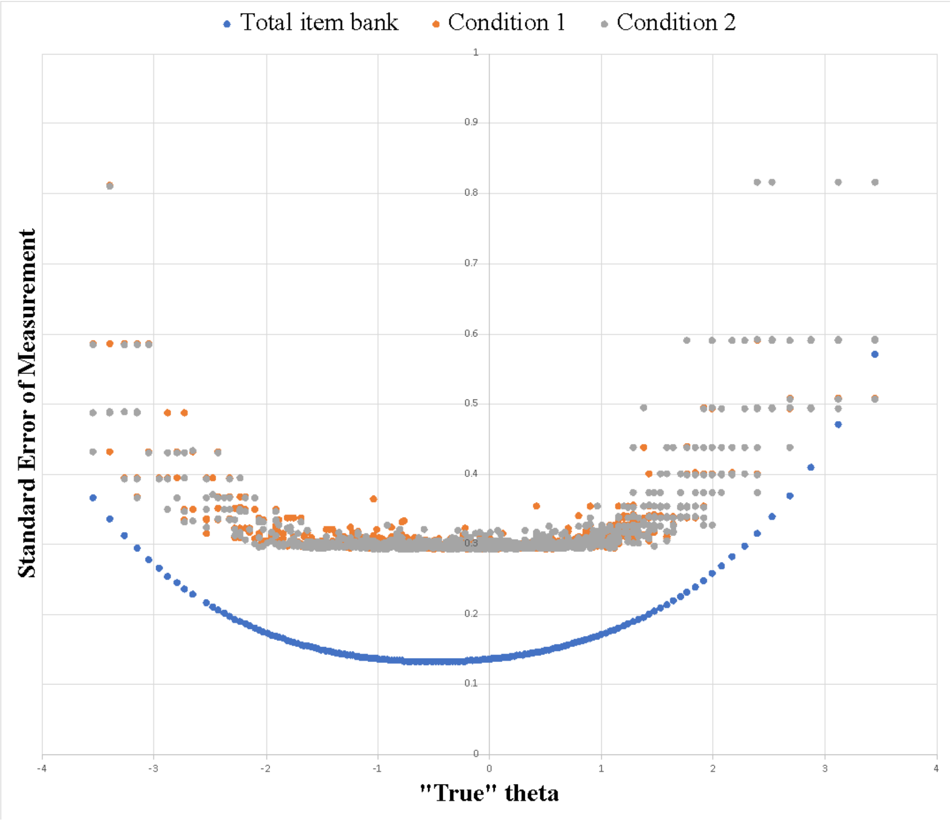

Figure 1. A graphical representation of the standard error of measurement associated with the full item bank and the two conditions.

The difference of the average SEM from the first condition (*M* = 0.336, *SD* = 0.091) and the second condition (*M* = 0.334, *SD* = .087) was not statistically significant, *t*(999) = 1.583, *p* = .114. Therefore, no evidence was found that the two CAT simulated conditions were associated with different levels of average SEM.

Research Question 2: Differences on Correlations with "True" Theta across Conditions

Table 2 includes the descriptive statistics associated with the "true" theta and the estimated theta under the two simulated CAT conditions.

Table 2

Means, SD's, and Correlations among "true" and CAT-generated thetas

	"True" Theta	Condition 1	Condition 2
"True" Theta	1		
Condition 1	0.968	1	
Condition 2	0.969	0.964	1
Mean	0.118	0.114	0.116
SD	1.395	1.442	1.44

Based on the Steiger's approach, the difference between the correlation based on thetas generated under Condition 1 and the "true" theta and thetas generated under Condition 2 and the "true" theta was not statistically significant, $z(999)=-.556$, p = .58.

Discussion

The primary aim for this study was to investigate potential advantages to developing computer adaptive software for confrontation naming tests that would allow for selection of an initial provisional ability score compared to the current standard in which computer adaptive tests begin at an empirically estimated common provisional ability score. It was hypothesized that starting test-takers at a provisional ability score closer to their "true" ability level would result in a more precise result, as this could potentially increase the number of maximally informative items presented to the test-taker. We addressed this aim by comparing current CAT practice standards against the limiting case and asked the following research questions: 1) Is there a difference between mean level of uncertainty for ability estimate scores for computer adaptive naming tests that begin at an average provisional ability score and those that begin at a provisional ability estimate set equal to the "true" ability score? and 2) Is there a difference between the mean concordance of computer adaptive naming test ability estimate scores and "true" ability estimate scores for computer adaptive naming tests that start at an average ability estimate and computer adaptive naming tests that begin at a provisional ability estimate set equal to the "true" ability score?

Simulated trials of fixed-length CAT PNTs using the *catIrt* package in R Studio were run under two conditions: the first under the current practice of starting the CAT at a provisional ability score equivalent to the mean, and the second under the limiting case, in which the CAT started at a provisional ability score equivalent to the simulees' "true" ability score. In theory, equating initial ability level estimates to the "true" ability level of the simulees would optimize the start point of the test and investigate the ideal conditions.

Despite assigning simulees their “true” ability estimates as their initial ability level estimate in Condition 2, according to the results of these simulations, there was a negligible improvement in the performance of the test between the two conditions. Condition 2, starting simulees at a provisional ability level equivalent to their “true” level, and therefore with a maximally informative first item, presented no real advantages for test precision over the current practice of starting a test taker at the mean ability level. Both conditions resulted in strong correlations for the final CAT ability level scores with the simulees’ “true” ability levels. Moreover, these strong correlations are not significantly different from each other. The CAT under Condition 1 was comparably precise in determining ability level. This suggests there might not be practical advantages to creating CAT software that would allow clinicians to adjust the provisional ability level estimate based on their clinical impressions or some alternative factor (e.g., a score from another standardized measure) without considering additional information. With no significant improvement occurring in the limiting case, ergo ideal conditions, we can assume there will be no advantage to this option in a realistic clinical setting where the client or patients’ “true” ability levels cannot be known and subjective impressions about a client or patients’ ability level will vary from clinician to clinician.

One potential reason the results demonstrate no significant improvement could be because the current CAT PNT is already performing at a high level of efficiency given the current item bank. As stated above, the CAT under Condition 1 demonstrated a strong correlation between the calculated ability level score and the simulees' "true" ability levels. The current practice of starting test-takers with an ability level estimate is resulting in overall good performance from the CAT. Such high precision for the CAT in this regard means there could be little space for improvement by adding the additional information of a starting point closer to the individuals' "true" score.

Although the non-significant results of this study do not support changing the current CAT set-up to include tailored provisional ability level scores, this study highlights the utility of simulations in making further advancements in this area. Simulations of CATs present several advantages over collecting such data from live participants. The first benefit being that with simulations, information that could never be known in a realistic setting, such as a latent trait, can be treated as known. In this study, simulations allowed the simulees' "true" ability scores to be known, and therefore a better understanding of the precision of the test could be obtained. Comparing this to testing the precision of the CAT on live participants, for whom we can never know their "true" ability, it is possible to know how accurately the test estimated the simulee's ability level.

Another advantage to simulations is that there is no test, re-test bias. Each condition was run in isolation for each simulee. Responses to items were predetermined, therefore the results of the CATs in each condition can be analyzed on their own. If this study had been conducted with live subjects, each subject would have had to have taken the CAT twice (once under each condition) which could alter the selection of informative items available for a participant or the results of the second condition could be impacted by a learning effect if the participant had seen an item in the previous condition or had become more comfortable with the format of the test over time. These factors that could impact the results are non-existent for simulations.

In addition to these positive implications, simulations offer a practical advantage. While designing and running simulations to create data for 1000 simulees was accomplished in a matter of weeks, obtaining this same data from live subjects would have been a much more cumbersome task. Collecting this same dataset from live people with aphasia presents numerous challenges. First, recruiting participants and administering two CAT PNTs to 1000 individuals would take an extensive amount of time. Additionally, conducting this research with live participants would have used excess resources, and therefore would have cost far more. Simulating CAT PNTs costs relatively little and could be completed in much less time, preserving resources.

Limitations

Although simulations present these advantages, there are limitations. Studies based in computer simulations are limited in external validity because they often represent the most controlled, ideal environments for data collection, conditions which are hardly ever reproduced in realistic settings. The results from studies based in computer simulations would have to be repeated with live participants to better be able to understand how the results transfer to realistic settings and conditions. While this appears to be a significant disadvantage, the ease with which simulations can be run means they can be an important step in prioritizing which investigations are worth pursuing with live participants in a clinical setting, thereby preserving resources and time. The results of this simulation study, although statistically insignificant, have provided important information that will guide future research in improving efficiency and precision of computer adaptive naming tests without requiring the use of extensive resources and excess time.

Future Directions

Although tailored provisional ability scores alone did not have a significant effect in this study, future research can look into additional ways to optimize CAT efficiency by finding ways to provide the CAT with additional information about the test-taker at the start. In this study, only one additional piece of information – an estimate of what the client or patient's "true" ability level may be – was given to the test; however, this is limited information. The CAT was provided with a tailored provisional ability score that estimates the test-taker's "true" ability level, but with a wide range of uncertainty around it. Next steps for optimizing CAT efficiency can investigate whether providing the CAT with a tailored start point and a confidence interval around that start point may increase the precision or efficiency of the test. In a clinical setting this could be done through the use of additional aphasia measures, such as the WAB-R, to inform the CAT PNT. The WAB-R has a naming subtest for which an estimated ability and 95% confidence interval can be inserted into the starting information of the CAT to better inform how the algorithm updates ability estimates and selects test items.

References

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

Embretson, S. E., & Yang, X. (2006). Item response theory. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research*. Lawrence & Erlbaum Associates.

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

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

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

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

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

Kaplan, E., Goodlgass, H., & Weintraub, S. (2001). *Boston Naming Test* (2nd ed.). Lippincott Williams & Wilkins.

Lee, I. A., & Preacher, K. J. (2013, September). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]. Available from http://quantpsy.org/corrtest/corrtest2.htm

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.

Mills, C. N., & Stocking, M. L. (1996). Practical Issues in Large-Scale Computerized Adaptive Testing. *Applied Measurement in Education*, *9*(4), 287. https://doi.org/10.1207/s15324818ame0904_1

Nydik, S.W. (2014). catIrt: An R package for simulating IRT-based computerized adaptive tests. R package version 0.5-0. http://CRAN.R-project.org/package=catIrt

Roach, A., Schwartz, M., Martin, N., Grewal, R., & Brecher, A. (1996). The Philadelphia Naming Test: Scoring and rationale. *Clin. Aphasiol.*, *24*, 121–133.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251.

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)

Appendix A

catIrt Simulation Code

```
theta <- rnorm(n=1000, mean=.1, sd=1.44)

params <- read.csv

b.params <- data.matrix(params, rownames.force = NA)

b.resp <- simIrt(theta = theta, params = b.params, mod = "brm")$resp

catStart1 <- list(init.theta = 0, n.start = 5,

              select = "UW-FI", at = "theta",

              n.select = 1, it.range = NULL,

              score = "EAP", range = c(-4, 4),

              step.size = 3, leave.after.MLE = FALSE)

catMiddle1 <- list(select = "UW-FI", at = "theta",

               n.select = 1, it.range = NULL,

               score = "MLE", range = c(-6, 6),

               expos = "none")

catTerm1 <- list(term = "fixed", n.min = 10, n.max = 30)

cat1 <- catIrt(params = b.params, mod = "brm",

            resp = b.resp, theta = theta,

            catStart = catStart1,

            catMiddle = catMiddle1,

            catTerm = catTerm1)

summary(cat1, group = TRUE, ids = "none")
```

```
catStart2 <- list(init.theta = theta, n.start = 5,
                  select = "UW-FI", at = "theta",
                  n.select = 1, it.range = NULL,
                  score = "EAP", range = c(-4, 4),
                  step.size = 3, leave.after.MLE = FALSE)
catMiddle1 <- list(select = "UW-FI", at = "theta",
                   n.select = 1, it.range = NULL,
                   score = "MLE", range = c(-6, 6),
                   expos = "none")
catTerm1 <- list(term = "fixed", n.min = 10, n.max = 30)
cat2 <- catIrt(params = b.params, mod = "brm",
               resp = b.resp, theta = theta,
               catStart = catStart2,
               catMiddle = catMiddle1,
               catTerm = catTerm1)
summary(cat2, group = TRUE, ids = "none")
```