7-14-2021

# Computer Adaptive Testing for the Assessment of Anomia Severity

Gerasimos Fergadiotis
*Portland State University*, gf3@pdx.edu

Marianne Casilio
*Vanderbilt University*

William D. Hula
*University of Pittsburgh*

Alexander Swiderski
*University of Pittsburgh*



## Citation Details

Published as: 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..

# Computer Adaptive Testing for the Assessment of Anomia Severity

Gerasimos Fergadiotis, Ph.D., CCC-SLP, Speech and Hearing Sciences, Portland State University, gfergadiotis@pdx.edu

Marianne Casilio, M.S., CCC-SLP, Hearing and Speech Sciences, Vanderbilt University, marianne.e.casilio@vanderbilt.edu

William D. Hula, Ph.D., CCC-SLP, VA Pittsburgh Healthcare System and University of Pittsburgh, William.Hula@va.gov

Alexander Swiderski, M.A., CCC-SLP, Department of Communication Sciences and Disorders, University of Pittsburgh, aswiderski@pitt.edu

Contact Information:
Gerasimos Fergadiotis, Ph.D., CCC-SLP
Associate Professor
527 SW Hall Street, RM430G
University Center Building
Portland, OR 97201
Phone: 503.725.2217
Email: gfergadiotis@pdx.edu

## Abstract

Anomia assessment is a fundamental component of clinical practice and research inquiries involving individuals with aphasia; and, confrontation naming tasks are among the most commonly used tools for quantifying anomia severity. While currently available confrontation naming tests possess many ideal properties, they are ultimately limited by the overarching psychometric framework they were developed within. Here, we discuss the challenges inherent to confrontation naming tests and present a modern alternative to test development called item response theory (IRT). Key concepts of IRT approaches are reviewed in relation to their relevance to aphasiology, highlighting the ability of IRT to create flexible and efficient tests that yield precise measurements of anomia severity. Empirical evidence from our research group on the application of IRT methods to a commonly used confrontation naming test is discussed, along with future avenues for test development.



**Keywords:** Assessment, Anomia, Item Response Theory, Confrontation Naming Tests


**Conflict of Interest Statement**
The authors have no relevant conflicts of interest.


**Funding Statement**
The authors gratefully acknowledge the support of NIH/NIDCD awards # R03DC014556 (PI: Fergadiotis) and R01DC018813 (PIs: Hula & Fergadiotis).

Anomia, or word-finding difficulty, is one of the cardinal features of aphasia[1], a disorder affecting 2.5 to 4 million people in the US[2]. Anomia treatment has received substantial research attention because of its ubiquitous and persistent nature. However, one major challenge in this area is the lack of psychometrically robust metrics to support treatment efficacy research, and to provide a unified framework for rehabilitation, reimbursement, and policy decisions. Addressing this challenge is the overarching goal of our research group. To this end, we have engineered a computer-adaptive system for assessing the severity of anomic deficits within a modern psychometric framework. In the following sections, we review the relevance and significance of anomia assessment and then outline two primary psychometric approaches to test development, highlighting the empirical evidence and implementation advantages in support of the modern framework we utilized. In the final section, we present a summary of persisting challenges and future areas for development.

## Naming Assessment is a Foundational Component of Aphasia Research and Clinical Practice

In aphasiology, professionals evaluate breakdowns in word access and retrieval using confrontation naming tests, where patients are presented with pictures of objects or actions and are asked to name them. Given that anomia is a primary diagnostic feature of aphasia, confrontation naming tests are routinely used for assessment across clinical and research settings. Confrontation naming tasks are integral to all well-established aphasia batteries, including the Boston Diagnostic Aphasia Examination–3rd Edition[3], Western Aphasia Battery–R[4], and the Comprehensive Aphasia Test[5]. Confrontation naming tasks also continue to be included in more recently developed tools, such as the Quick Aphasia Battery[6] and the Northwestern Assessment of Verbs and Sentences[7], and they are typically incorporated into bedside screening tools[8]. In addition, there are several standalone confrontation naming tests that are used extensively for

evaluation purposes and include the Boston Naming Test[9], the Philadelphia Naming Test[10], the Object and Action Naming Test[11], and the Test of Adolescent/Adult Word Finding[12].

The importance of standalone confrontation naming tests is reflected in the breadth and depth of research investigations that have used them. A variety of studies have used such tests to investigate the cognitive system underlying word production (e.g.,[13]), explore brain-behavior relationships (e.g.,[14]), quantify treatment efficacy with regard to both behavioral (e.g.,[15]) and neural[16] outcomes, and conduct meta-analytic research[17]. Further, such tests are among the most commonly used tools for tracking performance in randomized control trials in aphasiology[18,19].

The widespread use of confrontation naming tests can be attributed, at least in part, to their strong psychometric properties. First, confrontation naming tests possess a relatively straightforward administration and accuracy-based scoring procedure, given that target words are specified *a priori*[20,21]. Additionally, there is evidence in support of their construct validity. For example, the ability to access and retrieve words during confrontation picture naming depends on well-described cognitive processes that are also critical for successful language production in less constrained contexts (e.g.,[22]). Further, confrontation naming tests possess strong intercorrelations despite differing scoring procedures or specific stimuli (e.g.,[23]). Finally, confrontation naming tests have been shown to be strong predictors of both overall aphasia severity (e.g.,[24,25]) and discourse informativeness[23].

One aforementioned confrontation naming test—the Philadelphia Naming Test[10] (PNT)—is particularly prominent in the applied and theoretical research literature, as it was developed as part of a larger set of studies investigating models of lexical retrieval in normal processing and aphasia (e.g.,[26,27]). Though it shares robust psychometric properties with other confrontation naming tests discussed earlier, the PNT is weakly correlated with demographic

variables, including age, gender, and race[25] which allows for unbiased inference regarding a patient's naming ability skills. In addition, the PNT is an ideal tool for further measurement optimization because of the availability of a large archival dataset with item-level responses from approximately 300 patients with aphasia[28]. Such datasets are critical for the deployment of sophisticated, data-intensive techniques commonly used in modern test and scale development.

## Improving Efficiency and Utility Using Item Response Theory

While confrontation naming tests possess many advantageous properties, they were developed within a specific psychometric framework that precludes precise and efficient quantification of anomia severity. A more modern framework, which has been used by our group to develop a computer-adaptive system for anomia assessment, addresses several limitations and, as will be discussed in the following sections, has tangible benefits for both practicing clinicians and clinical researchers.

### Psychometrics at the level of the item vs. test

Most currently available anomia tests are developed under what is known as *classical test theory*[29,30] (CTT), a psychometric framework that was introduced in the field of psychological measurement in the early 1900s by Spearman[31–33]. At the heart of CTT lies the concept of the *observed total score* of a test, which is defined as the sum of (i) the *true total score* a patient would have obtained in the complete absence of measurement error and (ii) *measurement error*. This definition serves as the starting point for formalizing and quantifying the psychometric properties that clinicians are trained to recognize, such as test-retest and internal consistency reliability. However, it is critical to recognize—and perhaps begin to challenge the utility of—the fact that within the CTT framework, tests are conceptualized in a gestalt fashion. That is, under CTT, the emphasis is on the test as a whole, and the goal is to establish the psychometric

properties of the particular ensemble of items specific to that test. Given the need for whole-test validation, clinicians and researchers are faced with a plethora of confrontation naming tests, many of which contain redundant items with distinct psychometric properties from differing sources and forms of evidence. While existing tests would ideally be modified or enhanced to address item redundancy or reconcile contradicting forms of evidence, doing so would require additional whole-test validation for the new ensemble of items, ultimately resulting in a different set of psychometric properties. As such, the rigidity of the CTT approach serves as a barrier to both improving upon existing tests and generating new ones.

Further, scores from different published tests developed within a CTT framework cannot be directly compared. The difficulty stems from at least two related factors. First, tests include items that vary in number and difficulty, thus direct comparisons across tests using total number or percent-correct scores are not meaningful[34]. For example, 20% accuracy on a test with difficult words (e.g., stethoscope) may indicate more naming ability than 30% accuracy on a test with easier words (e.g., cat). Second, norm-referenced scores from standardized confrontation naming tests depend on the particular standardization sample of each test[34]. Therefore, *z*-scores or percentile ranks from different tests cannot be directly compared without assuming equivalent samples. This inability to capitalize on a wide range of pictures and items from different anomia assessments has direct consequences for both clinical practice and research inquiries. Regarding the former, this lack of test comparability: (i) disrupts the flow of diagnostic information across clinical settings; (ii) increases the likelihood that patients are re-evaluated unnecessarily depending on the availability of tools at each new setting, thus contributing to increased cost and testing burden; and (iii) hinders direct comparisons of performance across the care continuum. In reference to the latter, the use of different confrontation naming tests across studies (i) makes it

difficult to synthesize and evaluate treatment effects across studies that use different tools and (ii) limits the use meta-analytic studies to draw generalizable conclusions.

A modern alternative to CTT is item response theory[35] (IRT). Unlike CTT, which necessarily emphasizes test-level characteristics, IRT centers on item-level characteristics. In its simplest form, an IRT model seeks to explain the *probability of a correct response on a given item* on a test as a function of the two quantities: (i) the *item's difficulty* and (ii) the *patient's ability level*. Here, *item difficulty* can be understood to reflect the relative ease or challenge of producing a correct response on a given item. With regard to the latter, *ability* is operationalized as the degree to which an individual possesses a given skill or attribute. For our purposes, *ability* and any numerical estimates associated with it will be used to refer to the degree of naming impairment, or anomia severity. Figure 1 shows in graphical terms the model for three items from the PNT. In addition, we have developed an interactive web-based application to complement this manuscript (**https://aswiderski.shinyapps.io/IRTapp/**). Interested professionals can use the online tool to further explore key IRT concepts. Returning to Figure 1, if an item is calibrated using IRT methods, we obtain a sigmoid curve that allows us to predict the relative difficulty in terms of probability of correct response for patients with different levels of ability. In IRT applications, ability estimates and item difficulties are typically expressed on a logit scale that ranges from -4 to 4, but here have been re-expressed to *t*-scores, a metric more familiar to clinicians. Returning to the example at hand, a patient with moderate anomia (*t*-score = 50) has greater than 90% chance of naming the item "cat" but only a 10% chance of naming correctly "microscope". Considered in the context of confrontation naming tests, this model captures a phenomenon well-known to clinicians—the same item might be too easy for

individuals with a relatively mild anomia and too hard for individuals with a relatively severe anomia.

Figure 1 about here

Further, an important advantage of IRT-based ability estimates is that the resultant IRT scores behave similarly to interval-type data, where the distance between each unit of measurement is equal[36]; see Figure 2 for a conceptual representation). In contrast, raw-percent correct scores, which are commonly used by clinicians and applied researchers, do not behave in this manner. For example, a raw score improvement from 55 to 60 items correct (out of 100) has very different implications than a change from 95 to100 items correct on the same test. Clinicians intuitively know that the former is much easier to observe than the latter, despite both representing the same numerical difference. The reason for this disconnect is that, when using raw percent-correct scores, the magnitude of change depends on its location along the scale. In other words, a change of 5% for a patient who initially scored 10% on a test signals more improvement than a patient who achieved the same change but initially scored 50%.

Figure 2 about here

The first set of analyses necessary to developing an IRT-based test involve estimating the properties of the items. For example, prior to using an IRT-based test, item difficulties have to be estimated. Statistically, this is done by calibrating test items using data from a large sample of participants and evaluating how well those parameters explain the observed test performance. Then, test-level psychometric properties can be computed. As we will see in the following sections, one such property—the precision of a test (an IRT-based analogue to the CTT concept of reliability)—is estimated by combining the precision provided by each item on the test across different levels of impairment to arrive at a test's precision function. Conceptually, once items

are calibrated using IRT approaches, then items can be used as interlocking building blocks to develop tools tailored to testing applications, such as developing a screener or to detecting a particular effect size, which offers a significant advantage over CTT techniques.

Recently, we used IRT methods to derive difficulty estimates for each of the items on the PNT[37]. The goal was to calibrate a psychometrically robust test as a means of measuring anomia severity in a more precise and flexible manner. To this end, we obtained archival data from the Moss Aphasia Psycholinguistics Project Database [28], which consisted of complete administrations of the PNT from 251 individuals with aphasia. First, we assessed whether item-level accuracy on the PNT administrations selected met the assumptions necessary for IRT approaches. After confirming that these assumptions were adequately met, the items were calibrated using a 1-parameter logistic IRT model. Figure 3 shows the distribution of item difficulties and ability estimates for the sample on the left and right respectively. This calibration of the PNT using IRT methods yielded valuable psychometric information for all 175 items of the test, which can then be used in combination with other IRT-based metrics to draw inferences about an individual's or group of individuals' ability levels.

Figure 3 About Here

## Tests with valid measurement error estimates

One of the major advantages of IRT is that once a test is calibrated, we can derive valid measures of uncertainty around a patient's estimated ability. A central concept in IRT is *information*, which represents the degree to which a response can help us refine an ability estimate and reduce the uncertainty around it. In simple IRT models, the amount of information an item contains is dependent on its difficulty, with easier items being maximally informative for patients with correspondingly less ability and harder items being maximally informative for

patients with correspondingly more ability. This relationship between information and difficulty statistically formalizes what clinicians are already familiar with—items that are too easy or too difficult to name for a given patient do not contribute much to our understanding of the patient's anomia severity level. Figure 4 shows the information functions—a graphical representation of the amount of information an item contains across the ability continuum—for two items from the PNT. An item like "key" is maximally informative when administered to individuals with severe anomia but contributes progressively less information when administered to patients with relatively more mild anomia. Importantly, information is summed across items to create the test information function (solid line in Figure 4), which represents the certainty of an estimate provided by a test at each ability level. Importantly, the regions of ability for which this example test is maximally informative for is not uniform. Rather, information peaks where the majority of items are concentrated. Further, the inclusion of additional items (Panel B) enhances the amount of information the test contains, which is reflected in the height of the peak, of the test information function.

Figure 4 About Here

While clinicians may only be intuitively aware of *information* and its value in assessment, the notion of a standard error of measurement (SEM) and its role in constructing 95% confidence intervals (*CI*s) is a well-established concept inherent to most standardized behavioral tests. In IRT, a test's SEM is inversely related to information and conditional on level of ability. The SEM curve is also shown in Figure 5. Notice that the SEM is not constant across the ability continuum—the least measurement error is expected where the total test information function peaks with measurement error increasing progressively in regions that lack items of corresponding difficulty. Increasing the number of items in a test, as seen in Panel B of Figure 4,

lowers the SEM curve overall, resulting in increased confidence in the ability estimates. Consequently, the conditional nature of measurement error has important implications for generating confidence intervals around a given ability estimate, as the degree of confidence will vary depending on where an individual lies on the ability continuum. It follows that closer targeting of item difficulty to ability leads to better measurement precision[36]. Use of an IRT model, as in our work, provides SEMs that vary as a function of ability estimate and thus more precisely capture an individual's anomia severity. The solid curve in Figure 4 shows how the SEM of the PNT peaks for patients with aphasia at the extremes of the ability distribution and is almost three times lower for those with in the middle of the ability distribution[38].

Figure 5 about here

Assuming an average constant measurement error (dashed line in Figure 5), as is necessary in CTT approaches, leads to invalidly narrow CIs at the extremes of the ability continuum and overly wide CIs for those in the middle. This assumption of constant measurement error has serious implications for the assessment of change, as any CI and associated probabilities derived about the change score may be distorted. Specifically, the confidence intervals for mildly and severely impaired patients may be misleadingly narrow, leading to an inflated type I error rate. On the other hand, the CIs for moderately impaired patients may be overestimated, leading to decreased power to detect real change and an increased type II error rate.

## Developing more efficient and flexible tests

Perhaps the most advantageous aspect of IRT modeling is the ability generate item banks and computer-adaptive tools from a calibrated test[39]. An item bank is a set of items that measure a common trait and have been calibrated on a common scale. An IRT computer-adaptive test

(CAT) is an algorithm that selects and administers items from the bank that maximize information at a particular ability level. To this end, an iterative algorithm estimates a patient's ability and associated SEM after each item has been administered. Then, given the calibrated item difficulty, it decides which item to administer next to maximize information at the new ability estimate. This process is repeated until a specified number of items has been administered or when a specified level of precision (i.e., the SEM of the ability estimate) is reached. The technique quickly converges into a sequence of items bracketing the test taker's ability, thus efficiently shortening the test while maximizing its precision (Figure 6).

Figure 6 about here

Adaptive testing in an IRT framework fills a notable and critical gap in currently available assessment tools, as tests developed using CTT ultimately require the administration of "off-target" items, which increases testing burden on patients. As alluded to earlier, patients with mild deficits may be presented with items that are too easy whereas patients with severe deficits might be asked to name items that are too challenging. This situation is suboptimal because it contributes to fatigue, frustration, and a perception by the patient that the test might not be appropriate or relevant. As a result, a patient's performance might be contaminated and patient-clinician rapport may be negatively affected. Administration of items whose difficulty is mismatched with a patient's ability level also increases burden on the clinicians for at least two interrelated reasons. First, most clinicians face increasingly high productivity demands. Thus, reducing occupational strain is vital to minimizing adverse effects both in terms of quality of services and clinician well-being[40]. Second, to the extent that the assessment process can be viewed through a neuropsychosocial lens, it is incumbent on clinicians to investigate factors beyond the impairment level and to collect information on how a disorder affects daily activities

and participation in social aspects of life[41], which may not be feasible if available assessment tools are inefficient and provide insufficient information about baseline performance.

In the context of anomia assessments, static or non-adaptive short forms have been developed to address this limitation (e.g.,[25]). However, because these tests were developed using CTT practices, they contain a small number of items *a priori* that are designed to capture a wide range of ability. Consequently, they often lack the precision of an IRT-calibrated test, particularly for individuals at the extremes of the anomia continuum. Basal and ceiling criterions have also been implemented to improve testing efficiency and better tailor the item administration to an individual's ability level. Such criteria are necessitated on the assumption that items become progressively more difficult as a function of their presentation order, yet studies suggest that successive items do not necessarily exhibit a uniform increase in difficulty[42]. Neither of these alternatives, however, address the need for clinicians to tailor anomia tests for specific testing purposes. For example, a clinician may use an anomia test to screen for frank impairment in certain situations while also requiring an anomia test to gain a precise estimate of ability for benchmarking performance prior to and following a course of therapy in another. Regardless of the availability of a short form or testing criteria, no single anomia assessment developed under a CTT framework can address these varying situations, leaving clinicians in a similarly difficult position as before—administer multiple tests targeting the same ability or opt for a non-standard administration of a published or house-made test, both of which preclude precise estimation of anomia and valid comparison of performance across tests or administrations.

The IRT-calibrated PNT offers a viable solution to the aforementioned clinical challenges. Recognizing the need to capitalize on the advantages of IRT modeling, we first used

the CAT algorithm to generate a 30-item short form and a variable-length alternate form of the PNT via simulations based on archival data[43]. While such simulations are useful in establishing proof of concept and deriving variables necessary for further CAT development, they ignore a number of factors associated with live administrations that could affect the agreement between the scores of the two forms[44,45]. Consequently, CAT-generated forms based on simulations may fail to (i) capture common sources of measurement error, such as inaccurate online scoring on the part of clinicians or error stemming from inherent fluctuations in patient response patterns (e.g., attention lapses, fatigue), and (ii) provide only an idealized, upper bound of agreement between the CAT-generated forms and the full-length IRT-calibrated IRT.

In an effort to gain a more comprehensive view of CAT-generated forms of the PNT, we conducted a prospective study with 47 participants with aphasia[38]. Specifically, the study aimed to verify and further quantify the agreement of a 30-item CAT test, which we called PNT-CAT, with the full-length IRT-calibrated PNT, which was referred to as the full PNT. We assessed the absolute agreement (e.g., bias) and relative agreement (i.e., correlation) of the ability estimates generated by the two PNT versions. Overall, there was no evidence of significant bias and the relative agreement between the full PNT and the PNT-CAT was strong, as indicated by high correlation ($r$ = .95, 95% CI [.92, .97]). Notably, when we analyzed the average administration time of the two versions of the test, we found that, on average, the CAT version of the PNT required 8.26 min ($SD$ = 3.62) to administer whereas the full PNT was associated with substantially longer administration times (m = 31.13 min, $SD$ = 11.72). We concluded that the strong agreement between the full PNT and the PNT-CAT, as well as the relatively robust ability estimates, suggested that the 30-item CAT-generated form of the PNT was a suitable tool for (i) measuring anomia at the group level in research contexts and (ii) clinical assessment of anomia

of individual patients. In particular, the combination of the PNT-CAT's reduced administration time and increased confidence (i.e., narrower confidence intervals) around a patient's estimated ability level—the latter of which is the result of improved item targeting—translates to more efficient and accurate anomia testing overall, thereby freeing up resources that professionals can utilize for other clinically important tasks (e.g., evaluation of concomitant disorders, provision of patient education and counseling).

We then quantified the agreement between the 30-item PNT-CAT and a variable-length CAT-generated form of the PNT, called CAT-VL, in a follow-up study[46]. The CAT-VL was configured to avoid presenting overlapping items from the 30-item PNT-CAT administration for a given patient. In addition, for the CAT-VL, a stopping rule was set as precision equal to or lower than that obtained by the 30-item PNT-CAT. The two forms were administered to 25 patients with aphasia during two testing sessions separated by an interval of two weeks. Ability estimates from the two test versions correlated highly ($r$ = .89) and the estimated means and standard deviations were not meaningfully different from one another. Given these results, we concluded that both CAT-generated forms of the PNT may be productively used to assess anomia in clinical and research contexts. A copy of the software is available from the first author free of charge.

## Challenges & Future Work: A universal metric for anomia severity

While the CAT-generated forms of the PNT represent a substantial improvement over currently available anomia assessment tools, our work has yet to address the challenge of direct comparison across different published tests developed within a CTT framework. IRT approaches, however, are well-suited to address this problem via linking and equating procedures of existing tests[47] and by expanding the currently available item bank of IRT-calibrated PNT.

Specifically, IRT provides a technical path for adjusting differences in difficulty to ensure that all patients receive an interpretable ability estimate on a stable scale regardless of whether the estimate was derived based on the PNT or another anomia assessment tool (e.g., BNT, the naming subtest of the Comprehensive Aphasia Test).

Further, expanding the item bank of the IRT-calibrated PNT with hundreds of additional items from other commonly used anomia tests would ultimately result in the creation of a universal test for quantifying anomia severity, which may have significant implications for both clinical and research practice. For example, if the expanded IRT-calibrated item bank is integrated into a computer-adaptive framework, item sets could be more precisely tailored to patients of differing anomia severity with resulting scores rendered on a common scale with both norm-referenced and criterion-referenced interpretation. Further, a large IRT-calibrated item bank would enable the generation of additional short test forms with non-overlapping content that yield scores on the same metric while minimizing test-retest effects, providing increased experimental control for the applied researcher and enhanced care across the healthcare continuum for the treating clinician. Finally, this expanded bank would result in ability-dependent estimates of score precision at each time point, which are critical when evaluating treatment effects. These tailored estimates of score precision can be used to test statistically[48] a number of different questions such whether the initial or final status differs across two participants; after how many sessions participants start demonstrating a disassociation in performance compared to a matched control; or who is responding to intervention.

Overall, in clinical applications, a universal metric of anomia implemented via IRT-based CAT would (i) support accurate and meaningful comparisons of assessment results acquired with different testing instruments, (ii) facilitate the flow of diagnostic information

across clinical settings along the continuum of care, and (iii) minimize cost and testing burden by using adaptive short forms that minimize the loss of measurement precision. In research settings, it would allow us to (i) synthesize and evaluate treatment effects across studies that use different tools, (ii) greatly enhance our ability to conduct meta-analytic studies, and (iii) evaluate new and existing anomia treatments in a psychometrically robust manner.

## Conclusion

The value of confrontation naming tests in quantifying anomia severity is substantial, yet there remains a need to optimize currently available tools. The IRT psychometric approach presented here has the potential to significantly improve the efficiency and precision of anomia assessment with direct implications for clinical practice and research inquiries. Evidence from our research group regarding the development of a computer-adaptive system of the PNT lends credence to the advantages of the IRT framework and supports its potential as a psychometrically superior tool for anomia assessment, though incorporation of additional items from other commonly used confrontation naming tests is needed in order to create a robust, comprehensive, and universal metric for measuring naming performance in patients with aphasia.

## References


1. Raymer AM, Rothi LJG. Impairments of Word Comprehension and Production. In: Chapey R, ed. *Language Intervention Strategies in Aphasia and Related Neurogenic Communication Disorders*. Lippincott Williams & Wilkins; 2001:606-625.

2. Simmons-Mackie N. *Aphasia in North America*. Aphasia Access; 2018.

3. Goodglass H, Kaplan E, Barresi B. *Boston Diagnostic Aphasia Examination*. 3rd ed. Lippincott Williams & Wilkins; 2001.

4. Kertesz A. *Western Aphasia Battery – R*. Grune & Stratton; 2007.

5. Swinburn K, Porter G, Howard D. *Comprehensive Aphasia Test*. Psychology Press; 2004.

6. Wilson SM, Eriksson DK, Schneck SM, Lucanie JM. A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function. *PLOS ONE*. 2018;13(2):e0192773. doi:10.1371/journal.pone.0192773

7. Cho-Reyes S, Thompson CK. Verb and sentence production and comprehension in aphasia: Northwestern Assessment of Verbs and Sentences (NAVS). *Aphasiology*. 2012;26(10):1250-1277. doi:10.1080/02687038.2012.693584

8. El Hachioui H, Visch-Brink EG, de Lau LML, et al. Screening tests for aphasia in patients with stroke: a systematic review. *J Neurol*. 2017;264(2):211-220. doi:10.1007/s00415-016-8170-8

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

10. Roach A, Schwartz MF, Martin N, Grewal RS, Brecher A. The Philadelphia naming test: Scoring and rationale. *Clin Aphasiology*. 1996;24:121-133.

11. Druks DJ, Masterson J. *An Object and Action Naming Battery*. Psychology Press; 2000.

12. German DJ. *Test of Adolescent/Adult Word Finding*. DLM; 1990.

13. Nozari N, Dell GS. How damaged brains repeat words: A computational approach. *Brain Lang*. 2013;126(3):327-337. doi:10.1016/j.bandl.2013.07.005

14. Dell GS, Schwartz MF, Nozari N, Faseyitan O, Branch Coslett H. Voxel-based lesion-parameter mapping: Identifying the neural correlates of a computational model of word production. *Cognition*. 2013;128(3):380-396. doi:10.1016/j.cognition.2013.05.007

15. Kendall DL. Hunting Pompon, R., Brookshire, CE, Minkina, I., & Bislick, L.(2013). An analysis of aphasic naming errors as an indicator of improved linguistic processing following phono-motor treatment. *Am J Speech Lang Pathol*. 22:S240-S249.

16. Fridriksson J, Richardson JD, Fillmore P, Cai B. Left hemisphere plasticity and aphasia recovery. *NeuroImage*. 2012;60(2):854-863. doi:10.1016/j.neuroimage.2011.12.057

17. Quique YM, Evans WS, Walsh-Dickey M. Acquisition and Generalization Responses in Aphasia Naming Treatment: A Meta-Analysis of Semantic Feature Analysis Outcomes. *Am J Speech Lang Pathol*. 2018;0(0). doi:10.1044/2018_AJSLP-17-0155

18. Brady MC, Kelly H, Godwin J, Enderby P, Campbell P. Speech and language therapy for aphasia following stroke. *Cochrane Database Syst Rev*. 2016;(6). doi:10.1002/14651858.CD000425.pub4

19. Kiran S, Cherney LR, Kagan A, et al. Aphasia assessments: a survey of clinical and research settings. *Aphasiology*. 2018;32(sup1):47-49. doi:10.1080/02687038.2018.1487923

20. Herbert R, Hickin J, Howard D, Osborne F, Best W. Do picture-naming tests provide a valid assessment of lexical retrieval in conversation in aphasia? *Aphasiology*. 2008;22(2):184-203. doi:10.1080/02687030701262613

21. Mayer J, Murray L. Functional measures of naming in aphasia: Word retrieval in confrontation naming versus connected speech. *Aphasiology*. 2003;17(5):481-497. doi:10.1080/02687030344000148

22. Dell GS. A spreading-activation theory of retrieval in sentence production. *Psychol Rev*. 1986;93(3):283-321.

23. Fergadiotis G, Kapantzoglou M, Kintz S, Wright HH. Modeling confrontation naming and discourse informativeness using structural equation modeling. *Aphasiology*. 2019;33(5):544-560. doi:10.1080/02687038.2018.1482404

24. Schuell H, Jenkins JJ, Jimenez-Pabon E. *Aphasia in Adults*. Harper & Row; 1964.

25. Walker GM, Schwartz MF. Short-form Philadelphia naming test: Rationale and empirical evaluation. *Am J Speech Lang Pathol*. Published online May 2, 2012:S140-S153. doi:10.1044/1058-0360(2012/11-0089)

26. Dell GS, Schwartz MF, Martin N, Saffran EM, Gagnon DA. Lexical access in aphasic and nonaphasic speakers. *Psychol Rev*. 1997;104(4):801-838. doi:http://dx.doi.org/10.1037/0033-295X.104.4.801

27. Dell GS, O'Seaghdha PG. Stages of lexical access in language production. *Cognition*. 1992;42(1-3):287-314. doi:10.1016/0010-0277(92)90046-K

28. Mirman D, Strauss TJ, Brecher A, et al. A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. *Cogn Neuropsychol*. 2010;27(6):495-504. doi:10.1080/02643294.2011.574112

29. Traub RE. Classical Test Theory in Historical Perspective. *Educ Meas Issues Pract*. 1997;16(4):8-14. doi:https://doi.org/10.1111/j.1745-3992.1997.tb00603.x

30. Willse JT. Classical Test Theory. In: Salkind NJ, ed. *Encyclopedia of Research Design*. SAGE Publications, Inc.; 2010:150-153. doi:10.4135/9781412961288.n49

31. Spearman C. Correlations of sums or differences. *Br J Psychol*. 1913;5(4):417-426.

32. Spearman C. The proof and measurement of association between two things. *Am J Psychol*. 1904;15(1):72-101. doi:10.2307/1412159

33. Spearman C. Demonstration of formulae for true measurement of correlation. *Am J Psychol*. 1907;18(2):161-169. doi:10.2307/1412408

34. Embretson SE. The new rules of measurement. *Psychol Assess*. 1996;8(4):341-349.

35. Ayala RJ de. *Theory and Practice of Item Response Theory*. Guilford Publications; 2009.

36. Embretson SE, Reise SP. *Item Response Theory for Psychologists*. L. Erlbaum Associates; 2000.

37. Fergadiotis G, Kellough S, Hula WD. Item response theory modeling of the Philadelphia Naming Test. *J Speech Lang Hear Res*. 2015;58:865-877. doi:10.1044/2015_JSLHR-L-14-0249

38. Fergadiotis G, Hula WD, Swiderski AM, Lei C-M, Kellough S. Enhancing the Efficiency of Confrontation Naming Assessment for Aphasia Using Computer Adaptive Testing. *J Speech Lang Hear Res*. 2019;62(6):1724-1738. doi:10.1044/2018_JSLHR-L-18-0344

39. Wainer H, Dorans NJ, Flaugher R, Green BF, Mislevy RJ. *Computerized Adaptive Testing: A Primer*. Routledge; 2000.

40. Dollard M. Occupational Stress in the Service Professions. In: Dollard M, Winefield AH, Winefield HR, eds. *Occupational Stress in the Service Professions.* Taylor & Francis; 2003:1-42.

41. Chapey R, Duchan JF, Elman RJ, et al. Life Participation Approach to Aphasia: A Statement of Values for the Future. The ASHA Leader. doi:10.1044/leader.FTR.05032000.4

42. Pedraza O, Sachs BC, Ferman TJ, Rush BK, Lucas JA. Difficulty and Discrimination Parameters of Boston Naming Test Items in a Consecutive Clinical Series. *Arch Clin Neuropsychol*. 2011;26(5):434-444. doi:10.1093/arclin/acr042

43. Hula WD, Kellough S, Fergadiotis G. Development and simulation testing of a computerized adaptive version of the Philadelphia naming test. *J Speech Lang Hear Res*. Published online June 24, 2015:1-13. doi:10.1044/2015_JSLHR-L-14-0297

44. Makransky G, Dale PS, Havmose P, Bleses D. An Item Response Theory-Based, Computerized Adaptive Testing Version of the MacArthur-Bates Communicative

Development Inventory: Words & Sentences (CDI:WS). *J Speech Lang Hear Res*. 2016;59(2):281-289. doi:10.1044/2015_JSLHR-L-15-0202

45. Ware Jr. JE, Gandek B, Sinclair SJ, Bjorner JB. Item response theory and computerized adaptive testing: Implications for outcomes measurement in rehabilitation. *Rehabil Psychol*. 2005;50(1):71-78. doi:10.1037/0090-5550.50.1.71

46. Hula WD, Fergadiotis Gerasimos, Swiderski Alexander M., Silkes JoAnn P., Kellough Stacey. Empirical Evaluation of Computer-Adaptive Alternate Short Forms for the Assessment of Anomia Severity. *J Speech Lang Hear Res*. 2020;63(1):163-172. doi:10.1044/2019_JSLHR-L-19-0213

47. Lee W-C, Lee G. IRT Linking and Equating. In: *The Wiley Handbook of Psychometric Testing*. John Wiley & Sons, Ltd; 2018:639-673. doi:10.1002/9781118489772.ch21

48. Guo J, Drasgow F. Identifying Cheating on Unproctored Internet Tests: The Z-test and the likelihood ratio test. *Int J Sel Assess*. 2010;18(4):351-364. doi:10.1111/j.1468-2389.2010.00518.x

# List of Figure Legends

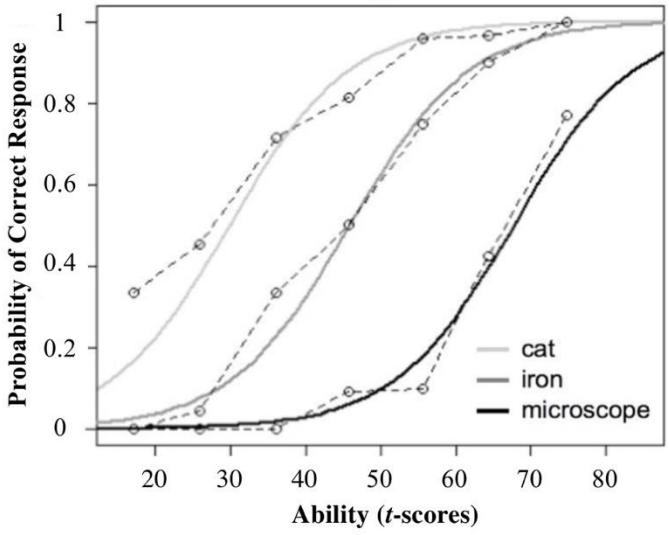


Figure 1. Item characteristic curves for three PNT items according to the IRT model (solid) and corresponding empirical data from 251 patients with aphasia (dashed).

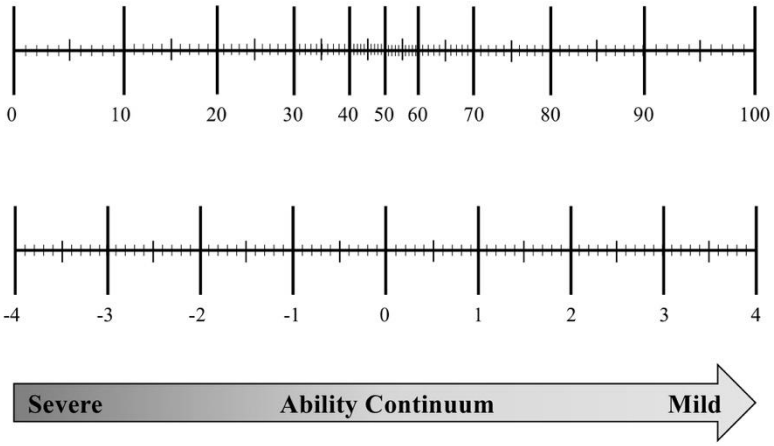


Figure 2. Conceptual representation of raw score scales (e.g., percent-correct; top) and IRT generated scales (bottom) across the ability continuum. The IRT scale is presented in the typical IRT metric (i.e., logit scale).

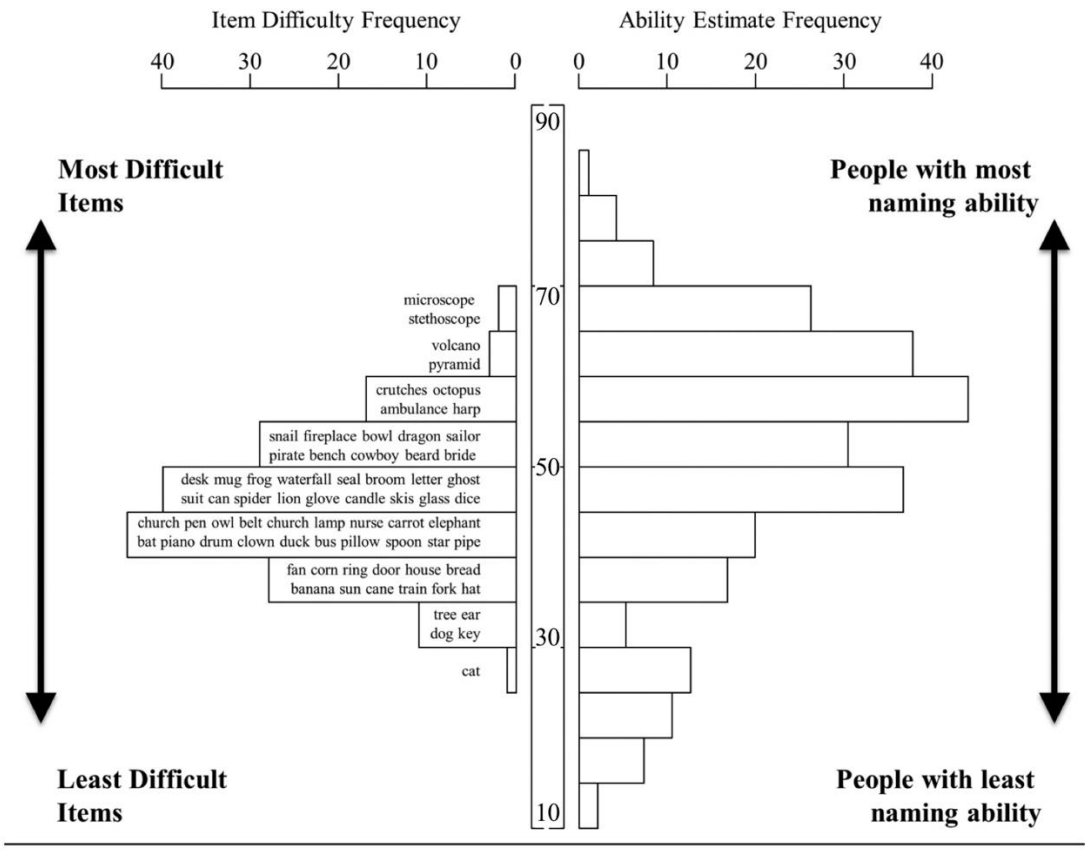


Figure 3. Map of items and persons showing select IRT-calibrated PNT items. Adapted from "Item Response Theory Modeling of the Philadelphia Naming Test" by G. Fergadiotis, S. Kellough, and W. D. Hula, 2015, *Journal of Speech, Language, and Hearing Research, 58*, p. 872. Copyright 2015 by the American Speech, Language, and Hearing Association.

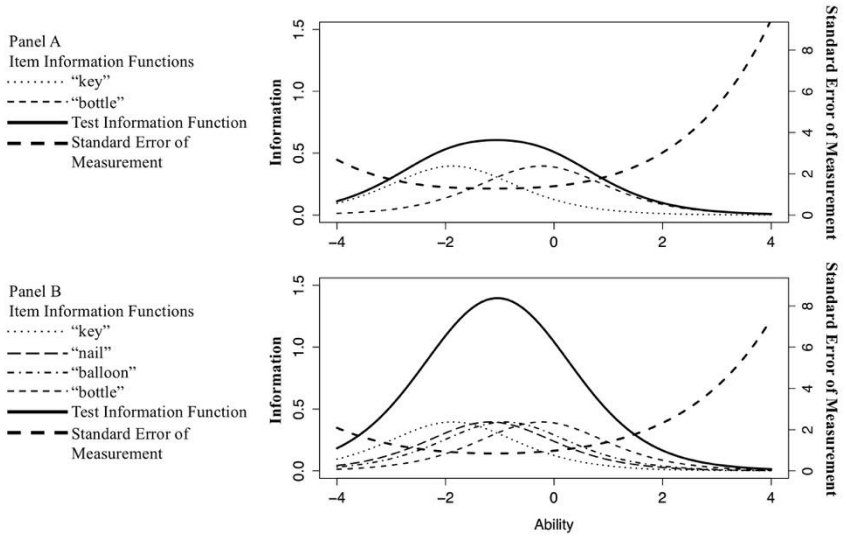


Figure 4. (A) Item and test information functions and the corresponding standard error of measurement for a two-item test. (B) Item and test information functions and the corresponding standard error of measurement for a four-item test. Ability is presented in the typical IRT metric (i.e., logit scale). Adapted from "Development and Simulation Testing of a Computerized Adaptive Version of the Philadelphia Naming Test" by W. D. Hula, S. Kellough, and G. Fergadiotis, 2015, *Journal of Speech, Language, and Hearing Research, 58*, p. 880. Copyright 2015 by the American Speech, Language, and Hearing Association.

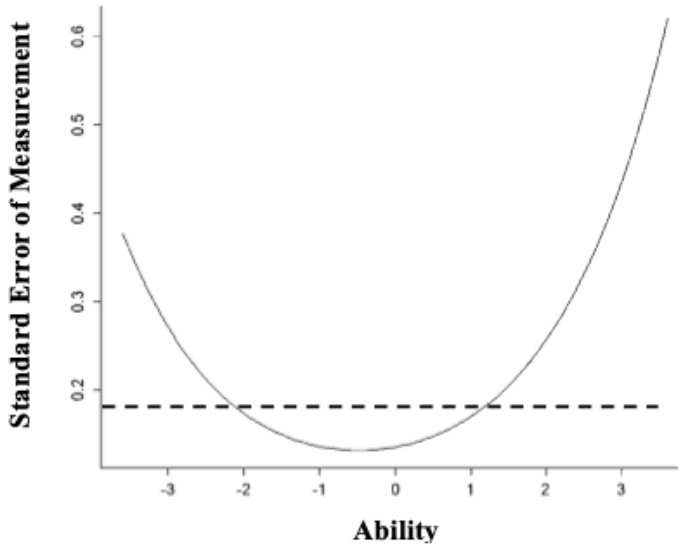


Figure 5. Conditional (solid) and average (dashed) standard error of measurement for naming ability estimates from the PNT. Ability is presented in the typical IRT metric (i.e., logit scale).

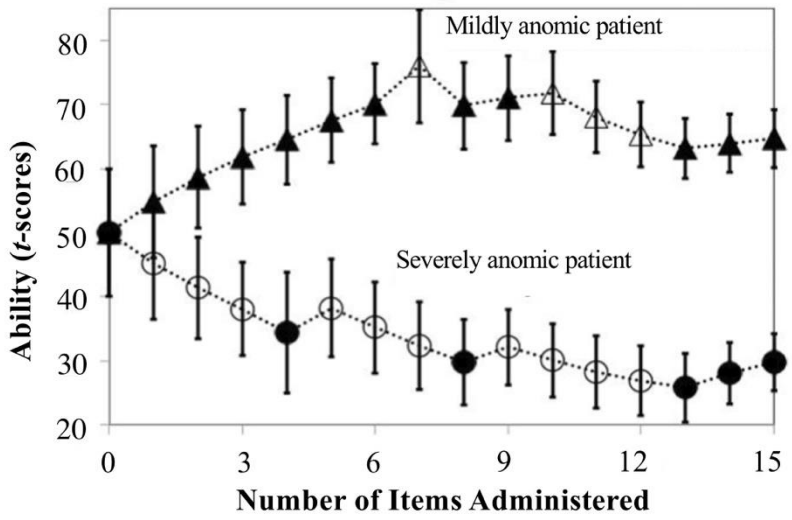


Figure 6. Incremental estimation of ability and standard error as a function of the first 16 items administered using the PNT-CAT for two hypothetical patients of differing anomia severity. Solid points represent correct responses; bars represent 1 standard error.

Biosketches

Gerasimos Fergadiotis, Ph.D., CCC-SLP is an associate professor in the Department of Speech and Hearing Sciences at Portland State University. His research interests are focused on the development of robust psychometric applications for the assessment of clinically relevant aspects of language production in aphasia.

Marianne Casilio, M.S., CCC-SLP is a doctoral student in the Department of Hearing and Speech Sciences at Vanderbilt University and former research speech-language pathologist with the Department of Speech and Hearing Sciences at Portland State University. Her research interests are centered on the measurement of language performance in aphasia as a means of (i) developing diagnostically informative assessment tools, and (ii) informing our understanding of linguistic processes and corresponding neural substrates.

William Hula, Ph.D., CCC-SLP is a research speech pathologist in the Audiology and Speech Pathology Service and the Geriatric Research, Education, and Clinical Center at the VA Pittsburgh Healthcare System and adjunct assistant professor in the Department of Communication Science and Disorders at the University of Pittsburgh. His research interests include the cognitive and neural bases of aphasia and its assessment and treatment.

Alexander Swiderski, M.A., CCC-SLP is a doctoral student in the Department of Communication Sciences and Disorders at the University of Pittsburgh and the Center of Neural Basis of Cognition at Carnegie Mellon. His research aims to understand the neural basis of language, how damage to neural systems impacts language processing in aphasia, and which brain structures support response to behavioral treatments. Specifically, he is interested in integrating findings from the neurobiology and cognitive psychology of language to inform the architecture of a tenable cognitive-psychometric model that supports item-level modeling commonly measured in aphasia assessments.

Learning outcomes

As a result of reading this article, the reader will be able to

1. summarize the importance of anomia assessment for patients with aphasia;
2. list the important psychometric properties of currently available confrontation naming tests;
3. discuss key conceptual differences between classical test theory and item response theory;
4. and summarize the ways in which anomia assessment tools developed using item response theory may improve clinical and research assessment practices.