ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Item Response Theory Modeling of the Verb Naming Test

Gerasimos Fergadiotis,[a] Marianne Casilio,[b] Michael Walsh Dickey,[c,d] Stacey Steel,[a] Hannele Nicholson,[e] Mikala Fleegle,[a] Alexander Swiderski,[c,d] and William D. Hula[c,d]

[a]Department of Speech & Hearing Sciences, Portland State University, OR [b]Department of Hearing and Speech Sciences, Vanderbilt University Medical Center, Nashville, TN [c]Department of Communication Science and Disorders, University of Pittsburgh, PA [d]VA Pittsburgh Healthcare System, PA [e]U.S. Department of Veterans Affairs, VA Minneapolis Healthcare System, MN

ARTICLE INFO

Article History:
Received August 5, 2022
Revision received December 19, 2022
Accepted January 23, 2023

Editor: Sarah Elizabeth Wallace

https://doi.org/10.1044/2023_JSLHR-22-00458

ABSTRACT

Purpose: Item response theory (IRT) is a modern psychometric framework with several advantageous properties as compared with classical test theory. IRT has been successfully used to model performance on anomia tests in individuals with aphasia; however, all efforts to date have focused on noun production accuracy. The purpose of this study is to evaluate whether the Verb Naming Test (VNT), a prominent test of action naming, can be successfully modeled under IRT and evaluate its reliability.
Method: We used responses on the VNT from 107 individuals with chronic aphasia from AphasiaBank. Unidimensionality and local independence, two assumptions prerequisite to IRT modeling, were evaluated using factor analysis and Yen's Q_3 statistic (Yen, 1984), respectively. The assumption of equal discrimination among test items was evaluated statistically via nested model comparisons and practically by using correlations of resulting IRT-derived scores. Finally, internal consistency, marginal and empirical reliability, and conditional reliability were evaluated.
Results: The VNT was found to be sufficiently unidimensional with the majority of item pairs demonstrating adequate local independence. An IRT model in which item discriminations are constrained to be equal demonstrated fit equivalent to a model in which unique discrimination parameters were estimated for each item. All forms of reliability were strong across the majority of IRT ability estimates.
Conclusions: Modeling the VNT using IRT is feasible, yielding ability estimates that are both informative and reliable. Future efforts are needed to quantify the validity of the VNT under IRT and determine the extent to which it measures the same construct as other anomia tests.
Supplemental Material: https://doi.org/10.23641/asha.22329235

Anomia, the impaired ability to access and retrieve words, is a hallmark symptom of aphasia (Goodglass & Wingfield, 1997; Kohn & Goodglass, 1985; Nickels, 2002). Although much research has centered on the production of nouns (e.g., Dell et al., 1997; Schwartz et al., 2006), verb production can be similarly disrupted in anomia (Berndt, Haendiges, et al., 1997; Berndt, Mitchum, et al., 1997; Nickels, 2014). Few options exist, however, for assessing verb production, and those that do are restricted by the psychometric framework within which they have been developed. The purpose of this study was to evaluate whether the Verb Naming Test (VNT), a relatively common subtest of verb production from the Northwestern Assessment of Verbs and Sentences (NAVS; Cho-Reyes & Thompson, 2012), can be adequately and reliably modeled using a modern psychometric approach, thereby creating a blueprint for future test and scale development for the assessment of verbs.

Verb Production in Aphasia

Models of spoken language production (e.g., Dell et al., 1997; Levelt et al., 1999) posit that successful word and sentence production depend on both retrieval of target

Correspondence to Gerasimos Fergadiotis: gfergadiotis@pdx.edu.
Disclosure: *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

lemmas (lexical processing) and selection of appropriate morphosyntactic frames for those lemmas (morphosyntactic processing). Notably, different word classes are thought to differentially engage one component of processing over the other. Content words (e.g., nouns, verbs, adjectives, adverbs) are traditionally thought to require greater recruitment of lexical processes, whereas function words (e.g., pronouns, articles) engage morphosyntactic processing to a greater degree. Verbs, which vary in both their morphological structure and semantic richness, can be understood as existing at the nexus of these two linguistic processes, playing a central role in sentence production (Chang et al., 2006; Gordon & Dell, 2003).

In aphasia, differential impairments in component processes of spoken language production are frequently observed, and these impairments may selectively affect certain word classes. Specifically, individuals with a primary morphosyntactic processing deficit (i.e., agrammatism) present as more impaired in the retrieval of function words relative to content words, whereas the opposite is observed in those with a primary lexical processing deficit (i.e., anomia; e.g., Bradley et al., 1980; Goodglass & Kaplan, 1983; Segalowitz & Lane, 2000). A similar division has also been observed in the production of verbs relative to other content word classes, specifically nouns. Here, individuals presenting with agrammatism perform worse at retrieving verbs relative to nouns, whereas those with anomia experience the reverse (e.g., Bates et al., 1991; Caramazza & Hillis, 1991; Laiacona & Caramazza, 2004; Miceli et al., 1984).

Some have suggested that such double dissociations are artifacts of study design (Bastiaanse & Jonkers, 1998) or analytic technique (Alyahya et al., 2018). Regardless, there is also evidence that verb impairments may be especially prominent in aphasia. As many as 75% of individuals with aphasia experience greater verb than noun naming impairments (Mätzig et al., 2009), and verb production impairments are more strongly correlated with functional communication impairments than noun production impairments are (Rofes et al., 2015). The relative prominence of verb impairments may reflect verbs' position at the nexus of lexical and morphosyntactic processing: Impairments in either process will negatively impact verb production. The high prevalence and functional impact of verb impairments have led to the development of multiple treatment programs designed to target verb production deficits specifically (Edmonds et al., 2009; Loverso et al., 1979; Thompson & Shapiro, 2005; Wambaugh & Ferguson, 2007). Thus, the assessment of verbs specifically is of value not only for the accumulation of evidence on the nature of aphasia and the cognitive architecture of language but also for the development of efficacious interventions for remediating language deficits in aphasia.

Current Assessment of Verb Production

Despite the need for rigorous measurement of verb production in aphasia, relatively few assessment tools have been developed or validated, at least compared to noun production batteries. One relatively prominent VNT is the 22-item VNT (Cho-Reyes & Thompson). Similar to confrontation naming tests for nouns, examinees are required to label a target with one action word in response to pictorial scenes. Items were selected specifically to vary in terms of their syntactic complexity (argument structure or valence) while holding lexical–semantic and phonological processing demands constant (see Cho-Reyes & Thompson, 2012, for further details). In their initial validation study, they found the VNT to possess strong psychometric properties (i.e., interrater reliability, discriminant validity) within a classical test theory (CTT) framework, thus making it an appropriate measure for further psychometric evaluation. With respect to the feasibility of this endeavor, the VNT is part of the AphasiaBank protocol (MacWhinney et al., 2011) and VNT data are readily accessible for research purposes.

The CTT framework, however, is limited by several prominent factors (Embretson & Reise, 2000). First, each test provides scores that reflect an individual's ability compared to the standardization sample. As a result, scores cannot be directly compared across tests without assuming that equivalent samples were used during their development. This hampers researchers' and clinicians' ability to compare the performance of individuals assessed using different tools intended to measure the same construct, even when the items used in these assessment tools overlap. Furthermore, tests developed under CTT typically ignore that measurement error varies as a function of ability level, often resulting in invalid confidence intervals (CIs). This has implications for determining the probability of true change as a function of spontaneous recovery or response to treatment. Third, available tools often need to be administered in their entirety to obtain relevant diagnostic information, and no standardized CTT-based approach currently exists for tailoring the selection of test items to an individual's ability level, thereby minimizing testing burden. Even when discontinuation rules are used to adapt the test's length (e.g., Boston Naming Test [BNT]; Kaplan et al., 2001), analytic and simulation-based studies have demonstrated that this seemingly intuitive approach can be complex with a propensity to generate biased ability estimates (von Davier et al., 2019). Finally, test–retest effects can potentially influence results when the same test is administered again at a later time. Thus, repeated administration of the same stimuli reduces the ability to measure treatment progress independent of confounding test variables.

Item Response Theory

An alternative to CTT is item response theory (IRT; Lord, 1980; Lord & Novick, 1968). In IRT, statistical models are used to describe how test takers' underlying latent trait, that is, ability level, determines observed patterns of behavior. IRT formalizes the notion that a latent (unobserved) trait of interest, such as naming ability, can be estimated via a statistical model based on test responses that are directly observed. 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: (a) the item's difficulty and (b) the patient's ability level.

For example, the one-parameter logistic (1-PL) IRT model defines the probability that an examinee responds correctly to an item (observed behavior), given an item's difficulty and a person's ability level (both estimated by the model). The 1-PL model can be represented mathematically as

$$P(x_i = 1|\theta_j) = \frac{e^{\alpha(\theta_j - \delta_i)}}{1 + e^{\alpha(\theta_j - \delta_i)}}, \quad (1)$$

where $P(x_i = 1|\theta_j)$ is the probability that response x on item i by examinee j is correct given their latent trait level θ_j, α is the item discrimination parameter, and δ_i is item i's difficulty parameter. Item difficulty describes the location of an item on the ability continuum and can be understood to reflect the relative ease or challenge of producing a correct response on a given item. Within the context of IRT, item difficulty parameters typically range from −4 to 4 and the higher an item's difficulty, the harder the item is. The discrimination parameter, α, describes how well an item can differentiate between individuals at different ability levels and can be also conceptualized as the magnitude of the nonlinear relationship between an item and the latent trait.[1] Discrimination can theoretically vary from $-\infty$ to $+\infty$ but typically ranges from 0 to 2. According to Baker (2001), discrimination can be classified as *none*, *low*, *moderate*, *high*, and *very high* based on the following ranges: 0 is *none*, 0.01–0.34 is *very low*, 0.35–0.64 is *low*, 0.65–1.34 is *moderate*, 1.35–1.69 is *high*, and > 1.7 is *very high*. The 1-PL model stipulates that all of the items are equally discriminating. That is, the 1-PL model assumes that all items of the same difficulty are equally informative for estimating a person's ability level (i.e., α is not indexed for item). The latent trait, θ_j, is a stochastic estimate of the construct measured by the test for the jth examinee. In other words, θ_j represents the degree of the underlying trait a person possesses relative to the difficulty of the items in the test. Even though θ_j is not directly observed, it can be estimated based on the test taker's observed responses on items of known difficulty. In the context of confrontation picture naming tests, ability and any numerical estimates associated with it are used to refer to the degree of naming impairment or anomia severity. For an introduction to IRT concepts and applications in the context of speech-language pathology, interested readers are directed to the works of Baylor et al. (2011) and Fergadiotis et al. (2021). For a more general and complete presentation, see the works of de Ayala (2013) and Embretson and Reise (2000).

IRT ability estimates are meaningful only to the extent that they meet certain psychometric assumptions. First, commonly used IRT models assume that the item set is unidimensional, meaning that all item responses are essentially a function of a single common underlying trait. Specifically, the observed variance of items can be decomposed into the sum of two parts: the variance accounted for by the underlying trait across all items, and any residual variance that is unique to each item and includes idiosyncratic variance and random error. To the extent that unidimensionality is violated when an IRT model assumes it, there is a negative effect on the accuracy of model parameters, as well as the ability estimates and their precision (Crişan et al., 2017). Importantly, interpretation of ability estimates becomes increasingly more challenging as the severity of model assumption violations increases, making it less and less clear what meaning one should assign to a given score. A second related assumption is that of local independence. Typically, responses on a test are correlated because the probability of responding correctly to the items is determined by a common factor that is assumed to be the person's ability level. However, when the effect of the common factor is partialed out (i.e., when responses are conditioned on ability level), all responses should be independent. Local independence can be violated in the presence of multidimensionality, as discussed previously, for example, when groups of items are based on a common stimulus, such as when several items refer to the same reading comprehension passage (Wainer et al., 2007) or when there are other sources of systematic influence not accounted for by the IRT model applied to the data (Levy et al., 2009). Finally, another assumption concerns the specific form of the model. As described earlier, the 1-PL model assumes that all of the items are equally discriminating, that is, that each item is related to the underlying trait with equal strength and, thus, equally informative for estimating ability. However, this can be an untenable assumption in many cases, and if an inappropriate model is used with the data,

[1]Relatedly, the IRT discrimination parameter bears a direct mathematical relationship to a factor loading estimated in the context of a common factor model and is also conceptually similar to a regression weight in the context of an observed-variable regression model. It indexes the amount of variance in the observed response variable accounted for by the latent trait. For interpretations of the discrimination parameter in the context of diffusion and race models, see the work of Tuerlinckx and De Boeck (2005).

not only can model parameters be distorted, but also the interpretation of a person's ability score becomes challenging. The two-parameter logistic (2-PL) model relaxes this assumption, allowing discrimination to vary across items. However, this complicates the model fitting process as the 2-PL model also requires larger sample sizes for stable estimation of item discrimination parameters (de Ayala, 2013). Varying discrimination parameters also complicate interpretation of the model because they can create situations in which items change their ordinal ranking of difficulty depending on the ability of the test taker.

When key assumptions of the chosen IRT model are well approximated, IRT provides a rigorous framework for addressing the CTT limitations discussed above. First, IRT defines a latent trait scale that is in theory independent of both the particular items that are administered and the item calibration sample. Therefore, when tests are calibrated using IRT methods, ability estimates based on different collections of items (e.g., different tests or different subsets of items from the same test) can be directly compared irrespective of the particular characteristics of the items and their standardization samples. A second benefit of IRT models is that they represent the precision of score estimates conditional on individual ability level (Embretson & Reise, 2000). This feature permits one to model the fact that an easy test given to a severely impaired patient provides a more precise and informative score estimate than the same test given to a mildly impaired patient. In contrast, CTT assumes that the precision of the total observed score is expressed as a single average value that is dependent on the variability in the sample at hand. Finally, 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 (Embretson & Reise, 2000). Thus, a change of 1 (in terms of the log odds of a correct response) on the latent trait scale has the same meaning regardless of where on the scale it occurs.

IRT Modeling of Aphasia Tests

While IRT is not new and dates back to the 1960s, it has seen relatively little use in the field of aphasiology and speech-language pathology more broadly, especially considering the number of assessment tools currently circulating in the field. Furthermore, as Baylor et al. (2011) discussed, even when IRT has been used, in some cases "... this work has either tended to be technical in orientation with limited impact on clinical practice or has not capitalized on some of the particular advantages of IRT" (p. 244). Early examples of the application of IRT to instruments relevant to aphasiologists include the Token Test (Willmes, 1981), the Test of Adolescent/Adult Word Finding (German, 1990), and the Aachen Aphasia Test (Willmes, 2003). More recently, IRT methods have been applied to the Western Aphasia Battery–Revised (Hula et al., 2010; Kertesz, 2007), the BNT (del Toro et al., 2011; Kaplan et al., 2001), and the Dutch Naming Test (Alons et al., 2022).

Perhaps the most productive application of IRT to enhance the psychometric properties of tests used has been to the development of IRT-based patient-reported outcome measures and the assessment of anomia. With respect to the former, modern psychometric approaches have been used to develop the Aphasia Communication Outcome Measure (Hula et al., 2015; Hula & Doyle, 2021), the Communicative Participation Item Bank (Baylor et al., 2021), and the Communication Confidence Rating Scale for Aphasia (Babbitt et al., 2011). With respect to the assessment of word retrieval, which is the primary focus of this study, the application of IRT techniques to improve clinical utility has been demonstrated in a series of recent studies. Specifically, capitalizing on the advantages of IRT, our research group has focused on applying IRT to refine the Philadelphia Naming Test (PNT; Roach et al., 1996), a confrontation naming test of noun production that is commonly used in research settings.

Our initial research efforts focused on fitting an IRT model to the PNT and evaluating its psychometric properties. After meeting the assumptions of unidimensionality and local independence, we found the PNT to demonstrate adequate fit to both 1-PL and 2-PL models, although the 1-PL model was ultimately selected for parsimony (Fergadiotis et al., 2015). Moreover, a regression analysis of three salient lexical variables (word length, age of acquisition, and frequency) showed all to be significant predictors of item difficulty parameters of the PNT, thus supporting the validity of the PNT under IRT (Fergadiotis, Swiderski, & Hula, 2019).

After establishing the utility of IRT as a psychometric framework for the PNT, we extended this work by developing a computerized adaptive test (CAT) version of the PNT (Fergadiotis, Hula, et al., 2019; Hula et al., 2015, 2020). The CAT version, or PNT-CAT, was designed to present items tailored to an individual's ability level, estimated from their previous responses. Our initial simulation study revealed that a 30-item PNT-CAT form yields equal or greater accuracy and precision of naming ability estimates than existing static short forms (Walker & Schwartz, 2012), showing the potential to reduce PNT test length, and thereby decrease test burden and administration time, without sacrificing measurement precision. We then validated these findings by comparing agreement in ability estimates from the 30-item PNT-CAT form with the original 175-item PNT using empirical data collected from 47 participants with aphasia. Here, agreement between the two test versions was almost perfect, as indicated by high correlation (r = .95, 95%

CI [0.92, 0.97]) with negligible bias, low variable and total error, and pairwise score differences not exceeding the Type I error rate (Fergadiotis, Hula, et al., 2019).

More recently, we compared two alternate PNT-CAT short forms of the PNT (30-item, variable length) with nonoverlapping content to evaluate potential differences in the effectiveness of various test lengths. As before, we found the two test versions were highly correlated (r = .89) with error variance that was low and in the range predicted by the IRT measurement model, suggesting that both PNT-CAT short forms can sufficiently assess anomia severity. Overall, the collective results of these findings across studies suggest that the PNT-CAT is a reliable and efficient system that can reduce administration time and provide accurate information of naming deficits.

The Value of Test Validation Under IRT

Quantifying the psychometric properties of existing aphasia tests under modern frameworks such as IRT has numerous advantages, as discussed in the preceding paragraphs. Because of their stronger claim to interval status, more realistic modeling of measurement error, and support for CAT and other flexible approaches to test administration, IRT-based tests have great potential to improve clinical decision making. Maximizing the validity of test scores in turn increases the validity of the inferences and decisions based on them, including (a) determination of referral for speech-language pathology services, (b) selection of a treatment program, (c) quantification of progress over time, (d) justification of continued provision of care, or (e) tailoring of education to individuals with aphasia and their caregivers. In addition, rigorously validated test scores are critical to applied clinical research, where sensitive and precise measures are needed to determine the efficacy of interventions. Finally, rigorously validated measures are necessary for developing or refining theories about the neural and cognitive architecture of language. In the absence of psychometrically robust measures, clinical decision making becomes inherently more variable, and research findings, whether applied or mechanistic, become more challenging, if not impossible, to interpret.

This Study

Given prior successful efforts at modeling the PNT under an IRT framework, coupled with the need to further refine the psychometric properties of the VNT, the aim of this study was to evaluate the fit of the VNT to an appropriate IRT model and its reliability under IRT. Our research questions were as follows: (a) Do the assumptions of unidimensionality, conditional independence, and equal discrimination hold for the VNT, and (b) what is the VNT's reliability under an IRT framework?

Method

Participants

A sample of 107 individuals with aphasia with complete audiovisual recordings of both the VNT and the short form of the BNT (Kaplan et al., 2001) was identified after screening all 296 participants within the AphasiaBank database (MacWhinney et al., 2011; aphasia.talkbank.org) on March 6, 2019, as part of a larger research project. Inclusion criteria were as follows: (a) aphasia due to a single left-hemisphere stroke, where aphasia was operationally defined as an aphasia quotient of < 93.8 on the Western Aphasia Battery–Revised (WAB-R; Kertesz, 2007) or < 11 on the short form of the BNT; (b) right-handed native English speakers; (c) adequate hearing and vision (aided or unaided) for testing purposes; and (d) no significant comorbid neurologic or psychiatric illness. Those with a concomitant clinical diagnosis of apraxia of speech or dysarthria were also included in the sample. Participant demographic and clinical data are shown in Table 1. A complete participant ID list is provided in Supplemental Material S1.

Transcription and Scoring

Participant responses on the VNT were phonemically transcribed by two research assistants at Portland State University in a pseudorandom order. Of note, participant responses included in this VNT data set contained a large number of multiword responses. Thus, the research assistants transcribed everything the participant said/gestured in response to the stimuli and used a set of transcription coding conventions adopted from the CHAT manual (MacWhinney, 2000) meant to capture elements of nonfluent speech (see Appendix A for more information). There were six missing data points all due to examiner error during administration (i.e., test items were unintentionally skipped).

Phonemic transcriptions were broad, and variations in dialect were transcribed as they were heard using a phonemic notation developed by our laboratory for the purposes of use with a computer algorithm (Fergadiotis et al., 2016). Refer to Appendix A for our phoneme conventions, a list of target phonemic transcriptions, and a list of transcription coding conventions. If a given production strayed from our lab's phonemic conventions, as was the case for some British dialects, that production was

Table 1. Participant demographic and clinical characteristics.

Characteristic	Value
Ethnicity	
African American	14%
Asian	1%
White	85%
Education (years)	
M (*SD*)	14.97 (2.37)
Min–max	11–20
Missing	4%
Age (years)	
M (*SD*)	61.52 (10.96)
Min–max	39–85.7
Missing	1%
Years after aphasia onset	
M (*SD*)	5.44 (4.78)
Min–max	0.25–25.75
WAB-R AQ	
M (*SD*)	70.02 (17.08)
Min–max	20.5–97.9
Number of people with AQ < 93.8	104
BNT-SF (% correct)	
M (*SD*)	6.97 (4.37)
Min–max	0–15
Number of people with score < 11	80
VNT (% correct)	
M (*SD*)	14.44 (6.48)
Min–max	0–22

Note. WAB-R = Western Aphasia Battery–Revised (Kertesz, 2007); AQ = aphasia quotient; BNT-SF = Boston Naming Test–Short Form (Kaplan et al., 2001); VNT = Verb Naming Test (Cho-Reyes & Thompson, 2012).

converted into Standard American English and transcribed in accordance with our conventions. Disagreements in transcription between the two research assistants were resolved by a licensed speech-language pathologist in a pseudorandom order.

The data set was scored following the VNT scoring protocol with minor modifications to resolve ambiguity about which production to select and score in the midst of a multiword response (see Appendix B). First, the final main lexical verb produced as part of the first complete response was scored for accuracy. Second, auxiliary verbs, verbs produced as personal commentary, and/or copula "to-be" verbs functioning as main lexical verbs were systematically ignored. Third, responses containing inflectional morphemes (e.g., *–ing*, *–ed*, *–s*) were recognized as correct verb approximations. Fourth, specific to and in accordance with the VNT scoring rules, phonemic paraphasias that were phonologically similar to the target verb (i.e., ≥ 50% of the phonemes were shared between the target and response) were scored as correct. Two research assistants scored and error coded the first responses within 10 s after stimulus presentation with or without a first verbal prompt, and disagreements were then resolved by a licensed research speech-language pathologist.

The VNT scoring protocol allows for an additional prompt from the test administrator and a second response following a first incorrect attempt. However, for the purposes of this study, only a participant's first response was considered for scoring and any subsequent response prompted by the test administrator, verbally or nonverbally, was ignored. This decision was made because we noted considerable variability in VNT testing administration across multiple sites that had contributed data to AphasiaBank. For example, oftentimes an additional prompt was not provided after an incorrect first attempt due to examiner error. When prompts were provided, they often deviated from the manual and included additional syntactic and semantic information that could influence the subject's word retrieval and responses. The examiner's prompts and cues were therefore annotated using a coding convention (see Appendix A) developed by our lab for this study. Then approximately 20% of the data were pseudorandomly selected to quantify and further analyze the types of testing variability and their impact on the accuracy of first and second responses. Results revealed that when the examiner followed the VNT rules and provided an additional prompt following an incorrect first attempt, it did not significantly affect the likelihood of a correct second attempt (only 9.43% correct second attempts produced). Hence, the subject's first responses were prioritized over second response transcriptions for resolution and scoring.

Modeling Approach

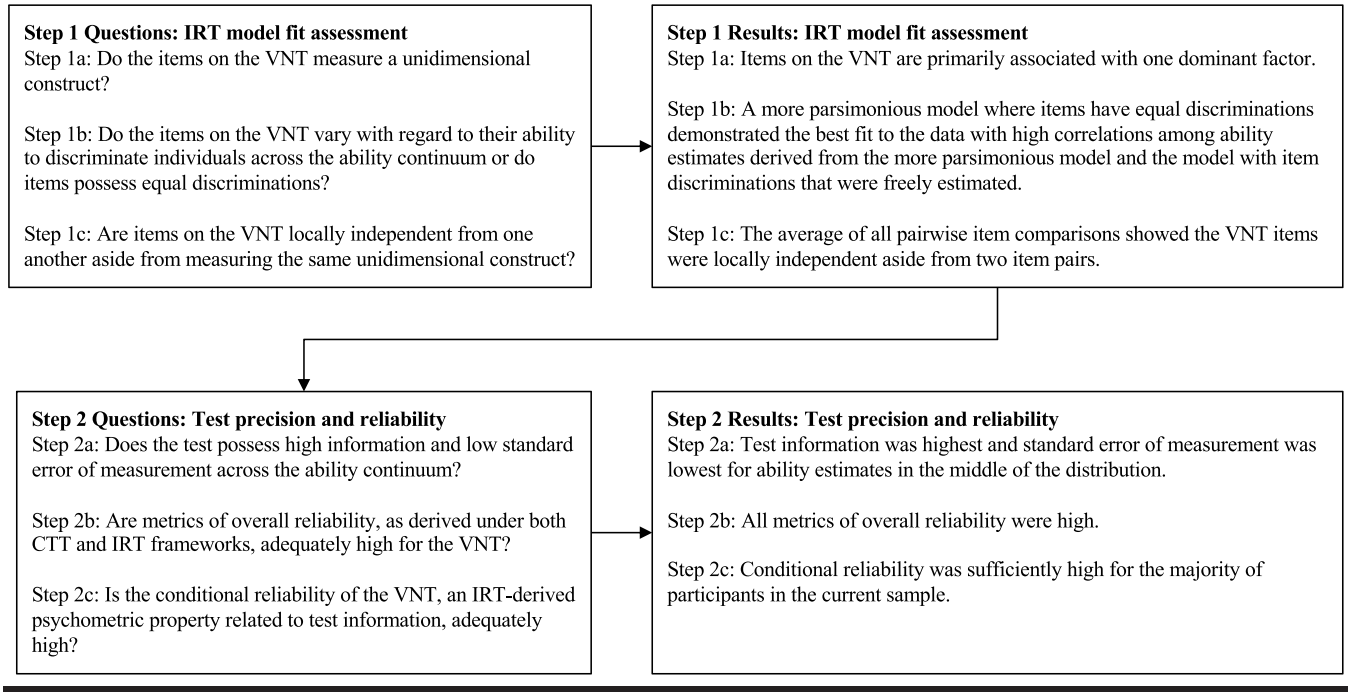
A graphical overview of our modeling approach can be found in the left-hand column of Figure 1.

IRT Model Fit Assessment

The following fit properties are well-established requisites for applying traditional IRT modeling to any test (e.g., Baker, 2001; de Ayala, 2013; Embretson & Reise, 2000; Lord & Novick, 1968; Wilson, 2005). In the case where assumptions are violated, multidimensional and multilevel IRT models may be appropriate for accounting for such data complexities (De Boek & Wilson, 2004). Notably, model assumption testing reveals important psychometric properties about a test, which may or may not be aligned with the intent of the test developers' and initial item design.

Unidimensionality. Unidimensionality was initially assessed using the modified parallel analysis proposed by Drasgow and Lissak (1983), as implemented in the *ltm* R package (Rizopoulos, 2007). The primary analysis of

Figure 1. A graphical overview of the item response theory (IRT) modeling approach and a summary of findings. VNT = Verb Naming Test; CTT = classical test theory.



unidimensionality was then conducted within a categorical confirmatory factor analytic framework in Mplus (Version 8; Muthén & Muthén, 2017). We specified and fit a unidimensional model for which covariances among residual terms were constrained to be zero, and loadings and thresholds were freely estimated. This specification corresponds to the 2-PL model, which assumes varying discrimination across items. For identification purposes, the factor variance was set equal to one and all loadings were freely estimated. Given the categorical nature of the items, the weighted least squares mean- and variance-adjusted estimator was used (Li, 2016). To evaluate global model fit, we used the mean- and variance-adjusted χ^2 statistic, the comparative fit index (CFI; Bentler, 1990), and the root-mean-square error of approximation (RMSEA; Steiger & Lind, 1980). On the basis of published guidelines, good fit was indicated by a nonsignificant χ^2 statistic, a CFI higher than .95, and an RMSEA value below 0.08 with the upper bound of the 95% CI below .10, although cutoffs are somewhat model dependent (Brown, 2015; Hu & Bentler, 1999; Kline, 2010). To assess for local strain in the models, modification indices with values greater than 3.84 were considered. Local strain refers to examining and identifying the different parts of the model for unnecessary parameters that hurt fit or missing parameters that might improve local fit. For all analyses, missing data were accommodated using maximum likelihood under the assumption of data missing completely at random (Rubin, 1976).

Equal discrimination. To assess the assumption of equal discrimination across items, first, a 1-PL model and a 2-PL model were compared with a nested model difference test based on the weighted least squares mean- and variance-adjusted χ^2 statistic using the DIFFTEST option in Mplus. A statistically significant result (i.e., $p < .05$) would suggest that the more restricted 1-PL model fits significantly worse compared to the more flexible 2-PL model (Bentler, 2000). In addition, modification indices from Mplus associated with loadings (which correspond to discriminations) were considered. Modification indices in general reflect the amount by which the chi-square statistic could be reduced if a specific single parameter restriction were to be removed from the model (Sörbom, 1989). We also evaluated fit at the level of each individual item using item-level χ^2 fit statistics using the *ltm* R package. Significant χ^2 values (i.e., $p < .05$) indicate items whose model implied functions may not be consistent with the data. Finally, the two models were evaluated in terms of their practical implications. Specifically, ability scores were generated under the two models and were compared in terms of their strength of association (Pearson product–moment correlation) and bias (average signed difference).

Local independence. The assumption of local independence was evaluated using Yen's Q_3 statistic (Yen, 1984), which focuses on the magnitude of the residual correlations of all possible pairwise item combinations. The

critical value for flagging potentially problematic residual correlations was determined based on Christensen et al.'s approach (2017) according to which the critical value is set equal to .2 but adjusted by the average Q_3 statistic across all pairwise residual correlations. Any residual correlations with values greater than the critical value suggest that the corresponding pairs of items have something more in common than all of the items have in common with each other. To estimate Q_3, the *sirt* R package (Robitzsch, 2022) was used to first fit a 1-PL model and then calculate the statistics of interest (average Q_3, pairwise Q_3s).

Precision and Reliability

Information and conditional standard error of measurement. An item information function quantifies the degree to which observing a response on an item decreases the uncertainty in an ability estimate. When information is summed across items, the resulting curve captures how informative the test is as a whole conditional on ability. A test's standard error of measurement (*SEM*) conditional on ability is inversely related to the square root of information. The test information function and the conditional *SEM* curve were derived using the *mirt* package in R (Chalmers, 2012).

Overall reliability: Categorical omega, and marginal and empirical reliability. First, the overall reliability attained by the VNT was estimated within a factor analytic framework using categorical omega. Categorical omega was estimated using the *ci.reliability* function from the *MBESS* R package (Kelley, 2022) based on Green and Yang (2009) using bias-corrected and accelerated bootstrap CI estimation (Kelley & Pornprasertmanit, 2016).

Then, the overall reliability was investigated within an IRT framework using the *mirt* R package. Specifically, two indices were computed: marginal and empirical reliability. The former calculates reliability through integration based on the model-implied test information functions and by assuming an a priori–selected probability density function of ability, which is in this case a standard normal distribution. On the other hand, the calculation of empirical reliability depends on computing the variance of ability estimates and their corresponding standards errors directly from the data. To the extent that the theoretical distribution of ability used in marginal reliability estimation is correctly specified, and the model parameters are not biased, then the resulting estimate is a sufficient estimator for overall reliability and in agreement with empirical reliability.

Conditional reliability. Conditional reliability estimates, which capture the precision of a test as a function of ability level, were extracted from the *plot(type = "rxx")* function in *mirt*, and were replotted against the empirical ability density to investigate the region of ability for which the VNT was maximally reliable. Expected a posteriori estimation was used to derive the ability estimates of participants in the sample.

Results

IRT Model Fit Assessment

A graphical overview summarizing our findings can be found in the right-hand column of Figure 1.

Unidimensionality

The initial exploratory analysis of unidimensionality using parallel analysis suggested that data were essentially unidimensional as the second eigenvalue in the observed data was equal to 1.58 and not substantially larger than the second eigenvalue of the permuted data sets (p = .90). A parallel analysis plot can be seen in Supplemental Material S2.

When evaluated within a confirmatory factor analytic framework, the unidimensional model assuming varying loadings (equivalent to a 2-PL model) converged to a solution with no out-of-range parameter values and its global fit indices provided evidence of adequate model fit, $\chi^2(209, N = 107) = 215.287$, p = .368; CFI = .994; RMSEA = 0.017, 90% CI [.00, .045]. Furthermore, no local model strain was noted (i.e., no modification indices with values > 3.84). The unidimensional model assuming equal loadings across items (corresponding to a 1-PL model) also converged to an admissible solution with evidence of adequate global model fit, $\chi^2(230, N = 107) = 254.147$, p = .131; CFI = .976; RMSEA = 0.031, 90% CI [.00, .052]. Modification indices suggested some model strain associated with the assumption of equal loadings across items. Based on these findings, the assumption of unidimensionality that is central to the specification of both the 1-PL and the 2-PL models was judged adequate for the purpose of IRT modeling. The full solutions (i.e., difficulty and discrimination parameter estimates across models) can be seen in in the left panel of Table 2 ("difficulty" and "discrimination").

Equal Discrimination

Overall, global fit indices associated with the 1-PL, which assumes equal discrimination across items, suggested adequate model fit. Model fit indices derived from Mplus suggested that model fit could improve by freeing the discrimination parameters of some items. However, as seen in the right panel of Table 2 ("item fit statistics"), the expected parameter change in each case was relatively small, with the exception of the item "give." In addition,

Table 2. Model parameters for 1-PL and 2-PL models and item fit indices based on the 1-PL model.

	Difficulty		Discrimination		Item fit statistics			
Item	1-PL	2-PL	1-PL	2-PL	χ^2	*p*	MI	EP
Cut	−0.26	−0.26	0.90	0.90	5.85	.76		
Bark	−0.71	−0.62	0.90	1.17	6.66	.67		
Put	1.81	2.00	0.90	0.76	3.63	.93		
Send	0.41	0.41	0.90	0.91	11.70	.23		
Drive	−0.52	−0.50	0.90	0.95	5.70	.77		
Wash	−0.41	−0.45	0.90	0.76	14.18	.12		
Read	0.26	0.33	0.90	0.62	5.56	.78	5.24	0.61
Laugh	−1.09	−1.21	0.90	0.75	11.88	.22		
Watch	0.47	0.41	0.90	1.19	13.85	.13		
Give	0.71	1.13	0.90	0.47	30.78	>.01	13.30	0.46
Swim	−1.23	−1.31	0.90	0.81	4.54	.87		
Stir	0.43	0.50	0.90	0.70	5.66	.77		
Pinch	−0.05	−0.06	0.90	0.75	11.74	.23		
Crawl	0.44	0.37	0.90	1.34	9.77	.37	5.95	1.38
Deliver	0.90	1.13	0.90	0.63	10.82	.29		
Pour	0.50	0.43	0.90	1.26	8.51	.48	4.08	1.28
Howl	0.41	0.33	0.90	1.52	8.52	.48	9.37	1.55
Throw	0.71	0.76	0.90	0.80	4.22	.90		
Bite	−0.09	−0.08	0.90	1.06	6.11	.73		
Shove	0.02	0.02	0.90	0.81	5.69	.77		
Tickle	0.37	0.41	0.90	0.78	6.53	.69		
Shave	−0.58	−0.49	0.90	1.31	5.82	.76	4.66	1.32

Note. PL = parameter logistic; MI = modification index; EP = expected parameter.

item level χ^2 tests estimated using the *ltm* package were nonsignificant with the exception again of the item "give" ($\chi^2 = 30.78$, $p < .01$). Finally, based on the DIFFTEST results from Mplus, despite the additional constraints imposed by the 1-PL model, model fit was not significantly worse compared to the 2-PL model, $\chi^2(21, N = 107) = 30.75$, $p = .07$.

In addition to the statistical analyses, we investigated the practical implications of selecting a 1-PL model versus a 2-PL model. Ability estimates generated under the two models were highly correlated ($r = .995$, $p < .001$). Furthermore, negligible bias was noted, which was not significantly different than zero (bias < .0001). Visually, the strength of the linear association of the two sets of scores can be seen in Figure 2. In addition, consistent with the statistical analysis results, the data travel through the point where the axes intersect, which also suggests negligible bias. Ability estimates and their standard errors can be seen in the Supplemental Material S3.

Local Independence

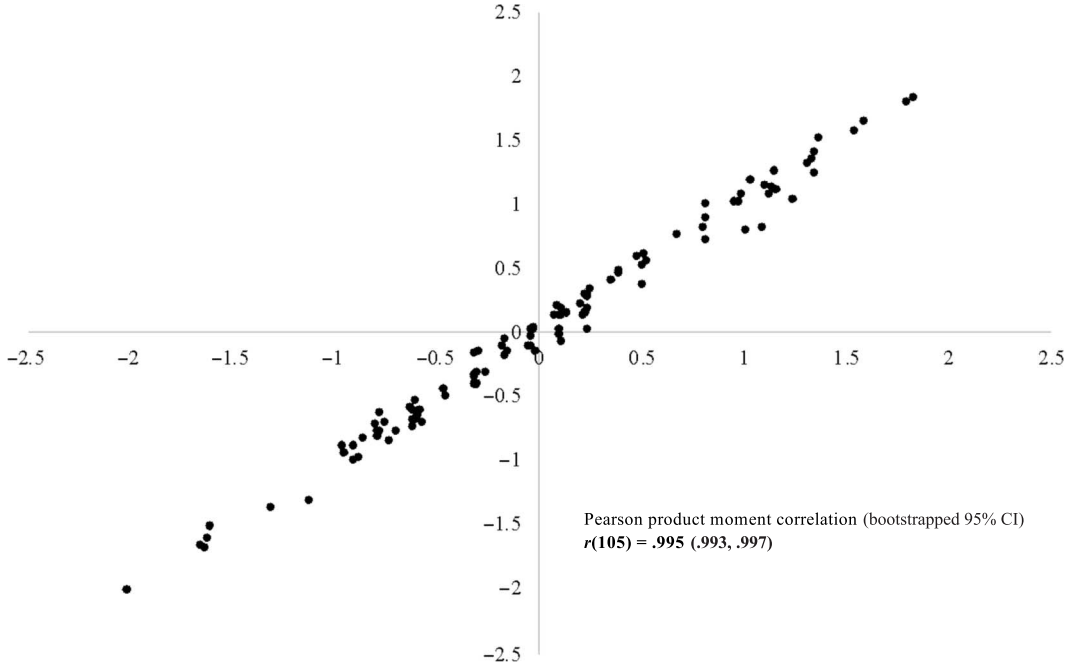
Two hundred thirty-one pairwise residual correlations were estimated. The average Q_3 statistic across all pairwise combinations was equal to .084, and the critical value for comparing residual correlations was therefore .284. Across 231 pairwise combinations, only two pairs of items were flagged as potentially problematic. The Q_3 statistic for the items "bite" – "shove" was equal to .36, and the Q_3 statistic for the items "send" – "wash" was equal to .31.

Precision and Reliability

Information and SEM. The test information function and the *SEM* curves can be seen in Figure 3. The test information function (solid blue line) peaks near the average of the ability continuum (i.e., $\theta = 0$) given the moderate difficulty of the VNT items and gradually decreases for regions of more extreme ability levels. The SEM curve (dashed red line) mirrors the test information function, given that the former is calculated as the reciprocal of the square root of the latter. As it can be seen in the figure, the uncertainty of ability estimates is a function of the ability level and ability estimates at the extreme of the ability distribution are associated with considerably more error.

Overall reliability. The estimated categorical omega using bias-corrected and accelerated bootstrap CIs suggested high internal consistency ($\omega = .9095$; 95% CI [.8011, .9105]). Similarly, marginal reliability was estimated to be equal to 0.8721, whereas the empirical reliability was estimated to be

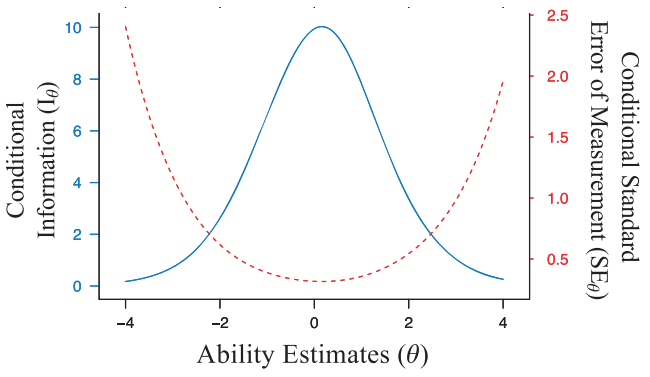
Figure 2. Item response theory (IRT) abilities generated under the 1-PL model that assumes equal discrimination across items (*x*-axis) and the 2-PL model that assumes unique discrimination parameters for each item (*y*-axis). Despite the different parameterization of the models, ability estimates were very similar.



equal to 0.8758. Both marginal and empirical reliability estimates were within the 95% CI of the categorical omega coefficient.

Conditional reliability. The conditional reliability curve that can be seen in Figure 4 (red solid line) follows the pattern of the test information function. Considering the empirical distribution of ability estimates shown in the histogram, the level of reliability of the VNT is above .80 for the vast majority participants in this study.

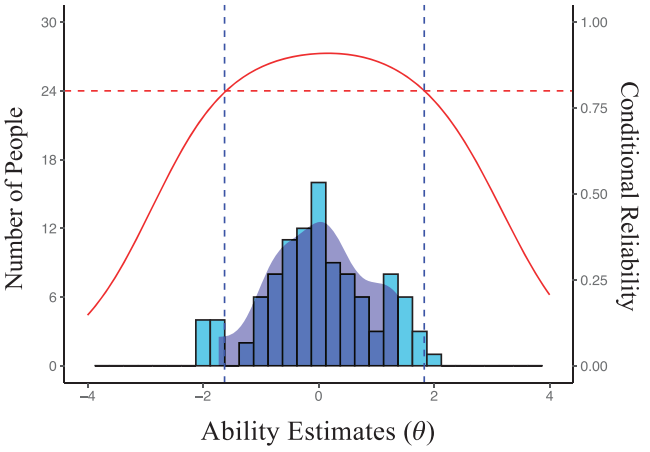
Figure 3. Test information function (I_θ; blue solid curve) and the standard error of measurement curve (SE_θ; dashed red curve) as a function of ability estimates (θ).



Discussion

The purpose of this study was to evaluate the fit and reliability of the VNT under IRT. We found that the VNT met all necessary assumptions for IRT modeling and that a 1-PL model, where all test items are assumed

Figure 4. Conditional reliability (red curve) plotted against the empirical distribution of the study sample (histogram). The truncated density area indicates the ability scores that fall between the 5th and 95th percentiles in the data. The blue dashed vertical lines indicate the range of scores for which ability estimates are calculated with conditional reliability above .80. The red dashed line indicates the conditional reliability equal to .80.



to be equally discriminating, demonstrated satisfactory fit to the data. Moreover, precision and reliability were high, suggesting that the VNT is a psychometrically robust test for the assessment of action naming in aphasia.

IRT Model Fit Assessment

Unidimensionality

As with successful application of IRT models in similar domains (e.g., naming of objects; Fergadiotis et al., 2015), this set of analyses supports the claim that the VNT, and presumably other action naming tests, can be usefully modeled with a unidimensional structure. Satisfying the unidimensionality assumption provides preliminary evidence of construct validity, as both observed and IRT-derived test scores appear to vary systematically along a single continuum. This is a necessary but not sufficient condition for most clinical and research applications: rank ordering individuals based on their overall ability to produce verbs, relating performance on the VNT to predictive and explanatory variables including group membership in clinical trials, and forming groups of individuals on the basis of their overall performance. Furthermore, given that unidimensionality is a necessary assumption for estimating the most common IRT models, satisfying this assumption confers several practical advantages. Most notably, from a test development perspective, it opens the door to applying a repertoire of robust and established psychometric tools to develop refined assessments, including efficient computer adaptive testing applications, in the future.

However, adequate fit to a unidimensional model should not be interpreted as an endorsement of the idea that a unitary psychological construct underlies action naming. Regardless of domain, behavioral responses to test items never rely on a single underlying psychological process. We recognize that naming of actions and objects depends on the complex interaction of cognitive–linguistic and sensorimotor processes engaged during word retrieval. However, our results strongly suggest that it is reasonable to assume that the overall ability to access and retrieve verbs can be approximated by a single quantity that reflects the relative ability of a person with aphasia to produce the target word. Note that this is the same assumption that is a prerequisite to claiming valid clinical inferences when using test-level scores (e.g., total score, percent correct) from a confrontation naming test without invoking the IRT machinery, a practice that is rarely questioned in practice given the utility of confrontation naming tests. Furthermore, even models that claim to measure distinct underlying cognitive processes involved in naming yield combined scores that are almost indistinguishable from ability estimates generated by unidimensional IRT models (Walker et al., 2022).

Equal Discrimination

Even though modification indices suggested that some discrimination parameters could be freed to improve model fit, the choice to proceed with a 1-PL model has several advantages. First, global indices suggested adequate model fit for the 1-PL, which minimizes the need for post hoc model modifications. Furthermore, the nested model comparison suggested that the constraints imposed by the 1-PL (i.e., equal discrimination parameters across items) did not significantly worsen model fit. Finally, in terms of the practical implications of using a 1-PL versus a 2-PL to model the VNT, the high agreement of ability estimates generated by the two models suggests that any misspecification is relatively benign. On the other hand, model revision based on modification indices is a data-driven approach, which makes it inherently susceptible to capitalizing on chance characteristics of the data. Therefore, model revisions resulting from such an approach may fail to generalize to other samples or to the population (MacCallum et al., 1992). Such risk increases with relatively small samples, as is the case in this study. Thus, while it is reasonable that the discrimination parameters of some items (such as "give") should be freed to better match the patterns in the observed data, it would be optimal to test this hypothesis in a larger independent sample of this population.

However, while the use of the 1-PL might be adequate for modeling the VNT and generating ability estimates, investigating the performance of the two model specifications in the context of computer adaptive testing may further motivate the selection of a 2-PL. Specifically, in computer adaptive testing, an algorithm is typically used to select the next item to administer based on response patterns from previously administered items. To do so, commonly used algorithms focus on the information curves of available items and select items for which information is maximized at the interim ability level that is calculated based on all of the previous responses. Given that the information function is a function of discrimination, it is plausible that refined estimates of item discriminations may lead to improved performance in terms of item selection.

Local Independence

The vast majority of pairwise combinations met the assumption of local independence, thereby justifying the use of standard IRT modeling procedures. A post hoc evaluation of the two items pairs (i.e., "bite" – "shove," "send" – "wash") with residual correlations > .284 revealed no overt concerns about dependencies. Given that the correlations (i.e., .31 and .36, respectively) were marginally above the prespecified cutoff, it may be that any dependence is not substantial enough to yield biased estimates of precision (i.e., standard errors on ability that are unrealistically small). The

results of the confirmatory factor analyses, where a unidimensional model was shown to fit the data adequately without specifying free correlations among items, further support the view that any item pair dependencies are likely negligible, as multidimensionality is a common cause for systematic residual pairwise item correlations. However, another cause may be the presence of distinct subgroups within our sample, where individuals within that group have response patterns that systematically differ from that of the reference group. If such subgroups were present, this would consequently lead to systematic subgroup differences in item parameter and ability estimates. Although outside the scope of this study, the quantification of systematic performance differences among subgroups is possible using differential item functioning and is a potential future direction for this line of work.

Precision and Reliability

The analyses based on the test information function, the *SEM*, overall reliability, and conditional reliability provide converging evidence that the VNT is a relatively reliable measure for the assessment of action naming.

Here, each analysis sheds light to unique measurement aspects and serves complementary purposes in understanding how the VNT test behaves, how it can be used, and how it can be further refined. First, the test information function forms the basis for the calculation of a test's precision and is integral for the estimation of standard errors of measurement and, by extension, CIs. These metrics are well-described in the clinical literature and inherent to most standardized behavioral tests. As can be seen in Figure 3, IRT-based *SEM* depends on the ability level, and unlike the *SEM* typically derived from CTT, it is not constant. The least amount of error occurs where the test information function peaks with measurement error increasing progressively in regions that lack items of corresponding difficulty. The CI of an ability estimate can be calculated using the formula 95% CI(θ) = Ability(θ) ± (1.96 × *SEM*(θ)). For example, using this formula, the 95% CI for the lowest ability estimate in our sample (θ = −2.01), which most likely corresponds to the most severely impaired individual tested, is [−3.02, −1.01] on a scale that typically ranges from −4 to 4. On the other hand, the 95% CI of an ability estimate of θ = 0.48, which is near the region for which the test information peaks, is almost half and equal to [−0.11, 1.07].

These findings establish the precision of the VNT but also illustrate its contingency on ability level. Put another way, the test information function can be used to understand the region of ability for which the test lacks informative items. Thus, it can serve as a map for further expanding or revising a test, to better measure performance among individuals along the whole ability continuum or within a targeted ability range. For example, in the context of the VNT, if one were to want to optimize precision at the extreme ends of the ability continuum, where precision for ability estimates was relatively lower than for those in the middle of the distribution, additional items could be added to better target performance within that ability range.

Beyond information derived from the test information function, reliability indices provide another layer of interpretation that, perhaps, may be the most meaningful for clinical audiences (Nicewander, 2018). Such indices are bounded quantities (0–1) and are on the same metric as the traditional reliability indices that clinicians are typically exposed to during their clinical training. Categorical omega, and marginal and empirical estimates of reliability can serve as summary statistics of the performance of a given test. Conditional reliabilities on the other hand would seem to be of even higher practical value since they can be interpreted as the familiar indicators of the precision of measurement yet still reflect the regions of ability for which measurement if more precise. On this metric, ability estimates for the vast majority of participants (100 out of 107) had a conditional reliability coefficient of > .8. From a practical standpoint, this finding, coupled with strong evidence of unidimensionality, suggests that IRT-derived test scores of the VNT are a psychometrically sound metric for quantifying action naming for applications that require at least that level of precision. This is for individuals with aphasia whose ability estimates range from −1.63 to 1.83. For those whose ability estimate falls outside this range, we recommend caution in drawing inferences based on VNT performance. In this study, which used participants from AphasiaBank, the majority of individuals outside this range fell on the more severe end of the aphasia continuum, although this may not necessarily be the case in another sample.

As alluded to previously, the same test modeled under both CTT and IRT can yield differing metrics of precision and reliability, which has significant implications for clinical practice and applied research. For example, the measurement of change is necessitated on specifying a score's 95% CI, as based on its *SEM*. Under CTT, where measurement error is assumed to be constant across the ability continuum, CIs become unrealistically narrow for individuals on the extremes and overly wide for those in the middle of the ability distribution. This in turn can invalidate assessments of change; specifically, change score estimates for very mildly and severely impaired individuals would have an inflated Type I error rate, whereas those for individuals in the middle would have increased Type II error rate.

Conceptual differences between CTT and IRT, as well as common statistical concepts of the latter (e.g.,

information), may not be familiar concepts to key stakeholders. This lack of familiarity may lead to a lack of awareness around potential psychometric limitations of behavioral measures and may additionally act as a barrier to the implementation of precise assessment tools or psychometric frameworks such as IRT in aphasiology. Report of IRT reliability indices (e.g., marginal and empirical reliability) in addition to more traditional IRT metrics, as done in this study, may help bridge this knowledge gap. Reliability indices are more likely to be meaningfully interpretable for clinical audiences (Nicewander, 2018), as they are bounded quantities (0–1) and are on the same metric as CTT-based reliability indices that clinicians are typically exposed to during their clinical training.

Limitations and Future Directions

In modern psychometrics, validation is a special case of evidentiary reasoning that requires evidence from multiple different sources to support the intended interpretation of test scores in specific measurement situations (Lissitz, 2009; Mislevy, 2006; Zumbo, 2007). In this study, we gathered initial evidence (e.g., unidimensionality) that supports the construct validity of the VNT. However, the interpretation of the VNT scores depends on, as of yet, untested premises about the cognitive processes used by test takers while responding to the VNT. According to the developers of the larger NAVS, all of its subtests, including the VNT, are purportedly an index of morphosyntactic processing (Cho-Reyes & Thompson, 2012). Some evidence for this claim comes from two sources. First, item development and scoring procedures were designed to hold constant other related subcomponents of word and sentence processing (lexical semantics, phonological encoding). Second, NAVS test scores were shown to discriminate a priori–defined groups of agrammatic and nonagrammatic speakers (Cho-Reyes & Thompson, 2012). Therefore, one future direction is to use cognitive and psycholinguistic theory to assess the extent to which the construct that is captured by VNT (i.e., the enacted construct; Cho-Reyes & Thompson, 2012) is the construct the test developers had in mind when they developed the test (i.e., intended construct, in this case morphosyntactic processing; Gorin & Embretson, 2006). Relatedly, another future direction is to evaluate the extent to response to items on the VNT is equivalent to response to items on picture naming tests of nouns, such as the PNT (Roach et al., 1996). Although prior research suggests that successful naming of verbs is contingent on item properties (i.e., imageability) distinct from item properties necessary to successful naming of nouns (i.e., frequency, age of acquisition, phoneme length; see Fergadiotis, Swiderski, & Hula, 2019; Mätzig et al., 2009; Szekely et al., 2005), both tests of verb and noun naming may ultimately be measures of anomia (i.e., word production impairments) and, as such, could potentially be equated within a single testing framework. However, further psychometric investigation using a rigorous framework such as IRT is needed in order to determine what construct underlies both verb and noun naming tests.

Conclusions

The VNT is a unidimensional measure that can be productively modeled under IRT and with a relatively high degree of precision. Reliability indices suggest test-level scores are reliable under both CTT and IRT, although the latter suggests that scores from individuals with notably mild or severe action naming impairment are relatively less reliable. Current users of the VNT should have confidence in interpreting scores derived under a 1-PL IRT model or a total-correct CTT model as valid estimates of verb naming ability for the purposes of determining overall severity and ranking individuals. Future work is needed to evaluate item properties of the VNT under IRT, as well as the degree to which an action naming task is reflective of morphosyntactic processing.

Data Availability Statement

This study relied on archival recordings from AphasiaBank (MacWhinney et al., 2011), which are freely available to the scientific community. Furthermore, the scored responses on the VNT used in this study are also available on AphasiaBank.

Acknowledgments

This work was funded by NIH/NIDCD Grant 1R01DC018813 (PIs: Hula & Fergadiotis).

References

Alons, E., Dijkhuis, L., van Tuijl, P., & van Ewijk, L. (2022). Development and diagnostic accuracy of a shortened Dutch naming test for people with aphasia using item response theory. *Archives of Clinical Neuropsychology, 37*(8), 1735–1748. https://doi.org/10.1093/arclin/acac057

Alyahya, R. S. W., Halai, A. D., Conroy, P., & Lambon Ralph, M. A. (2018). Noun and verb processing in aphasia: Behavioural profiles and neural correlates. *NeuroImage: Clinical, 18,* 215–230. https://doi.org/10.1016/j.nicl.2018.01.023

Babbitt, E. M., Heinemann, A. W., Semik, P., & Cherney, L. R. (2011). Psychometric properties of the Communication Confidence Rating Scale for Aphasia (CCRSA): Phase 2. *Aphasiology, 25*(6–7), 727–735. https://doi.org/10.1080/02687038.2010.537347

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

Bastiaanse, R., & Jonkers, R. (1998). Verb retrieval in action naming and spontaneous speech in agrammatic and anomic aphasia. *Aphasiology, 12*(11), 951–969. https://doi.org/10.1080/02687039808249463

Bates, E., Chen, S., Tzeng, O., Li, P., & Opie, M. (1991). The noun–verb problem in Chinese aphasia. *Brain and Language, 41*(2), 203–233. https://doi.org/10.1016/0093-934X(91)90153-R

Baylor, C., Eadie, T., & Yorkston, K. (2021). The Communicative Participation Item Bank: Evaluating, and re-evaluating, its use across communication disorders in adults. *Seminars in Speech and Language, 42*(3), 225–239. https://doi.org/10.1055/s-0041-1729947

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)

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (2000). Rites, wrongs, and gold in model testing. *Structural Equation Modeling: A Multidisciplinary Journal, 7*(1), 82–91. https://doi.org/10.1207/S15328007SEM0701_04

Berndt, R. S., Haendiges, A. N., Mitchum, C. C., & Sandson, J. (1997). Verb retrieval in aphasia. 2. Relationship to sentence processing. *Brain and Language, 56*(1), 107–137. https://doi.org/10.1006/brln.1997.1728

Berndt, R. S., Mitchum, C. C., Haendiges, A. N., & Sandson, J. (1997). Verb retrieval in aphasia. 1. Characterizing single word impairments. *Brain and Language, 56*(1), 68–106. https://doi.org/10.1006/brln.1997.1727

Bradley, D. C., Garrett, M. E., & Zurif, E. B. (1980). Syntactic deficits in Broca's aphasia. In D. Caplan (Ed.), *Biological studies of mental processes*. MIT Press.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Caramazza, A., & Hillis, A. (1991). Lexical organization of nouns and verbs in the brain. *Nature, 349,* 788–790. https://doi.org/10.1038/349788a0

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113*(2), 234–272. https://doi.org/10.1037/0033-295X.113.2.234

Cho-Reyes, S., & Thompson, C. K. (2012). Verb and sentence production and comprehension in aphasia: Northwestern Assessment of Verbs and Sentences (NAVS). *Aphasiology, 26*(10), 1250–1277. https://doi.org/10.1080/02687038.2012.693584

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q_3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement, 41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement, 41*(6), 439–455. https://doi.org/10.1177/0146621617695522

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

de Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer. https://doi.org/10.1007/978-1-4757-3990-9

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review, 104*(4), 801–838. https://doi.org/10.1037/0033-295X.104.4.801

del Toro, C. M., Bislick, L. P., Comer, M., Velozo, C., Romero, S., Gonzalez Rothi, L. J., & Kendall, D. L. (2011). Development of a short form of the Boston Naming Test for individuals with aphasia. *Journal of Speech, Language, and Hearing Research, 54*(4), 1089–1100. https://doi.org/10.1044/1092-4388(2010/09-0119)

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*(3), 363–373. https://doi.org/10.1037/0021-9010.68.3.363

Edmonds, L. A., Nadeau, S. E., & Kiran, S. (2009). Effect of Verb Network Strengthening Treatment (VNeST) on lexical retrieval of content words in sentences in persons with aphasia. *Aphasiology, 23*(3), 402–424. https://doi.org/10.1080/02687030802291339

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Erlbaum.

Fergadiotis, G., Casilio, M., Hula, W. D., & Swiderski, A. (2021). Computer adaptive testing for the assessment of anomia severity. *Seminars in Speech and Language, 42*(3), 180–191. https://doi.org/10.1055/s-0041-1727252

Fergadiotis, G., Gorman, K., & Bedrick, S. (2016). Algorithmic classification of five characteristic types of paraphasias. *American Journal of Speech-Language Pathology, 25*(4S), S776–S787. https://doi.org/10.1044/2016_AJSLP-15-0147

Fergadiotis, G., Hula, W. D., Swiderski, A. M., Lei, C.-M., & Kellough, S. (2019). Enhancing the efficiency of confrontation naming assessment for aphasia using computer adaptive testing. *Journal of Speech, Language, and Hearing Research, 62*(6), 1724–1738. https://doi.org/10.1044/2018_JSLHR-L-18-0344

Fergadiotis, G., Kellough, S., & Hula, W. D. (2015). Item response theory modeling of the Philadelphia Naming Test. *Journal of Speech, Language, and Hearing Research, 58*(3), 865–877. https://doi.org/10.1044/2015_JSLHR-L-14-0249

Fergadiotis, G., Swiderski, A. M., & Hula, W. D. (2019). Predicting confrontation naming item difficulty. *Aphasiology, 33*(6), 689–709. https://doi.org/10.1080/02687038.2018.1495310

German, D. J. (1990). *National College of Education Test of Adolescent/Adult Word Finding (TAWF)* [Test book]. DLM Teaching Resources.

Goodglass, H., & Kaplan, E. (1983). *The assessment of aphasia and related disorders* (2nd ed.). Lea & Febiger.

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

Gordon, J. K., & Dell, G. S. (2003). Learning to divide the labor: An account of deficits in light and heavy verb production. *Cognitive Science, 27*(1), 1–40. https://doi.org/10.1207/s15516709cog2701_1

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30*(5), 394–411. https://doi.org/10.1177/0146621606288554

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*(1), 155–167. https://doi.org/10.1007/s11336-008-9099-3

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hula, W. D., Donovan, N. J., Kendall, D. L., & Gonzalez-Rothi, L. J. (2010). Item response theory analysis of the Western

Aphasia Battery. *Aphasiology, 24*(11), 1326–1341. https://doi.org/10.1080/02687030903422502

Hula, W. D., & Doyle, P. (2021). The Aphasia Communication Outcome Measure: Motivation, development, validity evidence, and interpretation of change scores. *Seminars in Speech and Language, 42*(03), 211–224. https://doi.org/10.1055/s-0041-1730906

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–Second Edition*. Lippincott Williams & Wilkins.

Kelley, K. (2022). *MBESS: The MBESS R Package* (4.9.0) [Computer software]. https://CRAN.R-project.org/package=MBESS

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods, 21*(1), 69–92. https://doi.org/10.1037/a0040086

Kertesz, A. (2007). *Western Aphasia Battery–Revised*. Grune & Stratton.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). Guilford Press.

Kohn, S. E., & Goodglass, H. (1985). Picture-naming in aphasia. *Brain and Language, 24*(2), 266–283. https://doi.org/10.1016/0093-934X(85)90135-X

Laiacona, M., & Caramazza, A. (2004). The noun/verb dissociation in language production: Varieties of causes. *Cognitive Neuropsychology, 21*(2–4), 103–123. https://doi.org/10.1080/02643290342000311

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(1), 1–75. https://doi.org/10.1017/S0140525X99001776

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33*(7), 519–537. https://doi.org/10.1177/0146621608329504

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Lissitz, R. W. (2009). *The concept of validity: Revisions, new directions, and applications*. IAP.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

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

Loverso, F. L., Selinger, M., & Prescott, T. E. (1979, May 28-31). *Application of verbing strategies to aphasia treatment* [Paper presentation]. Clinical Aphasiology Conference, Phoenix, AZ.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Mätzig, S., Druks, J., Masterson, J., & Vigliocco, G. (2009). Noun and verb differences in picture naming: Past studies and new evidence. *Cortex, 45*(6), 738–758. https://doi.org/10.1016/j.cortex.2008.10.003

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed., Vol. 1). Erlbaum.

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). Aphasiabank: Methods for studying discourse. *Aphasiology, 25*(11), 1286–1307. https://doi.org/10.1080/02687038.2011.589893

Miceli, G., Silveri, M. C., Villa, G., & Caramazza, A. (1984). On the basis for the agrammatic's difficulty in producing main verbs. *Cortex, 20*(2), 207–220. https://doi.org/10.1016/S0010-9452(84)80038-6

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). American Council on Education/Praeger Publishers.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide*. Muthén & Muthén.

Nicewander, W. A. (2018). Conditional precision of measurement for test scores: Are conditional standard errors sufficient? *Educational and Psychological Measurement, 79*(1), 5–18. https://doi.org/10.1177/0013164418758373

Nickels, L. (2002). Therapy for naming disorders: Revisiting, revising, and reviewing. *Aphasiology, 16*(10–11), 935–979. https://doi.org/10.1080/02687030244000563

Nickels, L. (2014). *Spoken word production and its breakdown in aphasia* (Reprint edition). Psychology Press. https://doi.org/10.4324/9781315804620

Rizopoulos, D. (2007). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25. https://doi.org/10.18637/jss.v017.i05

Roach, A., Schwartz, M., Martin, N., Grewal, R., & Brecher, A. (1996). The Philadelphia Naming Test: Scoring and rationale. *Clinical Aphasiology, 24,* 121–133. https://doi.org/10.1037/t56477-000

Robitzsch, A. (2022). *sirt: Supplementary Item Response Theory Models* (Version 3.12-66) [Computer software]. https://CRAN.R-project.org/package=sirt

Rofes, A., Capasso, R., & Miceli, G. (2015). Verb production tasks in the measurement of communicative abilities in aphasia. *Journal of Clinical and Experimental Neuropsychology, 37*(5), 483–502. https://doi.org/10.1080/13803395.2015.1025709

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Schwartz, M. F., Dell, G. S., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and Language, 54*(2), 228–264. https://doi.org/10.1016/j.jml.2005.10.001

Segalowitz, S. J., & Lane, K. C. (2000). Lexical access of function versus content words. *Brain and Language, 75,* 376–389. https://doi.org/10.1006/brln.2000.2361

Sörbom, D. (1989). Model modification. *Psychometrika, 54*(3), 371–384. https://doi.org/10.1007/BF02294623

Steiger, J. H., & Lind, J. C. (1980, May 28). *Statistically-based tests for the number of common factors* [Paper presentation]. Annual Spring Meeting of the Psychometric Society, Iowa City, IA.

Szekely, A., D'Amico, S., Devescovi, A., Federmeier, K., Herron, D., Iyer, G., Jacobsen, T., Arévalo, A. L., Vargha, A., & Bates, E. (2005). Timed action and object naming. *Cortex, 41*(1), 7–25. https://doi.org/10.1016/S0010-9452(08)70174-6

Thompson, C. K., & Shapiro, L. (2005). Treating agrammatic aphasia within a linguistic framework: Treatment of underlying forms. *Aphasiology, 19*(10–11), 1021–1036. https://doi.org/10.1080/02687030544000227

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika, 70*(4), 629–650. https://doi.org/10.1007/s11336-000-0810-3

von Davier, M., Cho, Y., & Pan, T. (2019). Effects of discontinue rules on psychometric properties of test scores. *Psychometrika, 84*(1), 147–163. https://doi.org/10.1007/s11336-018-09652-3

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511618765

Walker, G. M., Basilakos, A., Fridriksson, J., & Hickok, G. (2022). Beyond percent correct: Measuring change in individual picture naming ability. *Journal of Speech, Language, and Hearing Research, 65*(1), 215–237. https://doi.org/10.1044/2021_JSLHR-20-00205

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)

Wambaugh, J. L., & Ferguson, M. (2007). Application of semantic feature analysis to retrieval of action names in aphasia. *Journal of Rehabilitation Research and Development, 44*(3), 381–394. https://doi.org/10.1682/jrrd.2006.05.0038

Willmes, K. (1981). A new look at the Token Test using probabilistic test models. *Neuropsychologia, 19*(5), 631–646. https://doi.org/10.1016/0028-3932(81)90001-4

Willmes, K. (2003). Psychometric issues in aphasia therapy research. In I. Papathanasiou & R. De Bleser (Eds.), *The sciences of aphasia: From theory to therapy* (pp. 227–244). Pergamon.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Erlbaum.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125–145. https://doi.org/10.1177/014662168400800201

Zumbo, B. (2007). Validity: Foundational issues and statistical methodology. *Handbook of Statistics, 26*(6), 45–70.

Appendix A (p. 1 of 5)

Verb Naming Test Transcription Information and Coding Conventions

Transcription Procedure

Participant responses were phonemically transcribed by two research assistants at Portland State University in a pseudorandom order. In addition, since it was thought that participant responses to a verb confrontation naming test may elicit more multiword responses than a noun confrontation naming test, research assistants transcribed everything the participant said/gestured in response to the stimuli and used a set of transcription coding conventions meant to capture elements of nonfluent speech.

Phonemic transcriptions were broad, and variations in dialect were transcribed as they were heard using the phonemic notation below. If a given production strayed from our lab's phonemic conventions, as was the case for some British dialects, that production was translated into Standard American English and transcribed in accordance with our conventions. Phoneme notation followed conventions developed by our laboratory for the purposes of use with a computer algorithm. Please see below for our phoneme conventions, a list of target phonemic transcriptions, and a list of transcription coding conventions.

Transcription Conventions

Phoneme Notation

See the table below for a list of the phoneme notations used by our laboratory, as well as lists of examples.

Table A1. Phoneme annotation.

IPA	Examples
/p/	"**p**at"
/b/	"**b**at"
/t/	"**t**en"
/d/	"**d**en"
/ɾ/	"bu**tt**er" (flap - allophone of /t, d/)
/k/	"**c**oat"
/g/	"**g**oat"
/f/	"**f**an"
/v/	"**v**an"
/θ/	"**th**in" (voiceless)
/ð/	"**th**en" (voiced)
/s/	"**s**ee"
/z/	"**z**oo"
/ʃ/	"**sh**oe"
/ʒ/	"occa**s**ion"
/tʃ/	"**ch**urch"
/dʒ/	"ju**dg**e"
/m/	"**m**an"
/n/	"**n**ose"
/ŋ/	"si**ng**"
/ɹ/	"**r**ed"
/l/	"**l**ate"
/w/	"**w**in"
/j/	"**y**es"
/h/	"**h**at"
/ʔ/	"co**tt**on" (glottal stop - allophone of /t/)
/i/	"sh**e**"
/æ/	"c**a**t"
/ɛ/	"r**e**d"
/ɪ/	"f**i**t"
/u/	"b**oo**t"
/ʊ/	"w**oo**d"
/ɔ/	"d**aw**n"

(*table continues*)

Appendix A (p. 2 of 5)

Verb Naming Test Transcription Information and Coding Conventions

IPA	Examples
/ɑ/	"n**o**t"
/ʌ/	"b**u**t" (stressed)
/ə/	"**a**lone" (unstressed)
/ɝ/	"h**ear**d" (stressed)
/ɚ/	"p**er**haps" (unstressed)
/aɪ/	"k**i**te"
/aʊ/	"c**ow**"
/ɔɪ/	"b**oy**"
/eɪ/	"st**a**te"
/oʊ/	"v**o**te"
/ɪɹ/	"d**eer**"
/ɔɹ/	"d**oor**"
/ɑɹ/	"d**ar**k"
/eɹ/	"d**are**"
/ʊɹ/	"c**ure**"

Note. IPA = International Phonetic Alphabet.

Target Transcriptions

See the table below for the phonemic transcription of the targets.

Table A2. Phonemic transcriptions of the targets.

VNT item	IPA target
Cut	kʌt
Bark	bɑɹk
Put	pʊt
Send	sɛnd
Drive	dɹaɪv
Wash	wɑʃ
Read	ɹid
Laugh	læf
Watch	wɑʧ
Give	gɪv
Swim	swɪm
Stir	stɝ
Pinch	pɪnʧ
Crawl	kɹɔl
Deliver	dɪlɪvɚ
Pour	pɔɹ
Howl	haʊl
Throw	θɹoʊ
Bite	baɪt
Shove	ʃʌv
Tickle	tɪkəl
Shave	ʃeɪv

Note. VNT = Verb Naming Test; IPA = International Phonetic Alphabet.

Transcription Coding Conventions

See the table below for a list of transcription coding conventions adopted from the CHAT manual (MacWhinney, 2000) for the purposes of this study.

Table A3. Transcription coding conventions.

Coding convention	Definitions/examples
Fillers &-	Fillers or filled pauses (e.g., "um," "uh," "hmm") were written orthographically preceded by &-.
Communicators	Communicators (e.g., "oh," "okay," "yeah") were orthographically transcribed without any additional notation. This list of communicators created by Brian MacWhinney and Mitzi Morris, accessed from talkbank.org, served as a reference for identifying communicators and their standardized spellings.
Phonological fragments &+	Phonological fragments or false starts, consisting of one or two phonemes, were written orthographically, preceded by &+.
Letter sequence @k	Letter sequences were denoted using @k following the string of letters produced. For example, the spelling of the verb cut was written as cut@k.
Gestures & = ges:	Any movement of a body part meant to express an action was written as & = ges:action. For example, & = ges:cut
Sound Effects &=	Any nonword vocalization meant to express an action was written as & = action. For example, & = laughs or & = cries.
Unintelligible utterances xxx	Unintelligible utterances were written as xxx in place of the unintelligible word/phrase/paraphasia.
Repetition of single words [x N]	All one-word repetitions were written once followed by the code [x N] where N is the number of times the word was produced in total.
Repetition of phrases <> [/]	All multiword repetitions were written out. All but the last repetition was included in <> followed by [/].
Retracing <> [//]	All retracings or revised utterances were written with the first phrase in <> followed by the [//] code.
Pause (.)	Silent pauses between utterances lasting more than approximately 1 s were denoted as (.).

Data Annotation Procedure

Participant responses and investigator prompts were annotated by two research assistants at Portland State University in a pseudorandom order at the time of transcription. Disagreements in transcription between the two research assistants are being resolved by a research speech-language pathologist in a pseudorandom order.

The following data annotation conventions (Table A4) were used to characterize participant responses and investigator prompts for the purposes of scoring the Verb Naming Test.

Table A4. Response and prompt annotations.

Notation	Definitions/annotation instructions
Response 1[a]	Any verbal response the participant gives after being presented with the test item/first prompt from the test administrator.
Response 2[b]	Any and all subsequent verbal responses from the participant following a second prompt[c] from the test administrator. **If no response 2, leave blank**.
Delay 1	1 = yes, 0 = no, Yes if the time between the initial item presentation/prompt and the participant's first response (excluding any initial fragments/fillers) is more than 10 s. **If no response 1, leave blank**.
Delay 2	1 = yes, 0 = no, Yes if the time between the initial item presentation/prompt and the participant's first response (excluding any initial fragments/fillers) is more than 10 s. **If no response 2, leave blank**.
Multiword 1	1 = yes, 0 = no, Yes if the participant verbalizes more than one word, excluding fragments and fillers. **If no response 1, leave blank**.
Multiword 2	1 = yes, 0 = no, Yes if the participant verbalizes more than one word, excluding fragments and fillers. **If no response 2, leave blank**.
Additional prompts 1	1 = yes, 0 = no, Yes if the administrator provides more than one prompt before any first participant response.
Additional prompts 2	1 = yes, 0 = no, Yes if the administrator provides more than 1 prompt following a first incorrect response. **If no response 2, leave blank**.

(*table continues*)

Appendix A (p. 4 of 5)

Verb Naming Test Transcription Information and Coding Conventions

Notation	Definitions/annotation instructions
No response	NR, If the participant has no verbal response (excluding fragments/fillers) input NR in the corresponding response column. **If no opportunity for response 2, leave blank**.
Facilitator prompt 1	0 = no, SE = semantic only, SY = syntactic and semantic, *P* = phonemic, G = gestures, E = sound effect, A = answer for any prompts given prior to the first response. If more than 1 type is given, separate the codes by a single space.
Facilitator prompt 2	0 = no, SE = semantic only, SY = syntactic and semantic, *P* = phonemic, G = gestures, E = sound effect, A = answer for any prompts following a first incorrect response. If more than 1 type is given, separate the codes by a single space. **If no opportunity for response 2, leave blank**.
Item not administered	NA, Input NA into Response 1 and **leave all other fields blank**.

[a]In instances where the test administrator did not provide a first verbal prompt, we defined Response 1 as any verbal response the participant gives after being presented with the test item. [b]In instances where the test administrator did not provide a first verbal prompt, we defined Response 2 as any verbal response the participant gives after completing their first response and after being presented with a first prompt from the test administrator. [c]Gestures or other nonverbal cues from the test administrator indicating the target action were treated as second prompts if they occurred after the participant's first response.

Facilitator Prompt Coding Conventions

See the table below for types of facilitating prompts or cues given by investigators with examples and the coding conventions used by our laboratory to annotate the data.

Table A5. Facilitator prompt coding conventions.

Code	Type	Example
EP	Exact prompt from the manual	"What's happening" or "Tell me what's happening" or "Can you tell me another word for what's happening?"
AP	Approximate prompt according to the manual	"What's he doing?" or "What's she doing to him?" or "What's she doing with it?" or "Can you think of another word for what he's doing?"
SE	Semantic only	"Here is a book. What is happening?"
SY	Syntactic and semantic	"What is the boy doing to the girl?" or "What is the dog doing?"
O	Orienting to picture or picture part	Points to picture and/or "Right here" or "This part" or "I'm sorry I didn't hear you"
P	Phonemic	"Mm Mm" or "It starts with an M"
G	Gestures	The investigator makes a pinching gesture with their index finger and thumb.
E	Sound effect	The investigator barks like a dog.
A	Answer	The investigator verbalizes the correct response.

Appendix A (p. 5 of 5)

Verb Naming Test Transcription Information and Coding Conventions

Transcription Resolutions

In the interest of developing a universal computer-adaptive naming assessment for both verbs and nouns, first response transcriptions were prioritized over second response transcriptions for resolution and scoring. Disagreements in first response transcription between the two research assistants were resolved by two research speech-language pathologists in a pseudorandom order.

Time-Limited Transcriptions

A research assistant applied a 10- and 15-s time limit to create two time-limited sets of transcriptions, one in accordance with the time allowed to name an item on the Verb Naming Test and one in accordance with the time allowed to name an item on the universal assessment under development, respectively. These time-limited sets of transcriptions were generated to be used for scoring such that any transcribed response that took place after the permitted time limit would be absent from the time-limited transcription and therefore not considered for scoring.

Time-Limit Procedure

As a general rule, the timed naming window started the moment after the picture was shown to the participant and the test administrator completed their initial verbal prompt, with both conditions having to be true in order to start the clock.

This rule was adapted for cases where there was no first prompt from the administrator and/or the first prompt came shortly after the naming attempt began such that it overlapped with or interrupted the naming attempt. For example, if there was no verbal prompt, the clock started at the moment the picture was shown to the participant. If the participant started naming immediately after being shown the picture and the test administrator's initial verbal prompt did not interrupt the participant's flow of speech, the clock started at the moment the picture was shown to the participant and any transcribed response that took place prior to the verbal prompt was considered for scoring. If the participant started naming immediately after being shown the picture and the test administrator's initial verbal prompt did interrupt the participant's flow of speech, the clock started at the moment the test administrator completed their initial verbal prompt and any transcribed response that took place prior to the prompt was considered for scoring.

Appendix B

Verb Naming Test Scoring Information

Verb Naming Test Scoring Protocol

The Verb Naming Test (VNT), a subtest of the The Northwestern Assessment of Verbs and Sentences (NAVS), was scored according to the protocol with some minor modifications/expansions made for clarity. See https://aphasia.talkbank.org/protocol/english/materials-aphasia/VNT.pdf for VNT administration and scoring protocol. In brief, a response produced within 10 s of item administration was considered for scoring and self-corrections were accepted. Correct responses included the target verb in any form (e.g., swim, swam, swimming), phonemic paraphasias of the target verb not resulting in a real word, and verbs semantically similar to and with the same argument structure as the target. Incorrect responses included: phrasal verbs and real word phonemic paraphasias of the target verb.

Minor Modification

VNT scoring protocol allows for a second response following a first incorrect attempt and prompt from the test administrator. For the purposes of this study, only a participant's first response was considered for scoring and any subsequent response prompted by the test administrator, verbally or nonverbally, was not considered for scoring.

In another related study assessing the variability in VNT administration, the results revealed inconsistent testing administration across various sites that participate in APhasiaBank, particularly with prompts that deviated from the manual and contained syntactic and semantic information. When the examiner followed the rules and provided an additional exact or approximate prompt following an incorrect first attempt, it did not significantly affect the likelihood of a correct second attempt (only 9.43% correct second attempts were produced). Oftentimes an additional prompt was not even given after an incorrect first attempt due to examiner error (e.g., phrasal verbs produced or the examiner waited until the subject self-corrected < or > 10 s). Thus, we determined that only the first responses within 10 s should be considered for scoring.

Scoring Protocol Expansions

An expanded scoring protocol was developed in order to dispel some ambiguity in the VNT scoring protocol, specifically to operationalize (a) selection of the scored attempt in the context of multiword/multiverb responses and (b) application of the VNT's phonological similarity rule (i.e., "50% of phonemes must be correct").

Scored Attempt

The final main lexical verb produced as part of the first response was selected for scoring. Auxiliary verbs, verbs produced as personal commentary, and/or copula "to-be" verbs functioning as main lexical verbs were systematically ignored. Paraphasic responses that were phonologically similar to the target or contained inflectional morphemes (e.g., *–ing*, *–ed*, *–s*) were recognized as verb approximations.

Phonological Similarity

A verb approximation or paraphasia was judged to be phonologically similar to the target verb only if the number of correct phonemes (i.e., phonemes shared between the target and response) comprised 50% of the total phonemes present in the response and 50% of the total phonemes present in the target. In other words, the 50% correct criterion had to be met for both target and response in order for the attempt to be deemed a phonemic paraphasia of the target. Only the lemma version of the target and response were used when applying this phonological similarity rule, and any inflectional morphemes present in the response (e.g., *–ing*, *–ed*, *–s*) were ignored. The schwa phoneme was included in the shared phoneme count, and rhotic vowels were treated as vowel plus /r/.

Scoring Procedure

Phonemically transcribed participant first responses were scored by two undergraduate research assistants at Portland State University in a pseudorandom order in accordance with the expanded VNT scoring protocol. Phonemic transcriptions were recorded of the participant's first full response after being presented with the test item/first prompt from the test administrator but before a second prompt. Only responses produced within the 10-s time limit were considered for scoring in accordance with the VNT scoring rules. If the participant self-corrected within 10 s, the final response was scored. For more information on VNT transcription procedures, conventions, and definition of terms, see VNT transcription information. VNT practice items were excluded from the VNT transcription study and the present VNT scoring study.

Research assistants were instructed to (a) identify the final main lexical verb attempt selected for scoring, if applicable; (b) identify which final verb(s) were ignored according to the expanded protocol, if applicable; (c) assign one of seven ancillary verb codes to scored verb attempts; and (d) score the response as correct (1) or incorrect (0).

Scoring Resolutions

Research assistant disagreements were resolved by an ASHA-certified speech-language pathologist at the level of scored attempt, ignored attempt, binary score, and ancillary verb code. Following the VNT protocol, Appendix B of the NAVS was used to determine correct alternate responses judged to be semantically similar with the same verb argument structure to the target. Verb attempts judged to be phrasal verbs were ultimately scored and resolved by two research speech-language pathologist s and two PhD-level linguists.