Published in final edited form as:

Aphasiology. 2019 ; 33(6): 689–709. doi:10.1080/02687038.2018.1495310.

Predicting confrontation naming item difficulty

Gerasimos Fergadiotis,

Portland State University, Portland, OR

Alexander Swiderski,

University of Pittsburgh and VA Pittsburgh Healthcare System, Pittsburgh, PA

William D. Hula

VA Pittsburgh Healthcare System and University of Pittsburgh, Pittsburgh, PA

Abstract

Background: Item response theory (IRT; Lord & Novick, 1968) is a psychometric framework that can be used to model the likelihood that an individual will respond correctly to an item. Using archival data (Mirman et al., 2010), Fergadiotis, Kellough, and Hula (2015) estimated difficulty parameters for the Philadelphia Naming Test (PNT; Roach, Schwartz, Martin, Grewal, & Brecher, 1996) using the 1-parameter logistic IRT model. Although the use of IRT in test development is advantageous, its reliance on sample sizes exceeding 200 participants make it difficult to implement in aphasiology. Therefore, alternate means of estimating the item difficulty of confrontation naming test items warrant investigation. In a preliminary study aimed at automatic item calibration, Swiderski, Fergadiotis, and Hula (2016) regressed the difficulty parameters from the PNT on word length, age of acquisition (Kuperman et al., 2012), lexical frequency as quantified by the Log10CD index (Brysbaert & New, 2009), and naming latency (Székely et al., 2003). Although this model successfully explained a substantial proportion of variance in the PNT difficulty parameters, a substantial proportion (20%) of the response time data were missing. Further, only 39% of the picture stimuli from Székely and colleagues (2003) were identical to those on the PNT. Given that the IRT sample size requirements limit traditional calibration approaches in aphasiology and that the initial attempts in predicting IRT difficulty parameters in our pilot study were based on incomplete response time data this study has two specific aims.

Aims: To estimate naming latencies for the 175 items on the PNT, and assess the utility of psycholinguistic variables and naming latencies for predicting item difficulty.

Methods and Procedures: Using a speeded picture naming task we estimated mean naming latencies for the 175 items of the Philadelphia Naming test in 44 cognitively healthy adults. We then re-estimated the model reported by Swiderski et al (2016) with the new naming latency data.

Results: The predictor variables described above accounted for a substantial proportion of the variance in the item difficulty parameters (*Adj. R*² = .692).

Conclusions: In this study we demonstrated that word length, age of acquisition, lexical frequency, and naming latency from neurotypical young adults usefully predict picture naming

item difficulty in people with aphasia. These variables are readily available or easily obtained and the regression model reported may be useful for estimating confrontation naming item difficulty without the need for collection of response data from large samples of people with aphasia.

Keywords

confrontation naming; item response theory; aphasia; assessment; psychometrics; reaction time

Anomia is a hallmark characteristic of people with aphasia (PWA) and is a manifestation of impaired cognitive processes that support word access and retrieval (Goodglass & Wingfield, 1997). Breakdowns in word access and retrieval can be evaluated with confrontation picture naming tests in which patients are presented with a picture of a noun and asked to name it. Such tests are routinely used for assessing anomic deficits because of their high reliability, their ability to yield rich diagnostic information, as well as their predictive utility in terms of lexical retrieval performance in conversational speech and overall aphasia severity (Herbert, Hickin, Howard, Osborne, & Best, 2008). Currently, confrontation picture naming tests are used extensively as part of large assessment batteries and standalone assessments in both research and clinical settings (Kaplan, Goodglass, & Weintraub, 2001; Kertesz, 1982; Roach, Schwartz, Martin, Grewal, & Brecher, 1996; Swinburn, Porter, & Howard, 2004). However, the use of multiple confrontation naming tests hinders accurate and meaningful comparisons of assessment results acquired with different testing instruments. Additionally, none of the currently available tools are well suited for multiple repeated measurements. Further, because these are static tests, each must be given in its entirety to obtain diagnostic information, thus contributing to increased testing burden, and patients' fatigue and frustration. This investigation is focused on the initial steps of a line of research whose overarching goal is to address these issues.

Item Response Theory

Item response theory (IRT) (Lord & Novick, 1968) is a psychometric framework expressed as a series of statistical tools that can be used to model the likelihood that an individual will correctly respond to each item of a test as a function of both the ability level of the individual and the derived difficulty parameter of an item (Ayala, 2009; Embretson & Reise, 2000). The 1-parameter logistic (1-PL) IRT model can be expressed as:

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

where $P(x_i = 1 | \theta_j, \alpha, \delta_i)$ is the probability of a correct response ($x_i = 1$) by examinee j , given person ability θ_j ; α is the discrimination parameter (constant across items); and δ_i is item i 's difficulty parameter. Person ability and item difficulty describe the location of the participant and items, respectively, along the same dimension. People with higher ability are more likely to correctly name a given item, while items of higher difficulty are more likely to elicit an incorrect response. Calibrating items (i.e. estimating their parameters) is typically done by fitting the model to a large dataset and evaluating its consistency with the empirical data.

Once stable difficulty parameters have been estimated, they can be used to convert a person's observed responses to a standardised metric, regardless of which specific items a person was administered. That is, IRT-based person score estimates are independent of the particular items administered and different sets of items (i.e. alternate forms) may be given to different examinees with scores that are expressed on a single common scale.

A central concept in IRT is information, which formalizes the notion that items that are too easy or too hard for a given patient provide very limited information for ability estimation. Item information for the 1-PL model is computed as $I_j(\theta) = \alpha_j^2 P_j(1 - P_j)$ and is maximized at the location of the difficulty parameter δ_j of each item. Then, information can be summed across the items of a test (i.e., $I(\theta) = \sum I_j(\theta)$) to reflect the total information of a testing instrument as a function of a person's ability (Ayala, 2009). One of the primary motivations behind the use of IRT is that information functions facilitate the calculation of standard error (*SE*). In the IRT framework, the *SE* of scores is conditional on a person's ability estimate. *SE* is defined as the reciprocal of test information at a given θ , $SE(\theta) = 1/\sqrt{I(\theta)}$ and with this formal expression a clinician can derive ability scores with accurate customized *SE*s contingent upon a person's score. This is in contrast to classical test theory (CTT) applications that assume measurement error is uniform regardless of the ability level of a patient. Therefore, when calculating *SE* with CTT, precision is typically either over- or underestimated depending on the composition of the items in a test and the ability level of the patient. Also, the fact that IRT models express the precision of score estimates as a function of individual ability level permits the test user to dynamically select the items that will contribute the most to the precision of a given examinee's score estimate thus effectively reducing the length of testing forms while minimizing the impact on reliability.

IRT in Aphasiology

IRT is widely utilized within rehabilitation measurement (Eadie et al., 2006) and patient reported outcomes (Cella et al., 2010). In the past decade, IRT has been implemented in clinical research and patient reported assessments (Babbitt, Heinemann, Semik, & Cherney, 2011; Beltyukova, Stone, & Ellis, 2008; Donovan, Velozo, & Rosenbek, 2007; Edmonds & Donovan, 2012) in communication science and disorders.

More recently, using archival data (Mirman et al., 2010) from the Philadelphia Naming Test (Roach et al., 1996), Fergadiotis, Kellough, and Hula (2015) estimated a 1-PL model and assessed the model's assumptions and its fit to the data. After concluding that the fit of the 1-PL model was adequate, Fergadiotis and colleagues (2015) derived the difficulty parameters of the items using traditional IRT methods. To collect validity evidence, they also accounted for a substantial proportion of variance of the difficulty parameters using lexical frequency measured by the Log10CD index (Brysbaert & New, 2009), Age of Acquisition (AoA; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), and length in phonemes, $R^2 = .63$, $p < .001$.

Although the use of IRT in test development is advantageous, it comes with costs that make it challenging to implement in aphasiology. To estimate item parameters and evaluate the

tenability of model assumptions, a sample size exceeding 200 participants is typically needed (Ayala, 2009; Embretson & Reise, 2000). The PNT item difficulty parameters (Fergadiotis et al., 2015) were estimated using a database with over 250 individuals' results on the PNT (Mirman et al., 2010) accumulated over a number of years. The Moss Aphasia Psycholinguistic Project Database (MAPPD; Mirman et al., 2010) was an essential contributor to the success of quantifying the difficulty parameters of the PNT; however, similar large databases are not readily available for other testing instruments. Calibration of other tests via traditional methods would require collection of item response data from large numbers of participants with aphasia and might be prohibitive in terms of both time and cost.

For the reasons suggested above, alternate approaches to estimating item difficulties warrant investigation. One of the recommendations by Fergadiotis and colleagues (2015), consistent with current trends in cognitive psychometrics (Embretson & Reise, 2000; Gierl & Haladyna, 2012), was to further extend the regression model they presented. This regression model established a starting point for augmented (Matteucci, Mignani, & Veldkamp, 2012) or automatic item calibration (Gierl & Haladyna, 2012) by predicting the difficulty parameter for any confrontation naming test item from its psycholinguistic characteristics. If the model could be improved to account for a larger proportion of variance in item difficulty, we could establish a foundation for the automatic calibration of test items. This would minimize the burden of IRT's test creation costs by eliminating the need for large calibration samples for estimating item difficulty parameters.

In a follow up study, Swiderski, Fergadiotis, and Hula (2016) assessed the ability of several variables to predict item difficulty. The initial set of variables included naming response time (Székely et al., 2003), log10 contextual diversity from the Subtlex database as a proxy for word frequency (Brysbaert & New, 2009), lexical density (Mirman et al., 2010), AoA (Kuperman et al., 2012), number of phonemes, name agreement (Snodgrass & Vanderwart, 1980), familiarity, and visual complexity (Székely & Bates, 2000) as predictor variables (Swiderski et al., 2016). Through backwards regression, a significant proportion of variance was accounted for with all eight predictor variables, $R^2 = .73$, $Adj. R^2 = .70$ $p < .001$. However, the inclusion of only lexical frequency, naming response time, length in phonemes and AoA yielded a more parsimonious model with similar predictive power, $R^2 = .73$, $Adj. R^2 = .71$ $p < .001$.

As noted below, the primary purpose of the current study is to develop a predictive model for naming item difficulty in the context of a unidimensional IRT model. However, to the extent that predictors can be explicitly related to a productive theoretical account of word production, it will facilitate the eventual development of a more elaborated and diagnostically useful measurement model. There is consensus in the literature that the cognitive system for word production operates across at least two distinct stages (Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Harley & Bown, 1998; Levelt, Roelofs, & Meyer, 1999; Rapp & Goldrick, 2000; Schwartz, Dell, Martin, Gahl, & Sobel, 2006). In the first stage, the system retrieves a semantically specified lexical representation based on a previously activated conceptual representation. In the most influential accounts (Foygel &

Dell, 2000; Levelt et al., 1999; Schwartz et al., 2006), this representation is also specified for syntactic and grammatical information. In the second stage, the system activates the word's phonological form. Below, we consider each of the four predictors retained in our previous study (Swiderski et al., 2016) in the context of this general two-stage model of word production.

Both rational consideration and empirical evidence suggest that word length is represented in the second stage of lexical retrieval. Length is a property of phonological representations and longer words are more likely to result phonological and nonword errors by people with aphasia (Gordon, 2008; Kittredge, Dell, Verkuilen, & Schwartz, 2008; Mirman et al., 2010; Nozari, Kittredge, Dell, & Schwartz, 2010) and tip-of-the-tongue states (Harley & Bown, 1998). Such examples are taken to indicate failed phonological retrieval following successful activation of a lexical-semantic representation (Gollan & Brown, 2006; James & Burke, 2000; Meyer & Bock, 1992).

Kittredge and colleagues (2008) argued that both semantic and phonological stages of word retrieval are sensitive to lexical frequency, based on the findings that frequency predicted the occurrence of both semantic and phonological errors in picture naming by people with aphasia. Other accounts have localized frequency effects to the phonological stage, based in part on observations that high and low-frequency homophone pairs were translated from English to Dutch with similar speed (Jescheniak & Levelt, 1994) and equally susceptible to the elicitation of phonologically related errors (Dell, 1990). Because homophones presumably share phonological but not lexical-semantic representations, these results were taken as evidence that frequency is encoded with the former. However, these findings proved difficult to replicate (Bonin & Fayol, 2002; Caramazza, Costa, Miozzo, & Bi, 2001) and a number of other studies have likewise produced evidence for encoding of lexical frequency at either the lexical-semantic stage (Vitkovitch & Humphreys, 1991; Wheeldon & Monsell, 1992), phonological stage (La Heij, Puerta-Melguizo, van Oostrum, & Starreveld, 1999), or both (Feyereisen, Der Borght, & Seron, 1988; Gollan & Brown, 2006; Knobel, Finkbeiner, & Caramazza, 2008).

As noted by Kittredge and colleagues (2008), there is even less agreement on the locus of AoA in two-stage models of word production than there is for lexical frequency. Effects of AoA have been observed in a wide range of lexical tasks (Johnston & Barry, 2006), with some theoretical accounts favoring an earlier locus, potentially including conceptual or lexical-semantic processing (Belke, Brysbaert, Meyer, & Ghyselinck, 2005; Brysbaert & Ghyselinck, 2006; Dent, Johnston, & Humphreys, 2008). Additional studies propose a later, phonological locus (Barry, Hirsh, Johnston, & Williams, 2001; Kittredge et al., 2008; Laganaro & Perret, 2011), and others favoring multiple loci (Ellis & Lambon Ralph, 2000; Johnston & Barry, 2005; Navarrete, Scaltritti, Mulatti, & Peressotti, 2013). Kittredge and colleagues (2008) based their support for a phonological locus of the AoA effect on their multilevel multinomial regression analysis of picture naming responses from 50 people with aphasia showing that AoA predicted the occurrence of phonological, but not semantic errors. They also argued that, even if AoA is encoded at all levels of a two-stage distributed connectionist network, its effects on phonological retrieval should be larger than on lexical-

semantic retrieval, for reasons having to do with the relative arbitrariness of mappings between stages.

In contrast to length, lexical frequency, and AoA, naming response time is a performance variable rather than a property of the lexical target and it reflects all of the processing that occurs between presentation of the picture stimulus and production of the response. As such, its association with item difficulty cannot be localized to any particular stage of word production. Indeed, given that naming response time depends in substantial part on the three remaining predictors, it is somewhat surprising that they are retained as independent predictors when it is entered into the model. Nevertheless, despite the fact that naming response time cannot contribute to the structure of a more elaborated measurement model, it is a useful and relatively easily obtained predictor of overall item difficulty.

Purpose of the study

Although the inclusion of naming response time was pivotal in regard to increasing the proportion of variance accounted for (Swiderski et al., 2016), it was derived from response times from the Center for Research in Language database (CRL; Székely et al., 2004), which does not have identical target pictures to those found on the PNT. Of the 175 items in the PNT, 69 (39%) have matching pictures in the CRL database, 86 (49%) have non-matching pictures, and 20 (11%) are not present. As noted in Gierl and Haladyna (2012), missing data is a common and unfortunate setback of utilizing predictor variables derived from other researchers. However, from a missing data theory perspective, depending on the missingness mechanism (Rubin, 1976), analyzing datasets with large proportions of missing data may lead to statistical parameter estimates that may include substantial bias (Enders, 2010; Rubin, 1976; Schafer & Graham, 2002). Further, excluding 30% of the items from the analysis can reduce the representativeness of the sample, and the amount of information that is available across the range of item difficulty values. Therefore, in order to develop a robust predictive model that would form a solid foundation for item difficulty modeling and automatic item calibration, we conducted a study to collect response times for the full set of the PNT items to build our predictive model. Then, using the full dataset, we determined the utility of psycholinguistic variables to predict the IRT difficulty parameters of the PNT items. The following research questions were addressed in this study: (i) What are the mean naming latencies for the 175 items on the PNT? (ii) Which psycholinguistic variables are strong predictors of the IRT difficulty parameters of the PNT items and what is the overall predictive utility of the model?

Method

Collection and analysis of naming latencies

Participants.—44 cognitively healthy adults (Table 1) recruited from the Portland State University community participated in the study. Mean age was 28.4 years ($SD = 8.50$), mean years of education were 13.6 ($SD = 2.8$) and there were 31 females in the group. All participants met the following inclusion criteria: (a) visual acuity within normal limits, as indicated by a score of 6/12 on the Tumbling E Chart (Lotery et al., 2000); (b) no history of stroke, head injury, progressive or acquired neurogenic disorders per self-report; (c) spoke

English as their first language per self-report, and (d) had no history of speech, language or hearing impairment.

Materials.—Picture stimuli were black and white line drawings of 175 items and 10 practice items from the PNT (Roach et al., 1996). The PNT consists of 175 nouns ranging from 1–4 syllables. The nouns are high-, medium-, or low-frequency as determined by Francis and Kucera (1982). The items were selected from a set of 277 items that were correctly identified by 85% of a control group (Roach et al., 1996).

Procedure.—For the collection of response times, three trained research assistants individually tested participants. The participants were instructed to name items in the most direct manner as possible while avoiding utterances such as *um*, *uh*, and *hm* as well as additional words that were not necessary (exp. “*a dog*” for “*dog*”).

The stimuli were presented on a 22-in ASUS monitor with a vertical and horizontal refresh rate of 75KHz and 80 KHz, respectively. To avoid item-ordering effects (e.g., fatigue and loss of interest) the items were presented in random order. Participants viewed the monitor from approximately 30 inches away. Each began with a fixation point (+) for 250 ms. A tone was presented 250ms after the onset of the fixation point. After 500ms presentation of the fixation point and tone, target pictures were displayed for 3000ms. Auditory and visual stimuli were presented via a Dell DCNE1F computer using E-Prime version 2.0 (Schneider, Eschman, & Zuccolotto, 2002). Only the exact target name was regarded as a correct response. For example, for the item *fish*, a category exemplar of salmon was not accepted. All audio samples were recorded on an Olympus VN-4100PC voice recorder.

To estimate naming latencies, spectrograms of the audio samples were manually analyzed with Audacity version 2.0.6 (Ash et al., 1999) recording and editing software. The latencies were estimated based on the acoustic characteristics of the initial phoneme of each item. As each item was preceded by a pure tone, measurements were made from the tone offset (and onset of picture stimulus) to the initial phoneme. The onset of items with word-initial plosives and affricates were identified by the most proximal edge of the burst (Reetz & Jongman, 2011). Fricative onsets were similarly identified by the leading edge of the frication noise. Items beginning with /l/, a lateral approximant, were measured to the beginning of their F1 formant, which is the earliest detectable acoustic energy produced by the consonant (Reetz & Jongman, 2011). To determine the latencies times for items with word initial /ɹ/, we identified the steep transition of the F3 formant into the subsequent vowel then traced the acoustic energy backward until it was no longer present. A measurement was then made to this point from the preceding prompt tone. Latencies for items with word-initial vowels and word-initial /w/ were measured similarly by identifying the F1 formant, tracing its acoustic signal backward to its beginning, and measuring from the tone to this point.

Analysis of naming latencies.—Outliers were trimmed by eliminating observations based on a nonrecursive moving criterion technique (Selst & Jolicoeur, 1994). This technique utilizes criterion values in terms of standard deviations from the mean, which increase as a function of sample size. Criterion cut scores were identified for our specific

sample sizes with linear interpolation based on the sample sizes suggested by Selst and Jolicoeur (1994) and are provided in the Supplemental Dataset (Dataset A). After outliers were trimmed, the mean and standard deviation of naming response time in seconds for each PNT item was computed.

Predictive modeling of item difficulty

Once the collection and analysis of the naming latencies was concluded, we conducted a multiple regression analysis with PNT item difficulty as the dependent variable and mean naming response time, lexical frequency (based on the Log10CD index), AoA, and length in phonemes as independent variables.

Item difficulty parameters.—The PNT item difficulty parameters were those estimated by Fergadiotis and colleagues (2015) from the sample of 251 PWA included in MAPPD (Mirman et al., 2010). The demographic and clinical characteristics of this sample are summarized in Table 2.

Predictors.—In addition to the mean response latencies collected in the present study, three explanatory variables were prepared to predict item difficulty parameters. Lexical frequency (Log10CD) was identified from the Subtlex corpus (Brysbaert & New, 2009) and were estimated from a corpus of 51.0 million words which were gathered from the subtitles of 8,388 films and television episodes (Brysbaert & New, 2009). Log10CD was chosen over wordform frequency from Subtlex based on evidence that Log10CD demonstrates higher explanatory utility across different experimental designs (Brysbaert & New, 2009, p. 985). Further, AoA (Kuperman, et al., 2012) and the number of phonemes were used. Except for naming latency, these are the same predictor variables that were used in the retained models reported by Fergadiotis and colleagues (2015). The model was updated to include the naming latencies collected as part of this study.

Results

Analysis of naming latencies

A total of 7,382 latencies were acquired. The total number of missing response times across all items was 318 (4.1% of all responses). Seventy-six percent of missing responses were due to subjects responding with names other than the target name (e.g., “shoe” for *slipper*), and the remaining 24% were null responses. There were no missing response latencies for 79 of the items. Among the remaining items, the item *slipper* was associated with the most missing data (27, or 61% of responses missing). The rest of the items had substantially less missing data, with *ambulance* having the next highest number of missing response times at 15 (34% of responses). Despite the missing data, we were able to obtain robust estimates for the mean naming latency for each item, particularly given that (1) the proportion of missing data overall was low, and (2) even for the item with the most missing data (*slipper*), the mean was based on 17 observations.

After outliers were identified and removed the minimum number of missing data per item remained at 0 (n=13) and the maximum number of missing data per item remained at 27 (for

slipper) with a total missing data proportion of 7.1% of the data set as a whole. Supplemental Dataset A includes the raw dataset, as well as the mean naming latencies, standard deviations, number of correct responses, and the interpolated nonrecursive moving criterion cut scores for all items.

As previously mentioned, 39% of the pictures utilized in a preliminary assessment of the predictive utility of naming latency to item difficulty matched those in the CRL database (Székely et al., 2004). To investigate the validity of our response times, first, we analyzed the relationship between our collected naming latencies and those reported in the CRL database for the subset of items that were elicited using the same pictorial stimuli (overlapping subset; $n = 68$ items). Then, we estimated the correlation between naming latencies based on the items that did not share the same picture (non-overlapping subset; $n = 107$). Finally, we estimated the correlation for all the items in the PNT regardless of whether they were elicited using the same picture (full set; $n = 175$). We found a strong, statistically significant correlation for the overlapping subset of items that were elicited using the same picture across the PNT and the CRL database (Pearson Product Moment Correlation $r = .77$, $p < .001$). For the non-overlapping items, the correlation was statistically significant but considerably lower (Pearson Product Moment Correlation $r = .54$, $p < .001$). When the full set of the PNT items was analyzed, the correlation between the naming latencies was statistically significant and fell between the correlation estimates for the overlapping and non-overlapping subsets (Pearson Product Moment Correlation $r = .64$, $p < .001$). See Figure 1 for scatterplots comparing response times for all the items, overlapping items only, and non-overlapping items.

Predictive modeling of item difficulty

Preliminary Analyses.—Once estimated, mean naming latencies were included in the dataset along with Log10CD, AoA, number of phonemes, and item difficulties. The data were prepared for statistical analysis following Kline (2010) and Tabachnick and Fidell (2007) in SPSS and R. All study variables were assessed in terms of the assumptions of normality, homoscedasticity, and linearity. First, distributions were visually inspected, and skewness and kurtosis statistics were estimated. The number of phonemes and naming response times were noted to be positively skewed and with various degrees of kurtosis. A close inspection of bivariate scatterplots (Figure 2) did not suggest significant violations of the assumptions of homoscedasticity and linearity. Because the order of item presentation on the PNT from the MAPPD database was fixed, we also investigated if item difficulty correlated with item order. The Pearson Product Moment Correlation between order and item difficulty was negligible ($r = .06$, $p = .40$), suggesting that only item order nor fatigue influenced the IRT ability estimates from Fergadiotis and colleagues (2015).

Multivariate outliers were assessed with Cook's distance and leverage statistics using the car R package (Fox & Weisberg, 2011). We flagged two cases as potential multivariate outliers: *nurse* and *corn*. The item *nurse* was identified as an outlier because its mean response time was substantially higher than what one would expect given its other psycholinguistic variables. We hypothesize that due to the image used, participants had difficulty naming the picture consistently as 'nurse' as evidenced by its high number of category exemplars (n

=10; e.g., ‘girl’, ‘woman’). Thus, *nurse* was eliminated from the dataset. No clear explanation was derived as to why *corn* was identified as an outlier and we chose a conservative approach and included it in the dataset. The full dataset analyzed, including descriptive statistics, Cook’s distances in Supplemental Dataset B.

Predictive Modeling.—The correlation matrix and descriptive statistics were estimated for item difficulty parameters and predictor variables and presented in Table 3. All predictor variables were significantly correlated with item difficulty. Number of phonemes and AoA were strongly related to Log10CD but less strongly related with each other. Naming latencies did not share a strong zero order correlation with any of the predictor variables

Multiple regression was utilized to assess the utility of Log10CD, AoA, length in phonemes, and naming response time for predicting the independent variable, item difficulty. All predictors were entered and evaluated simultaneously in the regression model using SPSS 24. The model was statistically significant ($F(4, 169) = 98.006, p < .011$) and it explained a substantial proportion of variance of item difficulty ($R = .836, R^2 = .699, Adj. R^2 = .692$). To further evaluate and gauge the predictive utility of the model on novel data, we performed repeated k-fold cross validation to estimate the out-of-sample model accuracy (Kohavi, 1995; Refaeilzadeh, Tang, & Liu, 2016). We used the caret R package (Kuhn, 2018) to create testing and training sets by randomly splitting the data into ten equally sized subsets. A single subsample was designated as the testing set, and the remaining data as the training set. We then ran and tested the model on all 10 datasets. Finally, this procedure was repeated 100 times and the predictive utility of the model in terms of R^2 estimates was averaged across runs. The average R^2 across replications was very similar to the $Adj. R^2$ from the regression analysis, Average $R^2 = .695$, supporting the model’s predictive utility in future applications and mitigating concerns for overfitting the model.

The full solution of the model can be seen in Table 4. All of the regression coefficients were statistically significant (for all variables $p < .001$). Using the unstandardized regression coefficients, an item’s predicted difficulty is equal to $-2.012 - .391(\text{Log10CD}) + .15(\text{AoA}) + .136(\text{Number of Phonemes}) + 1.482(\text{Response Time})$. As expected, higher Log10CD values were associated with lower item difficulty parameters. On the other hand, longer words, words acquired later, and words with longer naming response times were associated with higher difficulty parameters. Based on the magnitude of the standardized regression coefficients, which were comparable, the contribution of each predictor was relatively the same with the exception of number of phonemes which was slightly more influential than the rest of the predictors.

Discussion

The purpose of this study was to evaluate the utility of naming response times collected from healthy speakers as a predictor of IRT difficulty parameters of the items in the PNT. Using Log10CD, AoA, and length in phonemes, Fergadiotis and colleagues (2015) successfully explained 63% of the variance in PNT item difficulty parameters. Adding naming response times (Székely et al., 2004) increased the proportion of variance explained to 73% (Swiderski et al., 2016). However, naming latencies were available for only a subset of items

and given the potential for the missing data to bias the regression parameter estimates, it was necessary to collect a complete set of reaction times for all 175 PNT items to validate our preliminary results. Further, the pictures used by Székely and colleagues (2004) were not identical to those of the PNT, which could further contribute to error in predicting item difficulty from their latency data. Thus, in this study we collected naming latencies for the 175 PNT stimulus pictures from a sample of 44 healthy adult speakers. Use of these naming latencies resulted in regression parameter estimates and an overall R^2 estimate that were similar to those obtained using Székely and colleagues' data.

Naming Latencies

In the current study, we estimated mean naming latencies for all 175 items of the PNT. The strong correlations between our naming latencies and the latencies reported in the CRL database lends support to the validity of our estimates and allows for comparisons to earlier work. For example, in a previous investigation (Bates et al., 2003), naming latencies from the CRL database collected from 50 adult English speakers were compared to the 181 overlapping items from Snodgrass and Yuditsky (1996) and were found to correlate strongly ($r = .87, p < .01$). On the other hand, Snodgrass and Yuditsky (1996) reported a lower correlation ($r = .65$) when they compared naming latencies from 47 identical images to those utilized by Morrison and colleagues (1992). Further, Snodgrass and Yuditsky (1996) used 78 items from an investigation conducted by Humphreys and colleagues (1988) and reported a correlation of $r = .57$ between the two sets of latencies. Therefore, the magnitude of the correlations reported in this paper fall within the expected range established by previous investigations.

The correlational analyses between our naming latencies and the data reported in the CRL database were also informative for understanding the effect of using the same picture to elicit a target response. Specifically, we found that that response times acquired in our study correlated more highly with previously collected response times when the target image was the same. Based on the difference in the magnitude of the correlations, it appears that some variability in latencies can be attributed to whether the target was elicited using the same picture. However, as discussed later, the use of identical pictures for eliciting the targets of the PNT did not have a significant effect on the utility of the naming latencies for the purpose of predicting IRT difficulty parameters. Depending on the application and the nature of future research investigations, it might be preferable to use the same pictures to elicit targets for precision, comparison, and replicability purposes.

Though not definitive, the differences between our reported correlations and the results in previous studies could be attributed to a number of methodological differences across studies. For example, historically, voice keys have been used to collect naming latencies (Bates et al., 2003; Snodgrass & Vanderwart, 1980; Székely et al., 2004, 2005). However, the utilization of such devices forces the exclusion of responses preceded by nonlinguistic sounds. As such, throat clears, coughs, participant body movement, and noise from sources outside the testing room may trigger the voice key invalidly. In this study, naming latencies were collected and estimated using a computer, freeware software (Ash et al., 1999), and a microphone. Given the detailed description of the response time collection process in the

paper, one advantage of our approach is that the outlined process will not only allow for the replication of this study, but also determine the nature and source of discrepancies with future and past investigations.

It could be speculated that methods in data trimming could have influenced discrepancies in the magnitude of the correlations reported in this paper and previous investigations. Snodgrass and Yuditsky (1995) eliminated all response times from participants with mean response times outside 2 SDs of the mean of all response times collected. Székely and colleagues (2003) did not eliminate response times and suggested that data as a whole are reflective of the heterogeneous populations they will be applied to (e.g., neurotypicals, PWA, and children). We have chosen to utilize a nonrecursive moving criterion as this method aids in identification of outliers as a function of the number of observations. However, when we analyzed the relationship between the RTs from the CRL database and our trimmed and untrimmed response times they were nearly identical ($r = .7770$ and $r = .7772$, respectively) suggesting that discrepancies between the data sets are not likely due to data trimming methods.

Item Difficulty Modeling

The regression analysis provided meaningful results with respect to the explanation of item difficulty. The sign of the coefficients of the predictors were reasonable and consistent with our a priori theoretical expectations. As all predictors in the regression analysis were significant, it can be concluded that our approach did not include irrelevant parameters. Importantly, the inclusion of the four predictors in our model yielded a regression model that accounted for a considerable amount of variance in the item difficulty parameters ($Adj. R^2 = .699$). The strength of the predictive utility of our model was on par with several applications of IRT difficulty modeling in other research settings including figural matrix items for assessing general intelligence (Embretson, 1998; Freund, Hofer, & Holling, 2008), word problem variants for the quantitative section of the Graduate Record Examinations General Test (Enright, Morley, & Sheehan, 2002), and items for non-word repetition tasks (Morgan, Gorin, Gray, & Levy, 2011). This significant finding takes us a step closer to augmented or automatic item calibration by predicting IRT difficulty parameters for a large number of confrontation naming test items from their psycholinguistic characteristics (Embretson & Reise, 2000; Gierl & Lai, 2012; Matteucci et al., 2012). Although it does not ensure future success, the regression model forms a solid foundation based on which its' validity may be further assessed by estimating IRT difficulty parameters on existing/novel confrontation naming tests.

It is also important to emphasize that this regression model was not intended to be explanatory in nature as our purpose was not theory building or testing. Instead, our primary goal was to develop a statistical predictive tool that would maximize the proportion of variance explained in the item difficulty parameters. In turn, such a model would be used to explore more efficient ways for automatic item calibration that would not require large sample sizes from clinical populations. This study is situated within the realm of applied science as opposed to basic science and its goal was to maximize predictive utility not explanation (Simon, 2001). From that perspective, we can differentiate our predictors into

two sets of variables. The first set includes only naming latencies, which treat the naming process as a black box and reflect an amalgam of cognitive subprocesses from all levels of processing including visual perception, visual processing, word access and retrieval, as well as motor planning. As such, this variable is useful only for prediction of unidimensional latent-trait item difficulty. The remaining three variables, Log10CD, AoA, and number of phonemes, have been studied as explanatory variables linked to psychological processes in a number of studies that sought to establish causal relationships between these properties of the words and word access and retrieval (e.g., Bonin & Fayol, 2002; Dent et al., 2008; Jescheniak & Levelt, 1994; Kittredge et al., 2008; Laganaro & Perret, 2011).

Among these three predictors, only one, word length, can be confidently ascribed to a particular stage of word production. For both Log10CD and AoA, credible arguments place their effects at either or both steps of the general two-stage model described in the introduction. In attempting to go beyond the current goal of item difficulty prediction for a unidimensional model to development of a more diagnostically useful multidimensional model, identifying variables exclusively predictive of lexical-semantic access, and in particular those variables that share minimal variance with Log10CD and AoA, would be valuable. One candidate variable is semantic neighborhood density, which has been operationalized in a number of different ways (Bormann, 2011; Hameau, Nickels, & Biedermann, 2017; Kittredge, Dell, & Schwartz, 2007; Mirman, 2011; Mirman & Magnuson, 2008), and deserves exploration in this context. Additionally, as the PNT items were selected to have high name agreement, this variable was not likely to contribute to the predictive utility of the PNT item difficulty parameters. However, when estimating item difficulty for new items name agreement/disagreement may be a viable predictor that deserves exploration in the future. This variable was also found to be more sensitive to naming accuracy in a picture naming task (Britt, Ferrara, & Mirman, 2016) in older adults (range: 50 – 61) than younger adults (range: 18 – 25), which may increase its predictive utility in item difficulty estimation as PWA generally are older than 50 years of age.

The approach we followed in this paper can be applied to other languages because the predictors are either readily available (e.g., number of phonemes) or can be generated relatively easily. For example, AoA can be collected using crowdsourcing in an efficient and cost effective manner as demonstrated by Kuperman, et al. (2012). Kuperman et al. used Amazon's Mechanical Turk to collect 842,438 ratings from 1,960 responders over a period of 6 weeks for over 30,000 words. The overall cost of that study was reported to be less than \$4,000. Previous research has indicated that the results obtained from crowdsourcing platforms are comparable to conventional sources of data collection such as convenience and snowball sampling (Casler, Bickel, & Hackett, 2013; Buhrmester, Kwang, & Gosling, 2011). With respect to Log10CD, the approach introduced by Brysbaert and New (2009), that is based on film and television subtitles, has already been applied to British English (van Heuven, Mandera, Keuleers, & Brysbaert, 2014), Spanish (Cuetos, Glez-Nosti, Barbon, & Brysbaert, 2011), Dutch (Keuleers, Brysbaert, & New, 2010), Polish (Mandera, Keuleers, Wodniecka, & Brysbaert, 2015), and Chinese (Cai & Brysbaert, 2010). Finally, collecting naming latencies from neurotypical adults requires minimal resources especially when compared to the process of collecting data from neurologically impaired individuals.

One potential limitation of this study was that the age of the cognitively healthy adults ($mean = 28.4$, $SD = 8.50$) who contributed response latencies in the current study was quite different from the age of PWA in the general clinical population and the data utilized from the MAPPD database. In a literature review investigating age and aphasia severity, Ellis and Urban (2016) reported that on average, PWA were 55 years or older. Similarly, in the MAPPD sample (Mirman et al., 2010), which was utilized to estimate the IRT parameters of the PNT, the average age of the 251 PWA was 58.8 ($SD = 13.2$). Therefore, it is possible that collecting response latencies from older participants to match the age characteristics of the target population may increase the predictive utility of this variable. In general, older adults exhibit a general decline in processing speed that has been documented for naming tasks (Lima, Hale, & Myerson, 1991). However, given processing speed has a global effect on naming latencies, and that for the development of the predictive model we rely on the relative ranking of latencies and not their mean, we would not expect results to change substantially by introducing latencies from older adults. Yet it is possible that because of generational differences and different levels of familiarity with certain items, collecting data from older adults may yield more informative naming latencies and by extension increase the predictive utility of picture naming response time. However, inclusion of such a variable would greatly decrease the feasibility of this model as it was designed to estimate item difficulty only with readily available or easily acquired predictor variables. Our main reason for collecting data from neurotypical younger adults was to increase the predictive utility of the item difficulty estimation model initially reported by Fergadiotis and colleagues (2015) with such predictor variable. Because researchers have easier access to such populations (e.g., university students), demonstration of the predictive utility of their response time data to item difficulty may increase the feasibility of conducting future investigations when the number of target pictures increases substantially.

In conclusion, we demonstrated the utility and unique contribution of naming latencies for predicting IRT difficulty parameters for the PNT items. Our analysis suggests that the missing naming latencies of some items do not pose a significant threat to the validity of prior investigations, as our estimates remained practically unchanged when we analyzed sets with missing and complete data. Therefore, the regression parameters presented in this paper may be useful for developing psychometric applications for clinical populations, and identifying appropriate stimuli for treatment purposes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors gratefully acknowledge the support of NIH/NIDCD award # R03DC014556 to the first author and an ASHA SPARC Award to the second author, which supported presentation of preliminary results of this study to the 2016 Clinical Aphasiology Conference. The second author was affiliated with Portland State University when the data for this study were collected.

The authors also thank the Portland State University students who participated in the study and Lindsay Andreu and Erin Welp for their assistance with data collection.

References

- Ash R, Crook J, Dannenberg R, Drung B, Johnson V, Lucius L, & Shaw M (1999). Audacity (Version 2.06) [Computer software]. Retrieved from <https://www.audacityteam.org/>
- Ayala R. J.de. (2009). Theory and practice of item response theory. New York, New York: Guilford Publications.
- Babbitt EM, Heinemann AW, Semik P, & Cherney LR (2011). Psychometric properties of the Communication Confidence Rating Scale for Aphasia (CCRSA): Phase 2. *Aphasiology*, 25(6–7), 727–735. 10.1080/02687038.2010.537347
- Barry C, Hirsh KW, Johnston RA, & Williams CL (2001). Age of acquisition, word frequency, and the locus of repetition priming of picture naming. *Journal of Memory and Language*, 44(3), 350–375.
- Bates E, D'amico S, Jacobsen T, Székely A, Andonova E, Devescovi A, ... Tzeng O (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review*, 10(2), 344–380. [PubMed: 12921412]
- Belke E, Brysbaert M, Meyer AS, & Ghyselinck M (2005). Age of acquisition effects in picture naming: evidence for a lexical-semantic competition hypothesis. *Cognition*, 96(2), B45–B54. [PubMed: 15925568]
- Belyukova SA, Stone GM, & Ellis LW (2008). Rasch analysis of word identification and magnitude estimation scaling responses in measuring naïve listeners' judgments of speech intelligibility of children with severe-to-profound hearing impairments. *Journal of Speech Language and Hearing Research*, 51(5), 1124 10.1044/1092-4388(2008/06-0218)
- Bonin P, & Fayol M (2002). Frequency effects in the written and spoken production of homophonic picture names. *European Journal of Cognitive Psychology*, 14(3), 289–313.
- Bormann T (2011). The role of lexical-semantic neighborhood in object naming: implications for models of lexical access. *Frontiers in Psychology*, 2, 127. doi: 10.3389/fpsyg.2011.00127. [PubMed: 21713062]
- Brysbaert M, & Ghyselinck M (2006). The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual Cognition*, 13(7–8), 992–1011.
- Brysbaert M, & New B (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990. 10.3758/BRM.41.4.977 [PubMed: 19897807]
- Buhrmester M, Kwang T, & Gosling SD (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1), 3–5. 10.1177/1745691610393980 [PubMed: 26162106]
- Britt AE, Ferrara C, & Mirman D (2016). Distinct effects of lexical and semantic competition during picture naming in younger adults, older adults, and people with aphasia. *Frontiers in Psychology*, 7, 813 10.3389/fpsyg.2016.00813 [PubMed: 27458393]
- Cai Q, & Brysbaert M (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLOS ONE*, 5(6), e10729 10.1371/journal.pone.0010729 [PubMed: 20532192]
- Caramazza A, Costa A, Miozzo M, & Bi Y (2001). The specific-word frequency effect: implications for the representation of homophones in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(6), 1430–1450.
- Casler K, Bickel L, & Hackett E (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, 29(6), 2156–2160. 10.1016/j.chb.2013.05.009
- Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, ... Hays R (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, 63(11), 1179–1194. 10.1016/j.jclinepi.2010.04.011 [PubMed: 20685078]
- Cuetos F, Glez-Nosti M, Barbon A, & Brysbaert M (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicologica: International Journal of Methodology and Experimental Psychology*, 32(2), 133–143.

- Dell GS (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, 5(4), 313–349.
- Dell GS, Schwartz MF, Martin N, Saffran EM, & Gagnon DA (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, 104(4), 801–838. [PubMed: 9337631]
- Dent K, Johnston RA, & Humphreys GW (2008). Age of acquisition and word frequency effects in picture naming: A dual-task investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(2), 282–301.
- Donovan NJ, Velozo CA, & Rosenbek JC (2007). The Communicative Effectiveness Survey: investigating its item-level psychometric properties. *Journal of Medical Speech - Language Pathology*, 15(4), 433–447.
- Eadie TL, Yorkston KM, Klasner ER, Dudgeon BJ, Deitz JC, Baylor CR, ... Amtmann D (2006). Measuring communicative participation: A review of self-report instruments in speech-language pathology. *American Journal of Speech-Language Pathology*, 15(4), 307
10.1044/1058-0360(2006/030) [PubMed: 17102143]
- Edmonds LA, & Donovan NJ (2012). Item-level psychometrics and predictors of performance for Spanish/English bilingual speakers on an object and action naming battery. *Journal of Speech, Language, and Hearing Research*, 55(2), 359–381. 10.1044/1092-4388(2011/10-0307)
- Ellis AW, & Lambon Ralph MA (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(5), 1103–1123.
- Embretson SE (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3(3), 380–396. 10.1037/1082-989X.3.3.380
- Embretson SE, & Reise SP (2000). *Item response theory for psychologists*. Mahwah, N.J.: L. Erlbaum Associates.
- Enders CK (2010). *Applied Missing Data Analysis*. New York, NY: Guilford Press.
- Enright MK, Morley M, & Sheehan KM (2002). Items by Design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education*, 15(1), 49–74.
10.1207/S15324818AME1501_04
- Fergadiotis G, Kellough S, & Hula WD (2015). Item response theory modeling of the Philadelphia Naming Test. *Journal of Speech, Language, and Hearing Research*, 58, 865–877.
10.1044/2015_JSLHR-L-14-0249
- Feyereisen P, Der Borgh V, & Seron X (1988). The operativity effect in naming: A re-analysis. *Neuropsychologia*, 26(3), 401–415. [PubMed: 2453813]
- Fox J, & Weisberg S (2011). *An R Companion to Applied Regression* (Second). Thousand Oaks, CA: Sage.
- Foygel D, & Dell GS (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language*, 43(2), 182–216. 10.1006/jmla.2000.2716
- Francis W, & Kucera H (1982). *Frequency Analysis of English Usage*. Houghton Mifflin Company.
- Freund PA, Hofer S, & Holling H (2008). Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied Psychological Measurement*, 32(3), 195–210.
10.1177/0146621607306972
- Gierl MJ, & Haladyna TM (Eds.). (2012). *Automatic Item Generation: Theory and Practice*. New York, NY: Routledge.
- Gollan TH, & Brown AS (2006). From tip-of-the-tongue (TOT) data to theoretical implications in two steps: when more TOTs means better retrieval. *Journal of Experimental Psychology: General*, 135(3), 462–483. [PubMed: 16846276]
- Goodglass H, & Wingfield A (1997). *Anomia: neuroanatomical and cognitive correlates*. San Diego, CA: Academic Press.
- Gordon JK (2008). Measuring the lexical semantics of picture description in aphasia. *Aphasiology*, 22(7–8), 839–852. 10.1080/02687030701820063 [PubMed: 22399832]
- Harley TA, & Bown HE (1998). What causes a tip-of-the-tongue state? Evidence for lexical neighbourhood effects in speech production. *British Journal of Psychology*, 89(1), 151–174.

- Hameau S, Nickels L, & Biedermann B (2017). Influential neighbours? The role of semantic neighbours in word production. *Frontiers in Human Neuroscience Conference Abstract: Academy of Aphasia 55th Annual Meeting* 10.3389/conf.fnhum.2017.223.00117
- Herbert R, Hickin J, Howard D, Osborne F, & Best W (2008). Do picture-naming tests provide a valid assessment of lexical retrieval in conversation in aphasia? *Aphasiology*, 22(2), 184–203. 10.1080/02687030701262613
- Humphreys GW, Riddoch MJ, & Quinlan PT (1988). Cascade processes in picture identification. *Cognitive Neuropsychology*, 5, 67–103.
- James LE, & Burke DM (2000). Phonological priming effects on word retrieval and tip-of-the-tongue experiences in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(6), 1378–1391.
- Jescheniak JD, & Levelt WJ (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(4), 824–843.
- Johnston RA, & Barry C (2005). Age of acquisition effects in the semantic processing of pictures. *Memory & Cognition*, 33(5), 905–912. [PubMed: 16383177]
- Johnston RA, & Barry C (2006). Age of acquisition and lexical processing. *Visual Cognition*, 13(7–8), 789–845.
- Kaplan E, Goodglass H, & Weintraub S (2001). *Boston Naming Test* (2nd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.
- Kertesz A (1982). *Western Aphasia Battery test manual*. Grune & Stratton.
- Keuleers E, Brysbaert M, & New B (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, 42(3), 643–650. 10.3758/BRM.42.3.643 [PubMed: 20805586]
- Kittredge AK, Dell GS, & Schwartz MF (2007). Omissions in aphasic picture naming: Late age-of-acquisition is the culprit, not low semantic density. *Brain and Language*, 103(1–2), 132–133.
- Kittredge AK, Dell GS, Verkuilen J, & Schwartz MF (2008). Where is the effect of frequency in word production? Insights from aphasic picture-naming errors. *Cognitive Neuropsychology*, 25(4), 463–492. 10.1080/02643290701674851 [PubMed: 18704797]
- Kline RB (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.
- Knobel M, Finkbeiner M, & Caramazza A (2008). The many places of frequency: Evidence for a novel locus of the lexical frequency effect in word production. *Cognitive Neuropsychology*, 25(2), 256–286. [PubMed: 18568814]
- Kohavi R (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection (pp. 1137–1145). Presented at the 14th International Joint Conference on Artificial Intelligence (IJCAI).
- Kuhn M (2018). R package caret: classification and regression training (Version 6.0–80). Retrieved from <https://CRAN.R-project.org/package=caret>
- Kuperman V, Stadthagen-Gonzalez H, & Brysbaert M (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4), 978–990. 10.3758/s13428-012-0210-4 [PubMed: 22581493]
- La Heij W, Puerta-Melguizo MC, van Oostrum M, & Starreveld PA (1999). Picture naming: Identical priming and word frequency interact. *Acta Psychologica*, 102(1), 77–95.
- Laganaro M, & Perret C (2011). Comparing electrophysiological correlates of word production in immediate and delayed naming through the analysis of word age of acquisition effects. *Brain Topography*, 24(1), 19–29. [PubMed: 20938730]
- Lima SD, Hale S, & Myerson J (1991). How general is general slowing? Evidence from the lexical domain. *Psychology and Aging*, 6(3), 416–425. [PubMed: 1930758]
- Levelt WJ, Roelofs A, & Meyer AS (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(01), 1–38. [PubMed: 11301520]
- Lord FM, & Novick MR (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley Publishing Company.

- Lotery AJ, Wiggam MI, Jackson AJ, Refson K, Fullerton KJ, Gilmore DH, & Beringer TR (2000). Correctable visual impairment in stroke rehabilitation patients, *Age and Ageing*, 29(3), 221–222. 10.1093/ageing/29.3.221 [PubMed: 10855903]
- Mandera P, Keuleers E, Wodniecka Z, & Brysbaert M (2015). Subtlex-pl: subtitle-based word frequency estimates for Polish. *Behavior Research Methods*, 47(2), 471–483. 10.3758/s13428-014-0489-4 [PubMed: 24942246]
- Matteucci M, Mignani S, & Veldkamp BP (2012). Prior distributions for item parameters in IRT models. *Communications in Statistics - Theory and Methods*, 41(16–17), 2944–2958. 10.1080/03610926.2011.639973
- Meyer AS, & Bock K (1992). The tip-of-the-tongue phenomenon: Blocking or partial activation? *Memory & Cognition*, 20(6), 715–726. [PubMed: 1435274]
- Mirman D (2011). Effects of near and distant semantic neighbors on word production. *Cognitive, Affective, & Behavioral Neuroscience*, 11(1), 32–43.
- Mirman D, & Magnuson JS (2008). Attractor dynamics and semantic neighborhood density: processing is slowed by near neighbors and speeded by distant neighbors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1), 65–79. 10.1037/0278-7393.34.1.65
- Mirman D, Strauss TJ, Brecher A, Walker GM, Sobel P, Dell GS, & Schwartz MF (2010). A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. *Cognitive Neuropsychology*, 27(6), 495–504. 10.1080/02643294.2011.574112 [PubMed: 21714742]
- Morgan G, Gorin J, Gray S, & Levy R (2011). Nonword item generation: predicting item difficulty in nonword repetition, ProQuest Dissertations and Theses.
- Navarrete E, Scaltritti M, Mulatti C, & Peressotti F (2013). Age-of-acquisition effects in delayed picture-naming tasks. *Psychonomic Bulletin & Review*, 20(1), 148–153. [PubMed: 22893113]
- Nozari N, Kittredge AK, Dell GS, & Schwartz MF (2010). Naming and repetition in aphasia: Steps, routes, and frequency effects. *Journal of Memory and Language*, 63(4), 541–559. 10.1016/j.jml.2010.08.001 [PubMed: 21076661]
- Rapp B, & Goldrick M (2000). Discreteness and interactivity in spoken word production. *Psychological Review*, 107(3), 460–499. [PubMed: 10941277]
- Reetz H, & Jongman A (2011). *Phonetics: Transcription, production, acoustics, and perception*. Wiley-Blackwell.
- Refaeilzadeh P, Tang L, & Liu H (2016). Cross-Validation In Liu L & Özsu MT (Eds.), *Encyclopedia of Database Systems* (pp. 1–7). New York, NY: Springer New York
10.1007/978-1-4899-7993-3_565-2
- Roach A, Schwartz MF, Martin N, Grewal RS, & Brecher A (1996). The Philadelphia naming test: Scoring and rationale. *Clinical Aphasiology*, 24, 121–133.
- Rubin DB (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. 10.1093/biomet/63.3.581
- Schafer JL, & Graham JW (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. 10.1037/1082-989X.7.2.147 [PubMed: 12090408]
- Schneider W, Eschman A, & Zuccolotto A (2002). *E-Prime: User's guide*. Psychology Software Incorporated.
- Schwartz M, Dell G, 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. 10.1016/j.jml.2005.10.001
- Selst MV, & Jolicoeur P (1994). A solution to the effect of sample size on outlier elimination. *The Quarterly Journal of Experimental Psychology Section A*, 47(3), 631–650. 10.1080/14640749408401131
- Simon HA (2001). Science seeks parsimony, not simplicity: Searching for pattern in phenomena In *Simplicity, inference and modelling: keeping it sophisticatedly simple* (pp. 32–72). Cambridge: Cambridge University Press.
- Snodgrass JG, & Vanderwart M (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2), 174–215. [PubMed: 7373248]

- Snodgrass JG, & Yuditsky T (1996). Naming times for the Snodgrass and Vanderwart pictures. *Behavior Research Methods, Instruments & Computers*, 28(4), 516–536.
- Swiderski A, Fergadiotis G, & Hula WD (2016). Predicting item difficulty of confrontation naming tests within an Item Response Theory framework. Presented at the Clinical Aphasiology Conference, Charlottesville, VA.
- Swinburn K, Porter G, & Howard D (2004). CAT: Comprehensive Aphasia Test. Hove, UK: Psychology Press.
- Székely A, & Bates E (2000). Objective visual complexity as a variable in studies of picture naming. *Center for Research in Language Newsletter*, 12(2). La Jolla: University of California, San Diego.
- Székely A, D'Amico S, Devescovi A, Federmeier K, Herron D, Iyer G, ... Bates E (2003). Timed picture naming: Extended norms and validation against previous studies. *Behavior Research Methods, Instruments, & Computers*, 35(4), 621–633. 10.3758/BF03195542
- Szekely A, Jacobsen T, D'Amico S, Devescovi A, Andonova E, Herron D, ... Bates E (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language*, 51(2), 247–250. [PubMed: 23002322]
- Tabachnick BG, & Fidell LS (2007). *Using multivariate statistics* (6th ed.). Boston: Pearson/Allyn & Bacon.
- van Heuven WJB, Mandera P, Keuleers E, & Brysbaert M (2014). SUBTLEX-UK: a new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology* (2006), 67(6), 1176–1190. 10.1080/17470218.2013.850521
- Vitkovitch M, & Humphreys GW (1991). Perseverant responding in speeded naming of pictures: It's in the links. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(4), 664–680.
- Wheeldon LR, & Monsell S (1992). The locus of repetition priming of spoken word production. *The Quarterly Journal of Experimental Psychology*, 44(4), 723–761. [PubMed: 1615171]

Figure 1. Scatter plot comparisons of all, overlapping, and nonoverlapping pictures to the CRL database.

Figure 2.
Bivariate Scatterplots of Predictor Variables and Item Difficulty. Density plots are shown on the diagonal.

Table 1

Demographic and clinical characteristics of the participant for mean latency response times

Ethnicity, %	
Asian	2.5%
Filipino	5.0%
Hispanic	7.5%
Caucasian	82.5%
Unspecified	2.5%
Education, Years	
Mean	16.45
SD	1.28
Min	12.00
Max	18.00
Missing, %	0%
Age, Years	
Mean	28.4
SD	8.5
Min	19.0
Max	58.0
Missing, %	0%

Table 2

Demographic and clinical characteristics of the patient sample

Ethnicity, %	
African American	34.00%
Asian	0.40%
Hispanic	1.20%
Caucasian	44.00%
Missing	20.00%
Education, Years	
Mean	13.6
SD	2.8
Min	7.0
Max	21.0
Missing, %	20.0%
Age, Years	
mean	58.8
SD	13.2
Min	22.0
Max	86.0
Missing, %	20.0%
Months Post-Onset	
Mean	32.9
SD	51.0
Min	1.0
Max	381.0
Missing, %	20.0%
Western Aphasia Battery AQ	
Mean	73.4
SD	16.6
Min	27.2
Max	97.8
Missing, %	51.0%
Philadelphia Naming Test, % Correct	
Mean	61%
SD	28%
Min	1%
Max	98%

Table 3

Correlations, Means, and Standard Deviations for item difficulty and explanatory variables

	Variable	1	2	3	4	5
1	Item Difficulty	1				
2	Length in Phonemes	0.63	1			
3	Lexical Frequency	−0.62	−0.50	1		
4	Age of Acquisition	0.64	0.38	−0.49	1	
5	Mean response time	0.51	0.23	−0.16	0.39	1
	<i>M</i>	−0.45	4.50	2.71	4.94	0.86
	<i>SD</i>	0.75	1.77	0.53	1.30	0.14

* All correlations found to be significant on the .001 level

Table 4

Summary of Regression Analysis for Variables Predicting Item Difficulty.

Variable	<i>B</i>	<i>SD (B)</i>	β	<i>t</i>	Correlations		
					zero-order	partial	semipartial
Constant	−2.012	0.340		−5.916			
Lg10CD	−0.391	0.074	−0.280 [*]	−5.260	−0.623	−0.375	−0.222
Age of Acquisition	0.150	0.029	0.270 [*]	5.126	0.644	0.367	0.216
Phonemes	0.136	0.021	0.322 [*]	6.446	0.632	0.444	0.272
Response time	1.482	0.239	0.287 [*]	6.212	0.513	0.431	0.262

^{*}
 $p < .001$