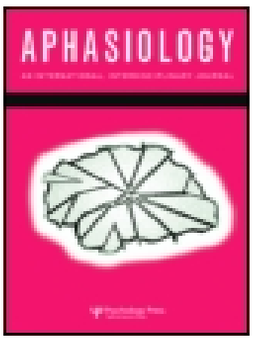

Aphasiology

ISSN: 0268-7038 (Print) 1464-5041 (Online) Journal homepage: http://www.tandfonline.com/loi/paph20

Modelling confrontation naming and discourse performance in aphasia

Gerasimos Fergadiotis & Heather Harris Wright

To cite this article: Gerasimos Fergadiotis & Heather Harris Wright (2015): Modelling confrontation naming and discourse performance in aphasia, Aphasiology, DOI: 10.1080/02687038.2015.1067288

To link to this article: http://dx.doi.org/10.1080/02687038.2015.1067288

Published online: 17 Jul 2015.

Aphasiology, 2015
http://dx.doi.org/10.1080/02687038.2015.1067288

Modelling confrontation naming and discourse performance in aphasia

Gerasimos Fergadiotis[a]* and Heather Harris Wright[b]

[a]*Department of Speech and Hearing Sciences, Portland State University, Portland, OR, USA;* [b]*Department of Communication Sciences and Disorders, East Carolina University, Greenville, NC, USA*

(*Received 17 December 2014; accepted 24 June 2015*)

Background: It is well documented in the literature that the ability to produce discourse is what matters most to people with aphasia (PWA) and their families. However, quantifying discourse in typical clinical settings is a challenging task due to its dynamic and multiply determined nature. As a result, professionals often depend on confrontation naming tests to identify and measure impaired underlying cognitive mechanisms that are also hypothesised to be important for discourse production.
Aims: The main goals of the study were to investigate the extent to which confrontation naming test (CNT) scores are predictive of (i) the proportion of paraphasias in connect speech, and (ii) the amount of information PWA can effectively communicate.
Methods & Procedures: Data from 98 monolingual PWA were retrieved from AphasiaBank and analysed using structural equation modelling. Performance in CNTs was modelled as a latent variable based on the Boston Naming Test, the Western Aphasia Battery–R Naming Subtest, and the Verb Naming Test. Performance at the discourse level was modelled based on the observed proportions of paraphasias in three discourse tasks (free speech, eventcasts, and story retell). Informativeness was quantified using the percentage of Correct Information Units based on story re-tell.
Outcomes & Results: For the first question, the regression coefficient between the two latent factors was estimated to be –0.52. When the latent factor based on the CNTs was regressed on informativeness, the estimated regression coefficient was equal to 0.68.
Conclusions: Performance on CNTs was not a strong predictor of the proportion of paraphasias produced in connected speech but was substantially higher for the amount of information communicated by PWA. Clinicians are cautioned not to predict a speaker's performance at the discourse level after establishing CNT performance. Other anomic behaviours (e.g., pauses) during discourse production may be associated with CNT performance and should be considered in future investigations.

Keywords: aphasia; word retrieval deficits; diagnostic procedures; structural equation modelling

The cardinal deficit of people with aphasia (PWA) is anomia (Goodglass & Wingfield, 1997). In single-word retrieval, this deficit is believed to be indicative of disruption in accessing a semantic description of the target concept, and/or retrieving a fully phonologically specified representation (e.g., Dell, 1986). Aphasic naming errors are generally considered to result from reduced or insufficiently persistent activation of target representations relative to competing non-target representations and/or noise in the system (Schwartz, Dell, Martin, Gahl, & Sobel, 2006).

*Corresponding author. Email: gfergadiotis@pdx.edu

Given the prevalence and importance of anomia, professionals typically include confrontation naming tests when evaluating the expressive abilities of PWA to quantify anomia severity and determine the underlying nature of the impairment.

The use of confrontation naming tests has many advantages (Herbert, Hickin, Howard, Osborne, & Best, 2008; Mayer & Murray, 2003). These tests are administered similarly in nature, and involve a manageable scoring design. When constructed through a sound psychometric process, they yield a cumulative score that reflects the severity of naming impairment. Furthermore, these tests have high test-retest reliability; and, moderate to high intercorrelations, suggesting a high degree of construct validity. In addition, testers know the target words, thus minimising ambiguity when scoring the responses. For the aforementioned reasons, utilising confrontation naming tests has become a common and reliable method among researchers and clinicians for measuring lexical retrieval at the single-word level. Typically administered tools include the Boston Naming Test (BNT; Kaplan, Goodglass, & Weintraub, 2001), the Western Aphasia Battery Naming subtest (WAB-R Naming; Kertesz, 2007), and the Verb Naming Test (VNT; Cho-Reyes & Thompson, 2012).

However, it is well documented in the literature that the ability to produce discourse is what matters most to PWA and their families (Cruice, Worrall, Hickson, & Murison, 2003; Mayer & Murray, 2003). For this reason, systematic approaches for studying discourse have been developed to capture multiple facets of discourse production (Marini, Andreetta, del Tin, & Carlomagno, 2011; Saffran, Berndt, & Schwartz, 1989; Shewan, 1988); and, to quantify the informativeness of connected speech (McNeil, Doyle, Fossett, Park, & Goda, 2001; Nicholas & Brookshire, 1993; Yorkston & Beukelman, 1980). Yet, measuring connected speech in typical clinical settings is a challenging task due to the dynamic and multiply determined nature of discourse (Armstrong, 2000; Prins & Bastiaanse, 2004; Spreen & Risser, 2003). In addition to being resource-intensive, current discourse measures often yield scores that are too noisy or may reflect constructs that were not intended to be captured by the researchers (i.e., construct irrelevant information) (Boyle, 2014; Cameron, Wambaugh, & Mauszycki, 2010; Fergadiotis, Wright, & West, 2013; Ross & Wertz, 1999). As a result, professionals often depend on confrontation naming tests to identify and measure impaired underlying cognitive mechanisms that are also hypothesised to be important for discourse production. This study investigates the implicit clinical assumptions that performance in single-word, picture-naming tasks is directly and strongly related to word retrieval and the ability to communicate informative content during discourse production.

Confrontation naming tests as predictors of discourse

During discourse production, in addition to the core processes that serve word retrieval of single words, production also depends on cognitive operations that extend beyond lexical processing (Wilshire & McCarthy, 2002, p. 169). These factors might influence the selection of lexical items based on syntactic, structural, and/or pragmatic criteria that can be either automatic or meta-cognitive. Therefore, the score a speaker may receive in a confrontation naming test may be different from the score (s)he may receive during discourse production. For that reason, several studies have investigated the relationship between performance at the discourse level and confrontation naming tests in PWA.

In general, even though there seems to be a general consensus that performance at the two levels is related, the strength of the relationship is such that it may not allow for direct generalisation from one level to the other. For example, Pashek and Tompkins (2002)

investigated the relationship between video narration and performance on confrontation naming. 20 people with mild anomic aphasia were presented with video segments without the soundtrack from the television shows *Leave It to Beaver, Popeye*, and *The Andy Griffith Show* and were asked to produce narratives. The authors estimated word-finding difficulty indices for each participant based on the number of pauses the participants exhibited (longer than two seconds) as well as the number of substitutions, circumlocutions, comments, instances of jargon, and attempts to spell a word. Additionally, participants completed the BNT following the administration and scoring criteria of Nicholas, Brookshire, MacLennan, Schumacher, and Porrazzo (1989). The analysis revealed a moderately strong relationship between the BNT and word-finding difficulty for nouns during discourse production (−0.76, $p < 0.05$). Based on this finding, which suggested approximately 58% overlap in variance between the two variables, the authors concluded that even though a relationship was found, performance on the BNT was not a reliable indicator of discourse performance.

Similarly, Mayer and Murray (2003) also noted the "divergence between single-word confrontation naming and discourse-level word retrieval" (p. 492) when evaluating aphasia severity and treatment planning for PWA. In their study, Mayer and Murray asked 14 PWA to describe several pictured scenes and participate in conversational topics initiated by the investigators. Half of the participants had mild aphasia and half had moderate aphasia according to the Western Aphasia Battery (Kertesz, 1982; AQ = 87.8 and 51.7, respectively). Using the elicited language samples, the authors estimated an index of word-finding difficulty for each participant that was based on pauses, comments indicating difficulty, self-corrections, paraphasias, deletions (Kemmerer & Tranel, 2000), and indefinite terms (Hickin, Best, Herbert, Howard, & Osborne, 2001). Then, they estimated the correlation between word finding during discourse and performance on the Test of Adolescent and Adult Word Finding (German, 1990). Significant associations between the two types of task emerged (r-values >0.70). However, even though the magnitude of the relationships was substantial, there a significant portion of variance remained unexplained.

Several others have investigated the relationship between performance at the discourse level and confrontation naming tasks. For example, Williams and Canter (1982) reported a strong positive correlation ($r = 0.80$) between word retrieval difficulties at the discourse level and picture naming in 40 PWA of various types. Also, Herbert et al. (2008) studied word production in free conversation and performance on a confrontation picture-naming task in ten PWA. They reported several significant findings, including that the proportion of content words out of substantive turns (conversation turns that included at least one content word) and the proportion of nouns out of substantive turns correlated with picture-naming performance ($r = 0.674$, $p = 0.016$ and $r = 0.789$, $p = 0.017$, respectively). Hickin et al. (2001) also reported similar results.

The most challenging aspect in regard to understanding the findings of the aforementioned studies is the interpretation of the discourse-based composite indices of word-finding difficulty. Aphasiologists typically conceptualise cognitive impairment as an underlying, unobserved variable that characterises individuals and *causes* PWA to exhibit a variety of behaviours including paraphasias, pauses, etc. Composite scores such as those used in previous studies are typically used in the measurement of formative constructs (e.g., socio-economic status) in which the exclusion of a single indicator has significant implications in regard to interpretation of the latent construct. To further complicate matters, the interpretation of the discourse-based construct is a function of the relative

weighting of each of the different behaviours that are included in the composite measures (Markus & Borsboom, 2013).

For these reasons, in the current study we chose to investigate two specific aspects of discourse production that are clinically relevant: (i) the number of paraphasias produced, and (ii) the amount of information conveyed during discourse production. The former represents a prominent aspect of language production and the current investigation can be thought of as an initial step in disentangling the interrelationships of confrontation naming testing and different behaviours exhibited in connected speech. The latter reflects an intuitively simple yet ecologically valid aspect of communicative transactions that can be defined as the amount of content produced by a PWA and that is perceived by a listener to be informative. In the following sections we present the validity theory framework within which we formulated our research questions.

Validity framework

Diagnostic assessment may be viewed as an argument that entails a logical inference from how individuals are perceived to perform under particular conditions (e.g., what they say or do) to the individuals' capacities more broadly (Mislevy & Yin, 2009). The focus of this study is on the everyday clinical practice that involves building arguments to reach conclusions about discourse production based on how speakers score on confrontation naming tasks. This process entails an often implicit, yet obligatory logical leap that professionals have to make to predict a speaker's performance at the discourse level after having estimated an individual score for confrontation naming performance. Within this framework, assessing the validity of such a diagnostic process becomes equivalent to assessing whether it is reasonable and justified to draw inferences about specific skills (i.e., paraphasias in discourse and capacity to convey information) based on a distinct set of data (e.g., performance in confrontation naming tests).

Toulmin's argument structure (1958) was used to formalise the research questions of the current paper (for a graphical representation of the first question see Figure 1).

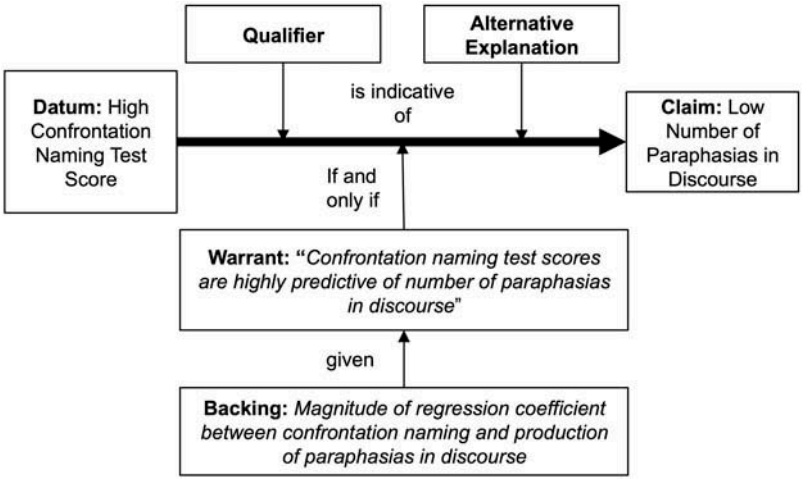

Figure 1. Toulmin's argument structure. The figure presents graphically one type of evidence necessary ("Backing") to draw a conclusion about a speaker's performance during discourse production based on a confrontation naming test score.

Toulmin created a system for judging the rationality of assertions consisting of several components: the claim; the data; the warrant; the backing; the alternative explanation; and the qualifier. The *claim* is what one wishes to support as a valid conclusion. Within the context of this paper, the claim may be that an individual will produce a given level of paraphasias in discourse. The claim is supported by the observed *data* collected in a confrontation naming task. In order to bridge the gap between the data and claim, an inference is made. Importantly, the validity of the inference is evaluated based on the argument's *warrant*. Theory, prior research, and experience provide *backing* for the warrant, which in turn, can be used to validate the inference. For example, in order to make the claim that an individual will produce a certain level of paraphasias in discourse after observing their performance in a confrontational naming task, an inference has to be made. A warrant that attests the relationship between the data and the claim is used to support the inference. In the current context, the warrant takes the form "A confrontation naming test score is highly predictive of the number of paraphasias a speaker exhibits in discourse." Moreover, it is possible to have *alternative explanations* for a particular set of data. Last, the argument structure may be enhanced with use of a *qualifier* that reflects the certainty with which we can draw inferences based on the observed data.

In the current study, we utilised structural equation modelling within a modern psychometric framework to investigate the following specific aims:

(i) to determine the strength of the relationship between performance on confrontation naming tests and the proportion of paraphasias in three different types of discourse when accounting for construct irrelevant variance (i.e., random noise and irrelevant systematic variance); and
(ii) to determine the strength of the relationship between confrontation naming testing and the amount of information PWA communicate successfully during narration.

Method

Participants

Data from 98 monolingual PWA were included in this study. Their data were retrieved from AphasiaBank (MacWhinney, Fromm, Forbes, & Holland, 2011), an online, shared database that collects and analyses digital recordings of discourse from PWA across a series of tasks. All participants had acquired aphasia secondary to a single left hemisphere stroke. Additional inclusion criteria included (i) chronic aphasia (min. = 6 months post-onset); (ii) aided or unaided normal hearing acuity; (iii) corrected or uncorrected normal visual acuity; (iv) English as their primary language; and (v) no reported history of psychiatric or neurodegenerative diagnosis. The sample's characteristics, including Western Aphasia Battery–Revised Aphasia Quotient classification (WAB-R AQ; Kertesz, 2007), years of aphasia duration, and aphasia severity are presented in Table 1. Age, gender, ethnicity, and education are presented in Table 2.

Discourse elicitation and data preparation

Stimuli and instructions

Discourse samples were collected in a single session and several tasks for eliciting discourse were employed including free speech, descriptions of sequential and single

Table 1. Aphasia Western Aphasia Battery–Revised (WAB-R) type and severity.

WAB-R classification	
Anomic	33
Broca	22
Conduction	22
Wernicke	14
Global	4
Transcortical motor	3
Duration post-onset (years)	5.59 (6.23)
Mean WAB-R AQ[a]	70.42 (17.03)

Notes: Standard deviations are shown in parentheses; [a]Aphasia Quotient.

Table 2. Participants' demographic information.

	PWA
Characteristic	(N = 98)
Gender ratio	50 M: 48 F
Age in years	64.18 (12.72)
Ethnicity	
African-American	6
Asian	2
Hispanic	3
Other	4
White	79
Education level completed	
Some high school	5
12th grade	23
Some college	15
Bachelor's or higher	55

Note: Ethnicity and education were unavailable for four individuals.

pictures, and retelling of the story of *Cinderella* (Grimes, 2005). Free speech samples were elicited by asking participants to (i) talk about their speech, (ii) provide information about their stroke, and (iii) talk about an important event in their lives. The sequential picture stimuli included a four-frame strip from Menn et al. (1998) and a six-frame cartoon strip. The first is referred to as *Broken Window* and depicts a boy kicking a ball; the ball goes through a closed window and knocks down a lamp before a man picks up the ball and looks out of the window. The second, referred to as *Umbrella*, depicts the story of a student who refuses to take the umbrella his mother offers him as he leaves for school; on his way to school it starts raining and he returns home to take the umbrella.

The single-picture stimulus was Nicholas and Brookshire's *Cat Rescue* (Nicholas & Brookshire, 1993). In this picture a little girl's cat is in a tree and her dad has tried to rescue the cat but got stuck in the tree; the fire department is arriving to rescue the cat and the man from the tree. Thus the picture shows a situation with a central theme and the interaction of animate and inanimate pictured elements. Also, the picture implied temporal sequencing of events prior to and after the pictured scene. The single and sequential picture stimuli can be found at: www.talkbank.org/AphasiaBank/protocol/pictures.

For the eventcasts, participants were first presented with the two sequential pictures and were asked to produce a story that was based on temporal sequencing. The tester used the following script: "Take a little time to look at these pictures. They tell a story. Take a look at all of them, and then I'll ask you to tell me the story with a beginning, a middle, and an end. You can look at the pictures as you tell the story." If the participants did not respond within 10 seconds, the tester prompted them to "Take a look at this picture [pointing to first picture] and tell me what you think is happening." If needed, the tester pointed to each picture sequentially, giving the prompt: "And what happens here?" Subsequently participants were presented with the single picture and were asked to produce a story with temporal sequencing using the following script: "Here is a picture. Look at everything that's happening and then tell me a story about what you see. Tell me the story with a beginning, a middle, and an end." Feedback was provided to avoid eliciting a simple description of objects, characters, and/or their physical characteristics.

Finally, participants also viewed and told the story depicted in a wordless picture book of *Cinderella*. They were told to look through the book to remember how the story goes, and were allowed as much time as desired to view it. Then, the book was taken away and they were asked to tell as much of the story as they could. The examiners used standard written scripts to keep verbal instruction and prompts consistent across testing sites. Furthermore, examiners were instructed to remain silent, as much as possible, during the administration of the task, while also providing as much non-verbal encouragement as possible.

Confrontation naming tests

PWA were administered three tests. These included the WAB—R Naming subtest (Kertesz, 2007), the Short Form of the BNT (Kaplan et al., 2001), and the VNT (Cho-Reyes & Thompson, 2012). Scoring for each of those tests was performed based on the published standardised guidelines (VNT: Cho-Reyes & Thompson, 2012, p. 1260; BNT: Kaplan et al., 2001, pp. 31–32; WAB-R Naming: Kertesz, 2007, p. 37). Confrontation naming scores were entered into the model as proportions correct.

Discourse-based paraphasia index

Samples were digitally recorded and orthographically transcribed in the CHAT format that is compatible with a set of programs called Computerized Language Analysis (CLAN; MacWhinney, 2000). Experienced speech-language pathologists affiliated with AphasiaBank tagged content words in CLAN to indicate the different types of paraphasias (please visit http://talkbank.org/AphasiaBank/transcribe/errors.doc for scoring criteria). The following types of paraphasia were included in the analysis: semantic, formal/phonological; neologisms; and mixed. All observed paraphasias were included but repetitions of paraphasias in attempts to self-correct were counted as a single occurrence unless the attempt led to a revised error. The number of words for each type of discourse (free speech, picture description, story retell) was estimated and proportions of paraphasias in each type of discourse were calculated and analysed.

Correct information units

To quantify informativeness, Correct Information Units (CIU; Nicholas & Brookshire, 1993) were identified for each story retell and the proportion of CIUs per number of

words was estimated (%CIU). CIUs are defined as words that are intelligible in context, accurate, and relevant to and informative about the content of the topic (Nicholas & Brookshire, 1993, p. 357, Appendix B). CIUs have well-defined criteria, and when applied to structured discourse tasks, studies suggest that CIU scores adequately reflect listeners' perceptions of informative discourse (Cameron et al., 2010; Carlomagno, Giannotti, Vorano, & Marini, 2011; Doyle, Tsironas, Goda, & Kalinyak, 1996). Furthermore, CIU analysis has been found to be an effective quantifier of communication efficiency (Cameron et al., 2010; Matsuoka, Kotani, & Yamasato, 2012). A research team (three paid research assistants and the first author) scored the discourse samples for CIUs. Eleven samples were randomly selected and CIUs were re-estimated for inter-reliability purposes. For 10 samples, the difference in %CIUs was less than 6 percentage points and for the eleventh there was an 8 percentage point discrepancy.

Statistical approach

Structural equation modelling (SEM) was applied to analyse the data (Bollen, 1989; Kline, 2010). SEM is a multivariate technique commonly used in psychometric investigations. Cognitive processes by their very nature cannot be directly observed nor measured the same way as other observed variables such as height or weight. However, they can be conceptualised as unobserved latent constructs and captured using SEM. Within the SEM framework, latent variables, also referred to as common factors, can be defined using behaviours that represent them. The observed behaviours are usually scores on manifest variables that reflect the influence of the common factors. By formally defining the relationship of the observed indicators and the underlying factor using a series of equations, one can assume "measurement" of the respective cognitive process. The variance of a manifest variable not accounted for by the common factors is referred to as the residual term or uniqueness. There are two sources of variance that combine comprise residual terms: *unique or specific factors*, that represent systematic variance associated with a specific indicator, and *random error* (e), often conceptualised as measurement error.

Conducting analyses to examine the strength of the relationship among groups of variables using SEM has several advantages over traditional approaches, because we can model explicitly the presence of latent common factors that are interpreted as the mathematical instantiations of the latent traits of interest. Therefore, when latent factors are correlated, the strength of the relationship that can take the statistical form of a correlation does not reflect error variance and systematic construct irrelevant variance.

Lexical access and retrieval in confrontation naming tests was conceptualised and modelled as an unobserved latent variable based on performance on the BNT, WAB-Naming, and VNT. Similarly, to answer the first research question, performance at the discourse level was also modelled as a latent variable based on the observed proportions of paraphasias from the three different types of discourse. Systematic common variance accounted for by the two unique constructs was estimated, separating these from construct-irrelevant noise and extraneous factors (e.g., test unreliability). Then, the factor scores formed based on confrontation naming tests were used to predict the factor scores that were based on the proportions of paraphasias during discourse production. The loadings of each factor indicated how strongly the latent variable influenced the observed variables, or, alternatively, how strongly an observed score (in a given test or based on a given type of discourse) reflected its corresponding underlying variable. Finally, the model stipulated that once the effect of the common factors was taken into account,

there was no systematic covariance among the residual terms of the observed indicators. To answer the second research question, the factor scores derived from the confrontation naming tasks were used to predict the observed CIU scores from the *Cinderella* story.

Four fit indices were taken into account to examine global model fit. Fit indices included the Satorra–Bentler scaled χ^2 statistic (Satorra & Bentler, 1994) to take into account the non-normality of the data, the comparative fit index (CFI; Bentler, 1990), the root-mean square error of approximation (RMSEA; Steiger & Lind, 1980), and the standard root-mean residual (SRMR; Hu & Bentler, 1999). A well-fitting model was expected to have a non-significant χ^2 at the 0.05 level; a CFI value >0.95; an RMSEA value <0.08, with the upper bound of the 90% confidence interval <0.10; and an SRMR value <0.05 (Brown, 2006; Hu & Bentler, 1999; Kline, 2010; Steiger, 2007). To assess for local strains in the solution, modification indices and normalised residuals were considered. When necessary, nested model comparisons were performed using the scaled difference χ^2 test statistic (Satorra & Bentler, 2001). Following Bollen (1989), the magnitude of the path coefficients and error covariances from the model were used to compare the relative influence of the factor on the manifest variables and to answer the substantive questions of the paper.

Results

Preliminary analysis

Data were prepared for statistical analysis following Kline (2010) and Tabachnick and Fidell (2007). After importing data in SPSS, they were screened for missing values and reasons for missingness were noted (e.g., some participants did not receive the full protocol, or were administered the full BNT instead of the short version). All discourse-based indices had complete data. Approximately 7%, 9%, and 5% of the data were missing in the WAB-Naming, BNT, and VNT, respectively. Missingness in the analyses was addressed using full information maximum likelihood under the assumption of data missing-at-random. Distributions were visually inspected and assessed in terms of the normality assumption. Several distributions were noted to be skewed and with various degrees of kurtosis. For this reason, the MLR estimator was used, which estimates parameters using maximum likelihood with standard errors and a χ^2 test statistic that are robust to non-normality. The correlational matrix of the key study variables can be found in Table 3.

Table 3. Correlation matrix of major study variables.

Variable	1	2	3	4	5	6	7
Confrontation naming tests							
1. BNT[a]	1						
2. WAB-Naming[b]	0.82	1					
3. VNT[c]	0.73	0.74	1				
Discourse-based indices							
4. Free Speech	−0.37	−0.39	−0.46	1			
5. Picture Description	−0.42	−0.45	−0.45	0.79	1		
6. Story Retell	−0.28	−0.27	−0.32	0.79	0.77	1	
7. CIU[d]	0.56	0.60	0.62	−0.37	−0.372	−0.136	1

Notes: [a]Boston Naming Test; [b]Western Aphasia Battery–R Naming Subtest; [c]Verb Naming Test; [d]Correct Information units.

Confrontation naming tests and paraphasias in discourse

A SEM model (Model A) with two latent variables and three indicators for each was fit to the data. The parameters of the cross-loadings and the covariances among the residual terms of the indicators were fixed to 0. To identify the model, the variances of the latent variables were set equal to 1. The model converged to a solution with no out-of-range parameter values. Even though the fit indices provided evidence of adequate model fit ($\chi^2(8, N = 98) = 15.338$, $p = 0.053$; CFI = 0.99; RMSEA = 0.097, 90% CI [0.00, –0.16]; and SRMR = 0.043) the assessment for local model strain suggested considerable misfit. Based on the modification indices, we re-specified the model and allowed the correlation between the residual variances of the BNT and the WAB-Naming tasks to be freely estimated. This decision was based on both statistical information and the fact that it was theoretically cogent given that these two tasks both use nouns whereas the VNT targets verbs. The re-specified model (Model B) converged to an admissible solution with improved model fit ($\chi^2(7, N = 98) = 12.013$, $p = 0.10$; CFI = 0.99; RMSEA = 0.083, 90% CI [–0.00, 0.16]; and SRMR = 0.036). Both factors were well defined given that the magnitude of the loadings was greater than 0.80 across all indicators. The regression coefficient between the two latent factors was estimated to be −0.52. The estimate had a negative value because higher scores on a confrontation naming test reflect better performance, whereas better performance in discourse was quantified by lower proportions of paraphasias. Thus, higher scores on confrontation naming tests tended to be associated with lower scores of proportions of paraphasias in discourse. The negative regression coefficient reflects this inverse relationship. The magnitude of the coefficient indicated that a one-unit increase in confrontation naming tests was associated with an expected decrease of 0.52 units in the proportion of paraphasias in discourse. The fully standardised solution of the model can be seen in Figure 2.

Confrontation naming tests and informativeness

An SEM model was specified to investigate the relationship between performance in confrontation naming tasks and informativeness (Model C). %CIUs were regressed on a latent variable that was formed based on the confrontation naming tasks. With the exception of BNT and WAB-R Naming, the covariances among the residual terms of the observed variables were fixed to 0. To identify the model, the variance of the latent variable was set equal to 1. The model converged to a solution with no out-of-range parameter values and its fit indices provided evidence of adequate model fit ($\chi^2(6, N = 98) = 0.299$, $p = 0.585$; CFI = 1.00; RMSEA = 0.00, 90% CI [0.00, 0.22]; and SRMR = 0.005). Furthermore, no local model strain was noted (highest normalised residual = –0.11; no modification indices with values >3.84). The fully standardised solution can be seen in Figure 3. The latent variable was well defined and the magnitude of the its estimated parameters was similar to the solution in Figure 2. The regression coefficient between the latent factor and %CIUs was estimated to be 0.68, which indicated that a one-unit increase in confrontation naming tests was associated with an expected increase of 0.68 units in the %CIU in discourse and the shared variance was 46%. The fully standardised solution of the model can be seen in Figure 3.

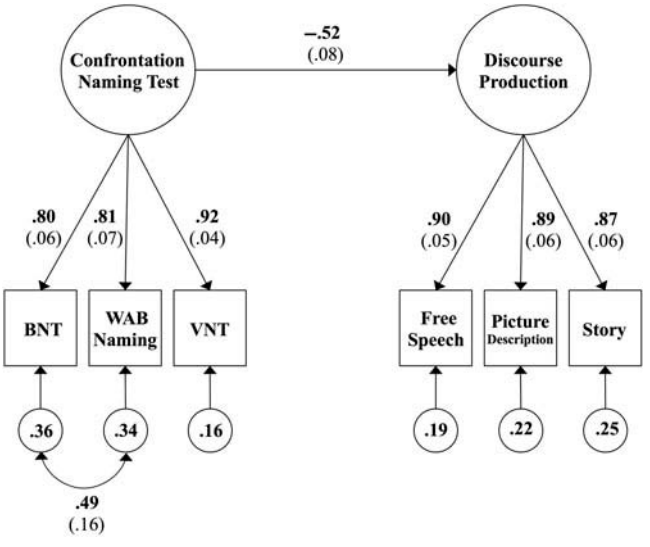

Figure 2. Model B Parameters and standard errors (SE). For each observed variable, the variance accounted for by the common factor, R2, is equal to (1 – residual variance) and is also equal to the λ2. For all parameter estimates, $p < 0.001$. –0.52 is a standardised regression coefficient such that (–0.52)2 is equal to the variance shared between the two factors.

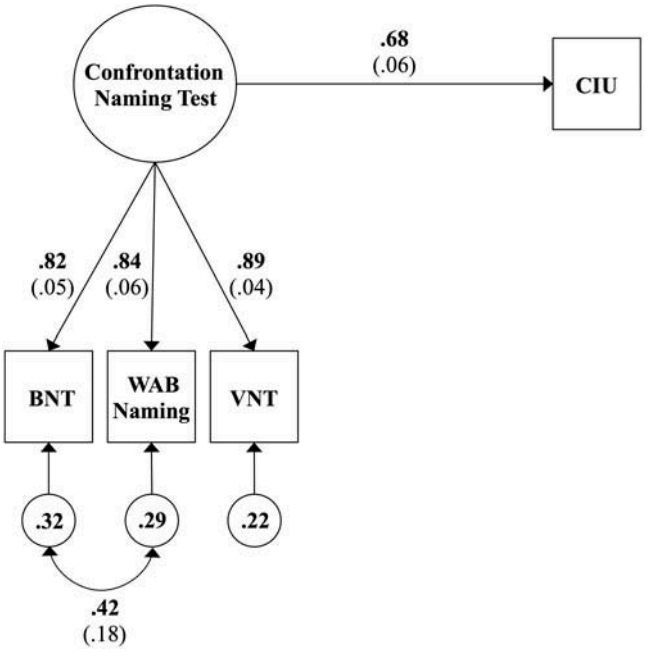

Figure 3. Model B Parameters and standard errors (SE).

Discussion

The purpose of the current study was to evaluate the inferential process involved in using confrontation naming tests as proxy measures to draw conclusions about the production of paraphasias in connected speech and discourse-based informativeness. Data from 98 monolingual PWA, retrieved from Aphasiabank, were included in this study. PWA were administered the WAB-Naming subtest, the Short Form of the BNT, and the VNT. Furthermore, proportions of paraphasias were estimated based on tasks for eliciting discourse including free speech, descriptions of sequential and single pictures, and story retelling. Informativeness, using %CIUs, was estimated based on story retelling. Using SEM, performance on confrontation naming tests was conceptualised and modelled as an unobserved latent variable based on the scores from the BNT, WAB-Naming, and VNT. Similarly, performance at the discourse level was modelled as a latent variable based on the observed indicators from the three different types of discourse. The variance-covariance modelling yielded several significant findings.

One of our main goals was to determine the strength of the relationship between performance on confrontation naming tests and the proportion of paraphasias in three different types of discourse when accounting for construct irrelevant variance. We found a significant regression coefficient (−0.52) between the two latent variables in the fitted model. This finding was consistent with previous reports that have investigated the relationship between performance at confrontation naming and discourse processing. When placed in the context of the argument structure from Figure 1, the magnitude of the regression coefficient represents a quantification of the strength of the warrant (i.e., its backing), and by extension of the overall validity argument. In this study, the regression coefficient was estimated to be equal to −0.52, which suggests a 27% overlap of the two latent variables. Consequently, over 70% of the variances across the two latent variables (reflecting the communalities of confrontation naming tests, and discourse indices, respectively) were unrelated. Therefore, based on the findings of this study, the inference involved in the argument in Figure 1 was not supported. These results are clinically relevant, because they suggest that it might not be useful or meaningful for professionals to use the overall performance on confrontation naming tests to draw inferences about the proportion of paraphasias a PWA might produce in connected speech.

The magnitude of the relationship between naming and discourse production was substantially lower than previous studies in the literature. This discrepancy may be attributed to the nature of the behaviours that were investigated in the current study. For example, Pashek and Tompkins (2002) reported a moderately strong relationship between the BNT and word finding difficulty during discourse production (−0.76, $p < 0.05$). Similarly, Mayer and Murray (2003) estimated the correlation between word-finding difficulties during discourse and performance on the Test of Adolescent and Adult Word Finding (German, 1990) to be higher than .70. In both studies, the researchers constructed composite indices reflecting anomic severity during discourse production that, in addition to paraphasias, included pauses, circumlocutions, self-corrections, and other observed behaviours of word-finding difficulty. In the current study, we included only paraphasias to evaluate a more narrowly defined diagnostic argument. Thus, we were able to investigate the specific inference with respect to paraphasias and disentangle it from other anomic behaviours. Our findings suggest that paraphasias produced in discourse in isolation are not well predicted by performance on confrontation naming tests.

An advantage of the current approach is that it can be elaborated in future studies to include additional behaviours of word-finding difficulty (e.g., pauses) to investigate how

each behaviour is uniquely associated with performance in confrontation naming tests over and above paraphasias. Furthermore, it could allow us to compare the utility of confrontation naming tests to predict different word-finding behaviours. This is not possible when behaviours are combined with equal weighting to form composites and then the composites are associated with another variable such as confrontation naming test performance. For example, in the Pashek and Tompkins study, it is not clear whether the high correlation is driven by the relationship of pauses with confrontation naming or the relationship between paraphasias and discourse production.

The inclusion of the CIU analysis allowed us to expand the focus on the study. CIU reflects the amount of information a PWA is able to communicate successfully during discourse production. It does not zoom in on a specific impaired aspect of discourse production but rather it reflects the abilities of a speaker at global level and it is useful in evaluating changes in response to treatment or experimental manipulations (Nicholas & Brookshire, 1993). The estimated regression coefficient in Model C (0.68) was moderate–high, suggesting a strong relationship between performance on confrontation naming tests and informativess at the discourse level. Furthermore, this relationship was stronger compared with that between confrontation naming performance and the proportion of paraphasic errors at the discourse level based on both zero-order correlations of the variables (Table 3) and the estimated path coefficients (0.68 vs. −0.52). Nonetheless, approximately half the variance in %CIUs was unaccounted for, which highlights that additional factors determine performance at the discourse level. Therefore, our results are consistent with findings in the literature that suggest that decontextualised tasks may not fully capture the communicative abilities of PWA.

It is important to note, however, that informativeness in this study was estimated based on a single observed indicator. Therefore, unlike the latent variable, the %CIUs reflect both (i) the underlying ability of a speaker and (ii) a combination of random measurement error and systematic, idiosyncratic variance associated uniquely with the single measurement. The latter source, consisting of random noise and construct irrelevant variance, is represented in the model as the residual variance associated with %CIUs (~54%). Therefore, theoretically it is possible that the relationship between the constructs reflected in confrontation naming tests and %CIUs would be higher if the residual variance was partialled out (e.g., by forming a latent variable) or reduced (e.g., by implementing procedures to further reduce measurement error). In practice, the first strategy may not be helpful for clinicians and researchers who do not have access to large samples and techniques such as SEM but rather rely on observed scores from a single individual at a time. On the other hand, our reliability suggests room for improvement that could be attained by designing scoring procedures that would decrease the uncertainty associated with the scoring CIUs.

With respect to the confrontation naming tests specifically, the results support those derived from previous studies. The zero-order correlations and the loadings of the confrontation naming tests were very high, indicating that performance across each of the tests could be accounted for primarily by a single factor. This factor can be thought of as a constellation of underlying cognitive processes, of which the level of impairment determines performance across the tests. In addition, the tests that assessed the ability to name nouns (i.e., BNT and WAB-R Naming) shared some additional variance (~8%) which could be attributed potentially to the nature of the stimuli. These effect sizes are consistent with previous studies in the literature that have reported the relationship among confrontation naming tests (e.g., Axelrod, Ricker, & Cherry, 1994; Yochim, Kane, & Mueller, 2009).

One limitation of the current study is that even though 98 participants is a fairly large sample for an aphasia study, it is smaller than typically seen in the SEM literature. Therefore concerns may be raised with respect to the sample size's adequacy for estimating the reported models and the likelihood of replicating the results in future studies. However, the robustness of the solutions is supported by two reasons. First, the parameter search of the maximum likelihood algorithm converged to an admissible solution with no Heywood cases (e.g., negative variances or standardised loadings greater than 1) that would have signalled misspecification (e.g., Kolenikov & Bollen, 2012). Furthermore, it has been demonstrated both analytically and via Monte Carlo simulations that the required sample size is a function of the model structure and population parameters. Specifically, the effects of sample size on the stability of the estimated loadings are diminished as communalities of the variables increase (e.g., Pennell, 1968; Velicer & Fava, 1998; Velicer, Peacock, & Jackson, 1982). Given that both factors were highly saturated and the estimated loading parameters were associated with small SEs (standard errors) (Figure 2), it is reasonable to assume that the population parameters were recovered accurately.

Furthermore, even though Model B demonstrated adequate fit based on the majority of the fit indices, the upper bound of the RMSEA associated with that model may be indicative of some model misspecification. Even though the RMSEA CI span for sample sizes $\leq$100 is often too wide to be used reliably when making rejection decisions about model fit (Chen, Curran, Bollen, Kirby, & Paxton, 2008), readers should use some caution when interpreting the results of Model B. For example, it is possible that some parameters should be freely estimated to account for the relationship between confrontation naming and picture description during which PWA sometimes label nouns instead of producing narratives with a story structure. However, even if that were the case, given the estimated loadings of the observed indicators to their respective factors and based on the modification indices of Model B, the additional parameters would have contributed a minimal increase in the explanatory power of the model. Nonetheless, it will be informative to explore this issue in future investigations when the sample size would be sufficient to detect minor misspecifications of the residual covariance structure (Kolenikov & Bollen, 2012).

Conclusions

The results of the study are promising as they contribute to the literature and understanding of the relationship between performance on confrontation naming tests and discourse tasks in adults with aphasia. The implicit clinical assumption that performance on confrontation naming tests is strongly related to the production of paraphasias during discourse was not supported. Given that the word-retrieval errors included in the model at the discourse level were restricted to paraphasias, it is possible that other anomic behaviours, such as pauses and circumlocutions, may be associated with confrontation naming test performance. Such word-retrieval errors at the discourse level should be considered in future investigations; and, applying the current approach will allow us to investigate how each behaviour is uniquely associated with performance on confrontation naming tests. Further, the CIU analysis suggested that confrontation naming tasks may be a better predictor of performance of the ability to communicate informative content efficiently. However, in both cases, our findings suggest that studying discourse may yield unique information that may be clinically relevant.

Acknowledgements

We are especially grateful to the study participants and AphasiaBank. We also thank the volunteers in the Aging and Adult Language Disorders Lab at Portland State University and especially Caroline Crone, Jade Horton, and Audrey Cohen for assistance with language analyses.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Armstrong, E. (2000). Aphasic discourse analysis: The story so far. *Aphasiology*, *14*, 875–892. doi:10.1080/02687030050127685

Axelrod, B. N., Ricker, J. H., & Cherry, S. A. (1994). Concurrent validity of the MAE visual naming test. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *9*, 317–321. doi:10.1093/arclin/9.4.317

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

Bollen, K. A. (1989). *Structural equation modeling with latent variables*. New York, NY: Wiley.

Boyle, M. (2014). Test–retest stability of word retrieval in aphasic discourse. *Journal of Speech, Language & Hearing Research*, *57*, 966–978. doi:10.1044/2014_JSLHR-L-13-0171

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Cameron, R. M., Wambaugh, J. L., & Mauszycki, S. C. (2010). Individual variability on discourse measures over repeated sampling times in persons with aphasia. *Aphasiology*, *24*, 671–684. doi:10.1080/02687030903443813

Carlomagno, S., Giannotti, S., Vorano, L., & Marini, A. (2011). Discourse information content in non-aphasic adults with brain injury: A pilot study. *Brain Injury*, *25*, 1010–1018. doi:10.3109/02699052.2011.605097

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, *36*, 462–494. doi:10.1177/0049124108314720

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

Cruice, M., Worrall, L., Hickson, L., & Murison, R. (2003). Finding a focus for quality of life with aphasia: Social and emotional health, and psychological well-being. *Aphasiology*, *17*, 333–353. doi:10.1080/02687030244000707

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*, 283–321. doi:10.1037/0033-295X.93.3.283

Doyle, P. J., Tsironas, D., Goda, A. J., & Kalinyak, M. (1996). The relationship between objective measures and listeners' judgments of the communicative informativeness of the connected discourse of adults with aphasia. *American Journal of Speech-Language Pathology*, *5*, 53. doi:10.1044/1058-0360.0503.53

Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology*, *22*, S397. doi:10.1044/1058-0360(2013/12-0083)

German, D. J. (1990). *Test of adolescent/adult word finding*. Allen, TX: DLM.

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

Grimes, N. (2005). *Walt Disney's Cinderella*. New York, NY: Random House.

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*, 184–203. doi:10.1080/02687030701262613

Hickin, J., Best, W., Herbert, R., Howard, D., & Osborne, F. (2001). Treatment of word retrieval in aphasia: Generalisation to conversational speech. *International Journal of Language & Communication Disorders*, *36*, 13–18. doi:10.3109/13682820109177851

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*, 1–55. doi:10.1080/10705519909540118

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston naming test* (2nd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Kemmerer, D., & Tranel, D. (2000). Verb retrieval in brain-damaged subjects: 2. Analysis of errors. *Brain and Language*, *73*, 393–420. doi:10.1006/brln.2000.2312

Kertesz, A. (1982). *Western aphasia battery test manual*. New York, NY: Grune & Stratton.

Kertesz, A. (2007). *Western aphasia battery—R*. New York, NY: Grune & Stratton.

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

Kolenikov, S., & Bollen, K. A. (2012). Testing negative error variances: Is a Heywood case a symptom of misspecification? *Sociological Methods & Research*, *41*, 124–167. doi:10.1177/0049124112442138

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

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

Marini, A., Andreetta, S., del Tin, S., & Carlomagno, S. (2011). A multi-level approach to the analysis of narrative language in aphasia. *Aphasiology*, *25*, 1372–1392. doi:10.1080/02687038.2011.584690

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

Matsuoka, K., Kotani, I., & Yamasato, M. (2012). Correct information unit analysis for determining the characteristics of narrative discourse in individuals with chronic traumatic brain injury. *Brain Injury*, *26*, 1723–1730. doi:10.3109/02699052.2012.698789

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

McNeil, M. R., Doyle, P. J., Fossett, T. R. D., Park, G. H., & Goda, A. J. (2001). Reliability and concurrent validity of the information unit scoring metric for the story retelling procedure. *Aphasiology*, *15*, 991–1006. doi:10.1080/02687040143000348

Menn, L., Reilly, K. F., Hayashi, M., Kamio, A., Fujita, I., & Sasanuma, S. (1998). The interaction of preserved pragmatics and impaired syntax in Japanese and English aphasic speech. *Brain and Language*, *61*, 183–225. doi:10.1006/brln.1997.1838

Mislevy, R. J., & Yin, C. (2009). If language is a complex adaptive system, what is language assessment? *Language Learning*, *59*, 249–267. doi:10.1111/j.1467-9922.2009.00543.x

Nicholas, L. E., & Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech Language and Hearing Research*, *36*, 338. doi:10.1044/jshr.3602.338

Nicholas, L. E., Brookshire, R. H., MacLennan, D. L., Schumacher, J. G., & Porrazzo, S. A. (1989). Revised administration and scoring procedures for the Boston Naming test and norms for non-brain-damaged adults. *Aphasiology*, *3*, 569–580. doi:10.1080/02687038908249023

Pashek, G. V., & Tompkins, C. A. (2002). Context and word class influences on lexical retrieval in aphasia. *Aphasiology*, *16*, 261–286. doi:10.1080/02687040143000573

Pennell, R. (1968). The influence of communality and N on the sampling distributions of factor loadings. *Psychometrika*, *33*, 423–439. doi:10.1007/BF02290161

Prins, R., & Bastiaanse, R. (2004). Analysing the spontaneous speech of aphasic speakers. *Aphasiology*, *18*, 1075–1091. doi:10.1080/02687030444000534

Ross, K. B., & Wertz, R. T. (1999). Comparison of impairment and disability measures for assessing severity of, and improvement in, aphasia. *Aphasiology*, *13*, 113–124. doi:10.1080/026870399402235

Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (1989). The quantitative analysis of agrammatic production: Procedure and data. *Brain and Language*, *37*, 440–479. doi:10.1016/0093-934X(89)90030-8

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. Van Eye & C. C. Clogg (Eds.), *Latent variable analysis in developmental research* (pp. 285–305). Thousand Oaks, CA: SAGE Publications.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507–514. doi:10.1007/BF02296192

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*, 228–264. doi:10.1016/j.jml.2005.10.001

Shewan, C. M. (1988). The Shewan Spontaneous Language Analysis (SSLA) system for aphasic adults: Description, reliability, and validity. *Journal of Communication Disorders*, *21*, 103–138. doi:10.1016/0021-9924(88)90001-9

Spreen, O., & Risser, A. (2003). *Assessment of aphasia*. New York, NY: Oxford University Press.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, *42*, 893–898. doi:10.1016/j.paid.2006.09.017

Steiger, J. H., & Lind, J. C. (1980, May 30). *Statistically-based tests for the number of common factors*. Presented at the annual spring meeting of the Psychometric Society, Iowa City, IA.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson/Allyn & Bacon.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Velicer, W. F., & Fava, J. L. (1998). Affects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*, 231–251. doi:10.1037/1082-989X.3.2.231

Velicer, W. F., Peacock, A. C., & Jackson, D. N. (1982). A comparison of component and factor patterns: A Monte Carlo approach. *Multivariate Behavioral Research*, *17*, 371–388. doi:10.1207/s15327906mbr1703_5

Williams, S. E., & Canter, G. J. (1982). The influence of situational context on naming performance in aphasic syndromes. *Brain and Language*, *17*, 92–106. doi:10.1016/0093-934X(82)90007-4

Wilshire, C. E., & McCarthy, R. A. (2002). Evidence for a context-sensitive word retrieval disorder in a case of nonfluent aphasia. *Cognitive Neuropsychology*, *19*, 165–186. doi:10.1080/02643290143000169

Yochim, B. P., Kane, K. D., & Mueller, A. E. (2009). Naming test of the neuropsychological assessment battery: Convergent and discriminant validity. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *24*, 575–583. doi:10.1093/arclin/acp053

Yorkston, K. M., & Beukelman, D. R. (1980). An analysis of connected speech samples of aphasic and normal speakers. *Journal of Speech and Hearing Disorders*, *45*, 27. doi:10.1044/jshd.4501.27