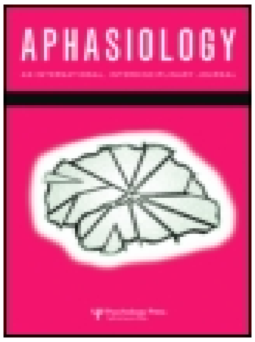

Aphasiology

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

Modeling confrontation naming and discourse informativeness using structural equation modeling

Gerasimos Fergadiotis, Maria Kapantzoglou, Stephen Kintz & Heather Harris Wright

APHASIOLOGY
https://doi.org/10.1080/02687038.2018.1482404

Routledge
Taylor & Francis Group

Check for updates

Modeling confrontation naming and discourse informativeness using structural equation modeling

Gerasimos Fergadiotis[a], Maria Kapantzoglou[a], Stephen Kintz[b] and Heather Harris Wright[c]

[a]Department of Speech and Hearing Sciences, Portland State University, Portland, OR, USA; [b]Department of Audiology and Speech Pathology, University of Arkansas at Little Rock, Little Rock, AR, USA; [c]Department of Communication Sciences and Disorders, East Carolina University, Greenville, NC, USA

ABSTRACT

Background: People with aphasia (PWA) and their families identify as their priority the ability to use language at the discourse level in order to meet their daily communicative needs. However, measuring connected speech can be a challenging task due to the complex and multidimensional nature of discourse. As a result, professionals often depend on confrontation naming tests to identify and measure impaired underlying cognitive mechanisms that are also hypothesized to be important for discourse production.

Aims: In the current study, we investigated the validity of making inferences about discourse performance based on scores from confrontation naming tests. Specifically, we investigated the strength of the relationship between word retrieval abilities, and the ability to convey information during discourse production.

Method & Procedures: Data from 118 monolingual PWA were retrieved from AphasiaBank and analyzed using structural equation modeling. Performance in confrontation naming tests was modeled 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 modeled based on indices of informativeness in three discourse tasks (free speech, eventcasts, and story re-tell). Informativeness was quantified using the percentage of Correct Information Units.

Outcomes and Results: Based on the fit statistics, the model exhibited adequate fit, indicating that the relationship between confrontation picture naming and informativeness was adequately reflected in the model. We found a strong relationship between confrontation naming test performance and discourse informativeness (standardized regression coefficient between the two latent factors = .79).

Conclusions: Performance on confrontation naming tests was a strong predictor of the amount of information PWA communicated during discourse production. However, our results also highlight that performance on the latter cannot be predicted solely from the former, as evidenced by the large proportion of unexplained variance in the informativeness latent variable.

ARTICLE HISTORY
Received 27 January 2018
Accepted 13 May 2018

KEYWORDS
Aphasia; discourse informativeness; word retrieval deficits; diagnostic procedures; structural equation modeling

CONTACT Gerasimos Fergadiotis gfergadiotis@pdx.edu Department of Speech and Hearing Sciences, Portland State University, PO Box 751, Portland, OR 97201

One of the hallmark characteristics of people with aphasia (PWA) is anomia (Goodglass & Wingfield, 1997), a deficit which is indicative of disruption in accessing a semantic description of the target concept and/or retrieving a fully phonologically specified representation (Dell, Chang, & Griffin, 1999; Schwartz, Dell, Martin, Gahl, & Sobel, 2006). Given their impact on effective communication, the nature and severity of anomia are typically considered for making a differential diagnosis and establishing the level of severity in PWA. The most commonly used method of assessment is confrontation naming tests that assess single-word production. Confrontation naming tests are routinely included in aphasia testing batteries in both clinical and research settings and their popularity can be attributed to their many advantages described in the literature (e.g., Herbert, Hickin, Howard, Osborne, & Best, 2008; Mayer & Murray, 2003). For example, the majority of confrontation naming tests follow a uniform administration procedure and are relatively quick and straightforward to score and analyze. Further, well-designed confrontation naming tests exhibit high test-retest reliability, and moderate-to-high intercorrelations suggesting high degree of construct validity (e.g., Fergadiotis, Kellough, & Hula, 2015; Walker & Schwartz, 2012). Another significant advantage is that the target words are known, and therefore, the ambiguity when scoring the responses is minimized.

At the same time, PWA and their families identify as their priority the ability to use language at the discourse level in order to meet their daily communicative needs (Cruice, Worrall, Hickson, & Murison, 2003; Wallace et al., 2017). Over the years, there has been an increasing emphasis on assessing discourse that is reflected by the number of systematic approaches that have been developed to capture multiple facets of discourse production (Marini, Andreetta, Del Tin, & Carlomagno, 2011; Saffran, Berndt, & Schwartz, 1989; Sherratt, 2007; Shewan, 1988). Further, there is an increase in the utilization of discourse measures in research applications for assessing the generalization of treatment outcomes to the discourse level (Bryant, Ferguson, & Spencer, 2016).

Measuring connected speech in typical clinical settings is a challenging task due to the multidimensional and dynamic nature of discourse (Armstrong, 2000; Prins & Bastiaanse, 2004; Spreen & Risser, 2003). Under certain conditions, discourse measures can yield unreliable scores or may reflect constructs that were not intended to be captured by researchers (Boyle, 2014, 2015; Cameron, Wambaugh, & Mauszycki, 2010; Fergadiotis, Wright, & West, 2013; Ross & Wertz, 1999). Additionally, analyzing discourse is resource intensive because in most cases it requires recording and transcribing language samples, which may be a prohibitive activity in clinical settings with high productivity requirements. As a result, professionals often depend on confrontation naming tests to identify and measure impaired underlying cognitive mechanisms that are also hypothesized to be important for discourse production. An important point is that this clinical inference is valid only to the extent that the performance in single-word, picture-naming tasks is strongly related to the ability to communicate effectively during discourse production (i.e., *informativeness*; Fergadiotis & Wright, 2015). The purpose of the current study, then, is an extension of our previous study (Fergadiotis & Wright, 2015) to determine the strength of the relationship between single-word production and discourse informativeness during narrative production in PWA. In the following sections, first we review discourse informativeness and measurement using correct information units (Nicholas & Brookshire, 1993). Then, we summarize our previous work on this topic that led to the current study.

Discourse informativeness

Due to the multidimensional nature of discourse, numerous measures have been developed for analyzing connected speech. According to Armstrong (2000), measures for analyzing discourse can be associated with one of three broad perspectives: structuralism, functionalism, or a mixture of structuralism and functionalism. Structuralism is concerned with describing discourse through its constituent parts. Functionalism is concerned with how discourse is successfully used to convey information. Measures that utilize both perspectives often quantify specific structures (e.g., words) to measure transactional success of the discourse (Pritchard, Hilari, Cocks, & Dipper, 2017). For example, a structuralist might want to know how many nouns or verbs a PWA produced. A functionalist might want to know how well a listener comprehended the discourse. Mixed measures may count the number of well-formed and grammatical words to determine how well the discourse could be comprehended.

It is these mixed measures that have proven to be the most reliable and valid. Correct information units (CIU) is a rule-based system for quantifying the amount of information produced in discourse samples (Nicholas & Brookshire, 1993), and it is the measure we used to quantify informativeness in the current study. The first attempt to quantify information was by Yorkston and Beukelman (1980) who produced content units for specific stimuli. CIUs extended the ability to quantify discourse samples regardless of genre (Nicholas & Brookshire, 1993). Bryant et al. (2016) conducted a review on the various linguistic measures used to analyze discourse within aphasiology, and the researchers found 536 different measures across 165 studies. The most commonly used measures fell under language productivity, with "number of words" being utilized in 55 studies and information content with CIUs utilized in 43 different studies. Pritchard et al. (2017) further analyzed the work of Bryant and colleagues by reanalyzing measures that directly measure the amount of information produced. The researchers found that CIUs and measures derived from CIUs had high test-retest reliability (>0.80), inter-rater reliability (agreement >0.80), and intra-rater reliability (agreement >0.80). Therefore, CIUs appear to be a reliable measure of communication efficiency as well as a valid theoretical construct that bridges the gap between structuralist and functional perspectives.

Discourse informativeness and single-word production

Recently, Fergadiotis and Wright (2015) conducted a study to evaluate the inferential process involved in using confrontation naming tests as proxy measures to reach conclusions regarding the ability of PWA to produce informative discourse. In that study, structural equation modeling was used (SEM; Bollen, 1989; Kline, 2010), which is a multivariate technique commonly applied in psychometric investigations. Using SEM to examine the strength of the relationship between variables is advantageous over other approaches, because we can model explicitly latent common factors that are interpreted as the mathematical instantiations of the latent traits of interest. Importantly, SEM allows us to estimate the strength of the relationship between latent variables while partialling out measurement error and systematic construct-irrelevant variance that typically obscure the true relationship of constructs.

Fergadiotis and Wright (2015) analyzed data from 98 monolingual PWA, based on language samples retrieved from AphasiaBank, an online, shared database that collects and analyzes digital recordings of discourse from PWA across a series of tasks

(MacWhinney, Fromm, Forbes, & Holland, 2011). Using SEM, single-word retrieval ability was conceptualized and modeled as an unobserved latent variable determining the performance on three confrontation naming tests. Informativeness was modeled with the percentage of CIUs in discourse elicited with story re-telling. Although informativeness is typically conceptualized as an unobservable latent variable (i.e., factor), similarly to word retrieval abilities, it could not be modeled as such, because there was a single measure for it (i.e., CIUs). Typically, three observable measures are required per factor for an SEM model to be statistically identified. As a result, unlike the word retrieval latent variable, the estimates of informativeness were contaminated by a combination of random measurement error and systematic, idiosyncratic variance associated uniquely with the single measurement. Nonetheless, the standardized regression coefficient was moderate-high (.68) suggesting a relatively strong relationship between word retrieval abilities based on confrontation naming tests and informativeness at the discourse level. The authors posited that the relationship could be stronger if error-free estimates of people's underlying ability to convey informative content in discourse were modeled using multiple indicators. The current study sought to address this limitation by including CIU estimates from multiple language samples per PWA so that an error-free estimate of informativeness could be computed. Specifically, in the current study, we utilized SEM to determine the strength of the relationship between underlying word retrieval abilities and discourse informativeness while accounting for construct irrelevant variance (i.e., random noise and irrelevant systematic variance). Word retrieval abilities were based on performance on three confrontation naming tests and informativeness was quantified by the proportion of CIUs in three different types of discourse.

Method

Participants

Data from 118 monolingual PWA were included in this study. Notably, the current sample included data from 98 PWA from the earlier study (Fergadiotis & Wright, 2015). The language samples and test scores for all participants were randomly selected and retrieved from AphasiaBank (MacWhinney et al., 2011). All participants had acquired aphasia secondary to a 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, aphasia severity, and demographic information are presented in Table 1.

Discourse elicitation & data preparation

Stimuli and instructions

Language samples were collected in a single session using a variety of elicitation techniques that included free speech, narratives based on sequential and single pictures (i.e., eventcasts), and re-telling of the story of *Cinderella* (Grimes, 2005). Free speech

Table 1. Participants' descriptive information.

Characteristic	
Gender ratio	66 M; 52 F
Age in years (*SD*)	63.23 (11.79)
Years post-onset (*SD*)	5.36 (5.30)
Mean WAB-R AQ[a] (*SD*)	67.07 (17.41)
WAB-R classification	
Anomic	39
Broca	26
Wernicke	19
Conduction	27
Global	2
Transcortical motor	4
Transcortical sensory	1
Ethnicity	
African-American	12
Asian	1
Hispanic	1
White	98
Other	2
Education	
Some high-school	2
12th grade	29
Some college	22
Bachelor's or higher	59

Ethnicity information was unavailable for four individuals and education information was unavailable for six individuals. Standard deviations are shown in parentheses.
[a]Aphasia Quotient.

narratives 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 elicitation script for free speech, as well as the remaining techniques can be found at http://aphasia.talkbank.org/protocol/instructions.pdf.

The sequential picture stimuli included a four-frame strip from Menn et al. (1998) referred to as "*Broken Window*" that 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 sequential picture stimulus was a six-frame cartoon strip referred to as "*Umbrella*" which 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) that shows 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. The single and sequential picture stimuli can be found at: http://aphasia.talkbank.org/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 s, 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 were presented with and were asked to tell 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. Further, 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 confrontation naming tests as part of a larger testing battery. These included the Western Aphasia Battery – Revised Naming subtest (WAB-R Naming; Kertesz, 2007), the Short Form of the Boston Naming Test (BNT; Kaplan, Goodlgass, & Weintraub, 2001), and the Verb Naming Test (VNT; Cho-Reyes & Thompson, 2012). Scoring for each of those tests was performed based on the published standardized 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.

Correct information units

To quantify informativeness, CIUs were identified for each type of discourse. 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). 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). In addition, 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. The raters were trained to criterion prior to analyzing the data presented in this study. The criterion was over 90% agreement. In addition, 64 samples (approximately 21 samples from each type of discourse) were randomly selected and CIUs were analyzed again for inter-reliability purposes. The average disagreement was 3.10% across all samples reanalyzed for reliability purposes. For the majority of the samples, the difference in %CIUs was less than six percentage points. For the remaining samples, one sample exhibited a seven-percentage point discrepancy and two samples exhibited an eight-percentage point discrepancy. Finally, we estimated the proportion of CIUs per number of words (%CIU) for each language sample elicited using the different elicitation techniques.

Statistical approach

Lexical access and retrieval in confrontation naming tests was conceptualized and modeled as an unobserved latent variable based on performance on the BNT, WAB-R Naming, and VNT within an SEM framework. Similarly, to answer our research question, informativeness was also modeled as a latent variable based on the observed proportions of CIUs from the three different types of discourse. Systematic common variance accounted for by the two unique constructs was estimated based on the communalities of the observed variables, separating them from construct irrelevant noise and extraneous factors (e.g., test unreliability). Then, factor scores formed based on confrontation naming tests were used to predict the factor scores that were based on the proportions of CIUs 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. For example, observed scores on each confrontation naming test were determined by the underlying naming ability of a person and a residual term that can be conceptualized as test unreliability. Thus, the model did not assume any other sources of systematic influence on the observed scores other than the ability to name pictures and produce informative content, for the two latent variables, respectively.

Four fit indices were taken into account to examine global model fit. Global fit statistics reflect how well the model replicates the observed data and specifically, the variance-covariance matrix of the observed indicators. Fit indices included the Satorra–Bentler scaled χ^2 statistic (Satorra & Bentler, 1994) to take into account the nonnormality 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 nonsignificant χ^2 at the .05 level; a CFI value >.95; an RMSEA value <.08, with the upper bound of the 90% confidence interval <.10; and an SRMR value <.05 (Brown, 2006; Hu & Bentler, 1999; Kline, 2010; Steiger, 2007). To assess for local strain in the models, modification indices were considered. Local strain refers to examining and identifying the different parts of the model for unnecessary parameters that hurt fit or missing parameters that might improve local fit. Following Bollen (1989), the magnitude of the path coefficients from the model were used to compare the relative influence of the factor on the manifest variables and to answer the substantive question of the paper.

Results

Preliminary analysis

Data were prepared for statistical analysis following Kline (2010) and Tabachnick and Fidell (2007). First, data were screened for missing values in SPSS. Approximately 2% of the data were missing across all variables as some participants did not receive the full protocol, or were administered the full Boston Naming test instead of the short version.

Missing values in the dataset were accommodated using full information maximum likelihood under the assumption of data missing-at-random (Enders, 2010). 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 maximum likelihood robust (MLR) estimator was used in MPlus, which estimates parameters using maximum likelihood with standard errors and a χ^2 test statistic that are robust to non-normality. Using SPSS, data were screened for outliers univariately (scores > 3.3 *SD*'s beyond the mean) and multivariately with Mahalanobis distance (p values < .001) taking into account the data points' leverage and influence. No outliers were noted in the dataset. The correlation matrix of the key study variables can be found in Table 2. Means and standard deviations are also provided as descriptive statistics and to allow interested readers to convert the correlation matrix to a variance-covariance matrix and replicate the analysis presented in this paper.

Main analysis

An SEM model was specified to investigate the relationship between performance in confrontation naming tasks and informativeness (Model A). A common factor formed based on performance on the three confrontation naming tests was used to predict a common factor formed based on the observed percentages of CIUs per type of discourse. To set the metric of each factor, all loadings were freely estimated and the variance of each factor was set equal to 1.

The covariances among the residual terms of the observed variables were fixed to 0.

The model converged to a solution with no out-of-range parameter values and its global fit indices provided evidence of excellent model fit, $\chi^2(8, N = 118) = 2.21$, $p = .974$; CFI = 1.00; RMSEA = .00, 90% CI [.00, .00]; and SRMR = .010. Further, no local model strain was noted (i.e., no modification indices with values >3.84). All parameter estimates were statistically significant at the .001 level.

The path diagram of the fully standardized solution of the model is presented in Figure 1. Consistent with SEM conventions for graphical model representation, squares indicate observed variables (e.g., percentage of CIUs based on free speech). Circles represent unobserved latent variables that are estimated as part of the model (e.g.,

Table 2. Correlation matrix of major study variables.

Variable	1	2	3	4	5	6
Confrontation naming tests						
1. BNT[a]	1					
2. WAB naming[b]	0.80	1				
3. VNT[c]	0.74	0.78	1			
Discourse-based %CIUs[d]						
4. Free speech	0.40	0.42	0.42	1		
5. Picture description	0.57	0.55	0.52	0.45	1	
6. Story re-tell	0.54	0.55	0.53	0.47	0.61	1
Mean	6.14	40.89	13.41	0.40	0.32	0.37
SD	4.70	17.02	6.56	0.24	0.20	.22

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

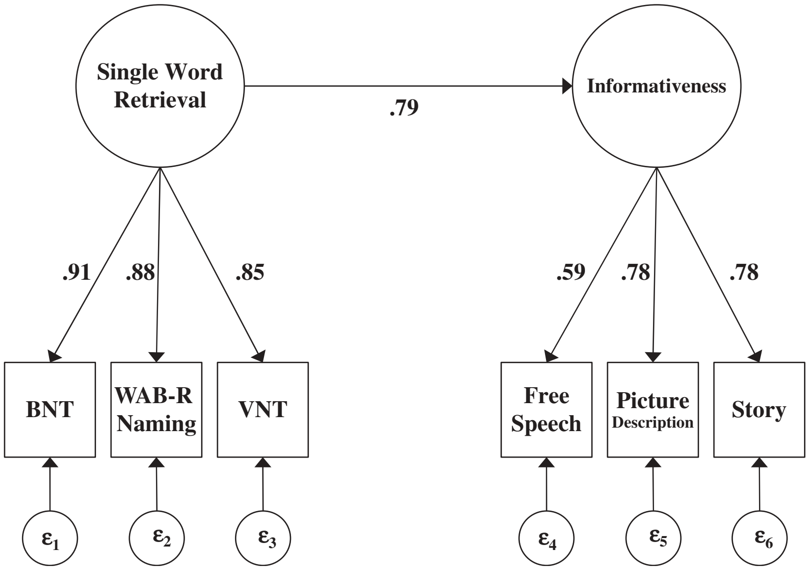

Figure 1. Two latent variables underlying the observed variables for (i) confrontation naming (left) and (ii) %CIUs during discourse production (right). For each factor, the observed variables are conditionally independent after accounting for their respective common factor. The regression coefficient (.79) captures the utility of performance in confrontation naming to predict proportions of CIUs during discourse production. The loadings reflect the strength of the relationship between the observed variable (e.g., BNT) and the construct it purports to measure minus any construct irrelevant variance partitioned in the error terms.

error- and bias-free estimate of single-word retrieval). In this model, straight arrows from the latent variables to observed variables are standardized factor loadings and represent the effect of latent variables on observed variables (or, alternatively, how strongly an observed indicator reflects the constructs of interest). The straight arrow between latent variables is a standardized regression coefficient. Both latent variables were well defined. Overall, loadings associated with the confrontation naming tests were higher than the loadings of the CIUs indicators. The regression coefficient between the two latent factors was estimated to be .79 which indicated that a one unit increase in confrontation naming tests was associated with an expected increase of .79 units in the %CIU in discourse. Further, based on the square of the estimate of the regression coefficient ($.79^2$), the shared variance between the two latent variables was approximately 63%. Figure 2 is a scatterplot of the factor scores across the sample.

Discussion

This study aimed to evaluate the implicit assumption professionals often make when using confrontation naming tests to reach conclusions regarding the ability of PWA to convey informative content during discourse production. Specifically, we analyzed data from 118 PWA who had scores on the WAB-R Naming subtest, the short form of the BNT, and the VNT. Further, %CIUs were estimated based on three tasks for eliciting discourse including free speech, descriptions of sequential and single pictures, and

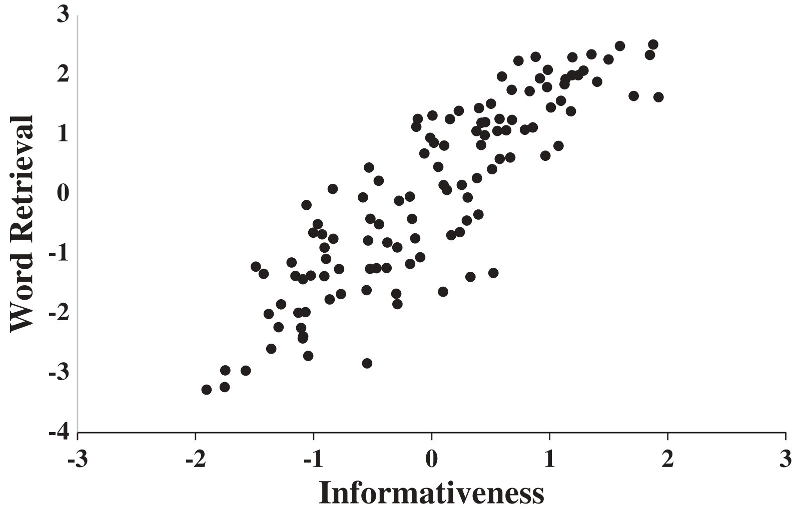

Figure 2. Scatterplot of factor scores based on confrontation naming tests (Y axis) and %CIUs (X axis).

story re-telling. Using SEM, single-word naming ability was conceptualized and modeled as an unobserved latent variable determining the scores on the three confrontation naming tests. Similarly, the ability to convey informative content during discourse production was modeled as a latent variable determining the scores on the observed measure (CIUs) from the three different types of discourse. The SEM approach allowed us to disentangle the theoretical constructs of interest from construct-irrelevant variance to study the strength of the relationship between single-word retrieval ability and the ability to convey informative content during discourse production.

Based on the SEM model, we found a strong relationship between single-word production abilities and the ability to convey information during discourse production. The regression coefficient between the factors representing single-word retrieval and informativeness was .79. These results suggest that word retrieval abilities accounted for approximately 63% ($.79^2$) of the variance in the ability to convey informative content in discourse. This finding was consistent with our previous investigation (Fergadiotis & Wright, 2015) which indicated that single-word production ability was also strongly related to the percentage of CIUs based on a single type of discourse (i.e., story telling; R^2 = .46). The regression coefficient in the current study was higher because the inclusion of additional types of discourse allowed us to partial out measurement error which had an attenuating effect (Spearman, 1904) on the regression estimate in the 2015 study. This is the first study that through the use of SEM quantified with accuracy the relationship between single-word production and an important aspect of discourse production in PWA (ability to produce informative content). Without the use of multiple indicators to triangulate the true ability of PWA to convey informative discourse, our prior investigation (Fergadiotis & Wright, 2015) as well as similar studies can only claim to have found the lower bound of the true relationship between to the two constructs of interest.

Importantly, the analysis presented in this paper was based on a larger sample which helped address another limitation from our previous investigation (Fergadiotis & Wright, 2015). The model in the 2015 study demonstrated adequate fit based on the majority of the fit indices, but the upper bound of the RMSEA suggested that the RMSEA fit index may not have been precise enough (90% CI [.00, .22]). In the SEM literature, it has been known that the RMSEA CI span for sample sizes ≤100 is often too wide to be used reliably when making decisions about model fit (Chen, Curran, Bollen, Kirby, & Paxton, 2008). However, in the current study, the RMSEA point estimate was consistent with adequate model fit, and its CI suggested adequate precision.

Even though the ability to access and retrieve single words was found to be a significant predictor of the difficulties PWA face when producing discourse, our results also highlight that performance on the latter cannot be predicted solely from the former, as evident by the proportion of unexplained variance in the informativeness factor. Similar findings have been reported in the literature before, across a variety of different measures of lexical retrieval including counts of content units (e.g., nouns and verbs), proportion of words correctly retrieved, and word retrieval errors (e.g., circumlocutions, paraphasias; Hickin, Best, Herbert, Howard, & Osborne, 2001; Mayer & Murray, 2003; Pashek & Tompkins, 2002; Williams & Canter, 1982). Conversely, there are several reports of PWA whose performance on confrontation naming tests differs substantially from their ability to produce discourse (e.g., Ingles, Mate-Kole, & Connolly, 1996; Manning & Warrington, 1996; Schwartz & Hodgson, 2002). The proportion of unexplained variance of the ability to convey informative discourse in our model could be attributed to a number of factors. During discourse production, in addition to the core processes for retrieval of single words (e.g., Dell et al., 1999), contextual information at the discourse level may influence lexical retrieval. Further, cognitive operations that extend beyond lexical processing (Berndt, Haendiges, Mitchum, & Sandson, 1997; Pashek & Tompkins, 2002; Williams & Canter, 1982; Wilshire & McCarthy, 2002) may influence discourse production, such as discourse grammar and coherence. Syntactic, structural and/or pragmatic elements are expected to influence the informativeness in discourse but are not typically measured with single-word naming tests.

Further, some PWA are more efficient with the strategic deployment of lexical items during discourse production. Depending on the nature of the deficits they exhibit and their self-monitoring skills, some PWA are better at anticipating, and sidestepping or reformulating utterances when they experience difficulties accessing and retrieving the intended words (e.g., Andreetta, Cantagallo, & Marini, 2012; Hartsuiker & Kolk, 2001). As a result, some speakers are more competent during discourse production to the extent that they can halt and repair ongoing productions either preemptively or after an erroneous production has been generated. However, the same individuals may not benefit from these compensation strategies during a confrontation naming task because significant delays or circumlocutions are typically marked as errors. The differential effect of these strategies across highly structured tasks (e.g., confrontation naming tests) and tasks with less constraints (e.g., narrative tasks) can dilute the strength of the relationship between single-word retrieval abilities and informativeness in discourse production.

An advantage of the SEM approach is that SEM models can be extended to build conceptual maps that represent the complex interrelationships of neuropsychological phenomena. In doing so, researchers can derive error free estimates of constructs of interest, specify the nature of their relationships, and test competing hypotheses to reach substantive conclusions. Multivariate techniques such as SEM can help us develop models that serve as blueprints for understanding the symptomatology and underlying cognitive deficits present in PWA (e.g., Nozari & Faroqi-Shah, 2017); study the links between abstract psychosocial variables and functioning and participation (e.g., Perrin, Heesacker, Stidham, Rittman, & Gonzalez-Rothi, 2008; Tatsumi et al., 2016); and assess the psychometric properties of tests and scales (Fergadiotis et al., 2015; Hula et al., 2015). One of the barriers for utilizing such approaches in aphasiology and other clinical fields has been that parameter estimation is typically based on techniques such as maximum likelihood. Maximum likelihood is only asymptotically efficient and it requires large sample sizes to achieve its desirable properties (Myung, 2003). However, large databases such as AphasiaBank that include data from hundreds of PWA can have a catalytic effect on our field by allowing us to use sophisticated techniques such as SEM and serve as a springboard for developing and testing hypotheses.

Clinical implications

We need to further differentiate between the two levels of analysis. The model in our study serves as a conceptual map that links the observed indicators (i.e., scores on confrontation naming tests and %CIUs measures) and the latent factors (i.e., single-word production abilities and informativeness). The magnitude of the regression coefficient between the two latent factors represents inference at an ability/conceptual level under the assumption that the two abilities are measured with perfect reliability. However, the magnitude of the relationship between observed measures (e.g., BNT and %CIUs based on story retell) and discourse informativeness is expected to be lower because the scores on the measures are affected by both the construct of interest and the unreliability in measurement. For example, based on the results, although single-word production abilities account for 63% of the variance in the ability to convey informative content in discourse, the BNT scores only account for 51% of the variance in informativeness as estimated by squaring the product of the path coefficients between the two variables (Bollen, 1989). Further, %CIUs in picture description or storytelling tasks account for approximately 61% of the variance in informativeness. Therefore, in practice, for clinicians and researchers that do not have access to large samples and techniques such SEM and rely on observed scores, inferences about discourse production based on a confrontation naming test may be even less justified.

The pattern of the factor loadings across the two factors was consistent with previous findings in the literature. Overall, the loadings of the confrontation naming tests were higher than the loadings of the %CIU measures based on the three types of discourse. This can be attributed to the higher reliability of confrontation naming tests (e.g., Walker & Schwartz, 2012). Because of the standardized administration and the constraints the tests impose, confrontation naming tests tap onto similar specific cognitive processes and are scored using uniform procedures. In addition, when scoring confrontation naming tests, the target is known and there is little ambiguity about the accuracy of a response (Herbert et al., 2008). However, based on the SEM solution, it is also clear that confrontation naming tests are not perfect in reflecting the underlying word abilities of a speaker. Recent advances in psychometric modeling of such tests can yield even higher reliability thus allowing clinicians to reach the maximum reliability possible in a given clinical situation (Fergadiotis et al., 2015; Hula et al., 2015).

For informativeness, picture descriptions and story re-tell yielded similar and high factor loadings, which suggests that both types of discourse are equivalent in reflecting informativeness and could also be used interchangeably for measurement purposes. Free speech, on the other hand, produced a weaker factor loading, suggesting that % CIUs based on this elicitation technique is not a reliable or valid indicator of informativeness. Specifically, our results suggest that %CIU scores based on free speech can be quite unreliable and therefore difficult to interpret as they reflect substantially more noise and construct-irrelevant variance (~65%) than actual information about the ability of a speaker to produce informative discourse (~35%). Similarly, a number of investigators have concluded that %CIUs is more susceptible to subjective judgments by the raters especially for conversational and semi-spontaneous discourse (Boyle, 2015; Linnik, Bastiaanse, & Höhle, 2016; Oelschlaeger & Thorne, 1999). Also, free speech samples often produce more variability in responses between different participants that make it difficult to perform certain analyses (Kintz, Fergadiotis, & Wright, 2016). Further, our findings are in agreement with previous research that found sequential pictures and storytelling tasks elicited similar results for lexical diversity for PWA (Fergadiotis, 2011), as well as coherence (Capilouto, Wright, & Wagovich, 2006). Further, Capilouto and colleagues (2006) claimed that free speech samples typically rely on previously recounted information that may allow for regularization and compensatory strategies that obscure the individual's actual discourse abilities. Therefore, while free speech elicitation tasks mimic real-world communicative interactions more so than other discourse types, they may provide less reliable information about the communication abilities of PWA when the intended analysis is CIUs.

On a related note, our model suggests that confrontation naming tests may be better predictors of informativeness compared to CIUs based on free speech tasks similar to the one used in AphasiaBank. This may appear as a counterintuitive conclusion but it is supported by standard path analytic procedures (Bollen, 1989) when applied to our model. The model implied magnitude of the relationship between the confrontation naming tests and the underlying discourse informativeness factor can be estimated by squaring the product of (i) their loadings to the single-word retrieval latent variable (.91, .88, and .85 for the BNT, WAB-Naming, and VNT, respectively), and (ii) the regression coefficient linking the two latent variables (.79; Figure 1). All three confrontation naming tests shared a greater amount of variance with the discourse informativeness factor (52%, 48%, and 45% shared variance for the BNT, WAB-Naming, and VNT, respectively) compared to the observed measure of %CIUs that was based on free speech (35%). Conceptually, the explanation is twofold. First, confrontation naming tests are quite reliable (as suggested by their loadings) and are good indicators of word retrieval; and, word retrieval skills are relatively strongly related with discourse informativeness. Second, as discussed earlier, the free speech task was found to be a quite unreliable (based on its loading), and hence its ability to reflect informativeness was considerably compromised.

Conclusions & future directions

The findings contribute to the literature by extending our previous work and providing a refined analysis of the relationship between performance on confrontation naming tests and discourse tasks in adults with aphasia. Though we provided evidence that word access and retrieval and discourse informativeness are strongly related, the strength of the relationship is such that it may not allow for direct generalization from the picture naming tests to the discourse level. That is particularly true for clinicians and researchers who are interested in drawing conclusions about one PWA at a time and rely on imperfect measures. Our findings suggest that studying discourse may yield unique information that may be clinically relevant.

Future investigations should consider a within-subject repeated measurements across time design to further examine the relationship between the picture naming tests and discourse measures. Additionally, to further explore the clinical utility of using confrontation naming tests to capture discourse informativeness, the measures could be included in discourse treatment studies as outcome measures. Finally, the relationship between confrontation naming tests and other discourse measures, such as core lexicon, should be considered to better understand the relationship between picture naming tests and discourse informativeness.

Acknowledgments

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 Jessica Jensen for assistance with language analyses.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Andreetta, S., Cantagallo, A., & Marini, A. (2012). Narrative discourse in anomic aphasia. *Neuropsychologia*, *50*, 1787–1793. doi:10.1016/j.neuropsychologia.2012.04.003

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

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

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

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

Boyle, M. (2015). Stability of word-retrieval errors with the aphasiabank stimuli. *American Journal of Speech-Language Pathology*, *24*, S953–S960. doi:10.1044/2015_AJSLP-14-0152

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

Bryant, L., Ferguson, A., & Spencer, E. (2016). Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical Linguistics & Phonetics*, *30*, 489–518. doi:10.3109/02699206.2016.1145740

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

Capilouto, G. J., Wright, H. H., & Wagovich, S. A. (2006). Reliability of main event measurement in the discourse of individuals with aphasia. *Aphasiology*, *20*, 205–216. doi:10.1080/02687030500473122

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., Chang, F., & Griffin, Z. M. (1999). Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science*, *23*, 517–542. doi:10.1016/S0364-0213(99)00014-2

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

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Fergadiotis, G. (2011). *Modeling Lexical Diversity Across Language Sampling and Estimation Techniques* (Doctoral Dissertation). Retrieved from ASU Electronic Dissertations and Theses.

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

Fergadiotis, G., & Wright, H. H. (2015). Modelling confrontation naming and discourse performance in aphasia. *Aphasiology*, *30*, 364–380. doi:10.1080/02687038.2015.1067288

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–S408. doi:10.1044/1058-0360(2013/12-0083)

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.

Hartsuiker, R. J., & Kolk, H. H. J. (2001). Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology*, *42*, 113–157. doi:10.1006/cogp.2000.0744

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

Hula, W. D., Kellough, S., & Fergadiotis, G. (2015). Development and simulation testing of a computerized adaptive version of the Philadelphia naming test. *Journal of Speech, Language, and Hearing Research*, 1–13. doi:10.1044/2015_JSLHR-L-14-0297

Ingles, J. L., Mate-Kole, C. C., & Connolly, J. F. (1996). Evidence for multiple routes of speech production in a case of fluent aphasia. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *32*, 199–219. doi:10.1016/S0010-9452(96)80047-5

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

Kertesz, A. (2007). *Western Aphasia battery – R*. New York: Grune & Stratton.

Kintz, S., Fergadiotis, G., & Wright, H. H. (2016). Aging effects on discourse production. In H. H., Wright (Ed.), *Cognition, language, and aging* (pp. 81–106). Philadelphia, PA: John Benjamins Publishing Company.

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

Linnik, A., Bastiaanse, R., & Höhle, B. (2016). Discourse production in aphasia: A current review of theoretical and methodological challenges. *Aphasiology*, *30*, 765–800. doi:10.1080/02687038.2015.1113489

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

Manning, L., & Warrington, E. K. (1996). Two routes to naming: A case study. *Neuropsychologia*, *34*, 809–817. doi:10.1016/0028-3932(95)00166-2

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

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

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

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*, 90–100. doi:10.1016/S0022-2496(02)00028-7

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 and Hearing Research*, *36*, 338–350. doi:10.1044/jshr.3602.338

Nozari, N., & Faroqi-Shah, Y. (2017). Investigating the origin of nonfluency in aphasia: A path modeling approach to neuropsychology. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *95*, 119–135. doi:10.1016/j.cortex.2017.08.003

Oelschlaeger, M. L., & Thorne, J. C. (1999). Application of the correct information unit analysis to the naturally occurring conversation of a person with aphasia. *Journal of Speech, Language, and Hearing Research*, *42*, 636–648. doi:10.1044/jslhr.4203.636

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

Perrin, P. B., Heesacker, M., Stidham, B., Rittman, M. R., & Gonzalez-Rothi, L. J. (2008). Structural equation modeling of the relationship between caregiver psychosocial variables and functioning of individuals with stroke. *Rehabilitation Psychology*, *53*, 54–62. doi:10.1037/0090-5550.53.1.54

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

Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. (2017). Reviewing the quality of discourse information measures in aphasia. *International Journal of Language & Communication Disorders*, *52*, 689–732. doi:10.1111/1460-6984.12318

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.

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

Schwartz, M. F., & Hodgson, C. (2002). A new multiword naming deficit: Evidence and interpretation. *Cognitive Neuropsychology*, *19*, 263–288. doi:10.1080/02643290143000187

Sherratt, S. (2007). Multi-level discourse analysis: A feasible approach. *Aphasiology*, *21*, 375–393. doi:10.1080/02687030600911435

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

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72–101. doi:10.2307/1412159

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). 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.). Pearson/Allyn & Bacon: Boston.

Tatsumi, H., Nakaaki, S., Satoh, M., Yamamoto, M., Chino, N., & Hadano, K. (2016). Relationships among communication self-efficacy, communication burden, and the mental health of the families of persons with Aphasia. *Journal of Stroke and Cerebrovascular Diseases*, *25*, 197–205. doi:10.1016/j.jstrokecerebrovasdis.2015.09.018

Walker, G. M., & Schwartz, M. F. (2012). Short-form Philadelphia naming test: Rationale and empirical evaluation. *American Journal of Speech-Language Pathology*, S140–S153. doi:10.1044/1058-0360(2012/11-0089)

Wallace, S. J., Worrall, L., Rose, T., Dorze, G. L., Cruice, M., Isaksen, J., … Gauvreau, C. A. (2017). Which outcomes are most important to people with aphasia and their families? an international nominal group technique study framed within the ICF. *Disability and Rehabilitation*, *39*, 1364–1379. doi:10.1080/09638288.2016.1194899

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. W., & 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

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