Minimizing variability in language sampling analysis: A practical way to calculate text length and time variability and measure reliable change when assessing clients

Language sampling has long been considered a useful tool in language research and clinical practice. By collecting and examining functional language use at the discourse level, clinicians and researchers gain valuable insight into a person's everyday communication abilities. Because of its reported relationship to functional language, Language Sample Analysis (LSA) has been considered an ecologically valid assessment method (Dietz & Boyle, 2018; Paul, Norbury, & Gosse, 2017). However, functional language can be highly variable as it differs between individuals and between instances of language use.

This paper reviews current applications of LSA across the range of clinical populations in both research and clinical practice. While these populations may differ, the principles of LSA and the issues that clinicians must address in applying LSA remain the same, and so language sampling is discussed within this paper as an approach to language assessment, rather than as method of assessing any specific group of clients, patients, students, or research participants. In particular, we highlight how a simple factor such as variability in language sample length can impact commonly used clinical measures in LSA, and the interpretation of clinical and research data. Variability can affect clinical decision making for individual clients by impacting the clinician's ability to detect meaningful change. We propose a simple but statistically sound method to help clinicians and researchers determine the length of language samples needed when conducting repeated sampling with individuals and to detect change as a result of intervention. This method can be applied to any of the commonly reported linguistic variables discussed below that are measured as a proportion of the total language sample. The paper has two parts: the main paper discusses variability in LSA and shows clinicians how they might use the Reliable Change Index (RCI) to

make decisions in their clinical practice; the supplementary digital material acts as a parallel resource that provides the details of the statistical calculation behind the RCI. Clinicians and researchers are invited to use this resource to replicate the methods outlined in this main paper, and apply the methods proposed herein to other proportional language measures.

The information LSA generates can be used to describe the linguistic ability of an individual at a given time, to look for features of disordered language or impaired cognition as reflected in language, as a source of comparative or normative data in clinical populations, as outcome measures for interventions, and to observe change over time (Boyle, 2014; Finestack, Payesteh, Rentmeester Disher, & Julien, 2014). As well as describing communication impairment (Cherney, Shadden, & Coelho, 1998; Muller, Guendouzi, & Wilson, 2008), LSA has been used to diagnose mood disorder (Rude, Gortner, & Pennebaker, 2004), cognitive functioning in relation to dementia (The Nun Study; Riley, Snowdon, Desrosiers, & Markesbery, 2005), and personality type (Mehl, Gosling, & Pennebaker, 2006; Pennebaker & King, 1999).

Recently, there has been debate in the research literature about various aspects of language sampling and analysis. These discussions occur across the range of practice areas including aphasia (see Aphasiology Vol 32, Issue 4; e.g., Dietz & Boyle, 2018), child language development and impairment (e.g., Finestack et al., 2014; Gillam et al., 2018; Pavelko & Owens Jr, 2017; Guo, Eisenberg, Bernstein Ratner, & MacWhinney, 2018) and adolescent communication (Nipplold, Vigeland, Frantz-Kaspar, & Ward-Lonergan, 2017).

While language sampling may be considered a “gold standard” for assessing spoken discourse (Dietz & Boyle, 2018, p. 461), there is debate about which measures at the macro-linguistic and micro-linguistic levels of discourse best represent meaningful change over time following intervention for an individual. Further, concerns have been raised about the

validity, reliability, and stability of LSA measures over time (Dietz & Boyle, 2018) and across different clinical populations. These concerns are important to consider as they impact the interpretation of findings and utility of LSA across cohorts, and the magnitude of change in individuals following intervention.

This paper discusses key issues associated with variability of language sampling measures and proposes a way of dealing with two aspects of measurement associated with variability when using them: sample length and change over time with repeated measures. The proposed method for dealing with the effect of length of samples relates to commonly used micro-linguistic measures that are calculated as a proportion of total discourse (i.e., number of words produced). This method can be applied to a range of proportional linguistic measures such as Type Token Ratio (TTR), a measure of lexical diversity, Correct Information Units (CIUs), a measure of relevant information as a proportion of total words (to learn how CIUs are calculated, see Nicholas & Brookshire, 1993), and Propositional Density (PD), a measure of informativeness calculated by identifying words in the language sample that carry meaning or propositional information (e.g., verbs, adverbs, adjectives, conjunctions) as a proportion of total words (see Turner & Greene, 1977).

RANGE, VALIDITY, AND RELIABILITY OF LSA MEASURES

Spoken or written language samples may be analyzed in a variety of ways depending on the purpose of clinical assessment, data collection, or research question(s). For example, a language sample may be analyzed for its information content (e.g., T-units, CIUs, or PD), semantic content (e.g., TTR, number of different words (NDW), moving average type token ratio (MATTR)), or syntactic structure (e.g., mean length of utterance (MLU), grammatical and complete sentences) at the micro-linguistic level, or for discourse structure and interaction features at the macro-linguistic level. A recent review by Bryant and colleagues

found that 536 different linguistic measures had been used in analysis of language samples in aphasia research over the past 40 years (Bryant, Ferguson, & Spencer, 2016). Likewise, in an investigation of linguistic features in the discourse of people with Alzheimer's dementia, Fraser, Meltzer, and Rudzicz (2016) analyzed 370 linguistic features through a variety of computerized methods. These studies highlighted the number of different measures available to clinicians and researchers when performing LSA.

Despite the variety of measures used in research and clinical practice, relatively little is known about the reliability of many measures reported in the literature (Armstrong, 2018; Dietz & Boyle, 2018). Establishing reliability of measures can assist with both diagnosis (i.e., differentiating features of typical linguistic behavior from impaired) and differential diagnosis (i.e., differentiating between different disorders), both of which are highly relevant to clinical practice. If measures are reliable, the validity of the measure can be determined. Once the reliability is quantified statistically, then it can be properly taken into account in determining the clinical significance of change over time for an individual or group of individuals; that is, determining what may constitute reliable (and supposedly meaningful) change over time. de Riesthal and Diehl (2018) argued that work on establishing and ensuring the validity and reliability of language sampling measures is of high need in order to develop the usefulness of LSA as an assessment method for intervention effects, both clinically and in research. To detect meaningful change following intervention, the sources of variation associated with each linguistic measure need to be acknowledged, discussed, and, if possible, quantified so that meaningful comparisons can be made across data sets or clinical assessment points, and then the effects of intervention can be measured. In the literature, this is referred to as reliable change – change in a measure for an individual (as opposed to change in the mean of a group) that can be

attributed to true growth in linguistic performance rather than variability inherent within the measure being applied. Unicomb, Colyvas, Harrison, and Hewat (2015) applied Jacobsen and Truax's (1991) Reliable Change Index to account for variability and detect change in single cases following treatment of stuttering (discussed further below). Variability, however, may stem from different sources within a sample, and some of the common sources also are discussed below.

Types of variability

Variability is inherent in language production due to sociolinguistic variables such as the topic and genre of the text or discourse produced by an individual, the audience and the mode of production (spoken or written), and the speaker/writer's stance in relation to what is being spoken or written about (Halliday & Matthiessen, 2004). Other sources of variation that impact on LSA measures and therefore affect test-retest reliability relate to fluctuations in the individual (e.g., cognitive status, wellbeing such as tiredness, levels of distraction) at the times of testing (Boyle, 2014). These intra-individual sources of variability are referred to as random variations rather than systematic or "special cause" variations (Karimi et al., 2013) and are commonly encountered in clinical practice. Cognitive factors in particular, such as working memory (Lalonde & Frush Holt, 2014) and phonological short-term memory (Newbury, Oetting & Stockman, 2015) in language development in children, are important sources of variation. Cognitive factors also are relevant to language changes in the aging population and those with acquired language disorders (Kemper, Greiner, Prenovost, & Mitzner, 2001; Kemper, Thomson, & Marquis, 2001). While these factors are all clinically relevant and impact day-to-day performance of an individual, there are two, test-retest reliability of measures (stability of the measure over time) and language sample length, that

can be easily assessed and controlled by both clinicians and researchers to allow reliable use and interpretation of LSA measures.

Test-retest reliability

Test-retest reliability, or stability of a measure over time, has been investigated for various measures and clinical populations (Altman, Goral, & Levy, 2012; Cameron, Wambaugh, & Mauszycki, 2010; Stark, 2010). Cameron et al. (2010) specifically investigated test-retest reliability (i.e., the stability of a measure over time) within-subjects with repeated sampling of a linguistic measure, CIUs, in adults with aphasia. In this study, CIUs were shown to have instability in sequential samples, and changing mood was hypothesised as a contributing factor. Using a different measure of informativeness, PD, for three large age cohorts of women participating in the Australian Longitudinal Study on Women's Health, Ferguson, Spencer, Craig and Colyvas (2014) found that PD calculated from written language samples remained stable over time (i.e., did not vary significantly with age) on repeated measures every three years over 16 years. PD is a measure that has been associated with cognitive functioning (Kemper, Thomson, et al., 2001). However, in a separate study involving a subset of the Australian Longitudinal Study of Women's Health cohort, Spencer and colleagues found a significant amount of within-subject variability over repeated sampling of 625 participants who wrote on each survey occasion. This variation was evident despite the language elicitation task remaining the same over the five survey periods across 16 years (Spencer et al., 2012). The authors identified text length variability to be one of the most prominent factors in test-retest reliability.

Text length variability

Text length of language samples has been widely recognized as a source of variability, particularly in measures of lexical diversity (e.g., Brookshire & Nicholas, 1994a;

Fergadiotis & Wright, 2011; Fergadiotis, Wright, & West, 2013). In 1994, Brookshire and Nicholas sought to establish the effect of text length variability on the use of CIUs as they were concerned that short language samples (100 words) might affect the test-retest stability of the measure (Brookshire & Nicholas, 1994a). In the area of adult aphasia, Brookshire and Nicholas (1994a) reported that a typical language sample task using single pictures or picture sequences for adults (e.g., the Cookie Theft Picture in the Boston Diagnostic Aphasia Examination; Goodglass, Kaplan, & Barresi, 2001) resulted in fewer than 100 words. Their empirical investigation found that the level of instability for repeated sampling was considerable. They discovered test-retest stability of CIUs improved as sample size increased to 300-400 words. On the basis of their analysis, they recommended that clinicians and researchers collect sample lengths (across a variety of stimuli, therefore not controlling for genre) of 300-400 words in people with aphasia (Brookshire & Nicholas, 1994b). Fergadiotis and colleagues (2015) also explored factors affecting stability of lexical diversity measures in 422 healthy adult speakers. They were interested in determining the validity of four commonly reported indices of automated (computer generated) lexical diversity. While two of the four measures were reported to be free of systematic effects of text length, the authors recommended that text length be reported in all research to aid evaluation and interpretation of results across studies. These reports add weight to calls by Finestack and colleagues (2014) in relation to LSA with children that researchers should consistently and systematically report details of language sampling including text length, contextual factors, elicitation task parameters, and transcription and coding procedures to allow results from the research to be more easily interpreted and applied in clinical practice.

The prominence of text length as an issue in LSA for both child and adult clinical populations relates to its impact on the work of the clinician who is confronted with the

time-consuming task of collecting, transcribing, and analyzing discourse. These factors are known to be barriers to implementing LSA in clinical settings (Armstrong, 2000; Bryant, Spencer, & Ferguson, 2017). Moreover, while eliciting longer language samples may be relatively straightforward for children and adults without language disorders, it can be challenging to obtain elaboration on topics from those with communication difficulties.

In summary, one of the main challenges to reliability in LSA comes from the variability inherent in natural language production. To illustrate the extent of the impact of variability and how it is accounted for in research, we conducted a scoping review of the literature.

A SCOPING REVIEW OF THE LITERATURE

This review focused on a few key measures that are frequently applied in studies employing LSA across pediatric and adult populations: MLU, TTR, NDW, CIUs (Nicholas & Brookshire, 1993) PD, and other measures of information content (see Pritchard, Hilari, Cocks, & Dipper, 2017). While PD is not as frequently used as the other measures, it employs similar methods to calculate information content; that is, it is a measure examining the proportion of one language structure as a factor of another, in this case propositions as a proportion of total words.

In May 2019, we searched four journal databases: CINAHL, Medline, Scopus, and Embase for peer-reviewed journal articles using LSA and the aforementioned measures to assess the language of children or adults. The search was completed using the terms “Mean Length of Utterance” OR “Type Token Ratio” OR “Number of Different Words” OR “Correct Information Units” OR “Main Concepts” OR “Content Units” OR “Propositional Density” OR “Idea Density” OR “Propositional Idea Density” AND “Language Sample Analysis” OR “Discourse Analysis”. The search was limited to original research studies (i.e., not reviews or

editorials) published in the last 10 years (2010-2019 inclusive), written in English (although analyzing any language), and published in peer-reviewed journals. Studies were excluded if they did not use LSA and at least one of the above measures to assess changes in language over time (e.g., as the result of intervention, or as a longitudinal assessment of language development, recovery, or decline) and/or did not report the results of these measures for groups or individuals.

The search process yielded 306 results. All studies were imported into an Endnote X9 library and were screened by the second author to remove duplicates. The titles and abstracts of each study were screened, and those that did not meet the inclusion criteria for review were excluded. The full texts of all remaining articles were retrieved so they could be examined. The search process is illustrated in Figure 1.

Thirty-two studies met criteria for inclusion in this review. Data were extracted relating to purpose of LSA (e.g., assess language change over time longitudinally or from intervention), study design, population, language sample elicitation procedure, results of analysis (including any reported values of specific language measures), and identified sources of variability and how they were addressed. The quality of included studies was not assessed. An integrative review method (Whittemore & Knafl, 2005) was employed to synthesize the findings across studies. A table summarizing the key information of included studies is available as supplementary digital content file 1 (docx document).

Characteristics of review studies

Language sample measures were used in the included studies to: measure change associated with intervention (*n* = 14), assess longitudinal change (*n* = 13), examine the test-retest stability of measures (*n* = 4), and determine if changes in the environment affected discourse production through repeated measurement (*n* = 1). The included studies used LSA

to assess the language of children (*n* = 17) and adults (*n* = 15). The majority of included studies analysed language in English (*n* = 28). Other languages included Spanish (*n* = 4), Norwegian, Farsi, and French (*n* = 1 each). Three studies analysed more than one language (all English and Spanish).

Most studies (*n* = 20) had 20 participants or fewer, and four studies had more than 100 participants. Participant numbers ranged from one to 1723. Change over time was reported in each study for the sample as a whole (*n* = 21), for individual participants (*n* = 8), or for both groups and individuals (*n* = 3). Studies used a range of methods to analyze change over time. Group studies predominantly used statistical comparisons to identify significance of change (*n* = 17). Studies also reported change descriptively (*n* = 6) or employed statistical modelling to identify factors affecting longitudinal change (*n* = 4). Two group studies used the results of previous studies to determine post-treatment intervention gains: one applied previously calculated Minimally Detectable Change (MDC) score, while another compared data to Reliable Change Index (RCI) scores. Studies that examined individual changes primarily did so using descriptive statistics (*n* = 5), though some also used effect sizes (*n* = 3) and statistical analysis with ANOVAs (*n* = 1). Three studies that examined individual growth determined change using predetermined scores: one used previously reported MDC scores, one compared to reported developmental norms, and one used a pre-set score of 10 percentage points.

Sources of variation

Most of the sources of variation reported within the included studies came from differences between individuals within study cohorts. However, when considering the application of LSA in clinical contexts, the within-subject sources of variation are of greater importance, so these sources were the focus of this review. Sample/text length and type

were each noted by a number of studies as within-subject sources of variation that have the potential to influence clinical or research findings. We have chosen to focus on sample length variability within this paper to examine this source of variability and propose a potential solution.

Sample length variation

Many studies included language productivity measures that provided some information to quantify the amount of discourse that was produced and analyzed. However, the reporting of sample length across the reviewed studies was largely inconsistent. These measures included the total number of words (*n* = 13), number of utterances (*n* = 7), and number of C-units (*n* = 1). Three studies reported sample size in terms of duration in minutes. The remaining eight studies did not report any information on sample length. The absence of sample length information in research studies is problematic as it impedes accurate evidence-based clinical applications of language sampling methods. As noted previously, text length affects the variability of measures. Clinicians and researchers therefore require knowledge of sample length to determine whether language measures provide reliable information about a client's communication.

In five studies, an effort was made to control for text length variability by analyzing samples of a set length. In four studies, the researchers collected an approximate number of utterances to analyze (Griffith et al., 2017; Preis & McKenna, 2014; Rice et al., 2010; Smith et al., 2014), and in one study a 200-word subset of the collected language samples was analyzed (Kirmess & Lind, 2011). Two studies performed pairwise comparisons of sample lengths at pre- and post-treatment and found no statistically significant difference in the length of samples (Medina et al., 2012; Silkes et al., 2019). These studies suggest one possible way that clinicians can control for text-length variability – by comparing samples of

the same length. However, this method can be problematic as it can lead to clinicians discarding meaningful and useful portions of a discourse sample that might inform clinical decision-making. In order to include full language samples, clinicians therefore require some other means of controlling text length.

Assessing changes in single case intervention studies

Of the 13 studies that assessed change in language samples as the result of intervention, four studies applied a multiple baseline approach to determine individual variation in measures prior to intervention. In these studies, the researchers collected between three and nine pre-treatment language samples, allowing them to assess natural variation in the language of individual participants. By quantifying natural variation for the individual, researchers determined what change post-intervention was attributable to the intervention and therefore showed the effectiveness of the treatment under investigation (Boyle, 2014). In contrast to this approach, most reviewed intervention studies collected language samples at three key time points: one sample prior to the intervention (pre-treatment), one sample immediately following intervention (post-treatment), and one sample sometime later that would demonstrate treatment effects were maintained (follow-up). Where multiple baselines were not used, researchers did not have individual data to determine variability or the degree of change attributable to the intervention. See Supplementary Data File 1 for details of treatment changes in the studies discussed above.

Within the intervention studies that used time-point comparisons, six studies assessed group change over time, and three studies measured change for individuals. In the studies that employed case-based designs, the assessment of variability was limited. Kirmess and Lind (2011) noted that the diagnosis of individual participants (e.g., type of aphasia, concurrent apraxia of speech) and the cause of the diagnosis (e.g., intercranial

hemorrhage versus vascular stroke) may have contributed to variability. Language modality (spoken versus written; Obermeyer & Edmonds, 2018) and text length (Kirmess & Lind, 2011; Wambaugh, Wright, Nessler, & Mauszycki, 2014) also were noted in discussions as factors that could contribute to variability in LSA. However, while Kirmess and Lind (2011) controlled for text length, no further procedures were employed to control or account for variability in the data. Indeed, the change resulting from intervention in these studies was determined descriptively (Kirmess & Lind, 2011; Wambaugh et al., 2014), with some reference to effect sizes (Kirmess & Lind, 2011), or by using a predetermined, arbitrary indicator of change (ten percentage points; Obermeyer & Edmonds, 2018).

The measurement of individual change from pre-treatment to post-treatment is more consistent with the type of comparison that a clinical speech-language pathologist might use when assessing a client to determine if intervention has been successful. While the collection of multiple baseline data points may offer the opportunity to measure individual variability, Boyle (2014) noted that "Clinicians rarely have the luxury of using such a practice" (p. 977). We are aware of one study that has investigated reliable change in individuals following intervention in single cases of pediatric stuttering treatment (Unicomb et al., 2015). However, given that these studies are limited, clinicians and researchers need to draw on data collected from studies conducted by researchers whose specific aim is to assess language sample variability, and that are appropriate for application to their own assessment contexts, to determine if actual change has occurred.

Ideally, these variability estimates would be drawn from studies employing repeated measures designs to determine the test-retest reliability of language sample measures. This review identified four such studies that examined individual variability over time and test-retest reliability using repeated measures. While these studies make a start towards

identifying if measures are stable over time using analysis of variance or correlation coefficients, only one calculated MDC for measures including CIUs and words per minute (WPM) that could then be applied to future studies (Boyle, 2014). This application was then evident in a later study by Rose et al. (2016) that applied Boyle's MDC scores to measure treatment outcomes.

Overall, our scoping review demonstrates that, while variability from various sources is widely acknowledged, there are few published studies available that estimate the total variability from various sources and model the effects of changing parameters in their experimental designs. We aim to address this with a method to assist clinicians and researchers to account for variability in LSA and reliably determine meaningful change.

TEXT LENGTH AND TIME-BASED VARIABILITY AND HOW TO ACCOUNT FOR IT

When speech-language pathologists assess clients or when researchers implement case-study designs (e.g., to measure individual change over time resulting from an intervention), they need access to methods that assess whether changes within an individual are likely to be real rather than just a result of variability within the measure itself. So how should clinicians and researchers assess variability to distinguish real change when they use LSA to measure change over time? To answer this question, we applied statistical methods to create a model for clinicians and researchers. This model is outlined in multiple steps to: (a) show how the variability of a measure is associated with text length; (b) demonstrate that the binomial distribution provides a good foundation for a statistical model for that variability; (c) adjust the binomial variability model to account for more or less variation than expected; (d) determine the variability associated with repeated measurements, i.e., time-based variability; (e) provide a formula for the Reliable Change Index (RCI) based on combining both text length and time-based variability; (f) create a table

based on the RCI formula to illustrate how the RCI varies and to use it to make clinical decisions (i.e., has a person's language ability actually changed or is it due to variation in the measure); (g) use RCI to make decisions about the size of a language sample needed for clinical applications of LSA; and (h) apply the RCI to other LSA measures. While the case we use to illustrate application of RCI focuses on the measure of PD, the spreadsheet in Supplementary Digital Content file 2 and instructions in Supplementary Digital Content file 3 illustrate how to adapt the method demonstrated here to discourse measures other than PD that may be more suitable to a clinician's or investigator's particular needs (e.g., TTR, CUIs).

Text length variability

Many discourse measures are reported as a proportion of a total text (p = x/n), for example, PD (p) reports number of propositions (x) as a proportion of the total number of words (n); e.g., 55 propositions in a 100-word sample gives PD = 0.55. The variability in the measure is exhibited clearly in Figure 2a which shows how the measure of PD varies less with an increased number of words in the sample. Any measure with a similar method of calculation would most likely be subject to the effect of text length (e.g., percent correct information units, which reports CIUs as a proportion of total words, and TTR, which reports number of different words as a proportion of total words).

To explore the effects of text length variability on PD, we analyzed the comments (n = 37,705 texts of 10 or more words) given as responses to a survey item from a longitudinal study (Australian Longitudinal Study of Women's Health). PD was determined for each text comment using the Computerized Propositional Idea Density Rater (CPIDR) software (Covington, 2007). Details of the survey, the response data, and analyses are described in Ferguson et al. (2014). Briefly, the survey involved women responding to an open question

asking them about their health or changes in their health over the past three years and this survey was repeated every three years (and is ongoing).

To illustrate the magnitude of variability in PD according to text length, PD for each language sample was plotted against text length for texts of 10 to 200 words (totalling 36,879 of the 37,705 texts in the data set), illustrated in Figure 2a. The figure shows that the amount of variability (i.e., the spread of data points around the mean) progressively declines as the sample size increases. This is shown more clearly in Figure 2b where the variability in PD is summarised as the standard deviation (*SD*) for all the comments within each text length and is plotted against text length. Again, it is clear to see the amount of variation in the measure decreases as text length increases.

Binomial distribution as a model for text length variability

As language sample measures such as PD examine linguistic features as a proportion of the total text (total number of words), the binomial distribution is expected to provide a useful model as its mathematical basis accounts for proportions where the text length varies. This is confirmed in Figure 2b where predictions based on the binomial model show a similar pattern to those from the data with varying text length. Further confirmation of this effect was obtained using a separate source of data, a book by Agatha Christie (1920; see Supplementary Digital Content file 3). These tests indicate that the binomial model would be a good foundation for determining variability in PD due to text length.

Does the binomial model need adjustment?

Before using the methods proposed here to assess reliable change for any language sample measure, the mathematically based binomial distribution needs to be checked for applicability to real-world application. Experience has shown the observed variability might be systematically more or less than expected for a given text length. In Figure 2b, the *SD*s for

the observed data are systematically smaller than those calculated from the binomial distribution for a given text length. Therefore, an adjustment needs to be made to the variability estimates predicted by the binomial model to ensure it correctly matches the particular language measure. From a detailed analysis of three text sources, an adjustment of 0.9 was determined as being suitable for use with PD, i.e., the variability for PD was less than expected and should be 90% of that obtained from the binomial model. See Supplementary Digital Content file 3 for the details behind this analysis. This check would normally be carried out by researchers to determine the correction factor that would then be used by clinicians as described below in assessing change. If this step was not done, significance tests would not be as accurate as possible.

Time-based variability for repeated measures

Additional factors examined for their relationship with the variability of PD were differences over time and between subjects. The study participants in the Australian Longitudinal Study of Women's Health were surveyed every three years (five times at the time of our study), so each person could provide up to five text comments. The study contained three factors: text length, variability due to multiple measurements over time, and different subjects. When repeating measurements over time, it is likely there will be some differences from time to time. It was necessary to fit a statistical model to determine the three components jointly so that correct estimates of each of the variability terms, including the time-based term, were obtained by adjusting for all the factors in the study. The variability for differences between time periods had *SD* = 0.017 (see the Supplementary Digital Content file 3 for additional explanation and the analysis).

Reliable Change Index (RCI)

The RCI is a measure of the likely range of change scores for a measure between two time points, if there was no true difference in scores. If a change in score was less than the RCI, then it is likely the difference does not reflect a real change, just measurement variability. This provides a simple test for change scores as those greater than (or equal to) the RCI are deemed statistically significant.

Jacobson and Truax (1991) presented an approach based on the normal distribution but that was not appropriate for LSA as the normal distribution assumes the variability is the same under all circumstances. As has been noted above, variability depends in part on text length. A new formula has been developed based on the binomial distribution to allow adjustment for text length. The development of the RCI formula and the details behind it can be found in the Supplementary Digital Content file 3.

The notion of clinically meaningful change also is important in a clinical context and is discussed in Jacobson and Truax (1991) and Jacobson et al. (1999) but is different from RCI. We do not propose to discuss this here, but once a clinician can define the size of change that is beneficial, the methods in this paper can be used to determine a suitable text length so that the RCI would be able to detect a clinically significant change. The approach is described below.

Using RCI to make decisions about changes in individuals

Table 1 below provides a series of worked examples that show how the RCI can be applied to determine if an individual's language sample shows real change in PD. The table demonstrates how RCIs vary with text length and time and how a change over time can be assessed. The table is available as an Excel workbook for download from Supplementary Digital Content file 2 so that clinicians or researchers can carry out their own calculations for language measures without having to do them manually.

The first six columns of the table provide the basic calculations to assess PD at two time points. For example, the second row (example A) in the table shows a comparison between two texts, both with length of 100 words, the first containing 50 propositions and the second 40. Hence the proportion at time 1 is p1 = 0.50 and at time 2 is p2 = 0.40. The difference between them (p1-p2) in the 7th column is 0.10. Now the question to be answered is whether this change is statistically significant at the level chosen, $\alpha = 0.05$ in this case, and hence reliable. As the difference 0.10 for (p1-p2) is smaller than $RCI_{LT} = 0.132$, i.e., the RCI incorporating both text length and time variation, the change is not significant and hence not a reliable change (i.e., a treatment has had no real effect on PD as the change in the linguistic measure from sample 1 to sample 2 is not large enough). In the 3rd row of the table (example B), the difference is larger, (p1-p2) = 0.20, and it also is larger than RCI_{LT} = 0.129, hence the difference is statistically significant, and it is a reliable change. If we look at example A and increase the text length to 200 words but keep the proportions the same such that the difference between the two testing periods is still 0.10, then this difference is significant as the RCI is 0.099. This illustrates the importance and effect of text length; the longer the text chosen, the greater the power to detect differences between samples.

To illustrate how to use the RCI, we will draw data from Table 1 using a fictional case. FC was a 78-year-old female taking part in a fictional research study investigating the effectiveness of a language intervention targeting informativeness in aging adults exhibiting symptoms of early cognitive decline. FC was a relatively healthy participant in this study who expressed concerns about her memory. A spoken language sample was taken as part of the assessment battery to explore pre-intervention levels of informativeness using PD. At this time (time 1), her language sample was 421 words in length and contained 198 propositions, giving p1 = 198/421 = 0.47. After 12 weeks of intervention (time 2), the

assessment battery was repeated. Another spoken language sample was analysed. This language sample contained 517 words and 311 propositions, p2 = 311/517 = 0.60. If we refer to the table, we can see that the change in PD from time 1 to time 2 is 0.13 and this value is greater than the RCI value of 0.074. This is an increase in informativeness as measured by PD and can be considered a significant and reliable change at the chosen alpha level of 0.05. Such a change suggests that the intervention was effective in improving FC's language informativeness, as measured by PD.

Using RCI to make decisions about the size of a language sample

Once the RCI has been determined for the measure and clinicians and researchers decide what constitutes a clinically meaningful change, the next step is to determine how large a language sample is needed to detect the clinically important change. This can be done using the spreadsheet in Supplementary Digital Content file 2 by varying the sample size while keeping the proportions the same until the RCI becomes smaller than the clinically meaningful change. For example, if the clinically meaningful change was 0.10 as in the example above, then a total sample length of 200 words would be sufficient because at this text length the RCI = 0.099, which is smaller than the set value that represents meaningful change.

Application of this method to other LSA measures

This method of using the binomial distribution to measure reliable change can be applied to any proportional language measure (e.g., CUIs, TTR) because they are affected by text length in the same way we have illustrated with PD. The two steps that need to be considered to make this most appropriate for a particular measure are the adjustment factor for the binomial distribution and the size of the time-based variability. However, it will take considerable time before researchers or diligent clinicians get around to

determining these two components. It is proposed that an immediate benefit could be obtained by clinicians by adopting an adjustment factor of 1.0 and a time-based variability setting of 0.02 in the RCI calculator presented in Supplementary Digital Content file 2 to create an estimate of reliable change. Although these settings might not be the most accurate for a given measure, it is expected they would provide a good working approximation until improved values can be determined. The benefit of doing this immediately would be that clinicians would begin to use an approach with good statistical underpinnings, and this might lead to greater uniformity in decision making. Subjective judgments by different clinicians (and researchers) about what constitutes a real change would be removed. Also, with greater use of this approach, there will be more motivation to do the development work to improve the variability estimates for each measure. Without using a tool like the RCI, clinicians will continue using their subjective judgement about what represents real change in language performance.

CONCLUSIONS

In this paper, we have demonstrated an approach to studying and accounting for variability in LSA, using PD as an illustrative example for assessing changes in individuals. Text length and time-based variability were quantified, and a suitable statistical model was proposed that could then be used by clinicians and researchers to assess changes over time. This approach could be applied to other LSA measures that examine a part of language as a proportion of total discourse. The simple method we have developed accounts for text length variation using the binomial distribution combined with time-based variability to estimate reliable change when clinicians or researchers repeat measurements to assess the language of an individual. The Excel workbook in Supplementary Digital Content file 2 has been provided so clinicians and researchers can calculate RCIs in their own practice and

make accurate decisions when they assess change. This can also assist clinicians and researchers in avoiding errors by assuming that a change in a linguistic measure is significant when it is not (i.e., when change is only due to variability in the measure, which may result from insufficiently large text lengths). This method also can help avoid the opposite error where clinically meaningful change may not be detected by other methods as the sample length is too small.

Accounting for the impact of text length would facilitate comparisons of some commonly used measures across research studies to build a more reliable picture of change over time in individuals with communication disorders as a result of treatment or with other time-based variables (e.g., aging). These outcomes also can provide clinicians with clearer guidance on assessing the discourse of their clients. However, to take advantage of this, studies need to report clearly all the details of their sample collection approach, especially text length at the individual level, if reliable changes are to be assessed. This method will also provide a foundation for clinicians and researchers to estimate how large a language sample is required in order to minimize some sources of variability and increase reliability of the measure(s) being used. In advocating for a reasonable and reliable way to compare samples, we acknowledge that language is a complex phenomenon and can be affected by a range of variables only mentioned briefly in this paper and commonly encountered in clinical practice. We have focussed on two sources of variability, text length and time-based changes, and their impact on commonly used language sampling measures. Collecting repeated measurements from an individual at appropriate time intervals is important in detecting response to intervention both clinically and in research. Reliability reporting of typical results for measures used in LSA is important to enable better interpretation of results which may (or may not) indicate meaningful change in an individual over time. We

echo previous calls for clinicians and researchers to accurately report language sample text length with all language sampling measures used and propose a simple calculator for RCI values to assist in determining how much language is needed from an individual in order to minimize variability of that measure in repeated sampling over time. In using this method, clinicians and researchers can more accurately determine change in language over time when making decisions about individuals with and without language impairments.

REFRERENCES

Altman, C., Goral, M., & Levy, E. S. (2012). Integrated narrative analysis in multilingual aphasia: The relationship among narrative structure, gammaticality and fluency. *Aphasiology, 26*(8), 1029-1052.

Armstrong, E. (2000). Aphasic discourse analysis: The story so far. *Aphasiology, 14*(9), 875-892.

Armstrong, E. (2018). The challenges of consensus and validity in establishing core outcome sets. *Aphasiology, 32*(4), 465-468.

Boyle, M. (2014). Test–retest stability of word retrieval in aphasic discourse. *Journal of Speech, Language, & Hearing Research, 57*(3), 966-978.

Brookshire, R. H., & Nicholas, L. E. (1994a). Speech sample size and test-retest stability of conected speech measures for adults with aphasia. *Journal of Speech & Hearing Research, 37*(2), 399-407.

Brookshire, R. H., & Nicholas, L. E. (1994b). Test-retest stability of measures in conected speech in aphasia. *Clinical Aphasiology, 22*, 119-133.

Bryant, L., Ferguson, A., & Spencer, E. (2016). Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical Linguistics and Phonetics, 30*(7), 489-518.

Bryant, L., Spencer, E., & Ferguson, A. (2017). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology, 31*(10), 1105-1126.

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*(6-8), 671-684.

Cherney, L. R., Shadden, B. B., & Coelho, C. A. (Eds.). (1998). *Analyzing discourse in communicatively impaired adults*. Gaithersburg, MD: Aspen.

Christie, A. (1920). *The mysterious affair at Styles* (Project Gutenberg EBook #863, 27 July 2008 ed.). Available: http://www.gutenberg.org.

Covington, M. (2007). CPIDR 3 User Manual (CASPR Research Report 2007-03). Athens, Georgia: Artificial Intelligence Center, University of Georgia.

de Riesthal, M., & Diehl, S. K. (2018). Conceptual, methodological, and clinical considerations for a core outcomes set for discourse. *Aphasiology, 32*(4), 469-471.

Dietz, A., & Boyle, M. (2018). Discourse measurement in aphasia research: Have we reached the tipping point? *Aphasiology, 32*(4), 459-464.

Fergadiotis, G., & Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology, 25*(11), 1414-1430.

Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical density in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology, 22 (2)*, 397-408.

Ferguson, A., Spencer, E., Craig, H., & Colyvas, K. (2014). Propositional Idea Density in women's written language over the life span: Computertized analysis. *Cortex, 55*, 107-121.

Finestack, L. H., Payesteh, B., Rentmeester Disher, J., & Julien, H. M. (2014). Reporting child language sampling procedures. *Journal of Speech, Language, and Hearing Research, 57*, 2274-2279.

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease, 49*, 407-422.

Gillam, S. L., Olszewski, A., Squires, K., Wolfe, K., Slocum, T., & Gillam, R. B. (2018). Improving narrative production in children with language disorders: An early-stage efficacy study of a narrative intervention program. *Language, Speech, and Hearing Services in Schools, 49*(2), 197-212.

Goodglass, H., Kaplan, E., & Barresi, B. (2001). *Boston diagnostic aphasia examination (3rd edition)*. Philadelphia, PA: Lippincott, Williams & Wilkins.

Griffith, J., Dietz, A., Ball, A., Vannest, J., & Szaflarski, J. P. (2017). An examination of changes in spoken productions within constraint-induced aphasia therapy. *Aphasiology, 31*(11), 1250-1265.

Guo, L., Eisenberg, S., Bernstein Ratner, N., & MacWhinney, B. (2018). Is putting SUGAR (Sampling Utterances of Grammatical Analysis Revised) into Language Sampling Analysis a good thing? A response to Pavelko and Owens (2017). *Language, Speech, and Hearing Services in Schools, 39*, 622-627.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (Vol. 3). London: Arnold.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67*, 300-307.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12-19.

Karimi, H., O'Brian, S., Onslow, M., Jones, M., Menzies, R., & Packman, A. (2013). Using statistical process control charts to study stuttering frequency variability during a single day. *Journal of Speech, Language, and Hearing Research, 56*, 17889-11799.

Kemper, S., Greiner, L. H., Prenovost, K., & Mitzner, T. L. (2001). Language decline across the lifespan: Findings from the Nun Study. *Psychology & Aging, 16*(2), 227-239.

Kemper, S., Thomson, M., & Marquis, J. (2001). Longitudinal change in language production: Effects of aging and dementia on grammatical complexity and propositional density. *Psychology & Aging, 16*(4), 600-614.

Kirmess, M., & Lind, M. (2011). Spoken language production as outcome measurement following constraint induced language therapy. *Aphasiology, 25*(10), 1207-1238.

Lalonde, K., & Frush Holt, R. (2014). Cognitive and linguistic sources of variance in 2-year-olds' speech-sound discrimination: A preliminary investigation. *Journal of Speech, Language, and Hearing Research, 57*, 308-326.

Medina, J., Norise, C., Faseyitan, O., Coslett, H. B., Turkeltaub, P. E., & Hamilton, R. H. (2012). Finding the right words: Transcranial magnetic stimulation improves discourse productivity in non-fluent aphasia after stroke. *Aphasiology, 26*(9), 1153-1168.

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology, 90*, 862-877.

Muller, N., Guendouzi, J. A., & Wilson, B. (2008). Discourse analysis and communication impairment. In M. J. Ball, M. R. Perkins, N. Mueller, & S. Howard (Eds.), *The handbook of clinical linguistics*. Oxford, UK: Blackwell Publishers.

Newbury, J., Klee, T., Stokes, S. F., & Morana, C. (2015). Exploring expressive vocabulary variability in two-year-olds: The role of working memory. *Journal of Speech, Language, and Hearing Research, 58*, 1761-1772.

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*(2), 338-350.

Nipplold, M., Vigeland, L. M., Frantz-Kaspar, M. W., & Ward-Lonergan, J. M. (2017). Language sampling with adolescents: Building a normative database with fables. *American Journal of Speech-Language Pathology, 26*(3), 908-920.

Obermeyer, J. A., & Edmonds, L. A. (2018). Attentive reading with constrained summarization adapted to address written discourse in people with mild aphasia. *American Journal of Speech-Language Pathology, 27*, 392-405.

Paul, R., Norbury, C. F., & Gosse, C. (2017). *Language disorders from infancy through adolescence: Listening, speaking, reading, writing and communicating* (5th ed.). St. Louis, MO: Mosby Elsevier.

Pavelko, S. L., & Owens Jr, R. E. (2017). Sampling Utterances and Grammatical Analysis Revised (SUGAR): New normative values for language sample analysis measures. *Language, Speech, and Hearing Services in Schools, 48*(3), 197-215.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*, 1296-1312.

Preis, J., & McKenna, M. (2014). The effects of sensory integration therapy on verbal expression and engagement in children with autism. *International Journal of Therapy and Rehabilitation, 21*(10), 476-486.

Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. (2017). Reviewing the quality of discourse information measures in aphasia. *International Journal of Language and Communication Disorders, 52*(6), 689-732.

Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., & Blossom, M. (2010). Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research, 53*(2), 333-349.

Riley, K. P., Snowdon, D. A., Desrosiers, M. F., & Markesbery, W. R. (2005). Early life linguistic ability, late life cognitive function, and neuropathology: Findings from the Nun Study. *Neurobiology of Aging, 26*, 341-347.

Rose, M. L., Mok, Z., Carragher, M., Katthagen, S., & Attard, M. (2016). Comparing multi-modality and constraint-induced treatment for aphasia: a preliminary investigation of generalisation to discourse. *Aphasiology, 30*(6), 678-698.

Rude, S., Gortner, E. M., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion, 18*, 1121-1133.

Silkes, J. P., Fergadiotis, G., Hunting Pompon, R., Torrence, J., & Kendall, D. L. (2019). Effects of phonomotor treatment on discourse production. *Aphasiology, 33*(2), 125-139.

Smith, J. M., DeThorne, L. S., Logan, J. A. R., Channell, R. W., & Petrill, S. A. (2014). Impact of prematurity on language skills at school age. *Journal of Speech, Language, and Hearing Research, 57*(3), 901-916.

Spencer, E., Craig, H., Ferguson, A., & Colyvas, K. (2012). Language and ageing - Exploring propositional density in written language - Stability over time. *Clinical Linguistics and Phonetics, 26*(9), 743-754.

Stark, J. A. (2010). Content analysis of the fairy tale Cinderella - a longitudinal single-case study of narrative production: "From rags to riches". *Aphasiology, 24*(6-8), 709-724.

Turner, A., & Greene, E. (1977). *The construction and use of a propositional text base* (Technical Report No. 63). Boulder, CO: University of Colorado, Institute for the Study of Intellectual Behavior.

Unicomb, R., Colyvas, K., Harrison, E., & Hewat, S. (2015). Assessment of reliable change using 95% credible intervals for the differences in proportions: A statistical analysis for case-study methodology. *Journal of Speech, Language, and Hearing Research, 58*, 728-739.

Wambaugh, J. L., Wright, S., Nessler, C., & Mauszycki, S. C. (2014). Combined aphasia and apraxia of speech treatment (CAAST): Effects of a novel therapy. *Journal of Speech, Language, and Hearing Research, 57*(6), 2191-2207.

Whittemore, R., & Knafl, K. (2005). The integrative review: Updated methodology. *Journal of Advanced Nursing, 52*(5), 546-553.

Supplementary Digital Content

1. Supplementary Digital Content 1.docx
2. Supplementary Digital Content 2.xls

3. Supplementary Digital Content 3.docx

Figure Captions

FIGURE 1 Search and inclusion process for reviewed studies.

FIGURE 2 Variability of PD as a function of text length for the survey comments from the Australian Longitudinal Study of Women's Health; (a) plot of PD (expressed as propositions per 1 word) for each comment against comment text length and (b) the plotted points are the standard deviations of the PDs for all comments at each text length in (a). The line in (b) is the standard deviation for the binomial distribution at each text length taking p as the mean of all the PD data.