IMPLEMENTATION OF PARTIALLY AUTOMATED CIU ANALYSIS FOR MEASURING READING COMPREHENSION IN A CLINICAL SETTING

by

GARRETT PORTER

A THESIS

Presented to the Department of Special Education and Clinical Sciences
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Master of Science

June 2018

THESIS APPROVAL PAGE

Student: Garrett Porter

Title: Implementation of Partially Automated CIU Analysis for Measuring Reading Comprehension in a Clinical Setting

This thesis has been accepted and approved in partial fulfillment of the requirements for the Master of Science degree in the Department of Special Education and Clinical Sciences by:

McKay Sohlberg	Chairperson
Stephen Fickas	Member
Samantha Shune	Member

and

Sara D. Hodges	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2018

THESIS ABSTRACT

Garrett Porter

Master of Science

Department of Special Education and Clinical Sciences

June 2018

Title: Implementation of Partially Automated CIU Analysis for Measuring Reading Comprehension in a Clinical Setting

There is a dearth of measures which evaluate reading comprehension in people with traumatic brain injuries returning to secondary level education. Existing standardized assessments do not accurately measure constructs of high level reading comprehension. Correct information unit (CIU) analysis can be a valuable tool for measuring reading comprehension in these more demanding contexts. However, the measure requires a significant amount of time to administer and score, leading practicing clinicians to use other measures.

This exploratory project sought to fill the gap by increasing the clinical feasibility of CIU analysis. Researchers implemented a human-in-the-loop automation of CIU scoring. A within rater comparison across three raters design was utilized to evaluate if the automation provided an increase in efficiency versus by hand scoring. Findings indicate a trend of increased efficiency across raters which was not statistically significant. This thesis supports further studies to continue development of the automated application of CIU analysis.

CURRICULUM VITAE

NAME OF AUTHOR: Garrett Porter

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Brigham Young University, Provo, Utah
Southern Utah University, Cedar City
Dixie State University, St. George, Utah

DEGREES AWARDED:

Master of Science, Communication Disorders and Sciences, 2018, University of Oregon
Bachelor of Science, Communication Disorders and Sciences, 2016, Brigham Young University
Associate of Science, General Science, 2011, Southern Utah University

AREAS OF SPECIAL INTEREST:

Cognition
Swallowing
Speech sound production

GRANTS, AWARDS, AND HONORS:

Rose Gross, University of Oregon, 2017-18

Ned J. Christensen Scholarship, University of Oregon, 2017-18

Janet Sant Scholarship, Brigham Young University, 2015-16

Brigham Young Grant, Brigham Young University, 2014-15

ACKNOWLEDGMENTS

I wish to express sincere appreciation to Professors Sohlberg and Fickas for their assistance, guidance and support throughout the development of this study and manuscript. Additionally, I wish to thank Priya, Kelsey, Randy, and Jason for helping me see the vision of the whole project. Finally, special thanks are due to my wonderful wife without whose support would this have been even remotely possible.

To my three little ladies…

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

Definition of Clinical Problem

Speech-language pathologists (SLPs) are faced with a paucity of measures for evaluating post-secondary level reading comprehension in people with cognitive impairments due to traumatic brain injury (Griffiths, Sohlberg, Kirk, Fickas, & Biancarosa, 2016; Keenan, Betjemann, & Olson, 2008; Kucheria, Sohlberg, Yoon, Fickas, Prideaux, in press). Correct information unit (CIU) analysis has been demonstrated to be a valuable tool for measuring reading comprehension in these more demanding contexts (Griffiths et al., 2016; Kucheria et al., in press; Matsuoka, Kotani, & Yamasato, 2012). However, CIU analysis holds limited clinical feasibility (i.e., easily and conveniently accomplishable in clinical settings) currently due to the significant amount of time required to administer and score the measure. This developmental project sought to fill the gap in higher level reading comprehension measures by decreasing the time necessary for a clinician to administer CIU analysis and consequently, increase the clinical feasibility of the measure.

CIU analysis is a rule-based measurement tool which can be used by SLPs to quantify the informativeness and efficiency of a speaking sample (Nicholas & Brookshire, 1993a). CIU analysis was originally introduced with the intent of measuring the discourse production of patients with aphasia. The measure provides counts of and analyses on the number of words and CIUs (i.e., words which are meaningful and correct in relation to the stimulus) in a spoken sample. It has primarily been applied in research settings with the same goal (Nicholas & Brookshire, 1993a; Wambaugh, Nessler, & Wright, 2013).

There are a number of barriers to the clinical feasibility of CIU analysis. One of the most impactful barriers is time. CIU analysis requires an extended amount of time to administer, transcribe the audio, and score the transcript (Nicholas & Brookshire, 1993a). Clinicians using the measure have a limited time window to complete assessments and are facing upward trends in productivity requirements (American Speech-Language-Hearing Association, 2017b; Centers for Medicare and Medicaid Services, 2017; Swanson, 2018). Finally, few clinicians have received in-depth training to administer, score, and interpret CIU analysis (Leslie, McNeil, Coyle, & Messick, 2011).

A potential solution for increasing the clinical feasibility of CIU analysis is to automate portions of the analysis. Using computer algorithms, the amount of time that SLPs need to administer an assessment can be significantly reduced (Long, 2001). Several of the rules involved in CIU analysis are linguistically complex. Current research in computer science recommends partial automation, providing for human supervision over computer tasks for these complex situations (Androutsopoulos & Malakasiotis, 2010; Gaur, Lasecki, Metze, & Bigham, 2016).

The hypotheses of this exploratory study operate under two assumptions: (1) that SLPs are not currently using CIU analysis because it requires an inordinate amount of time to administer, and (2) that SLPs would be more likely to use CIU analysis in clinical settings if it were more efficient. The hypotheses of this study are twofold: (1) that a partial-automation of CIU analysis scoring will demonstrate adequate interrater reliability, and (2) that a partial-automation of CIU analysis scoring will increase the efficiency with which CIU analysis can be administered.

CHAPTER II

REVIEW OF LITERATURE

What is CIU Analysis?

Overview. Nicholas and Brookshire's correct information unit (CIU) analysis was introduced in 1993 as a standardized, rule-based system used to measure discourse production in persons with aphasia. CIUs themselves were introduced previously, in the work of Busch and Brookshire (1985). Busch and Brookshire defined CIUs as words which carried meaning and were "intelligible, grammatical, and appropriate to the [stimulus]" (p. 256). Nicholas and Brookshire utilized CIUs and developed a measure to analyze the discourse of persons with impairments in language comprehension and expression, termed aphasia, as a result of stroke. CIU analysis provides a quantitative measure of the amount and accuracy of information a speech sample provides, as well as the speaker's efficiency in relaying that information (Nicholas and Brookshire, 1993a). The measures of the analysis include three counts (i.e., words, time, and CIUs) and three "calculated measures" (Nicholas and Brookshire, 1993a, p. 343). The three calculated measures are; words per minute (WPM), percent of words that represent CIUs (%CIUs), and CIUs per minute (CIUs/min or CIUrate).

The counts are raw measures of a speaker's discourse ability. The word count is the total amount of language the speaker was able to produce. The CIU count is the total amount of correct information related to the stimulus the speaker was able to produce. Finally, the time is the amount of time the speaker required to relay the information. The calculated measures provide a more comprehensive analysis of the accuracy and efficiency of the message relayed by the speaker. The WPM measure demonstrates a speaker’s efficiency in producing language. The %CIUs measure provides an analysis of how much of that language communicates accurate content with respect to the prompt or stimulus eliciting the discourse sample. The CIUrate measures the speaker’s efficiency in producing the message. Another major benefit of the calculated measures is that a speaker’s performance can be directly compared to that of other speakers. For example, by comparing transcripts of two speakers with word counts of 100 and 50, respectively, a clinician may assume that speaker one is a more informative speaker. After scoring the transcripts and finding the speaker’s CIUs (speaker 1 CIUs = 20, time = 2 minutes; speaker 2 CIUs = 40, time = 4 minutes), the clinician may assume that speaker two is more efficient. However, by analyzing the CIUrate the clinician will learn that the speakers are equally informative and efficient by producing 10 CIUs/min each. Overall, CIU analysis is a relatively simple, but revealing evaluation of a speaker’s discourse production.

Procedures. Nicholas and Brookshire (1993a) recommended a standard set of procedures for eliciting and collecting the speech sample, as well as rules for scoring. The clinician administering the measure should present speakers with a line drawing picture and ask them to describe what they see. The speech sample is then transcribed verbatim. Following transcription, the transcript is scored, by process of deletion, to count the total number of words and the total number of CIUs produced by the speaker. Finally, the time is calculated for each speaking sample. The three counts (i.e., word count, time, and CIU count totals) are then used to determine the three “calculated measures” (Nicholas and Brookshire, 1993a, p. 343).

Nicholas and Brookshire’s (1993a) original rule system incorporated two sections (i.e., counting words, and counting CIUs). Before beginning the counts, raters review the transcript to remove words and statements that are not related to the stimulus prompt, are said before beginning the discourse task, or are about beginning or finishing the task. Words that are removed are marked by drawing, with a pen, a horizontal line through the word. Following this preparation, raters evaluate the transcript to measure the word count. Words are evaluated for inclusion based on two rules; “Do not count the following”, and “Count the following” (Nicholas & Brookshire, 1993a, p. 348-350). Raters cross out with red X’s all the words that are not included in the word count. Then, raters count all the remaining words in the transcript based on sub-rules of rule two. After obtaining the word count, the transcript is evaluated to obtain the CIU count. The same two rules (i.e., “Do not count…”, and “Count…”) are applied with differing sub-rules. There are 12 sub-rules for rule one and nine sub-rules for rule two. To apply the rules, raters place a diagonal slash through words that should not be included (Nicholas & Brookshire, 1993a). Then, raters count the remaining words in the transcript based on the sub-rules of rule two. Time for the sample only includes time spent by the person being evaluated from the first to the last words included in the word count. With the advent of computers and computer-based text editing tools, raters now use analogous tools (e.g., delete, format with slashes, bold) within simple text-editing software to remove and mark-up the transcripts (see Table 1 for an example).

Original Paragraph	Paragraph with CIU mark-up
Oh, is it recording? Well, let's get started. Um the article that I read in the beginning was about ethics in public speaking. And uh again uh it talked about the pyramid of ethics which includes intent, means, and ends. Um and uh then it went on to explain different expectations and aspects of ethics in public speaking more specifically. Um, I'm done now.	~~Oh, is it recording? Well, let's get started.~~[a] ~~Um~~[b] the article that I read in the beginning was about ethics in public speaking. **And**[c] ~~uh~~ **again** ~~uh~~ it talked about the pyramid of ethics which includes intent, means, **and** ends. ~~Um~~ **and** ~~uh~~ then it went on to explain different expectations **and** aspects of ethics in public speaking more specifically. ~~Um~~ ~~I'm done now.~~

[a] ~~Words~~ formatted in black and strikethrough indicate words not included in sample.
[b] ~~Words~~ formatted in red and strikethrough indicate words not included in the word count.
[c] **Words** formatted in black and bold indicate words not included in the CIU count.

Table 1: Example of CIU analysis mark-up

Sensitivity. CIU analysis is sufficiently sensitive to differentiate between persons with and without neurocognitive-linguistic deficits (Capilouto et al., 2005; Carlomagno, Giannotti, Vorano, & Marini, 2011; Matsuoka et al., 2012; McNeil, Doyle, Fossett, Park, & Goda, 2001; Nicholas & Brookshire 1993a). The research studies to date have focused on disorders of an acquired nature rather than congenital disorders. In a study by Matsuoka et al., (2012), CIU analysis was used to assess differences between persons with a traumatic brain injury (TBI) and controls on narrative recall tasks. The study demonstrated a significant difference between the TBI and the control group in CIUs/min, or the efficiency in relaying information. A study conducted by Capilouto et al., (2005) revealed that, in narrative tasks, older adults produce significantly lower %CIU than their younger counterparts. These examples, along with Nicholas and Brookshire's (1993a) introductory study, demonstrate that CIU analysis can be a useful tool in correctly identifying changes in neurocognitive-linguistic function following an acquired brain injury.

Validity. Evidence of both concurrent and content validity for CIU analysis is provided in the literature. The sensitivity of the measure discussed above provides evidence of concurrent validity (Capilouto et al., 2005; Carlomagno et al., 2011; Matsuoka et al., 2012; McNeil et al., 2001; Nicholas & Brookshire 1993a). Evidence of content validity is provided by Nicholas and Brookshire (1993a) who determined that the word choices of the aphasic speakers "generally resembled" those of normal adult speakers (p. 344). Additionally, later studies have demonstrated that changes in CIU measures can be perceived by even naïve listeners as changes in informativeness (Doyle, Goda, & Spencer, 1995; Jacobs, 2001; Ross & Wertz, 1999).

Reliability. Within structured discourse tasks, CIU analysis has been shown to have acceptable inter-rater reliability. Based on a percent agreement calculation (i.e., "[total agreements/(total agreements + total disagreements)] x 100"), Nicholas and Brookshire (1993a) demonstrated inter-rater reliability exceeding 90% for CIUs (Nicholas & Brookshire, 1993a, p. 340). Intra-rater reliability exceeding 95% for CIUs was also reported (Nicholas & Brookshire, 1993a). Later studies have also demonstrated adequate to good inter-rater reliability despite changing the stimuli materials (Doyle et al., 1995; Edmonds, 2013). The lowest reported reliability (i.e., inter-rater of 56%, and intra-rater of 76%) occurred in one study which elicited the discourse in an unstructured conversation task (i.e., sharing personal information; Oelschlaeger & Thorne, 1999).

Applications of CIU Analysis

Practicing speech-language pathologists (SLPs) are required to evaluate language comprehension and production (Council for Clinical Certification in Audiology and Speech-Language Pathology of the American Speech-Language-Hearing Association, 2013). CIU analysis was initially used to measure language production during discourse (Nicholas & Brookshire, 1993a; Wambaugh et al., 2013). More recently, it has been used as a measure of language comprehension for listening and reading (Coelho, Lê, Mozeiko, Krueger, & Grafman, 2012; Kucheria et al., in press; Matsuoka et al., 2012). The potential to use CIU analysis to measure multiple domains of language makes it a desirable clinical assessment tool.

Discourse measurement. CIU analysis has been used extensively in research to measure discourse production in people with aphasia. The primary application has been to increase understanding of the effects of aphasia on discourse production (Nicholas & Brookshire, 1993a). Nicholas and Brookshire's (1993a) seminal study conducted a discourse analysis by measuring CIUs in order to evaluate the differences in the discourse production of aphasic and healthy controls. The researchers calculated cutoff scores for each of the five measures of CIU analysis (i.e., word count, CIU count, WPM, %CIUs, and CIUs/min). The calculated measures (WPM, %CIUs, and CIUs/min) most reliably distinguished between typical and aphasic speakers. Nicholas and Brookshire found that the most sensitive method for discriminating between people with and without aphasia was to concurrently evaluate WPM and %CIUs.

CIU analysis has also been used to assess the efficacy and effectiveness of discourse production treatments (Avent & Austermann, 2003; Fink, Bartlett, Lowery, Linebarger, & Schwartz, 2008; Jacobs, 2001; Linebarger, McCall, & Berndt, 2004; Marshall, Laures-Gore, DuBay, Williams, & Bryant, 2015; Oelschlaeger & Thorne, 1999; Peach & Reuter, 2010; Savage & Donovan, 2017; Wambaugh et al., 2013; Wambaugh, Wright, & Nessler, 2012). Researchers used measures of CIU analysis to measure the effectiveness of Modified Response Elaboration Training (MRET; (Wambaugh et al., 2013; Wambaugh et al., 2012). MRET was designed to increase discourse content production by eliciting productions in structured settings and with models. Following the speakers' responses, clinicians reinforced the original statement and requested elaborations to expand the speakers' language production. Speakers were then required to retell the original statements with elaborations. The researchers used CIU counts to demonstrate significant increases in speakers' discourse production following treatment.

Reading comprehension measurement. CIU analysis has also been used in several recent, seminal studies to provide a concrete measure of reading comprehension both in patients with TBI (Griffiths et al., 2016) and in typical readers (Keenan et al., 2008; Reed & Vaughn, 2012). The potential to use CIU analysis to measure reading comprehension is important, as there is a gap in the availability of reading comprehension measures designed to evaluate high level reading impairments (Griffiths et al., 2016; Kucheria et al., in press). Currently clinicians use standardized assessments to measure reading comprehension including the Nelson-Denny Reading Test (NDRT; Brown, Fishco, & Hanna, 1993), the Gray Oral Reading Test-Fifth Edition (GORT-5; Wiederholt & Bryant, 2012) or comprehensive batteries such as the Woodcock-Johnson IV (WJ IV; LaForte, McGrew, & Schrank, 2014). These assessments enable clinicians to quickly and easily screen individuals based on their reading levels. However, despite their feasibility, they are not designed to assess the constructs of high level reading comprehension required for post-secondary education (Keenan et al., 2008). As discussed by Kucheria et al., (in press), the limitations of current high-level reading assessments include short length and simplicity of reading stimuli, incongruent testing formats, and lack of theoretical grounding.

Current reading comprehension assessments do not provide expository reading stimuli that are comparable in length or reading difficulty of college level texts. (Kucheria et al., in press). The assessments utilize recognition testing formats such as multiple choice or mazes and allow the reading stimuli to remain in sight for readers to review. These formats are not in line with post-secondary testing procedures (Kucheria et al., in press). Finally, these assessments are based on theoretical frameworks that focus on fundamental reading processes such as phonological processing, word recognition, vocabulary, and sentence comprehension. However, readers at the postsecondary college level often have these skills intact (Hannon, 2012). Executive function, working memory, processing speed are more predictive of postsecondary readers' comprehension skills, and are constructs that are not evaluated in these assessments (Kucheria et al., in press; Macaruso & Shankweiler, 2010; Rapp et al., 2007; Sullivan et al., 2014).

In a series of studies evaluating reading comprehension in young adults returning to secondary education following a TBI, researchers identified CIU analysis as a useful tool in measuring reading comprehension in higher level reading impairments (Griffiths, 2013; Griffiths et al., 2016; Kucheria et al., in press; Sohlberg, Fickas, & Griffiths, 2011; Sohlberg, Griffiths, & Fickas, 2014). They argued that by using CIU analysis to measure comprehension, clinicians could evaluate "the mental representation of textual content constructed by the reader" as well as their working memory (Kucheria et al., in press, p. 7). Researchers utilized CIU analysis to measure "quantity of information comprehended" (Griffiths et al., 2016, p. 169) from expository texts with and without strategy use. In the strategy use condition, they demonstrated significant differences in reading comprehension as measured by efficiency of recall, or CIUrate (Griffiths, 2013; Griffiths et al., 2016; Sohlberg et al., 2014). This finding was corroborated by significant differences between conditions on another established assessment method (Griffiths, 2013; Griffiths et al., 2016; Sohlberg et al., 2014).

Challenges of CIU Analysis

Challenges in measurement. CIU analysis accurately measures an individual's ability to convey understanding via verbal statements that are transcribed in response to a structured comprehension task. Other factors, besides comprehension, may be responsible for poor performance on CIU analysis when attempting to measure reading comprehension. For example, speech sample elicitation tasks require some level of auditory language comprehension (Griffiths et al., 2016; Kucheria et al., in press; Nicholas & Brookshire, 1993a). This can be problematic because performance can be impacted by the speaker's understanding of the instructions. With limited comprehension of the verbally provided task instructions a speaker might provide a short summary instead of the detailed review that a clinician requested. While preliminary research into how reading comprehension measures can be affected by auditory language comprehension tasks has been conducted, this remains an area for research to examine (Coelho et al., 2012; Kucheria et al., in press). Additionally, CIU analysis requires that a speaker be able to formulate a verbal response, an ability which can be impacted following brain injury (Coelho et al., 2012; Matsuoka et al., 2012). As such, when measuring reading comprehension, results may be confounded by an individual's verbal ability (i.e., vocabulary, syntactic formulation; Kucheria et al., in press). It would thus only be valid if the individual did not have concomitant language challenges.

Nicholas and Brookshire noted that CIU analysis does not measure the relative importance of the information provided, or whether any important details have been neglected by the speaker. To meet this need they designed an additional, but separate, measure called main concept analysis (Nicholas & Brookshire, 1993b). Main concept analysis is an assessment of whether or not a patient makes a statement about each of a set of pre-determined ideas. For example, if the stimulus were a picture of a mother washing dishes then the patient should produce a statement conveying this idea. Wambaugh et al. (2013) further noted that CIU analysis does not provide a measure of the diversity of words, or novelty of utterances. These authors also used a supplemental measure which was a simple count of the unique or novel words produced. For example, if a patient used the word *comb* twice in one sample, it was counted only once if it was a noun in both cases. However if the patient used the word *comb* as a noun once and as a verb once then it was counted twice.

While there are limitations in interpreting CIU analysis, it remains an effective measure of language production (Nicholas & Brookshire, 1993a; Wambaugh et al., 2013). CIU analysis also remains a sensitive tool in discriminating between healthy controls and persons with aphasia (Nicholas & Brookshire, 1993a). Additionally, preliminary research has shown CIU analysis to be an effective tool in filling measurement gaps for persons with TBI both in their language expression and comprehension (Coelho et al., 2012; Griffiths et al., 2016; Kucheria et al., in press; Matsuoka et al., 2012).

Challenges in clinical implementation. If CIU analysis has potential to fill measurement gaps, why has it not been more widely adopted by SLPs in clinical contexts? Despite its clinical advantages, there are obstacles to clinicians using the measure more regularly in clinical contexts. These obstacles are largely due to implementation barriers.

Implementation science is defined as “the scientific study of methods to promote the systematic uptake of proven clinical treatments, practices, organisational, and management interventions into routine practice to improve service delivery” (Olswang & Prelock, 2015, p. S1819). The science can be used to help us understand how to reduce barriers to using CIU analysis. The Consolidated Framework for Implementation Research (CFIR) has particular relevance for trying to understand the factors that support or inhibit adoption of an assessment or intervention tool (Damschroder et al., 2009). The CFIR categorizes possible implementation barriers into the following five domains: intervention characteristics (IC), outer setting (OS), inner setting (IS), characteristics of individuals involved (CI), and the process of implementation (PI). The IC domain specifically focuses on aspects such as the source, evidence strength and quality, adaptability, trialability, complexity, design, and cost of the item being implemented. This domain seeks to quantify how extrinsically appealing the new intervention or measure will be to its targeted users. OS refers to aspects such as patient needs and resources as well as external policies and procedures of the organization doing the implementation. IS takes into account elements of the implementation process such as the structural characteristics, culture, climate, and readiness of the organization doing the implementation. The CI domain includes such principles as individuals’ knowledge and beliefs about the item being implemented, as well as individuals' self-efficacy. Finally, the PI domain takes into account the processes of planning, engaging, executing, and evaluating the implementation of the item. Specific to the implementation of CIU analysis in clinical contexts characteristics of the IS, OS, CI, and IC domains explain a number of potential reasons why the measure may not have been implemented in clinical settings.

A limiting IS feature for settings where CIU analysis could be used is the continued upward trend of productivity requirements for clinicians meaning less time for scoring assessments. The American Speech-Language- Hearing Association (ASHA) conducts a yearly survey of a representative subset of clinicians practicing working in healthcare settings (i.e., settings in which CIU analysis would be most likely to be used). The summary report of the 2005 survey indicates that 61% of clinicians reported having a productivity requirement with a mean productivity requirement of 74% (ASHA, 2005). The 2017 summary report indicates that 64% of clinicians had productivity requirements with a mean requirement of 80% (ASHA, 2017a). The majority of SLPs (i.e., 68% in 2017 versus 64% in 2015) reported that only time spent in direct patient care activities (i.e., with patient present) counted towards meeting their productivity requirements (ASHA, 2017a). It has become increasingly difficult for clinicians to engage in time consuming analysis of evaluations because it is not a direct patient care activity (Brown, 2017).

One example of a limiting OS demand pertains to the reimbursement guidelines for assessments performed by SLPs. Medicare is a trend-setting insurer for reimbursement guidelines in the United States. Under Medicare, the national average pay amount SLPs were allowed in 2017 for testing all aspects of language comprehension and expression was $201.60 (ASHA, 2017b) which can only be billed once (ASHA, n.d.). This pre-determined pay amount directly correlates to time allotments of 7 minutes for a clinician to familiarize themselves with a patient, 120 minutes to evaluate the patient one-on-one, and 30 minutes to review and document the evaluation (Centers for Medicare and Medicaid Services, 2017; Swanson, 2018).

A significant factor of the CI domain limiting the clinical uptake of CIU analysis is lack of training. Few clinicians receive in-depth training on administration and interpretation of CIU analysis. Due to the breadth of material required for entry into the field, new clinicians receive less depth of academic knowledge in each therapy domain (Council for Clinical Certification in Audiology and Speech-Language Pathology of the American Speech-Language-Hearing Association, 2013; Leslie et al., 2011). In our own training program at the University of Oregon, which is a highly ranked program within a large research university, students are not trained to administer CIU analysis and receive only a cursory overview of the principles behind the analysis. Without the opportunity for active learning provided during their initial training, each clinician must find the time to learn and implement CIU analysis, further compounding the time dilemma.

The ICs of CIU analysis are such that clinicians must spend a significant amount of time in completing the measures. Eliciting the speech sample may require clinicians spend as long as 1.25 hours for recalls of expository text (Kucheria et al., in press) and as long as 10 minutes for picture description tasks (Nicholas & Brookshire, 1993a). Transcription is the most time costly process of CIU analysis. As many as six to seven hours of transcription are required for every one hour of audio recording (Britten, 1995).

Typical recall times depend on the stimulus provided, the type of impairment, and characteristics of the individual being evaluated. Nicholas and Brookshire (1993a) reported recall times ranging from 28 seconds to 2 minutes 47 seconds in the context of their picture description task. Kucheria et al. (in press) recorded recall times ranging from 18 seconds to 20 minutes 25 seconds in their expository text recall task. This accounts for transcription times of 3 to 16 minutes for picture description, and from 1 minute to more than 2 hours for recalls (i.e., using a factor of 6 to 1 for time spent transcribing versus duration of audio sample; Britten, 1995). Finally, clinicians may spend anywhere from 3 to 25 minutes scoring a transcript and calculating measures. In summation, the time required to complete CIU analysis can range from approximately 20 minutes to 4 hours (see Table 2 for a summary of documented times).

Aspect of Administration	Discourse Measurement (minutes)	Reading Comprehension Measurement (minutes)
Elicitation & Recording	10[a]	75
Transcription	3 – 16[a]	1 – 124
Scoring	*	3 – 25
Counting & Calculation	*	*

[a]From Nicholas & Brookshire 1993a
*Time still unknown.

Table 2: Time to administer CIU analysis by hand

In summation, the administration of CIU analysis must be made more efficient to ensure the tool can be used to measure reading comprehension in clinical settings. The measure requires an extended amount of time to administer, clinicians have limited time to administer the measure, and clinicians have limited time to engage in the training needed to know how to administer CIU analysis.

Automation as a Solution

One possible solution to the time barrier is to automate the processes involved in applying CIU analysis by using computer algorithms. Researchers have demonstrated significant correlations between algorithm and hand-based measurement of speech rate in dysarthric speakers and normal speakers (Martens et al., 2015; Mujumdar & Kubichek, 2010). Martens et al., (2015) demonstrated correlations ranging from .80-.96 indicating their algorithm accurately measured speech rate. They neglected to report increases in the clinician’s efficiency simply stating that “manual [speech rate] calculation can be time-consuming” (Martens et al., 2015, p. 699). This suggests that their tool provides clinicians a more efficient method than calculating by hand. Long (2001) provided an in-depth study of time differences between manual and computer scoring of phonological and grammatical analyses. Long demonstrated increases in efficiency ranging from 17 to 35 times faster for computerized phonological analysis. Increases in efficiency for computerized grammatical analysis were more conservative, however 13 of the 14 analyses were significantly more efficient by computer. Finally, Long neglected to provide a point-by-point analysis of the scoring accuracy but reported that the computer analysis was as accurate as the hand scoring.

As Long (2001) demonstrated, certain analyses are considerably more difficult than others. Many of the computerized analyses developed for SLPs continue to require human input to complete the metalinguistic portion of the analysis (e.g., categorizing words grammatically, identifying utterance boundaries; Channel & Johnson, 2001). This is a phenomenon known in computer science as a human-in-the-loop approach.

CIU analysis incorporates both simple linguistic and complex metalinguistic rules. In order to automate the application of CIU analysis the process can be broken down into three major tasks: (1) transcription of the sample, (2) scoring and (3) analysis of the words.

Automatic transcription is known by varied terms (e.g., voice-to-text, automatic speech recognition, voice recognition software, etc.) and has become a mainstream technology. Automatic transcription powers digital assistants in wide commercial use and is being rapaciously developed across companies like Google, Amazon, Apple, and many more. A New York Times article dubbed the movement “The Great AI Awakening” (Lewis-kraus, 2016). Many of these companies are providing application programming interfaces (APIs) which allow any person to integrate automatic transcription tools into their software. Incorporation of these automatic transcription tools is a necessary prerequisite in automation of CIU analysis.

Scoring, the second challenge of automation, is more difficult to automate. As discussed above, scoring transcripts for CIU analysis involves applying rules to remove non-counted words, and phrases. These rules vary substantially in terms of their linguistic and metalinguistic complexity. An example of a simple rule is, “Remove non-word fillers such as um, er, or uh” (Kucheria et al., in press). These types of rules are easy to automate given the computer is highly proficient at finding linguistic patterns in text. Rules with the highest level of linguistic complexity include those which require linguistic analysis of one clause and comparison of that clause to all others in the transcript. One such complex rule requires that the clinician remove any text that is a paraphrase of a previous statement. Recognizing a statement which has been paraphrased is an open problem in computer science (Androutsopoulos & Malakasiotis, 2010).

A team of University of Oregon researchers from computer science and speech-language pathology disciplines collaborated to develop a partially-automated CIU analysis tool. The tool was created as part of the "RULE" training and assessment tool discussed by Kucheria et al. (in press). The researchers first identified two primary areas for development: transcription and CIU analysis (i.e., rule application). The researchers decided to prioritize automating the application of the CIU scoring rules given the widely available transcription tools discussed above.

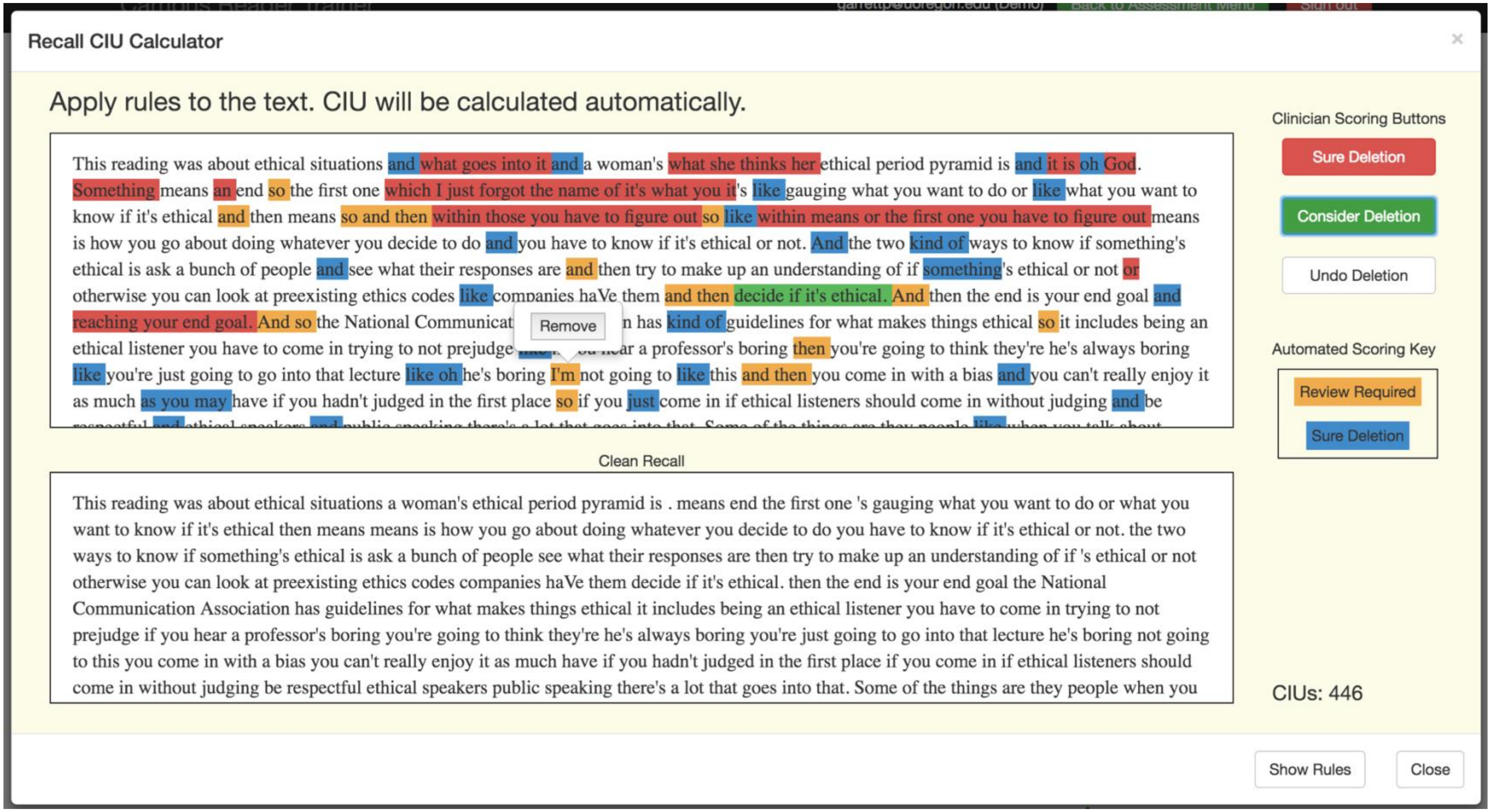
The researchers implemented the CIU scoring tool as a web-based user interface (UI) and a transcript analysis server (server). The server was designed to analyze the transcript to apply the CIU analysis rules (Kucheria et al., in press). The UI was designed to allow raters to review the automated scoring and finalize scoring of CIU analysis rules not implemented by the server.

The server was built using a widely-used set of Python tools called the Natural Language Toolkit (NLTK). The research team assessed each rule for the level of difficulty to automate. The rules were determined to be as simple, intermediate, or complex to automate. Using NLTK, the research team implemented all of the rules of easy, and intermediate difficulty (rules 1-8, and 10; see appendix A for a summary of the rules). The server marks words in the transcript using these rules and then returns the marked-up transcript to the UI. Both the server and the UI were designed and implemented using a collaborative approach known in computer science as agile software development (Beck et al., 2001). Functional software was quickly developed, and with researcher feedback and pilot testing, the team worked through several iterations of the server and UI before this thesis study was conducted.

Following the server's analysis of the transcript the server returned to the UI a list of objects which identified words to be removed from the transcript with the following attributes: word number from beginning, rule applied, and a binary indicator of the server's level of certainty. Two levels of certainty, high certainty and low certainty, were used to decrease the complexity of the tool and allow clinicians to quickly review the server's recommended deletions. Based upon the level of certainty, the UI displays the transcript with the words identified for deletion highlighted in either blue or yellow. The team subjectively identified which rules the server implemented with higher correlation to human scoring. An example of the UI version used in this study can be found in Figure 1 below. The blue highlight indicates a word identified under a rule where the server had higher reliability with human scoring. Words highlighted in orange denoted poorer reliability with human scoring and the clinician's need to review. The UI allowed raters to remove highlights from words and to add highlights. Rater's were given the option of red and green highlight colors. Any words highlighted in blue, yellow, or red were removed from the final CIU count calculated by the tool.

Summary and Purpose of Current Study

As a measure of the informativeness and efficiency of a person's expressive language, CIU analysis has a large and strong evidence base. By varying the stimuli for elicitation and instructions, researchers have established that the measure can be a powerful tool for measuring function across multiple linguistic domains and types of disorders. Specific to this study, CIU analysis provides a potentially useful tool to measure reading comprehension for individuals with TBI (Kucheria et al., in press). However, the time required to administer the measure is a significant limiting factor for



Apply rules to the text. CIU will be calculated automatically.

This reading was about ethical situations and what goes into it and a woman's what she thinks her ethical period pyramid is and it is oh God. Something means an end so the first one which I just forgot the name of it's what you it's like gauging what you want to do or like what you want to know if it's ethical and then means so and then within those you have to figure out so like within means or the first one you have to figure out means is how you go about doing whatever you decide to do and you have to know if it's ethical or not. And the two kind of ways to know if something's ethical is ask a bunch of people and see what their responses are and then try to make up an understanding of if something's ethical or not or otherwise you can look at preexisting ethics codes like companies haVe them and then decide if it's ethical. And then the end is your end goal and reaching your end goal. And so the National Communicat [illegible] n has kind of guidelines for what makes things ethical so it includes being an ethical listener you have to come in trying to not prejudge [illegible] ar a professor's boring then you're going to think they're he's always boring like you're just going to go into that lecture like oh he's boring I'm not going to like this and then you come in with a bias and you can't really enjoy it as much as you may have if you hadn't judged in the first place so if you just come in if ethical listeners should come in without judging and be respectful and ethical speakers and public speaking there's a lot that goes into that. Some of the things are they people like when you talk about

Clean Recall

This reading was about ethical situations a woman's ethical period pyramid is . means end the first one 's gauging what you want to do or what you want to know if it's ethical then means means is how you go about doing whatever you decide to do you have to know if it's ethical or not. the two ways to know if something's ethical is ask a bunch of people see what their responses are then try to make up an understanding of if 's ethical or not otherwise you can look at preexisting ethics codes companies haVe them decide if it's ethical. then the end is your end goal the National Communication Association has guidelines for what makes things ethical it includes being an ethical listener you have to come in trying to not prejudge if you hear a professor's boring you're going to think they're he's always boring you're just going to go into that lecture he's boring not going to this you come in with a bias you can't really enjoy it as much have if you hadn't judged in the first place if you come in if ethical listeners should come in without judging be respectful ethical speakers public speaking there's a lot that goes into that. Some of the things are they people when you

CIUs: 446

Figure 1: CIU Scoring Tool User Interface

its use in clinical contexts. The clinical feasibility of CIU analysis can be increased by partially automating the measure and thus reducing a clinician’s time to apply the measure. Automation should first focus on a human-in-the-loop (HITL) approach in order to increase efficiency while maintaining sensitivity, validity, and reliability. The purpose of this study was primarily exploratory with the goal of ascertaining the feasibility of partially automating CIU analysis in a clinical context (i.e., within a tool for measuring reading comprehension). The hypotheses of this study was that in a clinical context scoring of CIU analysis would be as accurate and more efficient when applied by HITL computer automation than by manual, hand scoring. Adherent to implementation science principles, the hope is that if CIU analysis can be made more efficient, it would be adopted by clinicians.

Specific research questions were:

(1) Is partially-automated, HITL scoring as accurate and reliable as scoring by hand?

(2) Does partially-automated, HITL scoring require less time than scoring by hand?

CHAPTER III

METHODS AND PROCEDURES

The study provided a descriptive analysis of the feasibility of partial automation of CIU analysis. The study focused on analyzing the partial automation of applying CIU scoring rules (Kucheria et al., in press; Nicholas & Brookshire, 1993a). The dependent variables were (1) time to apply all CIU scoring rules to a prepared transcript, and (2) intra-rater scoring accuracy. The independent variable was the scoring condition (i.e., scoring by hand versus scoring with the partial automation tool). The design of the study was a within rater comparison across three raters.

Recalls and Transcription

The study used a total of 68 transcripts with each rater scoring a different number of transcripts. The transcripts were generated from free recalls of reading passages. The recalls were gathered between November 2016 and March 2017 as part of a larger reading comprehension assessment project (Kucheria et al., in press). Participants were college students matching the following inclusion criteria: (1) not diagnosed with a disability or condition that affected their reading, and (2) not admitted to a hospital or outpatient program in the last 12 months for substance abuse or psychiatric issues.

Each participant provided one immediate and one delayed recall after reading two separate chapters of expository text for a total of 4 recalls. In two separate sessions, the participants read one of the two 2200- to 2400-word passages on a computer, with a 30-minute time limit. Reading passages were excerpted from open-source introductory-level, college textbooks and corresponded to a 12th grade difficulty level. There were two possible passages, about social psychology or ethics in public speaking. Directly after finishing the reading, readers were prompted to provide an immediate recall with the cue, “Retell the text as if you were telling someone all the information you learned and were helping them prepare for a test.” The recall was recorded using the RULE web-based assessment tool and downloaded to a MacBook Pro computer (Kucheria et al., in press). After recording the immediate recall, readers completed a recognition and reading comprehension subtest and either a comprehension subtest of the Nelson Denny Reading Test (i.e., during the first session) or an informal interview to gather demographic data and administer questionnaires about their reading (i.e., during the second session).

All of the recalls were transcribed by the transcription company, Focus Forward (Focus Forward, n.d.). This company uses home-based independent contractors to transcribe audio samples. The transcribers have passed a company designed test to ensure proficiency and accuracy in transcription.

Raters

A total of three raters participated in the study with a range of skill sets in CIU analysis. Based on the scale developed by Dreyfus and Dreyfus (1980), the raters represented an advanced beginner rater, a proficient rater, and an expert rater. The raters were one speech-language pathologist (SLP) undergraduate student research assistant who received compensation as part of a larger study (i.e., rater 1, advanced beginner rater), one SLP graduate student clinician (i.e., rater 2, author and proficient rater), and one SLP PhD student with 3 years of experience in the field including use of CIU analysis (i.e., rater 3, expert rater).

CIU Scoring

The transcripts were analyzed using both the typical manual approach to calculating CIUs (Nicholas & Brookshire, 1993a) and the previously described partially-automated, human-in-the-loop scoring system. Raters in this study followed a similar process as Nicholas and Brookshire's (1993a) original rules. In order to apply CIU analysis to the evaluation of verbal recalls of expository text, an adapted rule set of 12 rules, developed by Kucheria et al., (in press), was used to score the recalls both by hand and with a computer-based tool (see appendix A for a summary of the rules). For computer-based scoring raters both reviewed the computer analysis of the scoring as well as applying the rules which had not been automated. Following the adapted rules, the raters applied the rules to words and ideas in the transcripts and marked or deleted words and phrases that did not meet criteria. Words that were not deleted or marked were then included in the final CIU count.

Inter-rater reliability. To ensure that the raters were consistent in their interpretation of the scoring rules, they reviewed the adapted rule set together and collaborated to score a set of 4 transcripts. Following that training, the raters independently scored a practice set of 12 transcripts. Then they compared their scoring, discussed disagreements, and clarified misunderstandings. Initial inter-rater reliability as measured by an intra-class correlation coefficient (ICC; Koo & Li, 2016) was excellent (ICC estimate >.90).

Scoring procedures. After scoring the practice set and reaching adequate reliability, raters scored a subset of all the transcripts (rater 1 = 66/68, rater 2 = 68/68, rater 3 = 19/68) by hand using the adapted rule set. For example, raters one two and three each scored the first transcript under each condition (i.e., by hand and with the tool). Raters scored by hand first in order to enable them to provide subjective feedback for the automation process discussed above.The process of scoring each transcript was timed to account for changes in rating efficiency between scoring by hand and with the partially-automated tool. In order to model realistic clinical use, raters scored at most four transcripts at a time, followed by a wait period of at least 30 minutes between sittings. All hand scoring was completed between June and October 2017. Raters began timing after they had opened either a Google Docs, or Microsoft Word document and pasted the text of the transcript into the document. Raters stopped timing after applying all CIU scoring rules to the transcript. An explanation of scoring by hand is given above, and an example can be found in table 1 above.

In order to reduce any practice effect, the raters were required to take at least a one-month break after scoring the transcripts by hand. Following the break, the raters scored all the transcripts using the partial automation tool, following the same protocol described for hand scoring above. All scoring using the partial automation tool was completed between November and December 2017. In order to score transcripts with the partially-automated tool, raters used a computer with internet connection and the Google Chrome browser. Raters loaded the scoring tool in one tab, then copied and pasted the transcript to be scored into the tool. Raters began timing their scoring at this juncture to include time required by the transcript analysis server. Raters clicked the save button to send the transcript to the transcript analysis server. The server analyzed the transcript and sent back its results to the web interface. At this point, a rater could accept or reject a choice made by the server. A rater could also delete words that were not selected by the server. As discussed above, the web interface uses color coding to allow the rater to keep track of changes. An example of scoring with the partially-automated tool can be found in Figure 1 above.

Analysis

Intra-rater reliability. ICC estimates were used to compare the intra-rater reliability between hand scoring and partial automation conditions. Many studies on CIU have relied on percent agreement as a measure of reliability (Fink et al., 2008; Nicholas & Brookshire, 1993a; Oelschlaeger & Thorne, 1999; Wambaugh et al., 2013). Percent agreement was most often calculated using the equation [total agreements/(total agreements + total disagreements)] x 100. This measure provides a percentage of the number of words for which raters agreed to remove or not remove in relation to the total number of words. While this calculation provides an analysis of agreement word by word on each transcript, it is not a fine tuned statistical model for analyzing overall correlation in scores across multiple transcripts. ICC operates on data structured as groups and describes how strongly units in the same group resemble each other. Furthermore, ICC estimates have been used in various healthcare related studies to evaluate inter-rater, test-retest, and intra-rater reliability (Clare, Adams, & Maher, 2003; Houweling, Bolton, & Newell, 2014; Leach, Parker, & Veal, 2003; Owens, Hart, Donofrio, Haralambous, & Mierzejewski, 2004). Finally, ICC estimates reflect the degree of correlation between raters' scores as well as the agreement between raters' scores, allowing this study to more accurately depict intra-rater reliability with higher possible degrees of variance in original score numbers (Koo & Li, 2016). ICC estimates and their 95% confident intervals were calculated based on a single-rating, absolute-agreement, 2-way random-effects model.

Time difference. A paired t-test was conducted to examine the effect of scoring condition (i.e., hand scoring versus partial automation scoring) on the dependent variable of a given rater's efficiency of CIU calculation (i.e., scoring time). A p-value of < .05 was considered to be statistically significant.

Both the intra-rater reliability and time difference analyses were repeated for each rater. Additionally, both analyses were conducted using SPSS statistical package version 24 (SPSS Inc, Chicago, IL).

CHAPTER IV

RESULTS

Under each scoring condition, rater one scored a total of 66 transcripts, rater two scored 68 transcripts, and rater three scored 19 transcripts.

Accuracy and Reliability Across Conditions

Intra-rater reliability was evaluated using ICC estimates to evaluate potential differences in scoring content information units between the two scoring conditions, by hand versus with the partially-automated tool.

Rater one. The advanced beginner rater maintained stable reliability across scoring conditions. Rater one demonstrated an ICC estimate of 0.992 with a 95% confidence interval range from 0.976 - 0.996. Based on the ICC results, the intra-rater reliability between scoring conditions of CIU scoring is excellent (Koo & Li, 2016). See Table 3 below for overall results.

Rater two. The proficient rater maintained stable reliability across scoring conditions. Rater two demonstrated an ICC estimate of 0.948 with a 95% confidence interval range from 0.916 - 0.968. Based on the ICC results, the intra-rater reliability between scoring conditions of CIU scoring is excellent (Koo & Li, 2016). See Table 3 below for overall results.

Rater three. The expert rater maintained stable reliability across scoring conditions. Rater three demonstrated an ICC estimate of 0.974 with a 95% confidence interval range from 0.934 - 0.990. Based on the ICC results, the intra-rater reliability between scoring conditions of CIU scoring is excellent (Koo & Li, 2016). See Table 3 below for overall results.

Rater	Intraclass Correlation	95% Confidence Interval		Level of Significance	Interpretation of Reliability
		Lower Bound	Upper Bound		
One	0.992	0.976	0.996	<.001	Excellent
Two	0.948	0.916	0.968	<.001	Excellent
Three	0.974	0.934	0.990	<.001	Excellent

Table 3: Intraclass correlation coefficient estimates by rater

Time Effects Across Conditions

Rater one. Rater one demonstrated differences in scoring times between scoring by hand scoring (*M* = 1006 seconds/transcript, *SD* = 715 seconds), and scoring with the partial automation tool (*M* = 390 seconds/transcript, *SD* = 257 seconds). Scoring with the partial automation tool was significantly faster than scoring by hand ($t(65) = -10.84$, $p < .001$). Rater one demonstrated a large overall increase in efficiency with the partial-automation tool as demonstrated by the effect size (Sullivan & Feinn, 2012). See Table 4 for a summary of the paired t-test results. Of note is that rater one's mean scoring time by hand was nearly three times greater than with the partial automation tool.

Rater two. Rater two demonstrated mean differences in scoring times between scoring by hand scoring (*M* = 282 seconds/transcript, *SD* = 239 seconds), and scoring with the partial automation tool (*M* = 245 seconds/transcript, *SD* = 267 seconds). Scoring with the partial automation tool was significantly faster than scoring by hand ($t(67) = -4.92$, $p < .001$). Rater two demonstrated a large overall increase in efficiency with the partial-automation tool as demonstrated by the effect size (Sullivan & Feinn, 2012). See Table 4 for a summary of the paired t-test results.

Rater three. Rater three demonstrated mean differences in scoring times between scoring by hand (*M* = 303 seconds/transcript, *SD* = 219 seconds), and scoring with the partial automation tool (*M* = 220 seconds/transcript, *SD* = 140 seconds). Scoring with the partial automation tool was significantly faster than scoring by hand ($t(18) = -3.96$, $p < .001$). Rater three demonstrated a medium overall increase in efficiency with the partial-automation tool as demonstrated by the effect size (Sullivan & Feinn, 2012). See Table 4 for a summary of the paired t-test results.

	Paired Differences (seconds)			Effect	
	Mean	Std. Deviation	Significance	Size	Interpretation
Rater 1	[a]-615.985	461.446	<.001	0.909	Large
Rater 2	-37.221	62.446	<.001	1.335	Large
Rater 3	-82.579	90.865	.001	0.596	Medium

[a] Negative values indicate increased efficiency (time to score by hand subtracted from time to score with tool).

Table 4: Paired sample t-test results by rater

CHAPTER V

DISCUSSION

The purpose of the current study was to examine whether correct information unit (CIU) analysis scoring could be automated to achieve more efficient application. Specifically, the study focused on applying CIU analysis scoring to transcripts of free recalls in order to measure reading comprehension in a clinical setting. The experiment aimed to answer two questions: (1) Is partially-automated, human-in-the-loop scoring (HITL scoring) as accurate and reliable as scoring by hand?, and (2) If so, does HITL scoring require less time than scoring by hand? It was hypothesized that scoring accuracy would be at least maintained, and that the HITL scoring would require less time than hand scoring. To address these questions a partial-automation of CIU scoring was developed, and three raters participated in the current study.

Accuracy and Reliability

The results showed that raters were as accurate and reliable when using the HITL automation as compared to scoring by hand. The raters established an excellent level of inter-rater reliability before scoring by hand or with the HITL automation. Following this, the three raters maintained excellent intra-rater reliability while completing scoring by hand and with the partial-automation tool. This indicates that involving computer analysis does not negatively affect the accuracy and reliability of the scoring. It is important to note that this study did not provide an analysis of the reliability of the automated scoring without human review. All analyses were performed after a human rater had reviewed the computer's analysis and made any necessary corrections.

Efficiency

The results showed that raters required significantly less time to complete HITL scoring than scoring by hand. Given, the preliminary nature of the automation tool used for this study these results are extremely promising. However, the results should be interpreted with care. While the mean difference in scores was significant it was also relatively low for two of the three raters. Further improvements to the automated scoring and automation of other aspects of CIU analysis administration may lead to further gains in efficiency. The most time-consuming facet, transcription, was not part of the automation developed for this study. Additionally, automatic counting of CIUs and calculation of CIUrate were integrated as part of the tool developed for this study, but the consequential increases in efficiency were not studied.

Clinical Significance

If the automation created for this study is improved to perform better analysis and tackle other aspects of CIU analysis administration and scoring, it will have the potential to significantly impact clinical assessment of reading comprehension. This was a preliminary study that resulted in a statistically significant increase in efficiency for each of the three raters as a result of automated CIU scoring. This study operated under two assumptions: (1) that speech-language pathologists (SLPs) are not currently using CIU analysis because it requires an inordinate amount of time to administer, and (2) that SLPs would be more likely to use CIU analysis in clinical settings if it were more efficient. Following these assumptions, SLPs should be more likely to use CIU analysis in clinical settings with the automation tool developed for this study. Despite these facts the gains in efficiency demonstrated by the raters in this study were relatively small. Clinicians may not report that an increase in efficiency of 1-2 minutes would be enough to encourage their use of CIU analysis. However, the study does demonstrate that efficiency can be increased and that time is important. Additionally, the greatest impacts on efficiency may be produced by automating elicitation of speech samples and by incorporating automatic transcription of those speech samples. Overall, this study demonstrates that automation is the right path to follow to increase the efficiency, and thus increase the clinical utility, of CIU analysis.

Limitations

Research design. In this study, the raters each scored the exact same transcripts across each condition. This may have lead to a practice effect with raters increasing in scoring skill or familiarity with the transcripts as the study progressed. However, the concerns of this possible limitation were mitigated by fact that the number, length, and repetitive content of the transcripts did not allow the raters to be biased by previous experience with a transcript. Raters reported recalling only vague details about any transcript while performing HITL scoring. The only detail that raters reported recalling was the relative length of a transcript (i.e., uncharacteristically long or short). Additionally, the raters experience with hand scoring allowed them to provide essential feedback to the software engineers in charge of producing the partially-automated scoring tool. Finally, the number of raters involved in this study limited the statistical power of the results. However, the research question of this thesis was if it were possible to make CIU analysis more efficient through automation. A small number of raters, and a preliminary automation of CIU scoring allowed the team to rapidly evaluate and answer the research question.

Restricted automation. This study did not incorporate an evaluation of the time savings provided by automatic transcription. As demonstrated in Table 2, transcription is the most time intensive task of CIU analysis. The largest increase in efficiency may be provided by incorporating and studying the effects of automatic transcription. The automation used in this study was developed using an agile process (Beck et al., 2001) but the development occurred over a relatively short period of time (i.e., three months). Additionally, all feedback on the accuracy of the automation was provided by subjective judgements from the study's raters. This exploratory study was the first step in examining automation. Future development efforts will benefit from devising methods to increase the efficiency of transcription.

Rater training. Finally, this study did not evaluate the duration or difficulty of training raters. Subjectively, inter-rater reliability between raters two and three (i.e., the proficient and expert raters) was achieved much more rapidly and easily than reliability between raters one and three, or two and three. Rater one had less experience and training in speech-language pathology than the other raters at the time of the study, and this could be an influencing factor. However, given the small number of raters involved in the study it is difficult to ascertain the cause of difficulties in training raters. This preliminary study was intended to guide further research and development of a tool which would allow SLPs to apply CIU analysis in realistic clinical settings. By revealing this need the study accomplished its aim.

Future Directions

It is critical to determine the factors that strengthen or attenuate the effects of a partially-automated CIU analysis tool when considering the next steps for research and development of this tool. The version of the CIU analysis tool which was developed for this developmental study provides increases of efficiency ranging from one minute to 10 minutes. This increase may depend on the rater's level of experience. Further development of the partially-automated CIU analysis tool will lead to larger increases in its efficiency and less work for clinicians. Further studies on the effects of this automation will increase the likelihood of its uptake into clinical use by SLPs.

Automation development. Further development of the partially-automated HITL CIU analysis tool should focus in three areas: (1) integrate automatic transcription tools, (2) utilize machine learning tools for language analysis, and (3) apply the automation principles to provide an analysis of discourse and not just reading comprehension. The team of researchers which developed the automated tool used for this study is working to integrate the tool with Google's automatic transcription services. Additionally, the team is considering best methods to incorporate principles of machine learning into the language analysis algorithms. Machine learning can be defined as a set of algorithms which allow the computer to improve its analysis with experience and human feedback. The HITL nature of the tool provides for the human feedback and will allow the tool to continue to increase its efficiency and accuracy as clinicians and researchers use it. The more linguistically complex CIU scoring rules (e.g., paraphrasing analysis) are most easily automated through this process of machine learning. Finally, the automation developed for this study was specifically designed to evaluate recalls of text to measure a person's reading comprehension. Future research should apply the same principles of automation used with this tool to other applications of CIU analysis. In order to apply automation to CIU analysis for measuring discourse, a separate tool will have to be developed.

Studies of effects. As more of the CIU analysis tasks (e.g., transcription) are added to the automated tool it will be important to ensure each of them is in reality increasing clinician efficiency while maintaining the accuracy of scoring. For example, while adding automatic transcription has the potential to create a large time savings for raters it can also lead to more work if the accuracy of the transcription is less than ideal (Gaur et al., 2016). Therefore, further research should be conducted on the effect of automation on other CIU analysis processes (i.e., transcription, counting words and CIUs, calculating CIUrate and %CIU). Additionally, this study did not ask questions about the ease with which clinicans can administer CIU analysis. This factor of difficulty to score may also stand as a barrier to the clinical feasibility of CIU analysis. To mitigate this concern further automation may allow clinicians to be less involved in the tedious task of scoring transcripts for CIU counts. Further studies will be required to evaluate the impact difficulty may have on the clinical feasibility of CIU analysis.

Clinician training. Although scoring with the tool was more efficient, a considerable amount of time was spent training clinicians on principles of scoring for CIUs. Of note is that raters did not need to be re-trained to apply scoring with the automation tool beyond a cursory introduction to the user interface. Raters were trained to perform CIU scoring by reading the scoring rules, attending a 30-minute meeting to review the rules and score a transcript in unison, and scoring a set of 12 transcripts to determine an initial level of reliability. Future research should seek to provide a standard format for training SLPs. The training would need to be accessible, efficient, and targeted to the focus of CIU analysis (i.e., reading comprehension measurement versus discourse measurement). Finally, the training should incorporate a type of check to ensure clinicians can score CIUs with a determined level of reliability.

Summary

This study provides a reminder of the importance of conducting implementation science research when developing clinical tools in order to increase adoption and usability. It was specifically conducted to evaluate if CIU analysis could be automated and if the automation would increase the efficiency of administering the measure. The current study provides evidence to support the hypothesis that CIU scoring can be reliably automated with human-in-the-loop (HITL) automation and demonstrates statistically significat increases in efficiency for each rater. However, continued research is needed to improve the partially-automated tool for further increases of efficiency as well as to provide clinicians with proper training to apply the analysis. The goal of automating CIU analysis is to fill the gap in higher level reading comprehension measures and ensure clinicians can use the tool in real-world contexts. Thus, it will be important develop this tool with feedback from clinicians working in the contexts in which it will be used. If followed, these future research directions may ultimately produce a tool that clinicians can, and will, use to measure and understand higher level reading impairments in real-world settings.

APPENDIX A

READING COMPREHENSION CIU SCORING RULES

Unpublished table of rules used with permission from Kucheria et al., 2018.

RULE #	RULES
1)	a) Delete statements that are made before or after the speaker performs the task or suggest that the speaker is ready to begin or has finished the task and do not provide information about the chapter itself.
	b) Commentary on the task and lead-in phrases or words portraying subject's uncertainty or lack of conviction on the content that do not give information about the chapter(s) or topic and are not necessary for the grammatical completeness of the statement.
	c) Remove (Original rule states Keep) Commentary on the subject's performance or personal experience
2)	Words or partial words not intelligible in context to someone who knows the text, or topic being discussed.
3)	Attempts to correct sound errors in words except for the final attempt.
4)	Dead ends, false starts, or revisions in which the speaker begins an utterance but either revises it or leaves it uncompleted and uninformative with regard to the text(s) or topic. Select the revised phrase and delete the previous phrase
5)	a) Remove non-word fillers.
	b) Remove Filler words and phrases, interjections when they do not convey information about the content of the texts(s) or topic, and tag questions. Insert a period either before or after the filler depending on where it appears in the transcript
6)	Delete any phrases that are factually incorrect or do not match the content of the article/chapter/stimulus.
7)	Repetition of words or ideas that do not add new information to the utterance, are not necessary for cohesion or grammatical correctness, and are not purposely used to intensify meaning.
8)	The first use of a pronoun for which an unambiguous referent has not been provided. Subsequent uses of the pronoun for the same unspecified or ambiguous referent are counted as correct information units
9)	Vague or nonspecific words or phrases that are not necessary for the grammatical completeness of a statement and for which the subject has not

	provided a clear referent and for which the subject could have provided a more specific word or phrase.
10)	Conjunctive terms (particularly so and then) if they are used indiscriminately as filler or continuants rather than as cohesive ties to connect ideas. The conjunction "and" is never counted.
11)	Qualifiers and modifiers if they are used indiscriminately as filler or are used unnecessarily in descriptions of events, settings, or characters that are unambiguously mentioned (original rule: pictured). Remove qualifiers or modifiers that occur more than more than 2x in a row (i.e., 3rd instance onward they are deleted) OR are semantically similar.
12)	Additional rule: After going through these rules, read the entire transcript (only read through parts that you decided to retain). Delete any words that are grammatically incorrect.

REFERENCES CITED

American Speech-Language-Hearing Association. (2005). *SLP health care survey 2005: Frequency report*. Rockville, MD: Available from www.asha.org.

American Speech-Language-Hearing Association. (2017a). *ASHA 2017 SLP Health Care Survey: Hourly and per home-visit wage report.* Available from www.asha.org.

American Speech-Language-Hearing Association. (2017b). *2018 Medicare Fee Schedule for Speech-Language Pathologists.* Available from www.asha.org.

American Speech-Language-Hearing Association. (n.d.). Medically Unlikely Edits for Speech-Language Pathology Services. Retrieved March 30, 2018, from https://www.asha.org/Practice/reimbursement/coding/Medically-Unlikely-Edits-SLP/#1

Androutsopoulos, I., & Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, *38*, 135-187.

Avent, J., & Austermann, S. (2003). Reciprocal scaffolding: A context for communication treatment in aphasia. *Aphasiology*, *17*(4), 397-404.

Beck, K., Beedle, M., Van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., ... & Kern, J. (2001). Manifesto for agile software development.

Britten, N. (1995). Qualitative research: qualitative interviews in medical research. *BMJ*, *311*(6999), 251-253.

Brown, J. (2017). SLPs Report Continued Productivity Requirements, Employer Pressure: ASHA's biennial health care survey reveals workplace trends such as productivity emphasis and off-hours documentation. *The ASHA Leader*, *22*(11), 36-37.

Brown, J. I., Fischo, V. V., & Hanna, G. S. (1993). *Nelson-Denny reading test*. Rolling Meadows, IL: Riverside Publishing.

Busch, C. R., & Brookshire, R. H. (1985). Referential communication abilities of aphasic speakers. *Clinical Aphasiology*, *15*, 255-261.

Capilouto, G., Wright, H. H., & Wagovich, S. A. (2005). CIU and main event analyses of the structured discourse of older and younger adults. *Journal of communication disorders*, *38*(6), 431-444.

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*(10), 1010-1018.

Centers for Medicare and Medicaid Services. (2017). *CMS-1676-F: CY 2018 PFS Final Rule Physician Time*. U.S. Government Publishing Office. Retrieved March 30, 2018, from https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeeSched/PFS-Federal-Regulation-Notices-Items/CMS-1676-F.html.

Channell, R. W., & Johnson, B. W. (1999). Automated grammatical tagging of child language samples. *Journal of Speech, Language, and Hearing Research*, *42*(3), 727-734.

Clare, H. A., Adams, R., & Maher, C. G. (2003). Reliability of detection of lumbar lateral shift. *Journal of Manipulative & Physiological Therapeutics*, *26*(8), 476-480.

Coelho, C., Lê, K., Mozeiko, J., Krueger, F., & Grafman, J. (2012). Discourse production following injury to the dorsolateral prefrontal cortex. *Neuropsychologia*, *50*(14), 3564-3572.

Council for Clinical Certification in Audiology and Speech-Language Pathology of the American Speech-Language-Hearing Association. (2013). *2014 Standards for the Certificate of Clinical Competence in Speech-Language Pathology.* Retrieved March 20, 2018 from http://www.asha.org/Certification/2014-Speech-Language-Pathology-Certification-Standards/.

Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implementation science*, *4*(1), 50.

Doyle, P. J., Goda, A. J., & Spencer, K. A. (1995). The communicative informativeness and efficiency of connected discourse by adults with aphasia under structured and conversational sampling conditions. *American Journal of Speech-Language Pathology*, *4*(4), 130-134.

Dreyfus, S. E., & Dreyfus, H. L. (1980). A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition (A155480).

Edmonds, L. A. (2013). Correlates and Cross-Linguistic Comparisons of Informativeness and Efficiency on Nicholas and Brookshire Discourse Stimuli in Spanish/English Bilingual Adults. *Journal of Speech, Language, and Hearing Research*, *56*(4), 1298-1313.

Fink, R. B., Bartlett, M. R., Lowery, J. S., Linebarger, M. C., & Schwartz, M. F. (2008). Aphasic speech with and without SentenceShaper®: Two methods for assessing informativeness. *Aphasiology*, *22*(7-8), 679-690.

Focus Forward. (n.d.). Focus Forward: Qualitative & Quantitative Research, Coding, Transcription. Retrieved April 23, 2018, from http://www.focusfwd.com/

Gaur, Y., Lasecki, W. S., Metze, F., & Bigham, J. P. (2016, April). The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th Web for All Conference* (p. 23). ACM.

Griffiths, G. G. (2013). *Evaluation of a reading comprehension strategy package to improve reading comprehension of adult college students with acquired brain injuries* (Doctoral dissertation, University of Oregon).

Griffiths, G. G., Sohlberg, M. M., Kirk, C., Fickas, S., & Biancarosa, G. (2016). Evaluation of use of reading comprehension strategies to improve reading comprehension of adult college students with acquired brain injury. *Neuropsychological rehabilitation*, *26*(2), 161-190.

Hannon, B. (2012). Understanding the relative contributions of lower-level word processes, higher-level processes, and working memory to reading comprehension performance in proficient adult readers. *Reading Research Quarterly*, *47*(2), 125-152.

Houweling, T., Bolton, J., & Newell, D. (2014). Comparison of two methods of collecting healthcare usage data in chiropractic clinics: patient-report versus documentation in patient files. *Chiropractic & manual therapies*, *22*(1), 32.

Jacobs, B. J. (2001). Social validity of changes in informativeness and efficiency of aphasic discourse following linguistic specific treatment (LST). *Brain and Language*, *78*(1), 115-127.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, *12*(3), 281-300.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163.

Kucheria, P., Sohlberg, M., Yoon, H., Fickas, S., & Prideaux, J. (in press). Read, Understand, Learn, & Excel (RULE): Development and Feasibility of a Reading Comprehension Measure for Postsecondary Learners. *American Journal of Speech-Language Pathology*.

LaForte, E. M., McGrew, K. S., & Schrank, F. A. (2014). WJ IV Technical Abstract (Woodcock-Johnson IV Assessment Service Bulletin No. 2). *Rolling Meadows, IL: Riverside*.

Leach, R. A., Parker, P. L., & Veal, P. S. (2003). PulStar differential compliance spinal instrument: a randomized interexaminer and intraexaminer reliability study. *Journal of Manipulative & Physiological Therapeutics*, *26*(8), 493-501.

Leslie, P., McNeil, M., Coyle, J., & Messick, C. (2011). Clinical Doctorate in Speech-Language Pathology. *The ASHA Leader*, 16(9), 14-17. doi: 10.1044/leader.FTR2.16092011.14.

Lewis-kraus, G. (2016, December 14). The Great A.I. Awakening. Retrieved from https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html

Linebarger 1, M. C., McCall, D., & Berndt, R. S. (2004). The role of processing support in the remediation of aphasic language production disorders. *Cognitive Neuropsychology*, *21*(2-4), 267-282.

Long, S. H. (2001). About time: A comparison of computerized and manual procedures for grammatical and phonological analysis. *Clinical Linguistics & Phonetics*, *15*(5), 399-426.

Macaruso, P., & Shankweiler, D. (2010). Expanding the simple view of reading in accounting for reading skills in community college students. *Reading Psychology*, *31*(5), 454-471.

Marshall, R. S., Laures-Gore, J., DuBay, M., Williams, T., & Bryant, D. (2015). Unilateral Forced Nostril Breathing and Aphasia—Exploring Unilateral Forced Nostril Breathing as an Adjunct to Aphasia Treatment: A Case Series. *The Journal of Alternative and Complementary Medicine*, *21*(2), 91-99.

Martens, H., Dekens, T., Van Nuffelen, G., Latacz, L., Verhelst, W., & De Bodt, M. (2015). Automated Speech Rate Measurement in Dysarthria. *Journal of Speech, Language, and Hearing Research*, *58*(3), 698-712.

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*(13-14), 1723-1730.

McNeil, M. R., Doyle, P. J., Fossett, T. R., Park, G. H., & Goda, A. J. (2001). Reliability and concurrent validity of the information unit scoring metric for the story retelling procedure. *Aphasiology*, *15*(10-11), 991-1006.

Mujumdar, M. V., & Kubichek, R. F. (2010). A speech rate estimator using hidden markov models-biomed 2010. *Biomedical sciences instrumentation*, *46*, 392-397.

Nicholas, L. E., & Brookshire, R. H. (1993a). 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.

Nicholas, L. E., & Brookshire, R. H. (1993a). 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.

Nicholas, L. E., & Brookshire, R. H. (1993b). A system for scoring main concepts in the discourse of non-brain-damaged and aphasic speakers.

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*(3), 636-648.

Olswang, L. B., & Prelock, P. A. (2015). Bridging the gap between research and practice: Implementation science. *Journal of Speech, Language, and Hearing Research*, *58*(6), S1818-S1826.

Owens, E. F., Hart, J. F., Donofrio, J. J., Haralambous, J., & Mierzejewski, E. (2004). Paraspinal skin temperature patterns: an interexaminer and intraexaminer reliability study. *Journal of Manipulative & Physiological Therapeutics*, *27*(3), 155-159.

Peach, R. K., & Reuter, K. A. (2010). A discourse-based approach to semantic feature analysis for the treatment of aphasic word retrieval failures. *Aphasiology*, *24*(9), 971-990.

Rapp, D. N., Broek, P. V. D., McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific studies of reading*, *11*(4), 289-312.

Reed, D. K., & Vaughn, S. (2012). Retell as an indicator of reading comprehension. *Scientific Studies of Reading*, *16*(3), 187-217.

Ross, K. B. (1999). Comparison of impairment and disability measures for assessing severity of, and improvement in, aphasia. *Aphasiology*, *13*(2), 113-124.

Savage, M. C., & Donovan, N. J. (2017). Comparing linguistic complexity and efficiency in conversations from stimulation and conversation therapy in aphasia. *International journal of language & communication disorders*, *52*(1), 21-29.

Sohlberg, M. M., Fickas, S., & Griffiths, G. G. (2011). Reading comprehension strategies delivered via tablet for individuals with acquired brain injury. *Session presentation at American Speech Language and Hearing Association Annual Conference*. San Diego, CA: November, 2011.

Sohlberg, M. M., Griffiths, G. G., & Fickas, S. (2014). An evaluation of reading comprehension of expository text in adults with traumatic brain injury. *American journal of speech-language pathology*, *23*(2), 160-175.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of graduate medical education*, *4*(3), 279-282.

Sullivan, M. P., Griffiths, G. G., & Sohlberg, M. M. (2014). Effect of posttraumatic stress on study time in a task measuring four component processes underlying text-level reading. *Journal of Speech, Language, and Hearing Research*, *57*(5), 1731-1739.

Swanson, N. (2018). The Right Time for Billing Codes. *The ASHA Leader*, 23(3), 30-32. doi: 10.1044/leader.BML.23032018.30.

Wambaugh, J. L., Nessler, C., & Wright, S. (2013). Modified response elaboration training: Application to procedural discourse and personal recounts. *American Journal of Speech-Language Pathology*, *22*(2), S409-S425.

Wambaugh, J. L., Wright, S., & Nessler, C. (2012). Modified Response Elaboration Training: A systematic extension with replications. *Aphasiology*, *26*(12), 1407-1439.

Wiederholt, J. L., Bryant, B. R. (2012b). *Gray oral reading test: Examiner's record booklet; form-A (5th ed.).* Austin, TX: Pro-Ed.