

**COMPREHENSIVE MANAGEMENT AND ANALYSIS OF COMPLEX AND
LARGE RESEARCH DATASETS: AN APPLICATION IN COMMUNICATION
SCIENCE AND DISORDERS**

by

Mohammed D. Aldhoayan

Bachelor of Science, King Saud University, 2011

Master of Science, University of Pittsburgh, 2014

Submitted to the Graduate Faculty of
School of Health and Rehabilitation Science in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH
SCHOOL OF HEALTH AND REHABILITATION SCIENCES

This dissertation was presented

by

Mohammed D. Aldhoayan

It was defended on

March 28, 2018

and approved by

Malcolm R. McNeil, PhD, Distinguished Service Professor, Emeritus, Department of
Communication Science and Disorders

Valerie Watzlaf, PhD, RHIA, FAHIMA, Associate Professor, Health Information
Management

Qi Mi, PhD, Assistant Professor, Department of Sports Medicine & Nutrition

Gregory Marchetti, PT, MS, PhD, Associate Professor, Rangos School of Health Sciences,
Duquesne University

Dissertation Advisor: Leming Zhou, PhD, DSc, Assistant Professor, Health Information
Management

Copyright © by Mohammed D. Aldhoayan

2018

**COMPREHENSIVE MANAGEMENT AND ANALYSIS OF COMPLEX AND
LARGE RESEARCH DATASETS: AN APPLICATION IN COMMUNICATION
SCIENCE AND DISORDERS**

Mohammed D. Aldhoayan, Ph.D.

University of Pittsburgh, 2018

Today, healthcare is awash in data. With the increasing number of data resources and advancements in information technology, large and complex research datasets are at hand to be used and converted into knowledge, which is one critical driver in our journey toward effective, efficient, and safe healthcare. However, when these datasets are mismanaged or corrupted, they could produce low-quality and even misleading results. Thus, it is critical to maintain high quality standards of data management and promote for the concepts of data lifecycle and data curation due to their direct impact on the quality of research datasets.

This dissertation research is aimed at demonstrating a data management and quality assurance process through the implementation of a full research dataset lifecycle, building a centralized and secured database with a user-friendly interface and assess its usability, and testing a theory-based model of sentence comprehension using structural equation modeling for four types of sentences.

Results of this study have shown that automating the process of extracting, verifying, transforming, and storing large research datasets from their source files to a structured and analysis-friendly database increases the quality of the data, reduces the burden and time waste of researchers, and facilitates communications between the research team members. The conducted usability study has shown that users spend less time and make fewer errors when retrieving datasets using properly designed interfaces than using spreadsheets.

Results from the sentence comprehension analysis have shown that language processing, short-term memory, and working memory can be measured separately. The structural equation modeling analyses have also shown the importance of working memory and how it significantly predicts sentence comprehension of object cleft and garden path sentences.

The data processing techniques used in this dissertation have laid the cornerstone of a generic data processing and management system that gives clinical researchers, especially the ones with limited resources, the ability to manage and process their datasets by enabling them to define and execute different extraction rules. The results of the dissertation also provided clarification on the nature of the interactions between critical cognitive systems and their impact in sentence comprehension deficits.

TABLE OF CONTENTS

PREFACE.....	XV
1.0 INTRODUCTION.....	1
1.1 STATEMENT OF THE PROBLEM.....	1
1.2 SPECIFIC AIMS AND RESEARCH QUESTIONS.....	6
1.2.1 Research questions.....	7
1.2.2 Hypotheses.....	7
1.3 SIGNIFICANCE OF THE STUDY.....	8
2.0 BACKGROUND.....	11
2.1 DATA PROCESSING AND MANAGEMENT.....	11
2.1.1 Usability Testing.....	22
2.1.1.1 Models.....	23
2.1.1.2 Questionnaire.....	25
2.1.1.3 Remedy Usability Issues.....	28
2.2 SENTENCE COMPREHENSION.....	31
2.2.1 Working Memory in Language.....	32
2.2.2 Short-Term Memory in Sentence Comprehension.....	36
2.2.2.1 Phonological STM.....	38
2.2.2.2 Semantic STM.....	40

2.2.2.3	Syntactic STM	42
2.2.3	Attention in Sentence Comprehension.....	43
2.2.4	Measurements	45
2.2.4.1	Sentence Comprehension	45
2.2.4.2	Cognitive Systems and Functions	47
2.3	STATISTICAL ANALYSIS	52
2.3.1	Missing Data	55
2.3.2	Factor Scores	61
3.0	METHODS	64
3.1	RESEARCH DESIGN.....	64
3.2	PROCEDURES.....	67
3.2.1	Data Integration	67
3.2.2	Data De-Identification	67
3.2.3	Data Extraction	68
3.2.4	Data transferring	72
3.2.5	Procedures Evaluation.....	73
3.3	DATA DESCRIPTION AND MANAGEMENT	76
3.3.1	Measurements	76
3.3.1.1	Sentence Comprehension	76
3.3.1.2	Cognitive Systems and Functions	78
3.3.2	Database.....	83
3.3.2.1	Design Decisions	83
3.3.2.2	Database Design	86

3.3.2.3	Database Evaluation	89
3.3.3	User-Interface.....	90
3.3.3.1	Requirements Collection	91
3.3.3.2	Interface Design.....	92
3.3.3.3	Interface Implementation.....	93
3.3.3.4	Interface Testing	94
3.3.3.5	Usability Study	94
3.3.3.6	Usability Study Participants	98
3.3.3.7	Usability Study Tasks	98
3.3.3.8	Usability Study Data Processing.....	103
3.3.3.9	Usability Study Data Analysis.....	104
3.3.3.10	Usability Study Missing Data.....	104
3.3.3.11	Usability Study Data Screening	105
3.3.3.12	Usability Study Time and Error Difference Analysis.....	106
3.3.3.13	PSSUQ and ASQ Analysis.....	107
3.4	DATA ANALYSIS.....	108
3.4.1	Sample size.....	108
3.4.2	Data Screening	112
3.4.3	Descriptive Data	113
3.4.4	Data Modeling	114
3.4.4.1	Measurement model.....	116
3.4.4.2	Structural model	122
3.4.4.3	Multiple Linear Regression.....	126

4.0	RESULTS	128
4.1	PRELIMINARY RESULTS	128
4.2	USABILITY STUDY.....	131
4.2.1	Participants.....	131
4.2.2	Efficiency	132
4.2.3	Effectiveness	143
4.2.4	Satisfaction.....	144
4.3	DATA ANALYSIS (SENTENCE COMPREHENSION)	150
4.3.1	Participants.....	150
4.3.2	Data Descriptions	151
4.3.3	Missing data.....	154
4.3.4	EFA.....	154
4.3.5	CFA	161
4.3.6	SEM.....	167
5.0	DISCUSSION	176
5.1	DATA PROCESSING	176
5.2	DATA ARCHIVING	180
5.3	DATA DISTRIBUTION	181
5.3.1	Usability Study	183
5.4	DATA ANALYSIS (SENTENCE COMPREHENSION)	191
5.5	LIMITATIONS AND FUTURE WORK.....	198
6.0	CONCLUSION.....	202
	APPENDIX A	204

APPENDIX B	231
BIBLIOGRAPHY	236

LIST OF TABLES

Table 1. <i>Randomly generated data to demonstrate different missing data patterns.</i>	57
Table 2. <i>The scores of each item by the two members (1st and 2nd) of the usability testing.</i>	129
Table 3. <i>The mean and SD of performance of the two members (1st and 2nd) on the usability testing tasks.</i>	129
Table 4. <i>Participant characteristics (N=24).</i>	132
Table 5. <i>Participants' performance, in seconds, on each one of the ten tasks on both methods.</i>	134
Table 6. <i>Participants' performance on tasks using iRDMS Pairwise comparisons using Tukey and Kramer (Nemenyi) test with Tukey-distribution approximation.</i>	135
Table 7. <i>Participants' performance on tasks using Excel Pairwise comparisons using Tukey and Kramer (Nemenyi) test with Tukey-distribution approximation.</i>	136
Table 8. <i>Participants' performance on the first three tasks using Excel and iRDMS.</i>	138
Table 9. <i>Participants' performance, in seconds, on each one of the ten tasks categorized by the order of methods they used on both methods.</i>	142
Table 10. <i>Number of incorrect performances on each of the tasks using Excel and iRDMS.</i>	143
Table 11. <i>PSSUQ overall scores and each sub-scale with their correlations with different measures.</i>	144

Table 12. ASQ scores on each task and overall with their test of difference between Excel and the iRDMS.	146
Table 13. Significant correlations between ASQ scores with participants' performance on Excel and iRDMS, task order and Excel skills and experience.	147
Table 14. Correlations between ASQ questions on Excel and iRDMS for each task.	148
Table 15. Participants characteristics (N=100).	151
Table 16. Task Descriptive Statistics.	153
Table 17. Tasks Pearson Correlation matrix.	156
Table 18. Item loadings from the 5 and 2-factor EFA solutions (item loading $< 0.32 $ were suppressed).	160
Table 19. Scaled Chi Square Difference Test.	162
Table 20. Scaled Chi Square Difference Test.	163
Table 21. Scaled Chi Square Difference Test.	164
Table 22. Factor loadings and communalities of the 3-factor CFA model.	167
Table 23. Factor Correlations of the 3-factor CFA model.	167
Table 24. Test of significance of the regression coefficients for the Object Cleft sentences model.	168
Table 25. Test of significance of the regression coefficients for the Garden Path sentences model.	170
Table 26. Test of significance of the regression coefficients for the Compound sentences model.	172
Table 27. Test of significance of the regression coefficients for the Lexical Ambiguity sentences model.	174

LIST OF FIGURES

Figure 1. UK Data Archive model.....	13
Figure 2. DDI data lifecycle model.....	14
Figure 3. Engle, 2004 WM model.	35
Figure 4. Baddeley, 2003 WM model.....	38
Figure 5. Tests of Phonological Processing.....	50
Figure 6. Engle, Tuholski et al. 1999 SEM model.....	55
Figure 7. Database ER diagram.	88
Figure 8. iRDMS Data Flow Diagram (DFD).	93
Figure 9. A screenshot of the iRDMS.....	94
Figure 10. Cattell’s Scree test.	118
Figure 11. Scree plot by PA.....	119
Figure 12. The hypothesized measurement models.....	122
Figure 13. hypothesized structural models.	125
Figure 14. Parallel analysis.....	158
Figure 15. The 3-factor CFA model.	166
Figure 16. The Object Cleft sentences model.....	169
Figure 17. The Garden Path sentences model.....	171

Figure 18. The Compound sentences model.....	173
Figure 19. The Lexical Ambiguity sentences model.....	175
Figure 20. Engle, 2004 WM model.	195

PREFACE

This basis for this research originally stemmed from my passion for developing better methods of data management and processing. As the world moves further into the digital age, generating vast amounts of data and born-digital content, there will be a greater need to access legacy materials created with outdated technology. How will we access this content? It is my passion to not only find out but to develop tools to break down barriers to accessibility for future generations.

In truth, I could not have achieved my current level of success without a strong support group. First of all, my parents, my wife Ibtehaj, my two beautiful kids Ghena and Hisham and my entire family and friends who supported me with love and understanding. And secondly, my advisor Dr. Leming Zhou, my committee members, each of whom has provided patient advice and guidance throughout the research process. Thirdly, my colleagues who worked on the data collection, especially Dr. Malcolm McNeil and Dr. Wiltrud Fassbinder. Thank you all for your unwavering support.

This work was supported in part by Merit Review Award RX-001145-01A1 to Malcolm R. McNeil from the United States (U.S.) Department of Veterans Affairs Rehabilitation Research and Development Service. In addition, it was supported with resources and the use of facilities at the VA Pittsburgh Healthcare Center, VA Northern California (Martinez), and the University of Washington. The contents of this dissertation do not represent the views of the U.S. Department

of Veterans Affairs or the United States Government. Parts of this study were approved by the VA Central IRB Review Board, and the Institutional Review Boards of the University of Pittsburgh, the University of Washington, and Temple University. All participants provided verbal/signed and written informed consent prior to inclusion.

This work was supported in part by the School of Health and Rehabilitation Sciences Research Development Fund. Special thanks to James R. Lewis who developed and made available the After-Scenario Questionnaire and the Post-Study System Usability Questionnaire.

1.0 INTRODUCTION

1.1 STATEMENT OF THE PROBLEM

Health and rehabilitation studies seek to answer research questions by collecting data that can be used to test hypotheses. The quality and robustness of these tests are directly related to the quality of the collected data. Although the quality of the collected data is directly affected by the mechanisms of data collection, the used data management methods and the structure of the data have a significant impact on data quality (Krishnankutty, Bellary, Kumar, & Moodahadu, 2012).

With the advancement in technology and the evolution of data collection methodologies, research studies are now able to get access to more detailed data in a short period. Furthermore, many research studies now use mobile devices, sensors, and other advanced technologies to collect data from their subjects. This phenomenon is also observed in healthcare research where investigators now have access to “big data” databases that contain a massive amount of health data collected from a large number of subjects (Labrinidis & Jagadish, 2012).

This type of data is advantageous in answering research questions that require the analysis of comprehensive datasets that cover many aspects of the problem of interest. Therefore, these datasets are usually complex in their structure, because each aspect they cover has its own data characteristics and structure. Furthermore, these datasets are typically large in their sample size because they are collected from a large number of subjects or representing a combination of

many small datasets, which makes them more complex and challenging to be handled. Moreover, this challenge of managing the collected data increases when data collection is conducted in multiple sites because issues regarding inconsistency and standardization of the data may occur (Gerritsen, Sartorius, vd Veen, & Meester, 1993). This inconsistency challenges the data integration process where data must be integrated into one dataset before being analyzed. In addition, with large datasets that were collected at multiple sites, data exchange between the stakeholders of the research project is challenging due to the size of the data and regulations of the research project sponsors (Yin, 2015).

Although health and rehabilitation projects usually conduct some data management procedures, they do not implement the full research data lifecycle. These projects usually conduct data collection to gather the desired data, data processing to prepare the data for analysis, and data analysis to answer their research questions. However, before concluding the research project, there are more data management phases that must be implemented to ensure the data security, availability, and reusability. Unfortunately, it is common for researchers to ignore or drop these phases from the data management process in research studies (Surkis & Read, 2015). The importance of these phases has increased due to the fact that research nowadays is more data driven than before due to the advancement in data collection and analysis technologies. The impact of not implementing these phases is overwhelming as collected data might get lost or become undiscoverable, which kills any opportunity for data mining, secondary data analysis, or results reproducing. Furthermore, a gap between clinical data collection and the statistical analysis does exist (Stangl, 2005). The presence of this gap poses threats to the impact of the findings of these projects due to many reasons. This gap could lead to uninformative results-interpretation because of the miscommunication of the theory behind the analysis

between the two processes (Dawson & Trapp, 2001). Furthermore, this gap could cause a loss or reduction in the quality of data during the data transformation from their source files to the analysis procedures. This gap exists because of the poor management of data and the absence of a systematic data management process that ensures data quality. Poor data management means the absence of a plan or vision that looks at the possible uses of the data and chooses the proper data structure and storage (Krishnankutty et al., 2012).

Large research datasets are widely used in healthcare fields, such as public health, evidence-based medicine, and genomic analytics. For example, in the Framingham Heart Study, data is being collected since 1948 from 5,209 adult subjects from three different generations. This massive data contains more than 60 different exams that cover many aspects of heart, brain, bone, and sleep diseases. The National Heart Institute are storing this massive dataset in SAS databases. They are providing interested researchers with coding manuals, annotated forms and protocols that explain the content and variables of this dataset (Dawber, Meadors, & Moore Jr, 1951). Furthermore, this study is a perfect example that shows how research data management improves with time from using paper and keypunch forms to using programmed software and databases (W. Wilson, 1990). Therefore, without this development of proper data management and structure that satisfy the requirements of the data's nature, this dataset would have been a complete mess and using it in healthcare research would have been impossible.

In psycholinguistics research, investigators face many challenges that affect their ability to answer complex research questions. One challenge is having limited access to impaired populations due to the high cost of administrating the desired tasks and the needed accommodations (Dollaghan, 2004). Another challenge is in the complexity of the tasks that they use to assess the performance of their participants due to the convoluted nature of the human

brain (E. Chen, Gibson, & Wolf, 2005). Therefore, when conducting a study that involves larger sample size and multiple tasks that comprehensively measure the cognitive components of interest, issues regarding the extraction, processing, cleaning, management, and analysis of the collected data become overwhelming. Thus, researchers in this case usually face a trade-off between the quality of the collected and analyzed data and the generalizability power of their results (Germine et al., 2012).

Unlike other health research fields, psycholinguistics suffers a big-data resources scarcity, where finding open-source and ready for analysis data is challenging. Reasons for that, besides the lack of access to the impaired population and the expensive requirements of tests administration, could be that the collected data are usually structured in a format that satisfies the current purpose of the study (MacWhinney & Fromm, 2016). Also, these data are often kept private and not available to other researchers or health informaticians who seek to test research questions and hypotheses through primary or secondary analysis of the collected data (Mirman et al., 2010). These challenges along with other issues have influenced the psycholinguistics studies designs to be limited regarding what questions they can answer, which caused this field to be full of controversial theories, especially the ones related to the deficits in language processing and comprehension.

Sentence comprehension deficits in persons with aphasia (PWA) have been the target of psycholinguistics research for a long time. However, there is still a broad disagreement on the nature of sentence comprehension deficits and their hypothesized accounts. The main issue in this conflict is that each one of these hypothesized theories has its legitimate evidence that supports its claims, and no one theory can refute all the others. Nevertheless, recent theories have started to relate sentence comprehension deficits to impairments in cognitive functions and

working memory (WM) (Engle, Tuholski, Laughlin, & Conway, 1999; R. C. Martin, 1987; McNeil, Odell, & Tseng, 1991; Waters, Caplan, & Hildebrandt, 1991). Studies that investigated the role of WM in sentence comprehension have used the reduction in WM capacity as an indicator of impaired WM. Nonetheless, some studies indicated that WM capacity cannot explain sentence comprehension deficits and that WM must be divided into more detailed components (Caplan & Waters, 1999). Therefore, the investigation of WM components and their relationships to complex cognitive functions, such as reasoning, problem-solving, and language processing, have gained an increasing interest in the psycholinguistics research.

Similar to many studies in psychology, studies that investigate the role of WM and its components in sentence comprehension face crucial challenges due to the complexity of the brain structure. Human brains are similar to black boxes that take input and produce processed output but cannot be opened or investigated from inside. Therefore, in psycholinguistics research, subjects of interest are provided with tasks that try to understand their cognitive structure and functions based on their performance. However, it is extremely difficult to isolate the role of one particular cognitive component from the others, because human brains are structured as complex networks of neurons that perform cognitive tasks that overlap with each other (Clark, 2011). Thus, studies that investigate the role of WM and its components in sentence comprehension tend to measure cognitive components with tasks that are hypothesized to tap or load on a particular cognitive function. Although these types of measurements try to isolate the effect of other cognitive functions, it is almost impossible for one cognitive function to stand alone without any reliance on other functions. Therefore, investigating the role of one cognitive function or component in sentence comprehension by measuring it alone is biased and cannot be reliably used to build causal relationships. So, if one study found a correlation or relation

between short-term memory (STM) and sentence comprehension, it is not necessary that STM is affecting sentence comprehension since both of them can be experiencing the effect of unmeasured covariates. Furthermore, some studies have used treatment effect as an indicator that the improvement in one cognitive component leads to the improvement in sentence comprehension (B. A. Wilson & Baddeley, 1993). Although this procedure can be valid in other highly controlled environments, it is not reliable to be used here since sentence comprehension is affected by many things that were not all measured and observed, which can be the reason behind the improvement in sentence comprehension rather than the cognitive component of interest.

1.2 SPECIFIC AIMS AND RESEARCH QUESTIONS

The purpose of this study is to: 1) demonstrate a data management and quality assurance process through the implementation of a full research data lifecycle by using automated extracting, processing, transforming, and storing of large research datasets from their source files to a structured and analysis-friendly database; 2) build a centralized and secured database with a user-friendly interface that enables users of querying the data based on their authorized access. Also, to assess the usability of this interface by conducting a usability testing; 3) test a theory-based model of sentence comprehension using structural equation modeling for four types of sentences in relation to STM and conflict resolution (CR) while controlling for the effect of other language processing (LP) components, such as long-term memory.

1.2.1 Research questions

Q1: Is a web-based database with a user-interface more effective and efficient to use for data retrieval than flat files, such as Excel?

Q2: Are CR, STM, and LP separable and domain-specific components in PWA?

Q3: For PWA, do STM and CR predict comprehension success beyond the contribution of LP, on sentence structures that have been hypothesized to rely on these functions?

1.2.2 Hypotheses

In this dissertation, it was hypothesized that data retrieval and analysis will be more effective and efficient for researchers using the proposed database and its user-interface than using the original files. Since the terms effective and efficient are very broad and have many implications, it is referred to the International Organization for Standardization (ISO) definitions of usability with three aspects of usability: effectiveness, efficiency, and satisfaction. An alternative hypothesis could be that researchers feel more comfortable with the original files since they have been exposed to them for a longer time than to the proposed interface. However, this information was collected in the usability testing and this alternative hypothesis was analyzed.

Another hypothesis of this proposal is that deficits in STM and CR cause decrements in comprehension of sentences that require higher levels of STM or CR involvement. It was hypothesized that deficits in compound sentences (CS), such as “Touch the little red square and the big blue circle on this trial,” will be predicted by phonological and semantic STM (R. C. Martin, 1987; Miyake, Just, & Carpenter, 1994). Also, deficits in object cleft (OC) sentences,

such as “It was the blue circle that the green square touched on this one,” will be predicted by syntactic STM and CR (E. Chen et al., 2005; R. C. Martin, 1990; Vuong & Martin, 2011). Furthermore, deficits in garden path (GP) sentences, such as “The blue circle touched by the green square is above the green circle on this one,” will be predicted by syntactic STM and CR (E. Chen et al., 2005; R. C. Martin, 1990; Vuong & Martin, 2011). In addition, deficits in lexical ambiguity (LA) sentences, such as “He drank the port quickly,” will be predicted by CR (E. Chen et al., 2005; R. C. Martin & He, 2004). Finally, it was hypothesized that LP, STM, and CR are separate and separable domain-specific components.

1.3 SIGNIFICANCE OF THE STUDY

Data management and analysis are critical phases of any health and rehabilitation project as they have a direct impact on the quality and correctness of the findings (Krishnankutty et al., 2012). With the availability of advanced data collection methodologies, researchers now seek to answer questions that require more detailed and comprehensive datasets. However, because each project has its own data structure and analysis, managing data using general data structure files (e.g., Excel, Google Sheets) is not always feasible. In addition, although there are many Clinical Data Management (CDM) organizations and Clinical Data Management Systems (CDMSs) developers, issues regarding the high cost, lack of flexibility, and low usability prevent many health-related research projects from adopting them (Kuchinke et al., 2010).

In this dissertation, a comprehensive data management project that transforms data from complex and unstructured data files to an analysis-friendly database was conducted. This work

has demonstrated the implementation of a full research data lifecycle with emphasis on the data curation phases. This process has demonstrated how to overcome some issues that research projects usually face when using complex datasets. It illustrated the use of data mapping procedures in data extraction and data quality assurance. Furthermore, it showed how issues regarding privacy and confidentiality regulations compliance can be solved by implementing data de-identification procedures that do the task with minimum data loss and deletion. In addition, in this dissertation, the concept of data products, where collected data are made accessible and reusable by all the research team members anywhere and anytime, was demonstrated, which solves issues regarding data sharing and exchange. Furthermore, this dissertation showed how to conduct an integrated data management and analysis project that bridges the gap between the pure health sciences, statistics, computer science, and information systems.

Health informatics is generally viewed as a field that is related to healthcare data that are collected from electronic health records and patients' charts (Hovenga, 2010). This study demonstrated an important role that health informatics takes in health and rehabilitation research. It signified the impact that health informatics and health information management have on the quality of the findings of health and rehabilitation research. This effect is achievable by providing supporting tools and techniques that can be used to facilitate all the research phases starting from the study design and data collection and ending with data analysis and results interpretation. In addition, this project shows how health informatics could be used to answer research questions and test hypotheses in health and rehabilitation fields.

The data analysis that was conducted and the statistical models that were built in this project has helped to understand the nature of sentence comprehension in PWA. Increasing the

understanding of the deficits in sentence comprehension in PWA will open new opportunities to develop more accurate measurements of these impairments. These measurements will have a positive impact on increasing the accuracy of the diagnosis and lowering their potential cost. Furthermore, the understanding of sentence comprehension deficits will help clinicians to deliver more efficient treatments. These treatments will be more customized to each patient according to the accounts or causes of their impairments. Once these underlying cognitive causes of sentence comprehension deficits are understood, clinicians will be able to design a better treatment that only targets the impaired cognitive processes. This treatment design will decrease the consumption of time and cost because it only provides necessary interventions. Finally, this analysis will open opportunities for future research that further investigates sentence comprehension deficits in PWA and other populations.

2.0 BACKGROUND

2.1 DATA PROCESSING AND MANAGEMENT

Any project that involves data collection needs proper data management to ensure high data quality levels. The importance of data management increases in research projects where data is being used to understand health-related topics or in investigations that support treatment decisions (Bajpai, 2015). In recent years, an emerging research data management technique suggests that the outcome of the data analysis process should be a product that can be used to serve purposes beyond the one that is related to the primary research project. This universality nature cannot be achieved without having a plan for the lifecycle of the product starting from the data collection throughout the data processing and ending with data archiving (Hey, Tansley, & Tolle, 2009). Therefore, in recent years, the concept of data curation or data lifecycle has emerged to extend the data management process to cover the concept of having the collected dataset as a product. Data lifecycle is the process of managing data from its point of creation or collection to its final destination. Lord and Macdonald (2003) have defined data curation as “The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets, this may mean continuous enrichment or updating to keep it fit for purpose.”

The concept of data lifecycle is not new, in general, and have been practiced by researchers since the beginning of scientific research. The Data Curation Center (DCC) has developed a data lifecycle model that can be applied to any collected dataset. The model, in general, specifies three levels of data lifecycle management that should be followed by researchers. The first is data description and representation, which is the technical level that includes the processes of creating, collecting, processing, preserving, and analyzing data. The second level, which is the planning level and includes the creation of strategies and policies for the data lifecycle. The third level is the data maintenance, which includes the process of updating the data based on changes in the population and community (Higgins, 2008). Although this model has some specific data management recommendations on each level, it is not specific enough to be used for research data management.

Another data lifecycle model that was developed to be used on research datasets is the UK Data Archive model (Ball, 2012). This model consists of six phases. The first is data creation, which involves research design, the creation of plans and strategies, and conducting the data collection. The second is data processing, which involves digitalizing or entering the data, data extraction, data transformation, data validation, and data cleaning. The third stage is data analysis, which involves performing statistical analysis, interpreting the results, and prepare results for publication. The fourth stage is the data preservation, which includes loading the data to the most suitable storage and store the data in secure locations. The fifth stage is giving access to data, which involves sharing data, promoting data, and implementing access control procedures. The sixth stage is to provide researchers with opportunities to re-use the data for secondary analysis or to reproduce the originated results (Ball, 2012). Although this model covers all the aspect of research data lifecycle, the order of these stages implies that the data

processing stage only focuses on the primary investigation and that there should not be more than one data analysis being conducted in parallel (Figure 1).

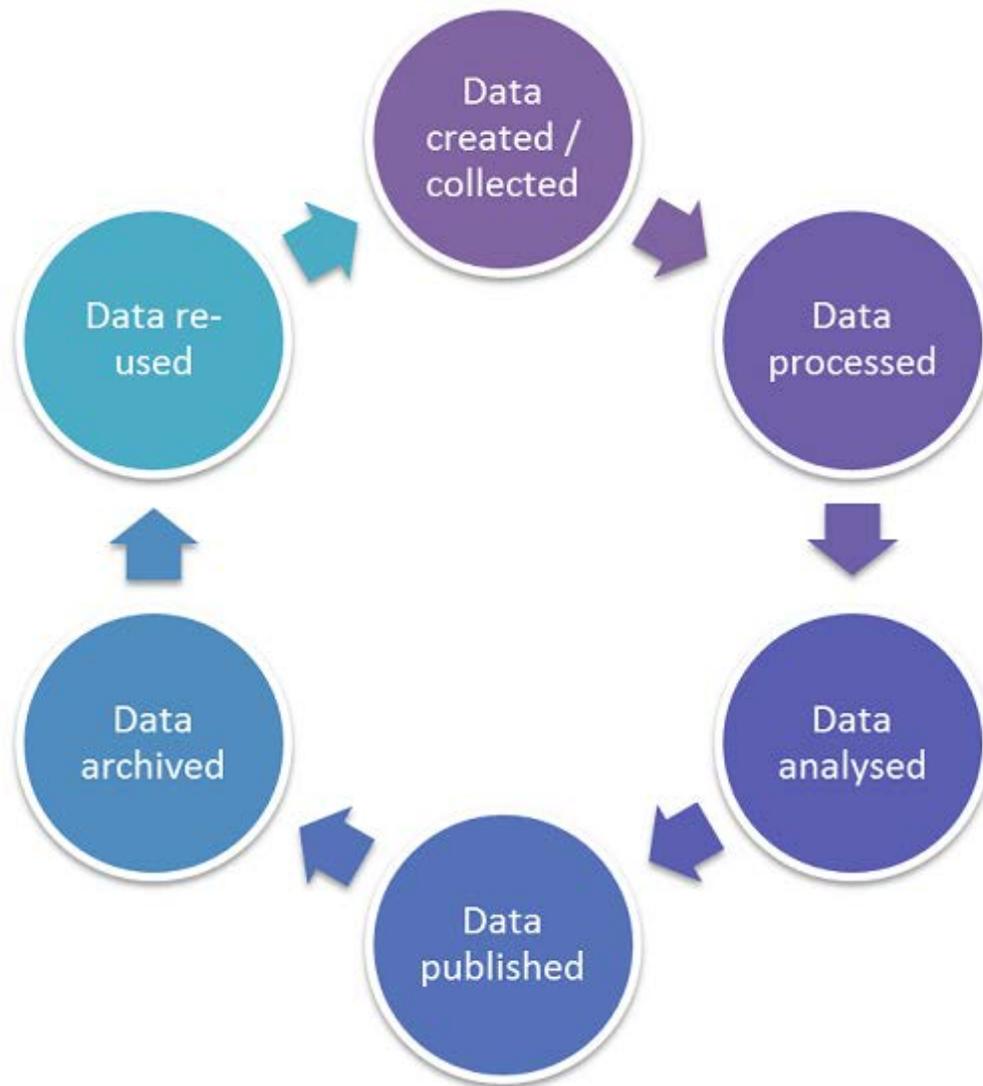


Figure 1. UK Data Archive model.

Downloaded from Lancaster University. Adapted from the UK Data Archive.

The Data Documentation Initiative (DDI) has introduced the third version of their data management and documentation model and called it “Data Lifecycle.” In this model, there are eight phases of data lifecycle (Figure 2). The first phase is the study concept, which includes the

planning for the data collection procedures and instruments and specifying the lifetime of the data to be collected. The second and third phases are the data collection and processing, which are similar to the UK archive model stages. The fourth phase is the data archiving phase, which includes transforming data to be stored in a format that is generic and not specific to one data analysis. The fifth phase is the data distribution phase, which is also similar to the giving access phase in the UK archive model. However, the DDI model suggests starting the data discovery and analysis after the distribution phase, which enables researchers to work in parallel on analyzing, interpreting and investigating the data. The last phase in the DDI model is repurposing, which involves conducting further data processing to satisfy the emergence of new data uses in case the initial data processing was not suitable enough. Furthermore, after the data repurposing or data processing the phases of data archiving and distribution should be updated as well (Vardigan, Heus, & Thomas, 2008).

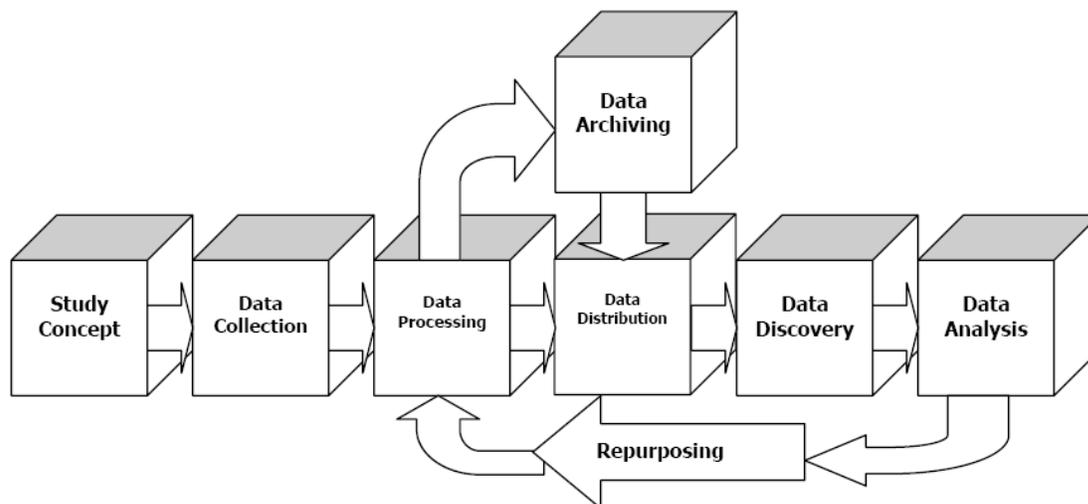


Figure 2. DDI data lifecycle model.
Adapted from (Vardigan et al., 2008).

The data management or data lifecycle task starts by specifying the data management plan that describes the source of all collected data, data extraction methodologies, data transformation from one form to another, and data quality assurance procedures. Therefore, the data manager or management team should be involved since the early study design phases to help the research team to pick the right data collection, handling and storage methods (Kumar & Arasu, 2014). Furthermore, a primary characteristic that contributes to the quality of this product is how discoverable it is. A discoverable data product is a one that can be located, identified, and accessed, through simple tools that are available to a broad audience of researchers (Parsons et al., 2011). Another characteristic of high-quality data products is the usefulness of the product and its suitability to be part of practical and adventurous research across disciplines. Moreover, this product should be secured by procedures that save it from loss, corruption, and any source of risk that threatens its quality throughout its lifetime (Simberloff et al., 2005). A critical choice that must be made in the development of data analysis plans is the structure of the storage of the collected data because the structure of the data is a significant factor to the feasibility of implementing a product with the mentioned characteristics.

Data processing is a task that goes beyond the statistical analysis procedures to include all the procedures that start with the data collection until the conduction of the statistical analysis itself (Miles & Huberman, 1994). Each research project requires a particular type of data analysis phases that fit its design as well as the nature of the collected data. In quantitative studies, the quality of the outcomes depends on the quality of the data that was used in the analysis. Although many factors affect the quality of the data, and although the design and implementation of the data collection methods have the highest quality impact among these factors (Grimes, 2010). However, sometimes it is not possible to collect data in a format or a structure that fits the

purpose of the research project. Therefore, data cleaning is another phase where researchers could detect and correct any abnormal data entities that might pose threats to the data quality (Hernández & Stolfo, 1998). Moreover, the removal of data continues throughout the data reduction phase where unnecessary data is excluded from the final dataset, and only the data that is going to be used in data analysis investigations is kept. The importance of this phase increases in projects where data is collected by multiple teams in multiple sites (Namey, Guest, Thairu, & Johnson, 2008).

CDM is a process of ensuring that the conclusions drawn from the research project are supported by high-quality data, where high-quality data is defined as “data that fit their intended uses in operations, decision making and planning” (Redman, 2008). This includes the transformation of data from unstructured and complex data files to files that are analysis-friendly. Furthermore, it includes the development of CDMS or the product of the collected data, which is a system that stores and manages the data in high-quality standards with acceptable costs (Lu & Su, 2010). This system serves as a data hub for all the collected data in the research project. This system should be able to perform all the tasks that data need to be ready for data analysis and use, such as data integration, data coding, and de-identification. It is also important for this system to be compliant with privacy and confidentiality regulations since it handles sensitive health-related data (Fu, Ding, & Chen, 2010). Finally, it is essential to conduct a User Acceptance Testing (UAT) to ensure the desired outcomes have been met in terms of the design and the functionality of the system (Leung & Wong, 1997).

CDM procedures have been used in clinical research in different formats and implementations because each research project has its unique data characteristics that require a special type of treatment. In particular, these different treatments are observable in the features

that each CDM chooses to implement in their CDMS. For instance, one data validation procedure that CDMSs use is that each data entry should be entered by two operators to ensure data consistency, in the case of manual data entry. Other CDM projects use special comparison algorithms that compare datasets in their final destinations to their source files to make sure that data is consistent between the two locations. Another difference in how CDMSs handle data is in the data acceptance, where some CDMSs allow the system users to enter data into the database while other CDMSs prohibit that as a security precaution and only allow data entry from the database administrator (Greenes, Pappalardo, Marble, & Barnett, 1969). Therefore, CDM can be viewed as a collection of data management tools and procedures from multiple disciplines that are formed and customized to serve the purpose of the research project.

Data structure describes the way of organizing data or arranging information in a storage space. In research, data structure should be selected based on the nature of the collected data and the technical tools that are available to use. The task of choosing the data structure becomes more important in projects that involve multiple teams in multiple sites, or when dealing with data that is distributed in separate files with different formats (Hellerstein, 2008). A commonly used data structures are flat files that are independent of each other. These files could be in various forms, such as sheets, text, etc. The popularity of these files is due to their availability to be accessed and manipulated with the most basic technologies. In addition, these files are relatively easy to distribute due to their small sizes in memory and compatibility with most of the sharing methods. However, when planning to keep the data for as much time as possible to be used in different projects, these files might not be the best choice for data storage. One main issue regarding these types of files is their vulnerability to risk, whether this risk is related to uncontrolled access, the chance of being lost, or any form of data corruption. Another issue

regarding these files is the slow process of information retrieval due to their independent structure and lack of proper indexing. Therefore, structured data types that offer multi-layers of data are the optimal choice that helps researchers to overcome these issues. Even more, with the proper design and implementation, these files could have a positive impact on the quality of the results of research projects as they offer more in-depth analysis and accuracy of collected data (Few, 2004).

Databases are one of the most common data files that are used in computer science and information systems. However, they are less used to store and manage data from research projects because they require some level of technical experience to deal with them. Nevertheless, these technical requirements vanish when building a Graphical User-Interface (GUI) that enables all the research team to access, query, and store data in the database. In research, database management systems (DBMS) should be built as a product that serves as the hub of data that all the collected data are managed through (Güting et al., 2005). Web-based DBMS serve this purpose by providing research team members with a portal that can be reached wherever the internet is available. Furthermore, when dealing with quantitative data where final scores must be calculated using multiple items, this technique increases the standardization of the data that has been collected by different teams by unifying the calculation methods and data structure, which minimizes any confusion or inconsistency in the data analysis. Furthermore, DBMS offer high levels of accessibility to these data due to the availability of a Structured Query Language (SQL) that provides standardized yet flexible statements that enable any form of data retrieval (Hellerstein, 2008). Even more, because web-based DBMS are centralized in one location that is connected to a network with multithreading capabilities, they offer faster and more frequent updates and higher data security. This increase in the data security is due to the ability to

implement an access control feature that allows each user to view limit portions of data depending on their privileges (Thomasian & Ryu, 1983). Lastly, databases and their management systems are perfect to be used in research because of their ability to transform data into any format or structure depending on the requirement of the data analysis procedures.

The design of databases and their management systems is a key player in achieving high levels of usability, accessibility, and practicality (Rumbaugh, Blaha, Premerlani, Eddy, & Lorensen, 1991). The process of designing databases starts by specifying the purpose and the use of the database. In research, this can be done by defining the scope of the collected data, and by identifying the type of research that can be done using the stored data (Teorey, Yang, & Fry, 1986). This step is important in the entire design process as it provides insights on what facts are necessary to be stored and viewed in the database. Therefore, a good practice in database design is to find and organize samples of the data that are going to be stored in the database and understand their structure and relationships. The next step in database design is to use these samples to divide the data into tables and entities with the proper fields and relationships (Batini, Ceri, & Navathe, 1992). However, because data in research is usually collected in an unstructured manner, transforming the structure of data from their source files into the database could pose design issues, such as improper indexing and data redundancy. Therefore, Codd (1970) has introduced the concept of database normalization, where databases are evaluated in several steps or forms to make sure that they are complying with the principles of database design. Normalized databases are the ones that have primary keys for each row on every table, which makes every detail in the database accessible and available. Furthermore, normalized databases do not include any columns or fields that cannot be linked to one primary key. Another characteristic of normalized databases is that they do not contain redundant data, whether it was

in the form of redundant columns, shared information by multiple rows, or redundant data in more than one table. The redundancy here is not the one that occurs when inserting one record more than one time, but it is the one that occurs in the design of the database where two fields are asking for the same information.

Although the concept of implementing DBMS as products of data collection and processing is relatively new, many “big data” databases have been developed and made available on the internet. The size of these databases is usually determined by the size of the population where data are collected from and how accessible that population is. For example, genomic databases could contain millions of records while psycholinguistic databases only contain few hundred records. Although these databases create tremendous research opportunities, many of these databases suffer from data quality and inconsistency. Therefore, research databases can be a double-edged sword where larger sample sizes could threaten the quality of the data (Boyd & Crawford, 2012). Research databases in the field of psycholinguistics are rare and small due to the limited accessibility to the population with psycholinguistic impairments and the cost of collecting data from this population. MacWhinney and Fromm (2016) have developed a standardized database that contains data from 290 PWA and 190 control participants; it is called “AphasiaBank”. Their primary goal is to provide researchers with a relatively large dataset that can be analyzed and explored as “big data”. Although the size of the AphasiaBank is comparatively small to other databases of other conditions or diseases, it is still considered large or big database giving that it contains data from a limited population. AphasiaBank has been already used by many studies to explore research topics, such as discourse, grammar, gesture, and lexicon (MacWhinney & Fromm, 2016).

Mirman et al. (2010) have also developed a web-based searchable database that contains data from over 240 PWA. This data that was obtained by multiple studies covered the performance of PWA on Philadelphia (picture) Naming Test battery and some other tests. The used sample of PWA comprises various subtypes of aphasia and different levels of severity. The primary goal of this project is to make this collected data available to any researcher who would like to test a hypothesis or theory and have no access to such dataset. Therefore, they provided a web-based portal with a user-friendly interface that can be used without heavy technical skills. Furthermore, they provided a description of each item they have in the database, which helps users to find what they are looking for and to make sure that they understand the data they are viewing. Their interface displays the data fields in the database and enables the users to choose the data that they want to view or download by clicking on the desired fields. Users can customize their selection by limiting the results based on many criteria, such as aphasia type, participants from a certain study, and model response codes. Users also have the choice to download the basic demographics of the participants and their clinical information, which provides more opportunities to investigate relationships between these factors and the subjects' performance on the linguistic tests. Nevertheless, with all these demographics and clinical information, data are still de-identified and safe to be shared with the public. Although this portal is available to the public, the data is protected with access control procedures, and the users are required to provide brief descriptions of their interests and uses of the data prior to gaining access (Mirman et al., 2010).

2.1.1 Usability Testing

The term usability is a vague term that can be defined by different ways depending on the context of the system or software. However, there are three general views of usability, product-oriented, user-oriented, and user performance views. The product-oriented view measures the usability of the system by its technical performance and how it complies with the system design standards. For example, heuristic evaluation is a method of evaluating the systems' usability based on pre-defined heuristics. This method, which is going to be discussed later in this section, does not take the end users' opinion or experience into account. It only measures the usability from a product point of view based on the experts' opinions who use the pre-defined heuristics as guidelines. The user-oriented view, on the other hand, does not look at the pre-defined heuristics or the system design standards when measuring the system usability. Although these pre-defined heuristics could be used to guide the design of the system, the user-oriented view measures the system usability by only asking the users about their experience with the system and measures their satisfaction. Furthermore, in this view, users are asked to self-assess the cognitive load or effort that they put into using the system. Therefore, the general idea behind this view is that if the users are satisfied with the system, then it does not matter what design standers or pre-defined heuristics were followed. Finally, the user performance view measures the practicality of the system and how well it serves the purposes that it was implemented for. This view measures the ease of use of the system and its accessibility by asking potential users to perform real tasks and measure their performance.

The importance of usability comes from the fact that each system is developed to serve a specific purpose. Whether this purpose is commercial, educational, research, or organizational,

the system would not serve its purpose unless it is usable. In fact, it has been shown that the systems' adoption, acceptance, or rejection mainly depend on their usability (Kushniruk, 2002). Furthermore, in the literature, most of the usability studies are being conducted to predict the users' adoption, acceptance, or rejection of the system (Peute, Spithoven, & WM, 2008). This importance has influenced the field of human-computer interaction to study how humans and computers interact from psychological, technical, organizational, professional, financial, legal, and political aspects (Alanazi, 2015). As a result, this has pushed the usability testing to adopt different models and techniques, where each technique measures a unique aspect of human-computer interaction.

2.1.1.1 Models

As mentioned in the previous section, the different views of usability and the many aspects of human-computer interaction have influenced the creation of different models of usability testing. These models mainly differ based on their definition of usability and, thus, how it should be measured. For example, ISO defines software qualities as “A set of attributes that bear on the effort needed for use, and on the individual assessment of such use, by a stated or implied set of users” (IEC, 2001). This definition describes the usability as product and user-oriented measure. Eason (1988) defines usability as “the degree to which users are able to use the system with the skills, knowledge, stereotypes and experience they can bring to bear.” This definition concentrates the concept of usability on how users can easily use the system without extensive training. ISO FDIS 9241-210 defines usability as: “The extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” (DIS, 2009). The definition of usability by ISO is a

user and contextual view of usability where it focuses on the goal and purpose of developing the system.

The ISO 9241-11 model of usability is one of the most common models for testing the usability of electronic systems due to its practicality and generalizability (Abran, Khelifi, Suryan, & Seffah, 2003). The ISO 9241-11 usability model has the advantage of providing guidance and principles that help to improve the usability of the system from the early phases of design. Furthermore, the ISO 9241-11 is a model that focuses on the practical side of the system and tackles that from a user stand of view. Although this approach has drawn some criticism to the model as it purely focuses on the process of using the system rather than its technical performance, the ISO 9241-11 philosophy is that the users' satisfaction is the end goal regardless of how the system was implemented. Thus, if a system that requires a high processing power but can execute the tasks quickly and in a satisfying style to the users, the ISO 9241-11 argue that this system is successful. Furthermore, another advantage of the ISO 9241-11 is that it provides a basis for comparing multiple system designs from different domains point of view. On the other hand, a major disadvantage of ISO 9241-11 is that it does not include the system learnability factor as a domain of usability. However, this can be solved by choosing a satisfaction survey or questionnaire that measures the learnability factor. Finally, the ISO 9241-11 is criticized by many experts for not considering the security of the system as a quality measure. Therefore, security must be assured by the system developers aside from the usability study (Abran et al., 2003).

ISO 9241-11 defines effectiveness as: "The accuracy and completeness with which users achieve specified goals" (ISO, 1998). Furthermore, effectiveness is defined by Oxford dictionary as "The degree to which something is successful in producing a desired result." Therefore, since

effectiveness is purely a goal driven measure, the main criterion to use in data management system effectiveness evaluation is whether or not participants can complete a set of specified tasks in a fashion that meets the user requirements and the goals of developing such a system (Harrison, Flood, & Duce, 2013). ISO 9241-11 defines efficiency as: “resources expended in relation to the accuracy and completeness with which users achieve goals” (ISO, 1998). Therefore, efficiency can be measured by the time that the participants consume to complete each task in the usability study (Frøkjær, Hertzum, & Hornbæk, 2000; Gil, Ratnakar, & Fritz, 2010). Furthermore, ISO 9241-11 defines satisfaction as: “The freedom from discomfort, and positive attitudes towards the use of the product” (ISO, 1998). Satisfaction reflects the attitudes of the user towards the software and is usually subjective and varies from one individual to another. Therefore, satisfaction questionnaires and other attitude rating scales are usually used to measure satisfaction.

2.1.1.2 Questionnaire

There are many satisfaction questionnaires that have been developed to measure the users’ satisfaction with electronic systems. The two main criteria for choosing and adopting questionnaires are their validity and reliability. Validity is a measure of whether the questionnaire is measuring what it intends to measure. Validity can be assessed by comparing the performance of the questionnaire of interest to another “gold-standard” questionnaire. Researchers use Pearson correlation to conduct this comparison, which does not have to be a perfect correlation to prove validity. Another way of validating the questionnaire is by assessing its content validity. This can be done by conducting a factor analysis on its items and confirm that they measure the underlying construct that they claim to measure. Reliability is a measure of

the consistency and responsiveness of the questionnaire. One way to assess reliability is by conducting test-retest to measure the change in the measure over time. Another way, and arguably the most commonly used, is the coefficient or Cronbach's alpha. Cronbach's alpha is a measure of how consistent a group of items is, and how they change together when one item changes. A researcher should choose questionnaires that have been validated and reliable.

Usability questionnaires are similar to usability models; each tackles usability from a unique point of view. There are usability questionnaires that require licensing to be used, and their license must be purchased. The Questionnaire for User Interaction Satisfaction (QUIS) by Chin, Diehl, and Norman (1988) is an example of these questionnaires. This questionnaire is developed by the Human-Computer Interaction Lab at the University of Maryland. It has multiple versions. The latest version has demographic questions, a satisfaction measure, and 11 specific interface factors (J. R. Lewis, 2006). Another questionnaire that is copyrighted is the Software Usability Scale (SUS) by Brooke (1996). This questionnaire has ten items on 5 point scale but is unidimensional and does not provide any subscales. Although the original author did not provide specific validity and reliability measures, the SUS questionnaire was consistent and correlated with other measures that target similar constructs. Although this questionnaire is copyrighted by Digital Equipment Corporation (DEC), Brooke (1996) indicated that any researcher is allowed to use it as long as they acknowledge the source of the measure (p. 194).

The Post-Study System Usability Questionnaire (PSSUQ) is a free to use questionnaire that was originally developed by a group of human factors engineers and usability specialists as an internal project at International Business Machines Corporation (IBM) (J. R. Lewis, 1995). PSSUQ was developed originally by J. R. Lewis (1992) with 18 items measured on a 7-point Likert scale to measure four out five system characteristics associated with usability, effective,

efficient, engaging, error tolerant, and easy to learn. However, a 19-item version of the PSSUQ was introduced with changes on the order of the items that captures all the five system characteristics associated with usability. J. R. Lewis (1995) has conducted a factor analysis to discover the measured subscales in PSSUQ. The factor analysis has revealed that PSSUQ measures four components of usability, The overall satisfaction score (Overall) with all the 19 items, subscale System Usefulness (SysUse) with 7 items, subscale Information Quality (InfoQual) with 6 items, and subscale Interface Quality (IntQual) with 3 items. The remaining three items were either highly cross loading on more than one factor or not loading on any factor. Therefore, they were part of the overall scale but not part of any of the subscales. The PSSUQ with its final version is intended to be used in scenario-based usability tests. Furthermore, the PSSUQ showed high reliability and validity when tested and validated by third parties (Fruhling & Lee, 2005).

Based on the factor analyses by J. R. Lewis (1995) to discover the measured subscales in PSSUQ, they developed the rules for calculating the scale and sub-scale scores for the PSSUQ. The Overall can be calculated by averaging the responses to items 1 through 19. The sub-scale SysUse can be calculated by averaging the responses to items 1 through 8. The sub-scale InfoQual can be calculated by averaging the responses to items 9 through 15. The sub-scale IntQual can be calculated by averaging the responses to items 16 through 18.

One important and unique questionnaire is the After-Scenario Questionnaire (ASQ) by J. R. Lewis (1991). The idea behind this questionnaire is to measure the participants' satisfaction after the completion of each task or each scenario. Therefore, this questionnaire is extremely short, and only have three questions measured on a 7-points scale. The three questions measure the: ease of task completion ("Overall, I am satisfied with the ease of completing the tasks in this

scenario.”), time to complete a task (“Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario.”), and adequacy of support information (“Overall, I am satisfied with the support information (on-line help, messages, documentation) when completing tasks.”) (J. R. Lewis, 2006). The overall score of this questionnaire is the average of the three questions. However, if one question was not answered by the participant or was not used in the questionnaire, the average of the two answered questions is the overall score of the ASQ (J. R. Lewis, 1995).

2.1.1.3 Remedy Usability Issues

After conducting the usability study, researchers or developers assess the results and compare the outcomes to their goals and standards. It is uncommon for researchers or developers not to find any usability issues with their systems. Therefore, in this case, remedy procedures must be taken. The first procedure to remedy low usability is to refer to the participants’ feedback via surveys, interviews, and while completing the tasks in case Think Aloud (TA) method was used (Ivory & Hearst, 2001). These insights from the users, especially from after tasks surveys, have been correlated with the errors they commit and with their performance (J. R. Lewis, 2006). Another remedy to low usability is by observing and investigating what type of errors each participant committed, and whether that error was a participant-related or a system-related error. Furthermore, the performance of the participants should also be observed and investigated to detect the barriers or the system features that consume the participants time (Olmsted-Hawala, Murphy, Hawala, & Ashenfelter, 2010). These two observations should provide significant insights on how to improve the usability of the system. This improvement will not be only

reflected on the efficiency and effectiveness measures of the usability study, but it will also improve the participants' adoption to the system in real life.

As mentioned in the previous paragraph, TA method can be a valuable source of participants' feedback and what they like or dislike. "Think aloud" is a usability technique where the participants, while performing tasks, are encouraged to say aloud what they are doing, thinking, liking, and disliking about the two systems while performing the tasks (Ericsson & Simon, 1980). This means that participants should mention the steps that they are taking to accomplish the task and to feel free to show their frustration or joyfulness while performing the tasks. However, there is still a huge debate in the literature of psychology and human-computer interaction fields on whether this technique could influence the participants' performance on the tasks. The main theory behind this criticism is that the human attention and memory have a certain capacity that becomes very limited when two tasks are competing for its resources. The influence of TA on tasks performance becomes, arguably, more damaging when conducting a scenario-based usability study with two systems comparison. However, Ericsson and Simon (1980) provided evidence that TA does not interfere with the participants' performance as long as the users are not required to think aloud of something that requires more cognitive processing than what is needed for the task. Furthermore, Ericsson and Simon (1980) have argued that any verbalization between the testing administrator and the participant is valid as long as it only requires the participant to visit their short-term memory. Therefore, Ericsson and Simon (1980) discourage any verbalization that draws on the participants' long-term memory, such as "Why did you click on that purple tab?" (Olmsted-Hawala et al., 2010). These arguments by Ericsson and Simon (1980) have been proven by many researchers who conducted double-blind experimental studies and did not find any significant difference between using and not using TA

(Bowers & Snyder, 1990; Olmsted-Hawala et al., 2010). Even more, Berry and Broadbent (1990) have shown some evidence that TA improves the tasks performance of the participants rather than degrading it. Furthermore, Wright and Converse (1992) have compared the performance of two groups, silent and TA, and concluded that the TA group had made fewer errors and had a faster completion time of the tasks. In the end, the TA is a legitimate technique in usability study even when the task performance is the major goal as long as all the tasks and scenarios are performed under the same conditions (J. R. Lewis, 2006).

Another method of fixing usability issues is by following the design heuristics. The design heuristics are broad rule-of-thumb or general principles for interaction design that should be followed during the system design and be used for design evaluation (Riel, 1996). The most common design heuristics and the most generalizable ones are those that were proposed by Nielsen and Molich (1990). Nielsen and Molich (1990) have proposed 10 heuristics that provide a fast, cheap, and easy to learn principles for interface design. These heuristics are: visibility of system status, match between system and the real world, user control and freedom, consistency and standards, error prevention, recognition rather than recall, flexibility and efficiency of use, aesthetic and minimalist design, help users recognize, diagnose, and recover from errors, and help and documentation. Although the original intention of Nielsen and Molich (1990) is to use these guidelines as evaluation methods, these heuristics can be used to guide the design during the entire lifecycle of the system (Molich & Nielsen, 1990). Finally, the same concept of using the design heuristics to guide fixing the usability issues can be applied to using the satisfaction questionnaires as a reference of what might or might not the users like to see in the system.

2.2 SENTENCE COMPREHENSION

The linguistic deficits in PWA can be divided into several general categories, such as naming, repetition, and comprehension. Each one of these linguistic tasks requires multiple processing steps that are supported by complex networks of neurons (Clark, 2011). Although each one of these deficits is equally important to the language system of PWA, sentence comprehension has been the target of many studies in the psycholinguistic field. Furthermore, many aphasia diagnostic tools are using sentence comprehension as a criterion of the severity and the nature of the impairments of PWA. Therefore, there have been many theories that hypothesized the accounts for sentence comprehension impairments in aphasia. There are many theories that build their argument on the belief that each language process or representation is located at a particular area of the brain. Thus, any damage to that particular area will result in a loss or reduction of the ability to activate that process or presentation (Maurer, Fromkin, & Cornell, 1993). Therefore, these theories explain the fact that PWA show deficits in the comprehension of non-canonical sentences while comprehending other sentences normally by arguing that the linguistic representations or neurons centers that support these types of sentences have been lost due to the damage of their area in the brain (Geschwind, 1979). These theories also support the classification of PWA into different types according to the nature of their linguistic impairments (Broca, 1861). Even more, Love and Oster (2002) have developed the subject-relative, object-relative, active, passive test (SOAP) measurement that can differentiate the types of PWA according to their sentence comprehension impairment. Another relatively recent theory has connected sentence comprehension deficits to impairments to the cognitive mechanisms that support them. This theory argues that the structure of the brain is a network of neurons instead of

centers, and therefore, losing or damaging a part of the brain does not necessarily mean losing any ability. According to this theory, impairments in WM and its underlying components can predict deficits in sentence comprehension (Engle et al., 1999; R. C. Martin, 1987; McNeil et al., 1991; Waters et al., 1991).

2.2.1 Working Memory in Language

The investigation of linguistic processing and its underlying cognitive mechanisms have been the focus of language and psychology research for a long time. The primary goal of this investigation is to gain a firm understanding of the disorders that are associated with each linguistic process. This has led researchers to develop linguistic models that represent the relationships and the interactions between cognitive constructs. The model that was proposed by Baddeley and Hitch (1974) was the cornerstone of subsequent extensive research on the role of WM in language processing. The general structure of this model consists of a central executive system (CES) and two temporary storage systems (Baddeley & Hitch, 1974). The role of the CES is to regulate the flow of thoughts, as well as, to allocate the resources and attention to the other components of WM. Therefore, the CES is considered to be the manager of the system that is responsible for dividing, focusing, and switching attention (Baddeley, Chincotta, Stafford, & Turk, 2002). The two temporary systems on the other hand act as language-based (phonological loop) and visuospatial-based (visuospatial sketchpad) temporary storage systems. Furthermore, there are rehearsal buffers that are associated with each one of these systems. In the phonological loop (which is the system of interest to aphasia research), the rehearsal buffers are responsible

for rehearsing and recycling the verbal material within a decay period, which is hypothesized by Baddeley to be 2 seconds, before the processed information get lost (Baddeley & Hitch, 1974).

Baddeley and Hitch's model has triggered the development of linguistic measures that aim at quantifying the capacity of WM. The first measurement of WM based on Baddeley and Hitch's model was developed by Daneman and Carpenter in 1980. In this measurement, they developed reading span tasks that load on language processing and storage, according to their proposal. The primary goal of their measurement is to differentiate between subjects' reading comprehension ability by measuring their WM capacity (Daneman & Carpenter, 1980). Basically, the subjects who use their processes efficiently will allow more storage and manipulation of information and subsequently have a higher WM capacity. Therefore, the processing of written information and the storage of intermediate products are assumed by Daneman and Carpenter to be two equally important roles of the single WM system.

Just and Carpenter (1992) came up with a theory that explains the WM capacity and how it is connected to language comprehension. They proposed that WM capacity can be defined as the maximum level of neural activation that is available to support either computation or storage in WM. Furthermore, they proposed that every language comprehension task needs some threshold level of WM capacity to be executed. This capacity threshold is set by the available amount of computational processes, such as encoding information from written or spoken words, and the available amount of information retrieval from long-term memory. This means, individuals with limited WM capacity will show a slower and poorer language comprehension performance when executing tasks that exceed their maximum threshold. On the opposite, Caplan and Waters (1999) have rejected Just and carpenter's theory and argued that different WM segments do not share the same neural activation capacity. They built their argument based

on the fact that Just and Carpenter's theory failed to explain the performance of individuals with brain damage. Furthermore, when they measured the WM capacity of individuals with aphasia using the Daneman and Carpenter's reading span task, they did not find a relationship between the individuals' performance on this task and their syntactic comprehension ability. This finding, according to Caplan and Waters, is a converse of Just and Carpenter's theory since it clearly shows that WM capacity does not predict the individuals' language comprehension.

McNeil et al. (1991) has presented some arguments regarding the functionality of the linguistic procedures and the WM components. Although they do not disagree with the view that WM components are separate cognitive sub-systems, they still provide some evidence that there are shared resources that these cognitive sub-systems rely on. Moreover, they used observations from PWA performances to challenge the linguistic theory that has been proposed by many linguistic models. Observations, such as the change in PWA's performance as a response to the change of environmental non-cognitive factors and the inconsistency of their performance even in controlled environments challenge the indication that subcomponents of the language get impaired. To answer these concerns, they provided five theoretical principles that describe how attention functions. First, they proposed a structure of three levels of attention, arousal, which serves as the fuel that energizes the linguistic processes, attention, which allocates this fuel to different processes, and the processes that are associated with each linguistic component. Based on this structure, the other four principles indicate that different linguistic components share the same pool of resources, these resources are distributed based on task demand, the impairments of PWA are related to the disturbance of the distribution of these resources rather than the reduction of their amount, and there is still a required overall threshold of available resources for each task to be executed (McNeil et al., 1991).

Engle et al. (1999) on the other hand have purposed a model that describes the structure of WM and combines the two views of Caplan and Water and Just and Carpenter, and line with the view of McNeil et al. (1991). They divided the WM into three main components: 1) domain-specific memory stores; 2) domain-general executive attention; and 3) rehearsal procedures that are associated with the domain-specific memory stores. They used explanatory and confirmatory factor analyses to measure the WM and STM separately. Their analysis showed that these two constructs are distinguishable yet highly related to each other. They proposed that the remaining unshared variance in the WM construct is related to the domain-general executive attention, which led to the assumption that WM consists of STM and executive attention (Figure 3).

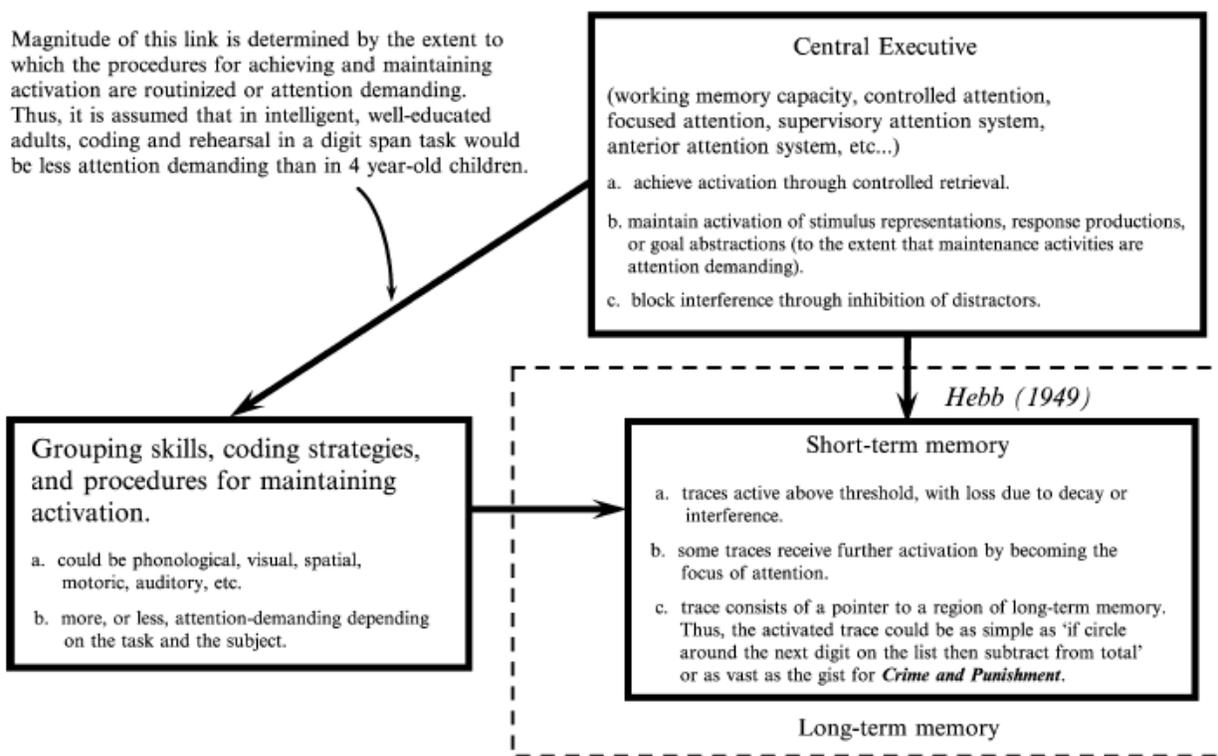


Figure 3. Engle, 2004 WM model.

Relationships of components of the working memory system as proposed by Engle et al. This diagram shows the three components of WM, attention, rehearsal procedures, and STM, which is an activated part of the LTM. (from Engle, 2004, p. 148).

2.2.2 Short-Term Memory in Sentence Comprehension

The structure of the memory and its functioning mechanisms have gained a considerable amount of discussion during the 1960s (Baddeley, 2003). The main issue was related to the nature of the memory system and whether it can be modeled as a two-component system or not. Hebb (1949), in “The Organization of Behavior” book, has distinguished between long-term memory (LTM) and short-term memory. He suggested that the LTM is a knowledge store that holds records of prior events, while the STM is related to temporary electrical activity (Hebb, 1949). Atkinson and Shiffrin (1968) also agree with the two-component memory model, where they proposed that STM serves as a temporary interface between the environment and the LTM. Even more, the two-component memory model was supported by the studies with impaired memory subjects (Baddeley & Warrington, 1970; Milner, 1966). The results from these studies show that some subjects could experience a reduction in their capacity for new learning while having an intact STM, which clearly indicates some independence between these two memory systems.

In Baddeley and Hitch (1974) WM model, they proposed two slave subsystems that served as an STM. The visuospatial sketchpad served as a visuospatial STM that retains visual information for a short period. The phonological loop consists of a short-term phonological store and an articulatory rehearsal component that revives the memory traces. In 2000, Baddeley updated his model to include a new component that is called the episodic buffer. He suggested that this component is a network that connects all the components within the WM together as well as linking them to LTM (Baddeley, 2000). This add-on was necessary to include a cross-domain component that explains the impairments that affect the WM as a unitary system. Therefore, the STM capacity in Baddeley’s view is defined by how much information it could

hold as well as the amount of time it could stay activated, which he suggested to be 2 seconds or more depending on the performance of the associated rehearsal component (Figure 4). Cowan (1999) has purposed a theoretical framework of STM that describes its relationships with other components and its functioning mechanisms. The STM in Cowan's model consists of four chunks that hold a "collection of concepts that have strong associations to one another and much weaker associations to other chunks concurrently in use" (p. 89). In his model, he did not disagree with Baddeley on the definition of STM capacity but did not include the time factor in his model either. Although he recognized that there could be some effect on STM capacity by time and interference, he stated that in his model, the effect of the maintenance rehearsal should be isolated before the calculation of the STM capacity. In fact, Cowan emphasized that the ST and LT-M have a lot in common regarding the structure but differ from each other regarding the capacity. This view was acknowledged by Engle et al. (1999) where they presented STM in their model as the activated part of the LTM. However, they proposed that STM activation strategy or mechanism may differ according to the processed stimulus or domain. This activation is maintained above a threshold level by rehearsal processes that are associated with the domain-specific STM stores.

As mentioned in the previous section, there has been a lengthy discussion on whether WM and its sub-systems are domain-general or domain-specific. This disagreement is a result of the different patterns that each study observes in their subjects' performance (Dennis, Agostino, Roncadin, & Levin, 2009; Swanson & Luxenberg, 2009). However, there is an active thread of research that shows a disassociation between the performances of PWA on STM spans (R. C. Martin, Vuong, & Crowther, 2007). To clarify, the performance of PWA on tasks that heavily require the maintenance of phonological information has no correlation with their performance

on tasks that involve the maintenance of semantic information in STM (Allen, Martin, & Martin, 2012; Nadine Martin, Kohen, Kalinyak-Fliszar, Soveri, & Laine, 2012). Furthermore, PWA who perform poorly on phonological STM tasks show relatively intact semantic processing (Majerus, Van der Linden, Poncelet, & Metz-Lutz, 2004). Also, this disassociation is observed when comparing the performance of PWA on phonological and semantic STM tests to their performance on syntactic STM ones, which indicates that syntactic STM is also a separate domain from the prior two (R. C. Martin & He, 2004).

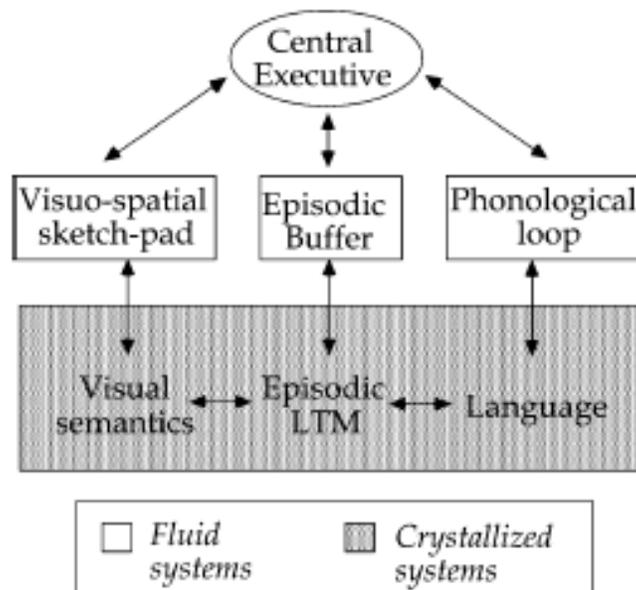


Figure 4. Baddeley, 2003 WM model.

Relationships of components of the working memory system after the addition of the episodic buffer in Baddeley's model. This diagram shows how episodic buffer interact with all the WM components and its interaction with LTM. (from Baddeley, 2003, p. 203).

2.2.2.1 Phonological STM

Since the recognition of the domain-specific nature of STM studies have been investigating the role of each domain in language processing in general and in sentence comprehension in

specific. However, the role of phonological STM in sentence comprehension is still an open debate. R. C. Martin and Feher (1990) have studied the relationship between the phonological STM and sentence comprehension for more than one type of sentences. They provided their subjects with these sentences under two modes, the unlimited mode, which means leaving the sentence in front of the subject until they provide an answer, and the limited mode, where they show the words of the sentence one by one. They claim that the difference in the subjects' performance between these two modes measures the role of STM in sentence comprehension. They concluded that phonological STM is not necessary for sentence comprehension for sentences with complex syntax. Furthermore, they argued that such an STM domain is needed for span tasks rather than sentence comprehension. These findings were supported by Waters et al. (1991) where they presented a case that had impaired STM in general and phonological STM in specific while showing an excellent sentence comprehension. They also support the argument that the articulatory rehearsal processes of STM or phonological STM have no role in sentence comprehension.

On the other hand, there are many studies that found an association between phonological STM and sentence comprehension. B. A. Wilson and Baddeley (1993) also presented a subject with an impaired phonological STM and used his test re-test scores to observe the relationship between the phonological STM and sentence comprehension. In his initial test, he had a memory span of two digits and a sentence span of three words, which clearly indicate that he has severe impairments in both, phonological STM and sentence comprehension. However, when they re-tested the subject after several years, he showed an improvement in the STM and almost a full recovery in sentence comprehension. They used this finding to conclude that the proposed link between phonological STM and sentence comprehension does exist. This finding was supported

by Hanten and Martin (2000), where they showed a reduction in sentence comprehension abilities of a child with impaired phonological STM compared to an age-matched control group (Hanten & Martin, 2000). One explanation for this inconsistency in the results of these studies is related to the nature of the sentences that have been used to measure the sentence comprehension ability. Furthermore, it is noticeable that the studies that found an association between phonological STM and sentence comprehension have used sentences that are overloaded with lexical items, such as the Token Test and tasks that require verbatim repetition. On the other hand, the studies that found no association have used sentences that are syntactically complex, such as relative clauses, passives, and garden path structures (Gvion & Friedmann, 2012). Another explanation could be related to the location of the complex construction in the used sentence. According to Martin (1987), when the complex construction occurs early in the sentence it poses more difficulty to subjects with impaired phonological STM than when it occurs late in the sentence (R. C. Martin, 1987). However, more comprehensive and novel research is needed to provide more confident results regarding this conflict.

2.2.2.2 Semantic STM

The role of semantic STM in sentence comprehension is somewhat clearer than the one related to phonological STM (Haarmann, Davelaar, & Usher, 2003). In general, the studies that investigate the role of semantic STM have reported two situations where semantic STM is necessary for a successful sentence comprehension. The first, when the target sentence is loaded with more than one lexical-semantic representation preceding the integration point. This situation was observed by R. C. Martin and He (2004) where they tested the sentence comprehension ability of a subject with impaired STM. Although this subject had impairments in phonological STM, the

researchers reported that this impairment had no effects on his performance in sentence comprehension tasks. Furthermore, they indicated that the subject's semantic knowledge is intact as he was able to access semantic information in tasks that do not require the semantic STM. Even more, his sentence comprehension was normal when sentences were shown throughout the task, and no maintenance of semantic information is needed. However, when they provided the sentences in a form that requires STM, the subject showed a poor performance compared to the control group. Also, the more distance between the lexical-semantic representation and the integration point the more his performance worsens. Their conclusion as many others who investigated the semantic STM (R. C. Martin & Freedman, 2001; R. C. Martin & Romani, 1994) was that the patients with impaired semantic STM have difficulty holding more than one lexical-semantic representation at a time, which affects their sentence comprehension ability (R. C. Martin & He, 2004).

The second situation where semantic STM found to be necessary for successful sentence comprehension is when the reader is required to hold multiple meanings of one lexical-semantic representation (homograph) at a time. Miyake et al. (1994) demonstrated this effect when they provided their subjects with two types of sentences that had homographs with a highly frequent meaning. The first type had this homograph in a meaningful context when used with its high frequent meaning, while the other required the subjects to find another less frequent meaning that fits the context. Subjects with low reading span abilities (low semantic STM capacity) performed poorly in the second type of the sentences compared to subjects with high reading span abilities (high semantic STM capacity). Therefore, Miyake et al. have concluded that subjects with lower semantic STM capacity hold only the dominant interpretation of the homograph, which affects their sentence comprehension accuracy and speed, while high semantic STM capacity has

multiple meanings of the homograph ready at the same time. Furthermore, they indicated that this difference between low and high span readers significantly drops when using sentences with homographs that do not have a high frequently meaning. However, since most of the studies that investigated semantic STM had small sample sizes, further analysis with relatively large sample size is needed.

2.2.2.3 Syntactic STM

Syntactic STM has probably the most obvious relationship with sentence comprehension in theory but the most difficult one to explain. The most type of sentences that have been used to express this relationship are the ones with center-embedded syntactic structures (Gibson, 1998). The center-embedded syntactic structures are the ones that contain constituent X embedded within constituent Y, where constituent Y is divided to the right and the left of constituent X. There are many theories that have been developed to explain why readers find center-embedded syntactic structures harder to process compared to the edge-embedded ones. Furthermore, many of these theories have used the capacity of STM or WM to explain this difficulty. E. Chen et al. (2005) have investigated this issue by providing center-embedded syntactic structures with different levels of complexity depending on how many embedded constituents they have. Their results have confirmed the existence of syntactic STM by observing slower performance and less accuracy with the increase of the sentences' difficulty. Furthermore, they explained this memory cost to be caused by having the parts of each constituent separated from each other, which requires the reader to hold parts of multiple constituents at the same time while waiting for the rest of their parts to be integrated. This phenomenon is absent in the edge-embedded sentences because each constituent is integrated before encountering the next one. While this theory is

being supported by many studies (Gibson, 1998; Hakuta, 1981), other theories link this difficulty to the syntactic processing and interference instead of memory (R. L. Lewis, 1996; Stabler, 1994). Therefore, the relationship between syntactic STM, syntactic processing, and interference requires further investigation and analyses.

2.2.3 Attention in Sentence Comprehension

Although WM models differ in their view of attention regarding its specific functions and structure, almost all WM models included it as one of the core components of the WM system. In Baddeley and Hitch (1974) model, they view attention as the CES that regulates the flow of thoughts and allocates the resources to the other components of WM. Moreover, the CES is considered to be the manager of the system that is responsible for dividing, focusing, and switching attention (Baddeley et al., 2002). Furthermore, Daneman and Carpenter (1980) have built their WM capacity span by measuring the efficiency of the CES. Their proposal is that they could differentiate good and poor readers by measuring their CES performance because good readers with efficient CES will be able to read, integrate, store information in LTM, and keep the STM activated for a longer time. This has led researchers to investigate the role of WM in sentence comprehension for multiple sentence types. Swets, Desmet, Hambrick, and Ferreira (2007) have used structural equation modeling to test this relationship, where they identified two factors that represent WM measurements, and investigated their relationship to relative clause ambiguity sentences. The first factor that they identified was representing the domain-general WM (verbal factor), and the second was representing the domain-specific WM, which they hypothesized to be measuring the attention. Although they found that the verbal domain-specific

factor has a stronger relationship to ambiguity attachment preferences, they stated that the domain-general factor was related to the ambiguity attachment preferences too, which shows that attention plays a significant role in sentence comprehension.

Engle and Kane (2004) have divided the central executive into subcomponents that explain the structure and the function of attention in WM. In fact, they believe that this central executive attention is the one that defines WM capacity rather than STM. Furthermore, they proposed that the two primary functions or capabilities that differentiate individuals are their ability to deal with effects of interference and their ability to avoid the effects of distractions. Therefore, they divided the executive attention into *goal maintenance* and *conflict resolution* factors, which are domain-general factors that function across all stimulus and processing domains. The first factor, the goal maintenance, is defined by the ability to maintain the goal of the task in the active memory and eliminate any distractions that could capture the attention away from it. The second factor, the conflict resolution, is needed when pre-potent or habitual behaviors conflict with behaviors appropriate to the current task goal (Engle & Kane, 2004). Moreover, Engle and Kane have hypothesized that individuals with higher WM capacity will perform quicker and more accurately than individuals with low WM capacity on tasks that load on these two factors (Kane & Engle, 2003).

Lim, McNeil, Dickey, Doyle, and Hula (2012) have investigated the resolution of competition in PWA using Picture-Word Interference Task. They reported that whenever an interference effect occurs in the task, PWA show significantly lower reading times, and higher error rates, which indicates that PWA have an impairment in CR. Even more, McNeil et al. (2012) have investigated the PWA's ability to resolve interference using a self-paced sentence-reading Stroop task. They found that PWA took a longer time to read sentences with incongruent

conditions, which shows that these types of sentences need higher CR ability to be processed correctly. Furthermore, January, Trueswell, and Thompson-Schill (2009) have used the functional magnetic resonance imaging to test whether subjects use shared prefrontal neural circuitry during the execution of the Stroop task, which taps on CR, and the comprehension of sentences with syntactic ambiguity. They found that these two distinct tasks show an overlap in neural responses, which also support the suggestion that CR plays an important role in sentence comprehension. Also, these findings agree with the previously mentioned studies that found a relationship between attention, in general, and the comprehension of sentences with ambiguity. Although all these studies point out the importance of CR in situations where one ambiguous word is biased toward one meaning but the subordinate meaning is needed instead, STM should also take a role in these situations. That is, when an ambiguous interpretation occurs in a sentence, the reader is required to keep multiple possible interpretations active in memory, and use attention or CR to block the irrelevant interpretations and only use the relevant ones. Therefore, the investigation of the particular role that STM and CR play in the comprehension of multiple types of sentences is needed to understand the nature of the deficits in such procedures.

2.2.4 Measurements

2.2.4.1 Sentence Comprehension

The development of sentence comprehension measures has been the target of linguistic and psychology research for a long time. The goal of this investigation is to develop sensitive measures to the different types of deficits in sentence comprehension. These measures have been used in clinical settings as diagnostic tools and as treatments planning aids. Each one of these

measures has its unique sensitivity to a certain aspect of the sentence comprehension process. The general structure of these measurements is to provide the subjects with sentences then ask them to perform actions based on their understanding of these sentences. Chen et al. (2005), for example, have provided their subjects with four types of sentences that describe a situation and asked their subject about the subject, object, verb, etc. Because these sentences were presented in a different format each time, the difference in the subject's performance detects any impairments in sentence comprehension (E. Chen et al., 2005). R. C. Martin and He (2004) on the other hand asked their subjects direct questions, such as "Which is soft, sandpaper or cotton?", and used the subjects' response as an indicator of the presence or absence of sentence comprehension deficits. Although these types of measurements have been used in many research studies, issues, such as non-standardization of the stimuli, the requirement of decision making abilities along with the cognitive ones, and loading on cognitive processes that are independent from the one under investigation, raise some concerns regarding the validity and reliability of these measurements (Adani & Fritzsche, 2015).

McNeil and Prescott (1978) have proposed a Revised Token Test (RTT) as a standardized test regarding its presentation, response stimuli, administration, and scoring. Although this test asks the subjects to perform tasks in response to commands, the used stimuli are simple enough that these tasks do not require high decision-making abilities. Moreover, this test is scored on a 15-point multidimensional scoring system that captures subtle differences in language performance and provides continuous descriptive data that can be used in almost any data analysis procedure. Therefore, their measurement has been used in clinical and research settings since its development in multiple languages with different cognitively impaired populations. In (2007), Eberwein and colleagues have developed a computerized version of the

RTT (CRTT). They used this version of the test to measure elderly and hearing-impaired participants' sentence comprehension ability and found that it is as highly sensitive as the paper and pencil RTT (Eberwein et al., 2007). Even more, Sung et al. (2011) used a reading version of the CRTT (CRTT-R), which manipulates the number of adjective phrases, to demonstrate its ability to provide sufficient task demand on STM by comparing the performance of PWA to a control group. Furthermore, Sung et al. (2009) presented the CRTT-R in a full-sentence version, a self-paced-reading version, and a self-paced moving-window version, to test the relationship between WM capacity and sentence comprehension. McNeil et al. (2010) have investigated the role of CR in sentence comprehension among PWA by incorporating a Stroop component into the CRTT-R. This was accomplished by providing commands, such as "Touch the red circle," with manipulating the color of the font to be either congruent, incongruent, or neutral. All these studies are an emphasis that the CRTT is a standardized test but yet highly customizable to fit any purpose while keeping high levels of sensitivity and reliability. Also, these demonstrations show that the CRTT can be used to study the relationship between sentence comprehension and most of the WM spans.

2.2.4.2 Cognitive Systems and Functions

Measurements of LP and STM have been used extensively in the linguistic and psychology research. These measurements are powerful tools for detecting impairments in LP and STM as they show how PWA's performance differ from the matched groups. Furthermore, they serve as assessment tools that provide insights to clinicians regarding their clients' response to certain treatments that target these cognitive domains. Also, these measurements provide continuous scales that can be used in research to investigate the relationships of LP and STM to any other

component of interest. Each one of these measurements is designed to probe a particular linguistic ability or domain depending on the used stimuli and the presentation technique. Therefore, most the measurements that load on phonological STM use rhyme judgments to capture the phonological ability of a subject but differ in the presentation technique. For example, Harris, Olson et al. have investigated the link between STM and sentence comprehension by observing the effect of treatments that target STM on sentence comprehension ability of their subjects. In order to measure the phonological STM, they provided their subjects with lists that contain six words, for example, “fair, shoot, purse, boot, hearse and share”, and asked them to point out the words that rhyme with “stare” (L. Harris, Olson, & Humphreys, 2014). The difficulty of this task increases by increasing the number of items in the list, as more items mean more consumption of phonological STM. Another measurement that probes the semantic STM was designed by Roach, Schwartz, Martin, Grewal, and Brecher (1996) and called “Philadelphia Naming Test” (PNT). In this test, subjects are provided with pictures and asked to describe them with one word as fast as possible. The scoring of this test was designed to capture the first three answers, in case there were more than one provided, which helps to have a scale of scores instead of dichotomous one.

N Martin, Kohen, and Kalinyak-Fliszar (2010) have developed the Temple Assessment of Language and Short-term Memory in Aphasia (TALSA), which is a battery of measurements that target LP and STM. The TALSA battery contains assessments of phonological and semantic LP abilities along with measurements that probe STM. The STM measurements vary in their complexity by having a different number of items, different interval lengths, such as 3,5 or 7 seconds, and different fillings between the tasks’ stimuli. One interval filling could be asking the participant to name numbers that appear between the stimuli, which helps to investigate the role

of STM under interference. On the other hand, another filling could be silent, which helps in isolating any interference with STM. In addition, these measurements vary in the particular ability that they test by requiring the subject to perform different judgments (synonymy and rhyming judgments) depending on the targeted ability. Because TALSA can be presented with single or multiple word stimuli, it is powerful for testing PWA with different levels of impairments. This is especially important with people with mild aphasia who tend to have an almost normal processing ability of single words but still struggle with LP (N Martin et al., 2010). The theoretical motivation behind the development of TALSA is the suggestion that impairments in LP and sentence comprehension could be a result of the failure of the processes that are supposed to maintain the STM activated above a certain threshold during the LP tasks (Nadine Martin, Saffran, & Dell, 2013). Therefore, TALSA has been used in multiple studies that aim at measuring the STM and LP of PWA. Kalinyak-Fliszar, Kohen, and Martin (2011) have used TALSA to investigate the efficiency of treatments that target STM. They indicated that TALSA was a good choice for their research since it can reliably measure the improvement of participants in response to treatments. Furthermore, N Martin et al. (2010) have demonstrated how to use TALSA by testing thirty individuals with aphasia and ten aged-matched controls and then reported some guidelines and principles based on their results (Figure 5).

Subtest			Interval Condition		
			1-sec UF	5-sec UF	5-sec F
Phoneme Discrimination (n=30)	Words	Mean	0.95	0.93	0.87
		Range	.80 - 1.00		
	Nonwords	Mean	0.93	0.91	0.80
		Range	.70 - 1.00		
Rhyme Judgments (n=29)	Words	Mean	0.88	0.87	0.79
		Range	.60 - 1.00		
	Nonwords	Mean	0.84	0.82	0.76
		Range	.35 - 1.00		
Repetition (n=29)	Words	Mean	0.80	0.83	0.68
		Range	.33 - 1.00		
	Nonwords	Mean	0.48	0.42	0.17
		Range	0 - 1.00		

Figure 5. Tests of Phonological Processing.

Mean proportion correct at each interval condition and range for 1 sec interval. This table shows how accuracy decreases with the increase of interval time and with filling the interval with counting numbers aloud (from Martin et al., 2010, p. 6).

Attention measurements have lately been used in the linguistic and psychology research. Their emergence in this field is a result of the growing evidence that attention has a much important role in LP than what researchers used to believe before. However, measuring attention is a very challenging task since most evidence indicate that it is a domain-general component, which makes the isolation of its effect from the effects of other cognitive components very difficult. Therefore, attention measurements usually designed to capture a specific hypothesized effect of attention on LP or sentence comprehension. As discussed in the attention section, according to Engle and Kane, (2004), CR is one of the two tasks of attention that have important roles in LP (Engle & Kane, 2004). The role of CR in LP and sentence comprehension have been shown by studies that investigated its relationship to sentence comprehension in PWA

populations (Lim et al., 2012; McNeil et al., 2012). The CR measurements mainly include the Stroop effect in their stimuli as it is arguably the most sensitive task to CR effect (January et al., 2009). For example, Lim et al. (2012) have used a Picture-word interference (PWI) task to investigate the differences between the CR ability of PWA and a control group. In their PWI task, they introduced their participants with the stimuli in three modes, neutral, congruent, and incongruent. In each one of these modes, they provide the participant with a picture with a word written on it and ask the participant to perform semantic judgments on whether the word inside the picture is describing an animal or non-animal. However, in the congruent mode, the picture always match the word on it, in the incongruent mode, the picture is paired with a word from a different semantic category, and in the neutral mode, the word is presented on a polygon, which eliminated any interference caused by lateral masking. Salthouse and Meinz (1995) used another measurement with the Stroop effect, where they tested relationships between aging, inhibition, and WM. Their measurement was in a Number Stroop format, where they provide their participants with blocks that contain numbers (from 1 to 4) and ask them to response with how many numbers there are in each block. Similarly to the PWI, the Number Stroop task is presented in neutral, congruent, and incongruent modes. In the congruent mode, the numbers will match their quantity (e.g., the digit 3 is displayed three times – as in 333), in the incongruent modes, the quantity will not match the numbers shown, and in the neutral mode, the numbers will be substituted by the character “X”, which eliminates any confliction. Although all these tasks share the Stroop effect in their stimuli, each one of them taps on the CR in a different way. Therefore, using more than one CR task is important to capture the CR effect and isolate any noise that is usually presented in these tasks.

2.3 STATISTICAL ANALYSIS

Choosing which statistical analysis to use in research is usually derived from the type and structure of the collected data as well as the theory of interest. Factors, such as the distribution of data, number of groups, and the existence of latent variables, are the primary criteria of narrowing the possible statistical analyses to be used. Statistical analyses, such as univariate and multivariate regression and analysis of variance ANOVA could produce very robust and informative statistical models that can test many research hypotheses. However, when conducting a statistical analysis that includes latent or unobserved variables, these statistical methods become less helpful since their weak handling of latent variables. Therefore, Structural Equation Modeling (SEM) can be described as a diverse set of mathematical and statistical models that investigate the hypothesized relationships between latent constructs. Again, the decision of the specific mathematical and statistical methods to use in SEM should be influenced by the type of the hypothesis of interest.

The main goal of SEM is “to determine the goodness of fit between the hypothesized model and the sample data” (Byrne, 1994, p. 7). When hypothesizing a relationship between a group of constructs, this hypothesis will be accepted when having a model with a good fit and will be rejected when having a one with a bad fit. Therefore, SEM is considered a good modeling technique to test relationships using both cross-sectional and longitudinal data (Yuan & Bentler, 2006). Every SEM analysis consists of two models that test two general relationships, a measurement model, and a structural model. The measurement model presents the relations between the observed variables and their hypothesized underlying constructs. The structural model represents how these latent or unobserved constructs interact with each other and how

they interact with other external factors (Bagozzi & Yi, 2012). General steps of SEM are model specification, estimation, evaluation, and modification (Hox & Bechger, 1998). In the model specification, the hypothesized relationships are described and converted from theory into statistical properties. In model estimation, the model parameters (regression coefficients in structural equations and the variances and (co)variances of independent variables) are estimated using a fitting function that minimizes the discrepancy between the observed (co)variances and these estimated parameters. In the model evaluation or assessment of fit, the statistical fit is evaluated by calculating the absolute fit indices, such as T statistic, and practical fit using parsimonious fit indices, such as the Root Mean Square Error of Approximation (RMSEA). Finally, when the original proposed model does not provide a good fit, model modifications are made to try to obtain a better fit model (Bagozzi & Yi, 2012). However, if the hypothesized model does not provide a good fit, this might indicate that the underlying theory should be rejected.

SEM has been used in psycholinguistics to investigate how cognitive domains interact with each other or with other psychological components. Engle et al. (1999) have used SEM to study the relationships between WM, STM, and fluid intelligence (gF). The theory behind their research is to try to use measures that are hypothesized to load on each one of these components of interest to define latent variables that can be used to study their relationships. Their prior hypothesis was that factor analysis would be able to identify one construct for WM and a separate one for STM. In addition, they hypothesized that both constructs would have a relationship with the third construct, the gF. However, they hypothesized that the relationship between STM and gF would be diminished if they statistically controlled for the effect of the WM. On the other hand, they predicted that the relationship between WM and gF would not be

affected even if they statistically controlled for the effect of STM. Moreover, they wanted to find out whether this relationship will remain significant after removing the shared variance between WM and STM, which will represent the relationship between attention and gF. Therefore, their measurement model was representing the relationships between the observed tasks that load on these cognitive domains and the hypothesized WM and STM latent variables. Their structural model, however, was representing the relationships between WM and STM constructs and the gF construct.

They included 10 measures in an explanatory factor analysis (EFA) to test the hypothesized measurement model. These measures were: operation span (OSPAN); reading span (RSPAN); counting span (CSPAN); backward span (BSPAN); forward span, dissimilar (FSPAND); forward span, similar (FSPANS); keeping track (KTRACK); Immediate Free Recall Secondary Memory (IFRSM); verbal reasoning task (ABCD); and continuous opposites (CONTOP). As they hypothesized, their EFA revealed two distinct but highly correlated constructs with three observed measures for each one of them. Even more, the model validation using confirmatory factor analysis (CFA) confirmed the findings of the EFA by having nonsignificant chi-square test, an estimate of RMSEA that is equal to .05, and estimates of GFI, AGFI, TLI, and CFI were all above .90. In the model specification phase, they blogged the two construct from the factor analysis with a gF latent variable that is measured by two observed tasks, Cattell's Culture Fair Test (CATTELL) and Raven's Progressive Matrices (RAVENS). After the model estimation and model evaluation or assessment of fit phases, they found that this model has a good fit according to the previously mentioned tests. Furthermore, they found a significant relationship between WM and gF and a nonsignificant relationship between STM and gF. This finding led them to conduct model modifications by removing the link between STM

and gF, which showed a better goodness of fit. Therefore, they added a common variance component in the model to answer the question whether there would still be a significant relationship between the residual of WM and gF (Figure 6). They found that the relationship between the residual of WM and gF was significance after removing the shared variance with STM. The conclusion of this study was that WM capacity have an effect on gF, and this effect is related to the attention (the residual of WM) rather than to STM, which confirms that WM consists of STM and central executive system (Engle et al., 1999).

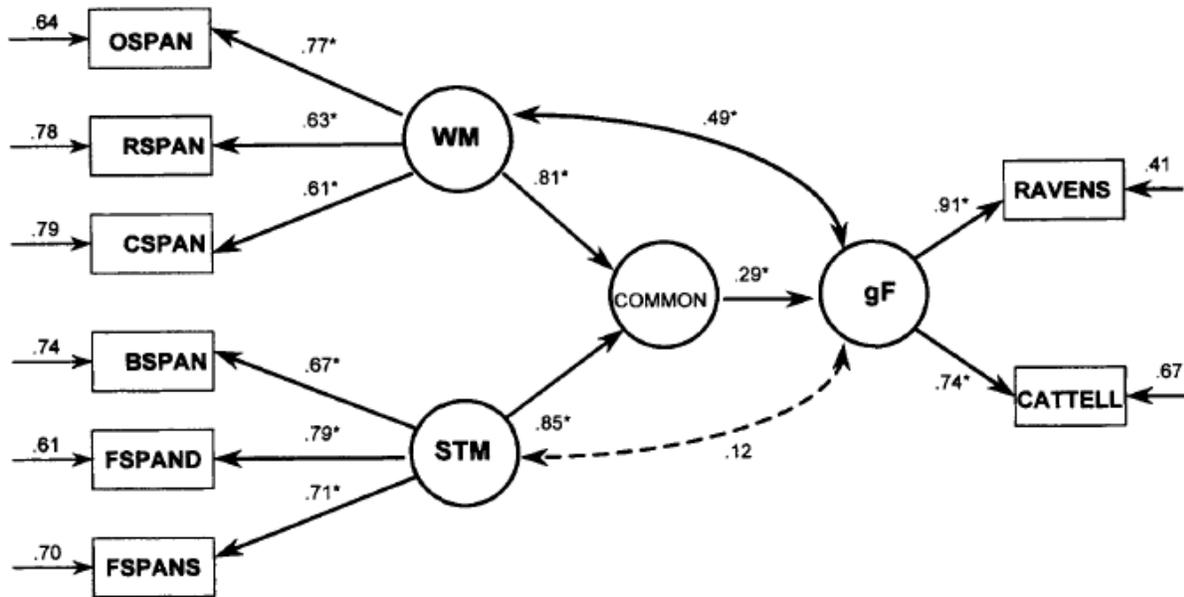


Figure 6. Engle, Tuholski et al. 1999 SEM model.

The final SEM model where relationships of WM, STM and their common variance are presented in relation to gF (Engle, Tuholski et al. 1999, p. 324).

2.3.1 Missing Data

Missing data or missing values occur when there are no observations available for one or more cells in a dataset. Data can go missing due to different reasons and situations. One is when

participants or subjects do not provide any response for various of reasons, for example, when asked about sensitive personal data, such as income or age, some subjects prefer not to provide these data. Another example is when studies ask subjects to perform some tasks, and some subjects refuse to do such tasks because they are difficult or demand physical or mental effort beyond their abilities. Another situation where data go missing is when the study either have a longitudinal design where subjects are required to have follow-ups for a long period of time, or when a study requires multiple sessions to collect the subjects' data. In either situation, dropouts, where subjects stop to show up or decide to stop participating, can cause some missing data in the dataset. Even more, a subject could attend all sessions and perform all the tasks that they were asked to perform and still have some missing data due to poor data management or poor tasks presentation.

The missing data issue is critical due to its effect on the dataset's analysis outcome. The more missing data in a dataset, the more effect they have on the outcome. This effect becomes even more critical in large datasets that have missing data in almost every record or in small datasets that cannot afford to delete any records. The effect of missing data could cause serious damage if missing data were not handled correctly. For example, in medical trials, studies have shown that missing data have caused incorrect conclusions about drugs safety, the effectiveness of treatments for many diseases, and limited the ability to draw conclusions on weight loss trials (Kemmler, Hummer, Widschwendter, & Fleischhacker, 2005; Lagakos, 2006; Ware, 2003). The latter situation usually occurs in experimental clinical trials where subjects in the treatment group that do not benefit from the treatment, dropout from the study, which makes the treatment looks effective since the only subjects with complete data are the ones who benefitted from the treatment.

Since data go missing for various of reasons, missing data can form different patterns based on their distribution in the dataset. The first missing data pattern is Missing Completely At Random (MCAR). This pattern means that there is no pattern for the missing data and that the missingness is not related to the value itself nor to the other responses or variables in the dataset. For example, in (Table 1), missingness in Y have no relationship to the values that are supposed to be in the observation nor to the other responses in X. The second pattern of missing data is Missing At Random (MAR). This pattern occurs when the missingness is not related to the missing values themselves but to the other responses in the dataset. For example, when the income data for subjects who are older than a certain age are missing, then this is probably because individuals after a certain age tend not to share their income and is not related to how much they make. The third pattern of missing data is Missing Not At Random (MNAR). This pattern occurs when the missingness is related to the missing value itself. To extend the last example, when individuals with income higher than a certain level refuse to share their income. This pattern is the most problematic one since there is not much that can be done to predict or remedy for the missing values (Kang, 2013).

Table 1. Randomly generated data to demonstrate different missing data patterns.

Complete		MCAR		MAR		MNAR	
X	X	X	Y	X	Y	X	Y
1	2	1	2	1	2	1	2
4	8	4	NA	4	NA	4	8
5	7	5	7	5	NA	5	7
3	9	3	9	3	9	NA	NA
2	10	2	NA	2	10	NA	NA

MCAR= Missing Completely At Random, **MAR**= Missing At Random, and **MNAR** = Missing Not At Random.

Preventing missing values is a task that should start from the study design (Council, 2011). A simple trick that could significantly reduce the number of missing data is to shorten the follow-up period. Furthermore, another study design precaution is to conduct the study in populations that are currently not well served with trials to increase the chance of having more incentive participants. Also, another precaution is to run a practice trial where subjects perform tasks that are similar to what they are going to perform in the actual study and only include the participants who tolerate these tasks (Little et al., 2012). Even more, missing data can be reduced by developing data collection and management policies and using validated and secured data management systems. One example of data collection policies that could reduce the number of missing values is not to release the participant until their data is double-checked and validated, so they can be asked to provide some information or perform some tasks again in case some data were missing (Graham, 2009).

However, if there are still some data missing after the data collection have been completed or after the participant with missing data has been released, there are multiple procedures that can be performed to remedy or regenerate the missing data. These procedures and what they can do, in general, depend on the pattern of missing data. For example, when data are MNAR, all techniques that handle missing data performs poorly in estimating the missing data. Although there are some analysis models that have been developed for MNAR data, such as, selection models and pattern mixture models, these models still soften some pattern assumptions and are not yet well-suited for widespread use (Baraldi & Enders, 2010).

The first option for handling missing data and the most common one is listwise or case deletion. When using this technique, any record with a cell with a missing value will be completely excluded from the analysis. Although this technique seems to be a quick and easy

solution, especially when having enough sample size to reach the desired statistical power, it still can produce bias results (Donner, 1982). For example, using the experimental clinical trial example that was used before, if this technique was used to delete all the records of the subjects that dropped out of the treatment group, the analysis could falsely conclude that the treatment is very effective because the only remaining subjects in the treatment group are the ones that benefitted from the treatment. Therefore, this technique should only be used when the sample size is large enough, and the data are MCAR.

The second option for handling missing data when the sample size is not large enough is pairwise deletion. When using this technique, data records with missing values will still be used in the analysis and only the fields with missing data will be excluded. Although this technique provides a solution for small sample size datasets and MAR data, it still could produce biased parameters since each parameter was estimated with a different sample size and standard error (Kim & Curry, 1977). In this case, the analysis will be deficient and meaningless (Kang, 2013). However, pairwise deletion is still an option if there are not many missing data points.

The third option for handling missing data is to substitute the missing values with the mean value from all subjects for the missing field. Although the mean is theoretically a valid estimation of any randomly drawn observation, there are still many assumptions to be fulfilled before using this technique. To be able to use the mean substitution, the data must be normally distributed, and the missing data should be MCAR. Even more, because using the mean substitution will not add any new information and will only make the sample size larger, the mean substitution is generally not acceptable (Malhotra, 1987).

The fourth option for handling missing data is by conducting regression imputation. When using this technique, missing values are predicted by a regression equation that includes

the other available observations. Like the mean substitution, this technique will not add any new information and will only inflate the sample size and reduce the standard error. However, building on the idea of predicting the missing values from the available ones, the fifth technique, which is called multiple imputations (MI), predicts the missing values from various combinations and probabilities of the available ones. MI is an iterative process of conducting missing values predictions using regression equations along with some variation introduced by randomly selecting values from a standard normal distribution (Rubin, 2004). MI produces multiple datasets from each iteration and then conduct the standard analysis for complete data. Next, it combines the estimated parameters in the analysis results into a final set of estimated parameters. Another MI method is MI by Chained Equations (MICE). MICE in an iterative method that estimates each column with missing values separately using all the other columns. The method stops and produces the imputed dataset after no more changes in the values occur (Buuren & Groothuis-Oudshoorn, 2011).

The sixth option for handling missing data, especially in SEM framework, is using Maximum Likelihood (ML) and Expectation-Maximization (EM). ML basically estimates the parameters of the model using the available observations. Next, a reverse engineering is conducted to predict the missing values based on the parameters that were just estimated. EM is an extension of ML by making the ML process more iterative. EM has two estimation steps, Expectation (E) and Maximization (M). In the expectation step, EM estimates the means and the covariance matrix of the dataset using the available data (Dempster, Laird, & Rubin, 1977). In the maximization step, EM uses ML to maximize the probability of predicting the missing values using the estimated parameters. These two steps are repeated until there is no change in the means and the covariance matrix, which are used to perform the SEM analysis. Although this

algorithm has some issues with its stability in estimating standard errors for all parameters, Yuan and Yuan and Bentler (2000) have proposed methods for getting consistent standard error estimates that are also robust to departures from normality.

In conclusion, choosing which technique to use in case of having missing data depends on the pattern of the missing data, the desired analysis to be performed on the dataset, and the sample size. Listwise and pairwise deletion methods should only be used when there are few missing data, and their missingness is random. MI has shown its ability to produce predicted values that accurately reflect the true observations. The main advantage of MI and its different methods is that they reproduce the data set with some variability that reflects the uncertainty of the estimation method. Furthermore, MI can be used in any data analysis and is robust against the departure of normality. However, in SEM framework, ML and its algorithms are a better choice to deal with missing data. The only situation that MI should be used in SEM is when not using ML as an estimation method. ML main advantage over MI is that it produces determinate results and performs the missing data estimation and the model fitting simultaneously (Allison, 2003). This has enabled the development of robust methods to implement ML and its algorithm to account for the assumption of normality (Yuan & Bentler, 2000).

2.3.2 Factor Scores

EFA is the first step in identifying and detecting the underlying constructs of a certain set of variables. After EFA, researchers have the option on how to proceed with the results of the EFA. The first, and arguably the best option to go with, is to conduct CFA to construct a model with latent variables. The relationships between these latent variables can be then investigated by

conducting a path analysis or SEM. However, this option is not always feasible, especially when having a limited sample size. Therefore, researchers who seek to investigate the relationships between unobserved variables have the option to calculate a score for each factor depending on the EFA results. Although the statistical analyses that can be conducted using these factor scores, such as ANOVA and multivariate linear regression, usually require smaller sample size than SEM, there are some sacrifices to be made when using these techniques. The main one is that SEM controls for the measurement error of each item in the model (Bentler & Weeks, 1980). This advantage will be lost or reduced to some degree when using factor scores, depending on the chosen factor scores calculation method. Nevertheless, factor scores provide many features that SEM have and can serve as a good alternative to SEM (DiStefano, Zhu, & Mindrila, 2009).

Generally, there are two types of methods for calculating factor scores. The first type is the non-refined methods, which are the methods that are simple and only perform cumulative procedures to create a score for each individual in the analysis. The second type is the refined methods, which are the methods that use more sophisticated computations and produce scores that represent a linear combination of the items. The main advantage of these methods is that they consider the common variance between each item and each factor as well as the measurement error in the items.

The simplest non-refined method is to sum all the items that load on each factor. This heavily depends on the cutoff value that the researcher uses to determine whether an item loads on a factor. By using this method, the scale and variability of the original raw scores will be preserved. However, by using this method, all the items will have the same weight in the factor score, which means that an item with a loading close to the cutoff value will have the same effect on the factor score as an item with a loading close to one. Another approach to implementing this

method is to standardize the items raw scores prior to calculating their summation. This approach is beneficial when the scales of the raw scores are different. Another non-refined method is to give each item a weight before calculating their summation. This method can be implemented by multiplying each item score by its loading on the factor. The researcher here can choose whether to include all the items in the dataset in calculating each factor since each item is weighted by its loading or only to include the items that load higher than a cutoff value on each factor. This method should be used with caution since the used extraction and factors' rotation methods in the EFA have a direct impact on the item loadings and thus have an impact on the factors scores.

In the refined methods, the calculated factor scores usually are standardized in a scale similar to a Z-score metric, where scores range between approximately -3.0 and +3.0. These methods tend to maximize the validity of factor scores and retain the relationships between the factors. This means that if the used rotation method in EFA was orthogonal, factors scores should be uncorrelated. The first refined method is the Regression Scores. This method creates a regression model where all the items are predicting the factor scores (Thurstone, 1934). The main advantage of this method over the weighted item sum is that it considers the correlation between factors in the case of oblique rotation and the correlation between the items themselves by multiplying the correlation matrix of the raw variables, the items loading matrix, and the factors correlation matrix to find the regression coefficients for each item. The second refined method is the Bartlett factor scores. This method uses maximum likelihood estimation to estimate the factor scores based on the row vector of observed variables, the diagonal matrix and the factor pattern matrix of loadings (Bartlett, 1937). The main advantage of this method is that it produces unbiased factors scores that truly represent the unique and common score of a factor from the set of variables that load on it (Hershberger, 2005).

3.0 METHODS

3.1 RESEARCH DESIGN

One aim of this dissertation is to demonstrate comprehensive data management and analysis procedures of large, complex, and unstructured research datasets. This demonstration will provide solutions to issues that research projects face due to using complex and large datasets. In order to get access to such datasets, a collaboration was established with McNeil et al. (2014) on their research project, that was sponsored by the U.S. Department of Veterans Affairs (VA) where they collected data from PWA on many tasks. They have collected these data from multiple PWA populations in multiple sites around the United States of America. A main benefit of this collaboration is the minimization of threats to validity by using data that is collected by reliable clinicians who have been working with this population for a long time and by using established and valid instruments to assess the study variables. Furthermore, this dataset fits the purpose of this health informatics dissertation since it contains all the required tasks to demonstrate a comprehensive and informative data management and analyses project.

This collected dataset is complex and unstructured due to the use of advanced software to collect it. Also, because McNeil et al. (2014) project was sponsored by the VA, there were many regulations that had to be met before any data sharing or analysis. This allowed to demonstrate how complex, large, and unstructured datasets can be de-identified and made ready for data

sharing and analysis. Furthermore, data extraction procedures that extract and transform data from their original unstructured files to a structured and analysis-friendly database were developed. These extraction procedures were used to demonstrate the manipulation and extraction of complex and unstructured datasets using novel data mapping techniques that can be customized based on the structure of the original data. In addition, novel data-quality assurance procedures that flag any abnormalities in the data, whether they were caused by inconsistency in data collection or errors in data extraction, were developed. Such quality assurance procedures cannot be conducted without the use of computerized algorithms and special reference materials.

As an alternative to the unstructured data files that the dataset is initially stored in, an internet-based research data management system (iRDMS) that stores the dataset in a structure that is suitable for data analysis was developed. This database is storing each task independently from other tasks, which gives it the flexibility to accept newly collected data from new populations without changing the schema of the entire database. Also, this flexibility is enabling and will enable researchers who are interested in only a portion of the dataset to access this database and get data related to their desired tasks only.

This database is stored on an online server and is accessible to researchers by a user-interface that comes with access control and user privileges procedures. This user-interface was designed to be easy to use by researchers and was evaluated by conducting a usability study. This study was used to gain insights on whether the suggested database and its interface are easier, faster, and more accurate methods for data retrieval than the source data files. The centralization of this database and the protection techniques that it is protected by have provided evidence that the suggested data management technique ensures high levels of data security for the collected dataset. The centralization of this database gives more control on the quality of the

stored data since all the entered data go through quality assurance procedures prior to their entrance. Even more, this centralization enables users to always access the latest version of the database instantly whenever it is updated, which also solves issues regarding data exchange between the project's investigators.

To demonstrate that the conduction of complex and detailed data analysis on complex and unstructured datasets is feasible using the proper data extraction and management techniques, a statistical model was developed and used to test complex hypotheses and answer detailed research questions using extracted data from the original dataset. Because of having unobserved theoretical variables, SEM was used to build statistical models. SEM is a collection of statistical techniques that powerfully represent the relationships between unobserved latent variables and the observed variables that they represent. Furthermore, SEM is the ideal analysis choice for eliminating the irrelevant noise that each measured variable carry, because it could use factor analysis which captures the common variance between the observed variables and ignores the unrelated ones (Joliffe & Morgan, 1992). In addition, because of the complex and nested nature of the theoretical cognitive model, SEM can handle that by estimating all the parameters in the model simultaneously, which helps it to detect the effect of a change in one variable on all the other variables (Byrne, 1994). Even more, SEM is a procedure with multiple steps where each step was used to answer some of the research questions. Also, in case the prior hypotheses were not accepted by the SEM model, SEM will help to understand the relationships between the variables with insights from the literature.

3.2 PROCEDURES

3.2.1 Data Integration

As described in the research design section, the data collection was conducted in multiple locations. The collected data was synchronized and stored in a shared location. However, the files from each site were separated from the files with the same type that came from the other locations. Therefore, procedures that integrate the files with the same type from multiple locations into manageable files that can be scanned and accessed easily were developed. Also, these integrated files were converted into formats that were suitable for the next data processing procedures.

3.2.2 Data De-Identification

The first step in the data analysis was to make sure that the data did not violate any regulations regarding the privacy of health-related data. Due to the VA's restrictions regarding data storage and sharing, data de-identification procedures that remove any identifiers in the data set were developed. These de-identification procedures scan through the whole dataset and identify all the fields that have dates as their content. Since there was more than one file in this dataset, each file had a de-identification procedure that was developed to access the dates fields in that file. However, since there were some tasks that produce their results as databases, de-identification procedures that identify the dates' fields by screening through the schema of the database were developed. In these de-identification procedures, the first day or the first session where the

subject started the experiment to be day number one for that subject was used. This information was provided in a screening file where all the subjects' information and session dates were provided. Then, a dictionary that consists of subjects' identifiers as keys and their first-day dates as values were generated and used to retrieve the first day of each subject throughout the whole de-identification process. Next, all the dates for that subject were representing the number of days between them and the first day. This de-identification method was chosen to be conducted instead of removing the dates entirely from the dataset to keep information about the time and order of each item and session in the collected data, because the existence of these information would open opportunities for future research that investigate the time factor and its effect on the subjects' performance. Finally, the de-identification process did not affect the time of the items and session since time without a date cannot be used as an identifier. After the de-identification of the dataset, it was moved to data processing and extraction.

3.2.3 Data Extraction

In this research project, participants' information and performance were either collected manually by the clinicians or automatically by special software. In the first scenario, the files that have been used to collect the participants' data were formatted in a structure that was friendly to the clinicians. However, the structure of these files was not suitable for hierarchal storage in databases since they contain irrelevant data, such as instructions on how to use the files, expected values in each field, and empty cells that separate one field from another. For example, the pre-screening exams were collected manually and each subject had their data in a separate file with a standardized format across all subjects and all sites. Therefore, an extraction algorithm that was

trained to scan each file and extract the data from specific fields had to be developed. Demographic data, on the other hand, were collected in an organized structure and were ready to the data quality assurance and transferring procedures.

In the second data collection method, clinicians have deployed special packages of software to present the experiment's stimuli and collect the responses from the participants. The first software is called "E-Prime", which was developed by Psychology Software Tools, Inc. to be used in computerized behavioral experiments and research. E-Prime is a widely-adopted software in more than 60 countries and has been used to present, collect, and analyze experimental data. However, when conducting advanced statistical analysis, E-Prime does not have the capability to analyze the data using such methods. Therefore, data had to be collected from E-Prime and prepared to be analyzed using some external analysis software. Nevertheless, when E-Prime presents the stimuli to the participants, it collects data regarding their performance along with data that are for internal uses or for files tracking purposes. This latter type of data is irrelevant to the data analysis and can be a source of confusion and unnecessary space occupation. Even more, when E-Prime is being used to present experiment stimuli of multiple tasks, it collects the data for all the tasks and outputs them into one file. For some tasks, it even outputs part of their scores in separate files. In the E-Prime file, each row represents the participant response to one item that belongs to a certain task. For the dataset in this dissertation, E-Prime created this file with over 648 columns where each task had a group of these columns dedicated to store its data. As a result, the output file from E-Prime was full of empty spaces or null values since one column can only have data from one task but not another, and some tasks were missing some of their values because they were produced into separate files, which poses data extraction and preparation challenges.

The main extraction challenge in this situation was that columns cannot be directly copied to the database because this would have caused a massive data loss and many empty data fields or null values. Therefore, the extraction of the data from E-prime file into new files or database that is suitable for data analysis was impossible to be done manually or using the traditional data extraction methods, such as Visual Basic for Applications that comes with Excel or the import features in most DBMS. As a solution to this issue, a novel mapping technique that acted as a map was developed to guide the extraction algorithm that was developed for this specific purpose. This map was created as an Excel sheet that can be accessed by the extraction algorithm that was developed using Python programming language. The basic idea behind this map is to fill each task table in the database with its data from the E-Prime output file. Therefore, each column in this map is representing a new column in the database, and each row acts as an extraction guide for one specific task. Thus, the content of each cell in the map is indicating where to get the data from the E-Prime file for the new columns for one specific task. Furthermore, there are many different contents that cells could hold which makes this approach flexible for multiple types of data extraction. The extraction algorithm recognizes each one of these cells and acts differently based on their content.

The algorithm was designed to read through the E-Prime file row by row, and use the map to learn what data to extract and where to store them in the new file or database. In the map, there is a column that is used to guide the extraction algorithm to a specific row in the map based on the processed row, which is called the “hook” column. For example, if the algorithm is processing an item from the STMsem2 task, it will search for the row in the extraction map that have STMsem2 in its hook column and uses that row to guide the extraction procedure. There are six different contents that the extraction algorithm can recognize: Column Name, Position,

Condition, External File, Equation, and Text. In the case of Column Name, the algorithm finds a column name in the content of the cell, which leads it to go directly to that column in the E-Prime file and gets its data for the currently processed row. In the case of a Position, the algorithm finds directions to a specific cell in the E-Prime file. For example, “Fixation1.OnsetTime[Block] for (Running[Block] = Practice1 & Practice1.Sample[Block] = 1)“, means get the data from column Fixation1.OnsetTime[Block] where column Running[Block] equal to “Practice1” and column Practice1.Sample[Block] equal to “1”. In the case of Condition, the algorithm processes a certain condition to determine the content of the new record. For example, ““Prac” if Running[Trial] = “PracTrialList” or “Exp” for Running[Trial] = “TrialList””, means put “Prac” as a value of the new field if the value of Running[Trial] for the processed row is “PracTrialList”, if not, the algorithm goes to the next condition. In the case of External File, the algorithm is asked to get the data from an external file, because E-Prime outputs some values for some tasks in external files. For example, “ReadT from SubjectXXXX_SessionX_LPsyn2a.txt”, means get the content in column “ReadT” from the file that belongs to subject “XXXX” and session “X” for task “LPsyn2a”. In the case of Equation, the algorithm is asked to perform a simple calculation on the values of more than one column in the E-Prime file and put the results into the content of the new field. For example, a content as “Stimd1.OnsetTime - Waitc1.OnsetTime“, tells the algorithm to put the result of Stimd1.OnsetTime - Waitc1.OnsetTime in the new field. Finally, there were some new fields that should be filled with fixed values that were not from the E-Prime file. The Text content tells the algorithm to copy the exact content of the cell to the new field. The content of these cells can be changed to almost anything to serve most of the extraction and mapping purposes. The output of this extraction procedure was a file that was smaller and had significantly fewer spaces and

null values than the E-Prime file. In addition, tasks that were divided into separate files were all organized and integrated into this new file. This file, then, was used to perform data evaluation and transferring to the database.

Another data that was collected automatically was the one related to the CRTT tasks. As mentioned in the background section, these tests were provided through a special software that was developed by Eberwein et al. (2007). The software stores the data in an SQLite database that contains more than 35 tables. Some of these tables contain commands that show the software how to represent the data to the participants, which were not collected in this project since they were not useful in any data analysis. Other tables, however, contain data related to the subjects' performance and descriptions of the items in each one of the tasks. Since the CRTT database schema and structure differ from the proposed database schema and structure, some algorithms that access the CRTT database and extract data from several tables to get information regarding subjects' performance and descriptions of the items for each task were implemented. The table [Scoring], which contains data about each item, was accessed and extracted item by item along with data about the participants who performed these items, description of these items, the tasks that these items belong to, and the scores on these items, each information from its separate table. Next, all the acquired data were treated the same as the E-Prime file and were passed to the next step.

3.2.4 Data transferring

One of the main challenges that large research projects face is the transformation of data from their original source files to well-structured databases. This task was started in the extraction

phase where data were extracted from their source files and organized into more structured files to be used in the quality assurance procedures. After data was checked and assured to be correct and complete, they were passed to transferring algorithms that acted as a bridge between the database and the processed files. These transferring algorithms and scripts were created to access each line or record in each file and transfer it to its appropriate position in the database. Since the newly processed files were organized, and since the transferring procedure was standardized for each task, they did not need dynamic maps that specify a special treatment for each task. The first data that were transferred were the demographic data, where each subject had a row in the subjects table and a unique identification (SubjectID) number. In the E-Prime file and the CRTT data, an algorithm, first, scanned the records to capture data that describe the tasks' items, which were the same for all subjects. Therefore, the items for each task were captured and stored in tables with standardized names "[TaskName_Items]". Then, data regarding subjects' sessions and tasks order were transferred to the database and connected with their unique ID. Next, the algorithm scanned through the E-Prime file and the CRTT and transferred data regarding each item in each task to their appropriate table in the database, which has standardized names "[TaskName_Scores]". After transferring all the records to the database, the database was ready for final scores calculations and analysis. Each one of these steps was evaluated to assure the accuracy of the data transferring.

3.2.5 Procedures Evaluation

Each one of the data preparation and processing procedures was evaluated to make sure that the desired outcomes were met. These evaluation methods did not only ensure that the extraction

methods were working properly, but they also ensured that the data collectors (the clinicians and the systems) were collecting the data in the correct and standardized format. In addition, these evaluation methods helped to reduce missing data since they flag empty data items when it is expected for them to have some data.

In the de-identification phase, there was a generated warnings file that stores any errors during the de-identification process. The de-identification algorithm used this file to provide warnings when there were unexpected errors in the process or if an inconsistency in the dates between the first-day dictionary and the encountered date has occurred. For example, when the algorithm encounters a date for the fourth day but cannot find the first-day date, it writes in the warnings file that this subject has a fourth session date, but his first date is missing. Therefore, this algorithm was built not to output any data unless it was de-identified. Otherwise, it flags this datum and removes it from the new de-identified output. This warning system serves many purposes. For example, it ensures that the privacy of the data was not violated even in the case of errors. Furthermore, it helps the clinicians to ensure that all files were filled up to date and that no fields were left behind empty. In addition, random samples of the de-identified data were selected and examined manually to ensure their date calculation accuracy and their clearness of any actual dates. The de-identification algorithm was refined and corrected based on the insights from these evaluation procedures.

The evaluation of the extraction procedures is highly important since these procedures have a high potential impact on the data quality. Therefore, the evaluation of these procedures has examined every detail in the extracted data automatically. These evaluation procedures started during the extraction process where the extraction algorithm inserts the word “Missing” whenever it encounters an empty field that should contain some values. Furthermore, since there

were multiple tasks that were extracted from the E-Prime file and other files, a double-checking map that had a similar structure as the extraction map but had different contents was developed. Thus, this map was created as an Excel sheet and was accessed by an evaluation algorithm that was developed using Python programming language. Each column in this map represents one of the new columns in the new file, and each row acts as an evaluation guide for one specific task. Therefore, each cell in the map contains specific conditions that indicate whether something is wrong with the extracted data or not. These conditions were developed based on the values or the output that each column for each task (each cell) is expected to have. This technique is highly customizable to accept any conditions and could be applied to different types of data.

In this map, there are eight conditions that test whether the content of a certain cell is what it is expected to be. These conditions are: the values of the cell should fall between two values, all larger than or smaller than one value, all equal to one or range of values, half of the values are equal to one number and the other half is equal to another one, all values should be the same, the count of the items in the cell are equal to a specific number, cell could contain empty or missing values, and cell should not be empty or missing in case the content of that cell was not numeric. Each cell can contain one or more conditions; one satisfied condition means that this cell is correct. Since there was a massive amount of data, the most convenient way to view numeric data was by plotting them. Therefore, all the cells that did not satisfy at least one condition were plotted as stem and leaf plot along with details on what violations have occurred. However, when the cell contains non-numeric data, its content was output in a text file along with details on what conditions were violated. Again, samples of data were selected randomly and were traced back to their source file to ensure their accuracy and correctness. All these evaluation procedures were conducted on all the extracted data before inserting them into the

database. Furthermore, the outcomes of these evaluation procedures were used to refine the extraction algorithms in case the errors were found to be related to them.

The main goal of the transferring algorithms was to transfer data from their source files to the database without missing or repeating any of them. Therefore, these algorithms were evaluated by selecting random information from the source files and matching them with the data from the database, where a match means that the content is identical, and the position in the database is appropriate. Also, this randomly selected data should occur only once in the database, which ensures that there was no redundancy in the transferring procedure. The same evaluation procedure was conducted multiple times and in both directions, from source files to the database and the other way around to ensure that there were no data in the database from outside the source files.

3.3 DATA DESCRIPTION AND MANAGEMENT

3.3.1 Measurements

3.3.1.1 Sentence Comprehension

In the SEM four tasks that measure the sentence comprehension success for four sentence types were included.

Compound sentences:

- **CRTT-R subtest 4 (CS):** Participants read 20 sentences, such as “Touch the big blue circle and the little green square on this trial,” and were asked to perform the request in

the sentence. A score on a 15-point multidimensional scale was given to each one of the constituents in the sentence depending on whether the participant performed the request accurately in terms of the color and the shape of the objects. A final score was calculated by measuring the average of all the sentence constituents' scores.

OC sentences:

- **CRTT-R_{OC} (OC):** Participants read 20 OC sentences, such as “It was the blue circle that the green square touched on this trial,” and were asked to perform the request in the sentence. A score on a 15-point multidimensional scale was given to each one of the constituents in the sentence depending on whether the participant performed the request accurately in terms of the color and the shape of the objects. A final score was calculated by measuring the average of all the sentence constituents' scores.

GP sentences:

- **CRTT-R_{GP} (GP):** Participants read 20 GP sentences, such as “The blue circle touched by the green square is above the green circle on this trial,” and were asked to perform the request in the sentence. A score on a 15-point multidimensional scale was given to each one of the constituents in the sentence depending on whether the participant performed the request accurately in terms of the color and the shape of the objects. A final score was calculated by measuring the average of all the sentence constituents' scores.

Lexical Ambiguity sentences:

- **Lexical Ambiguity (LA):** Participants read 28 simple sentences followed by a picture and were asked to judge whether it relates to the sentence or not. Half of these sentences included an ambiguous object toward its subordinate meaning with a verb that disambiguates it, such as “He drank the port quickly.” The other half included

unambiguous objects, such as, “He drank the wine quickly.” The score was calculated by measuring the response times and accuracy.

3.3.1.2 Cognitive Systems and Functions

In the SEM twenty one items or tasks that measure LP, STM, and CR were included.

Phonological LP: three tasks that are hypothesized to load on phonological LTM and language processes were included.

- **Rhyme Judgment Words & Non-words (LPphon1)**: Participants hear 2 non-words in a row and decide if they rhyme. The score will be calculated by measuring the response accuracy.
- **Rhyme Judgment Written Non-words (LPphon2)**: Participants see two written non-words in a row, and decide if they rhyme. The first word is displayed for maximally 4 seconds, but participants were instructed to push a button as soon as they were ready for the next word. The score will be calculated by measuring the response accuracy.
- **Rhyme Judgment Pictures (LPphon3)**: Participants see two pictures in a row, and decide if the words for those pictures rhyme. The first picture is displayed for maximally 4 seconds, but participants were instructed to push a button as soon as they were ready for the next word. The score will be calculated by measuring the response accuracy.

Semantic LP: three tasks that are hypothesized to load on semantic LTM and language processes were included.

- **Category Judgment (LPsem1)**: Participants hear 2 words in a row and decide if they are in the same category. The score will be calculated by measuring the response accuracy.

- **Pyramids and Palm Trees (LPsem2):** a stimulus word is displayed on the top of the screen, two probe words on the bottom. The words on the bottom are more or less associated with the word on the top. Participants decide which of the two on the bottom go with the one on the top. The score will be calculated by measuring the response accuracy.
- **Neutral items from Word/Picture interference (LPsem3):** participants see words in neutral shape and judge if it is a living thing or not. The score will be calculated by measuring the response accuracy.

Syntactic LP: three tasks that are hypothesized to load on syntactic LTM and language processes were included.

- **Grammaticality Judgment (LPsyn1):** Participants will listen to 120 sentences and decide whether these sentences well-formed or have grammatical violations (hierarchical syntactic structure or morphosyntax). The score will be calculated by measuring the response accuracy.
- **Sentence/Picture matching (LPsyn2):** Participants hear sentences (stimulus) and decide if a subsequent picture (probe) matches the sentence. The score will be calculated by measuring the response accuracy.
- **Anagram Test (LPsyn3):** Participants will be presented with 30 sentences where each word is written on a small card, and they are asked to arrange these cards to form well-formed sentences. The score will be calculated by measuring the response accuracy on canonical sentences.

Phonological STM: three tasks that are hypothesized to load on phonological STM were included.

- **Rhyme Probe Span (STMphon1):** Participants will listen to lists of words starting with two words per list and go up to 8 words per list depending on the participants' performance. After listening to the list, participants will listen to a word and decide whether it rhymes with any word in the list. If subject scores below 75% correct rate, the tasks will stop. The score will be calculated by measuring the number of correct lists other than list one.
- **Rhyme Judgment with a filled interval (STMphon2):** Participant will listen to 40 pairs of words. Words in each pair is separated by 5 seconds interval where participants are required to say numbers that are viewed on the screen out loud. This interval eliminates the rehearsal of the first word and forces the use of STM. Participants will decide whether the second word rhymes with the first one. The score will be calculated by measuring the response accuracy.
- **N-back phonological (STMphon3):** Participants will listen to a list of words. They will decide whether the current word rhymes with the one two items back. The score will be calculated by measuring the response accuracy.

Semantic STM: three tasks that are hypothesized to load on semantic STM were included.

- **Category Probe Span (STMsem1):** Participants will listen to lists of words with starting with two words per list and go up to 8 words per list depending on the participants' performance. After listening to the list, participants will listen to a word and decide whether it matches the category of any word in the list. If subject scores below 75% correct rate, the tasks will stop. The score will be calculated by measuring the number of correct lists other than list one.

- **Category Judgment with a filled interval (STMsem2):** Participant will listen to 40 pairs of words. Words in each pair is separated by 5 seconds interval where participants are required to say numbers that are viewed on the screen out loud. This interval eliminates the rehearsal of the first word and forces the use of STM. Participants will decide whether the second word matches the category of the first one. The score will be calculated by measuring the response accuracy.
- **N-back semantic (STMsem3):** Participants will listen to a list of words. They will decide whether the current word matches the category of the one two items back. The score will be calculated by measuring the response accuracy.

Syntactic STM: three tasks that are hypothesized to load on syntactic STM were included.

- **Sentence Probe Span (STMsyn1):** Participants will listen to lists of sentences starting with 2 sentences per list and go up to 8 sentences per list depending on the participants' performance (e.g., 'The woman kissed the man'). After listening to the list, participants will see a picture and decide whether it fits any of the sentences in the list. All sentences will hold the same semantic words but will differ in the used verb and the syntactic role, which reduces the burden on semantic memory. If subject scores below 75% correct rate, the tasks will stop. The score will be calculated by measuring the number of correct lists other than list one.
- **Sentence Picture Matching with a filled interval (STMsyn2):** Participants will read 40 sentences and decide whether the picture following the sentence matches the sentence or not. After the 3rd word in the sentence, there will be a 5 seconds interval where participants are required to say numbers that are viewed on the screen out loud. This

interval eliminates the rehearsal of the first word and forces the use of STM. The score will be calculated by measuring the response accuracy.

- **N-back syntactic (STM_{syn}3):** Participants will listen to a list of active or passive SVO sentences (e.g., ‘The doctor kissed the banker’ or ‘The banker was kissed by the doctor’). They will be asked to decide whether the current sentence is similar in syntactic structure to the one two items back. The score will be calculated by measuring the response accuracy.

CR: three tasks that are hypothesized to load on CR were included.

- **CRTT-R Stroop (CR1):** Participants will read 60 sentences such as “Touch the red circle,” where they are required to touch the red circle among 10 items with different shapes and colors. In the congruent mode, the word “red” will be written in a red font. In the incongruent mode, the color name will be written in a different color, and subjects will be required to touch the item that is similar in color to the font color. In the neutral mode, the whole sentence will be written in a black font. The score will be calculated by measuring the accuracy on the color adjective in the incongruent compared to the control condition and the CRTT-R score.
- **Picture-Word Interference (CR2):** Participants will be shown pictures with words written on them and will be asked to perform semantic judgments on whether the word inside the picture is describing the semantic category of the picture or not. In the congruent mode, the picture always matches the word on it, in the incongruent mode, the

picture is paired with a word from a different semantic category, and in the neutral mode, the word is presented on a polygon, which eliminates any interference caused by lateral masking. The score will be calculated by measuring the response accuracy for the recognition of the word representing the predetermined semantic category.

- **Number Stroop (CR3):** participants will be shown blocks that contain numbers (from 1 to 4) and will be asked to respond with how many numbers there are in each block. In the congruent mode, the numbers will match their quantity (e.g., the digit 3 is displayed three times – as in 333), in the incongruent modes, the quantity will not match the numbers shown, and in the neutral mode, the numbers will be substituted by the character “X”, which eliminates any confliction. The score will be calculated by measuring the number of accurate responses.

3.3.2 Database

3.3.2.1 Design Decisions

During the development of the database, many functional and non-functional factors that may influence its design were considered. In this dataset, all tasks or measurements were independent of each other, which means that each task can stand alone and can be used in a separate data analysis. In addition, each task has its own characteristics in terms of the number of items needed to complete that task and the number of columns needed to score each item. Therefore, in order to eliminate having empty cells by design, separate tables for each task that satisfy its required data fields were created. For example, if each item in task A requires the [Reading-Times] and

the [Response-Accuracy] to a sentence, and each item in task B requires only the [Reading-Times] of a sentence, all items that belong to task B would have empty cells under the [Response-Accuracy] column in case both tasks were stored in one table. However, if these two tasks were separate into two tables, the table for task B would only have [Reading-Times] column and no empty cells, and task A would have a table that captures all its required data fields.

Furthermore, all subjects in the dataset went through the same items for each task, which means that the items' description data (e.g., provided stimuli, item order) were the same for all subjects. However, the scores for each item were different from one subject to another as each subject perform differently. If the scores of the items and their descriptions were stored in one table, there would have been many redundant and unnecessary data in the database. Therefore, this has influenced the database design to create separate tables for each task, where one stores the scoring data for each item, and the other stores the items' description data while connected by [ItemId] column.

One of the main advantages of the tasks in the dataset is that each task can be scored by different methods, each method captures a different aspect of the subjects' performance. However, not all tasks can be scored by the same set of methods. For example, the final score for task A can be either the average of the [Reading-Times] and the [Response-Accuracy] (calculation method RT+RA) or only the average of [Reading-Times] (calculation method RT). The final score of task B, on the other hand, can be calculated by either the average of [Reading-Times] on positive sentences (calculation method +RT) or the average of [Reading-Times] on negative sentences (calculation method -RT). One table for both tasks would mean that four columns are needed to capture the scores, half of which are empty for each task. This issue was

solved by having a separate “Final Score” table for each task where each column presents one possible calculation method for that task. Furthermore, this design decision along with the previous decision regarding the separate tables for items and scores gave the database the flexibility to accept new tasks with new characteristics without altering or affecting its original schema or design.

Since the collection of the dataset was sponsored by the VA, some of their privacy and confidentiality regulations have affected the database design decisions. Part of the VAs privacy regulations is that only researchers who were part of the data collection project have the rights to access the full dataset. This has influenced the database to be secured with user access privileges, where each user can only access the data that they were permitted to view. In fact, this feature might help to store data from more than one data collection project, in the future, without affecting the privacy of each data set. Another regulation is that the collected dataset must be stored in a server that belongs to the University of Pittsburgh. For this reason, MySQL was chosen to be used as the database management system, and the PHP platform to build the database user-interface because the University of Pittsburgh server well supports them. Furthermore, since the University of Pittsburgh server does not allow database alteration on the server and is limited in terms of the supported technologies, this database was built on a local server and then uploaded to the server when it was completed. Also, all the entries to the database were entered using python scripts that got the data from their original files. In addition, in order to keep the data consistent in the database and to protect the database from low-quality data, the database does not accept any entries from the online users except data from the database administrator (details in the user-interface section).

3.3.2.2 Database Design

In database design, there were three main phases that were followed: conceptual design, logical design, and physical design. In the conceptual design, the general relationships between the entities (tasks in this case) in the database and what attributes each table should have were specified (P. P.-S. Chen, 1976). Also, data were analyzed to find out the attributes that were the same for all subjects and the ones that were unique for each one and the tables were designed accordingly. In the logical phase, attributes to use to represent the relationships between the entities were specified (e.g., primary keys, foreign keys) and the database model was specified (e.g., relational database, NoSQL database) (Storey, 1991). As was expected, at least one full record as an example of the collected dataset was available at this stage. Therefore, this record was analyzed and used to specify the data types that each attribute expects and to add the fields constraints accordingly (e.g. non-null variables). Furthermore, at this stage, database normalization was performed by following the normalization forms to reduce redundancy and maintain the quality of the data (Codd, 1970; Dogac, Yuruten, & Spaccapietra, 1989). In the physical design, the database engine or software that was used to build the database was chosen (e.g., MySQL, SQLserver, MongoDB), the site where the database is going to be stored was selected and the requirements that site demands were collected. These decisions were critical since this dataset contains data that is owned by the US Department of Veterans Affairs (VA) and that their requirements for the host server must be followed. Also, at this stage, the database administrative roles of the research team were specified and the access control privileges were specified accordingly (P. P.-S. Chen, 1976).

As mentioned before, each task in the database has its three tables that store its related data. These three tables are connected, together, and to the other unique tables with database

relationships. However, each task and its tables are independent of the other tasks and not connected to any tables that belong to the other tasks. The first table for each task is [TaskName_Items], which holds data that describe the items in that task with a unique identifier for each item as the primary key. This identifier along with the “SubjectID” are the primary keys in the [TaskName_Scores] table, which holds data regarding subjects’ performance on each item. Lastly, the final scores for that task were calculated in the [TaskName_FinalScore] table, where each column represents one method of final score calculation and each row represents a single subject.

In addition to the tables of each task, there were four unique tables that store general data that are unified across all tasks. The [Subjects] table stores all the data regarding the subjects, such as age, education level, and date since the condition acquisition. The [Sessions] table stores data regarding sessions that each subject attended, their date, start time and end time. The [Tasks] table stores all the tasks that are in the database with their unique identifiers. The [TaskOrder] table connects the two previous tables by specifying which tasks were taken in which sessions. The below entity relationship diagram (ER) represents the relationship between the unique tables (Subjects, Session, Tasks, and TaskOrder) and the tables that belong to the task “STMsyn2” (STMsyn2_Scores, STMsyn2_Items, and STMsyn2_FinalScores) as an example of the relationship between the tasks’ tables and the unique ones (Figure 7).

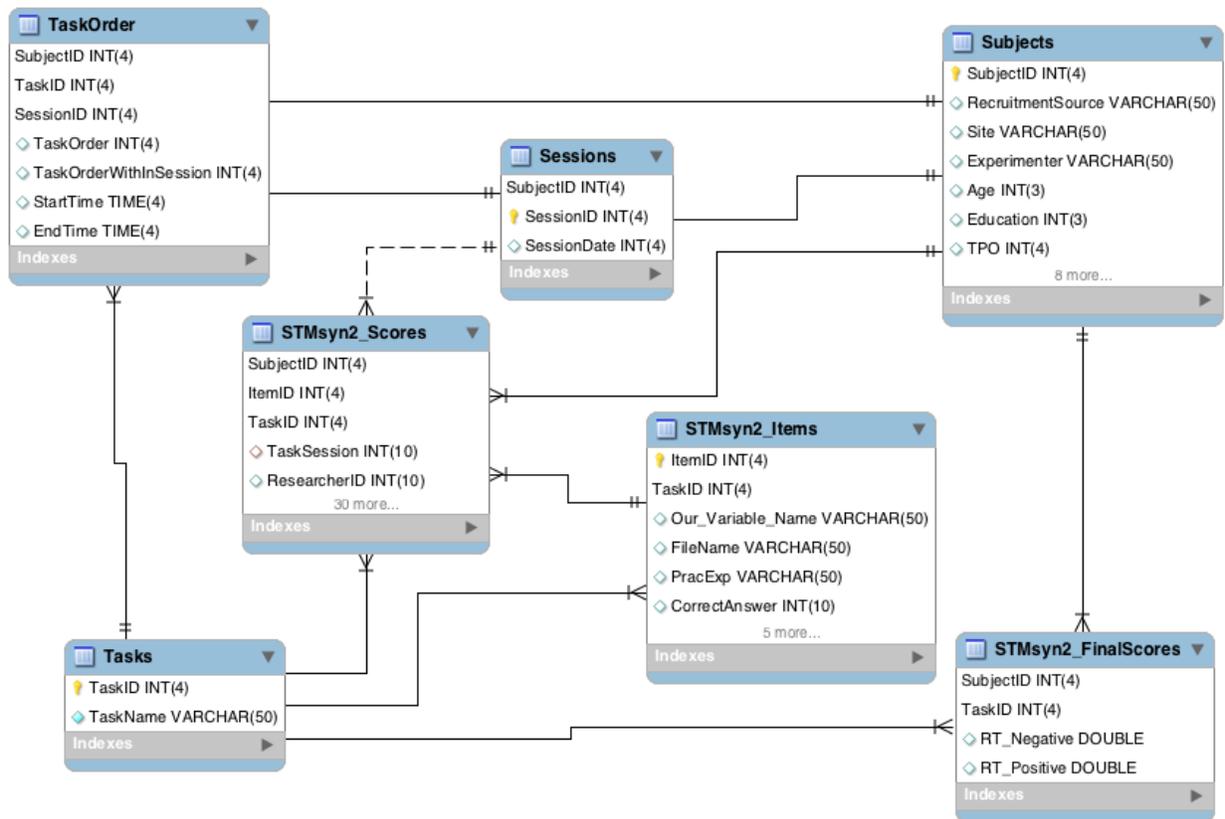


Figure 7. Database ER diagram.

This physical ER diagram shows the relationships between four unique tables (Subjects, Session, Tasks, and TaskOrder) and one task (STMsyn2). The tables of each task in the database have the same relationships with the unique tables but have different columns and types. Therefore, in order to avoid large, complex, and confusing diagrams, only this diagram is presented.

The database was created by python scripts that execute the SQL create-statements using the “mysql.connector” python library. Statements that created the unique tables, such as [Subjects], [Sessions], and [Tasks], was developed and stored in text files manually. However, statements that created the tasks’ tables, such as [TaskName_Items] and [TaskName_Scores], were developed by special algorithms. These algorithms accessed the extraction maps (that were described in the procedures section) and identified the columns that each task needed. Next, these algorithms went over a subset of the E-Prime file and learned the data types that each

column could have. Finally, these algorithms wrote the create-statements for all the tasks' tables into a text file that can be accessed by the create algorithms.

3.3.2.3 Database Evaluation

The database was evaluated by testing the schema of the database and the stored data in the database. Testing the schema included testing primary keys by evaluating their suitability to be used as primary keys, testing foreign keys by ensuring that they map to primary keys, testing that tables were connected by the attributes that truly represent the relationship, and testing the appropriateness of fields constraining. This test was conducted by running “DESCRIBE <table_name>” command to get the representation of the table in the database, and by generating schema diagrams based on these descriptions (Haraty, Mansour, & Daou, 2001). Furthermore, the schema of this database was evaluated using five normalization forms (Kent, 1983). These normalization forms were suggested as evaluation methods because they ensure that the database is properly organized. The main benefit of following the guidelines of these normalization forms is having a database with no redundant data across its tables. Redundant data in the database occur when having the same data in more than one table or in the same table in the database. Although this can be helpful in reducing the complexity of the query statements and in increasing their execution speed, data redundancy could increase the complexity of the database management, exposes the data to the risk of corruption, and increase the size of the database (Lee, 1995). Furthermore, because PHP technology and automated algorithms to create and execute the query statements were used, the complexity of the statements was not an issue. Moreover, one of the essential design features of the database is the independence of each task, de-normalizing the database would take away this feature and cause the tasks to overlap in their

data fields. Therefore, the initial design of the database was evaluated and reformed when one of the normalization forms was violated.

Testing the data, on the other hand, involved running multiple queries, that were progressive in complexity and the number of records they required, on the database and comparing their results to the data in the original data files. Although this technique is less common in database testing, it is extremely important in this case since data regarding subjects' performance were separated from data regarding the study stimuli, and this testing has confirmed that they were connected correctly (Mishra, Koudas, & Zuzarte, 2008).

3.3.3 User-Interface

One of the primary goals of this dissertation was to demonstrate practical and convenient data sharing, accessing and retrieval methods. This goal was achieved by developing a user-interface that enables users to access the database in a secure and convenient way.

Data management system design can be defined as “the process of capturing the relevant information and the processing requirements of an enterprise and mapping them onto an underlying database management system” (p. 479) (Dogac et al., 1989). The concept of data management system design or data modeling process started when American National Standards Institute, Standards Planning And Requirements Committee (ANSI-SPARC) introduced their first design standard (Tsichritzis & Klug, 1978). In their report, they divided the data management design process into three layers: external layer, where requirements and expectations are captured from the users of the system, conceptual layer, where entities within the system are represented and their relationships are specified, and the internal layer, where

logistics are planned for. This concept has evolved over time by breaking down these phases into more specific ones to reflect the advancements in technologies and workflows. In fact, nowadays, data management system design approaches consist of phases that can be added or dropped based on the nature of the project. In this system development process, the waterfall model was followed as a system development life cycle of this software. In this project, the process was divided into several phases: requirements collection, database conceptual design, database logical design and normalization, database physical design, interface design, interface development, testing, and finally prototyping (Royce, 1970). Since the waterfall model was adopted, the process has moved forward from one phase to another, with the exception of two cases: an error in one phase that was not detected until the next phase and a user feedback that requires going back to a certain phase and make the necessary changes (Connolly & Begg, 2005).

3.3.3.1 Requirements Collection

In requirements collection, several meetings with clinicians and data collectors who were expected to be the end users of the system were conducted. The users were asked to specify what functionalities they expect the interface to perform. Furthermore, in these meetings and interviews, information about the data that will be stored in the database, the possible uses of that data, and the expectations from the graphical user-interface were collected (Batini, Lenzerini, & Navathe, 1986). Also, once a functioning prototype of the system was available, a small internal validity testing (like the testing that was conducted in the usability study which will be discussed in the testing section) was conducted to give the end users a general idea of what the interface looks like, and to capture their feedback regarding the design. Insights from this phase were used

to fix usability issues and to improve the functionality of the system (Weitzel & Kerschberg, 1989).

3.3.3.2 Interface Design

The final design of this interface was a click-and-choose design that enables users to query the data without the need to be familiar with SQL. In fact, this interface does not accept any written SQL statements from users as a security procedure to protect the database from any unwanted changes. Although this interface does not accept written SQL statements, it was designed in a way that enables the users to access every data in the database. This was accomplished by making the content of this interface dynamic so that it reflects the content of the database. For example, if there were three tasks in the database, the interface would show three check boxes each represent one task. Also, users can customize their selection of data by specifying a range of values for one parameter or only retrieving the data for a subset of subjects. Furthermore, the interface offers the users the option to acquire some descriptive statistics (Figure 8). Finally, although some waterfall models inhibit moving backward between the phases, the outcome of this phase was presented to the potential users and was changed depending on their feedback as a design strategy rather than a requirements collection process.

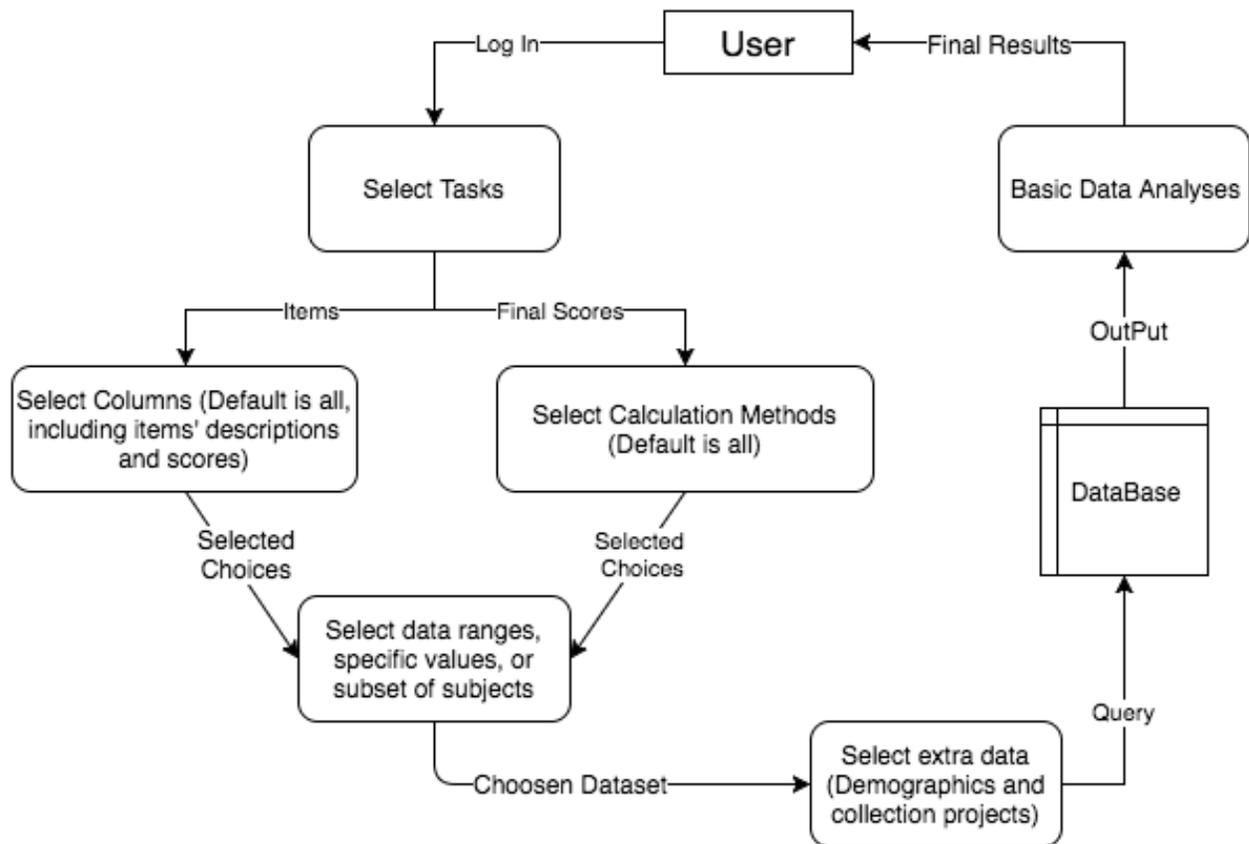


Figure 8. iRDMS Data Flow Diagram (DFD).

This Data Flow Diagram (DFD) shows the steps that users will take to get the final results from the database.

3.3.3.3 Interface Implementation

In the interface implementation, the user requirements for the graphical user-interface were implemented. Also, the requirements of the hosting server were considered by choosing a supported programming platform to implement the interface. Furthermore, this interface was developed and tested on local devices and then was made available on an online server. This interface was developed using “PHP 7.1.1” platform, and was integrated with “MySQL 5.6” server to access the database (Figure 9).

Select a Task To Retrieve

2 out of 5 steps

Pre Screening

- E-prime RT baseline
- SOAP

LP

- LPphon1-Rhyme Judgment Nonwords
- LPphone2-Rhyme Judgment written nonwords
- LPphon3-Rhyme Judgment Pictures
- LPsem1-Category Judgment
- LPsem2-Written Pyramids and Palm Trees
- LPsyn1-Grammaticality Jdgments (Linebarger)
- LPsyn2a-Sentence Picture Matching
- LPsyn2b-Verb Lexical Decision

Sentence Comprehension

- OUT4-CRTT-R Subtest IV
- CR1a-CRTT-R-fade
- OUT2-CRTT-R Garden Path
- CRTT-L
- OUT3-CRTT-R Relative Clause
- OUT1-Sentence-picture matching/semantic ambiguities

CR

Back Next

Figure 9. A screenshot of the iRDMS.
This was taken from the second step (Task choosing) under the detailed scores mode.

3.3.3.4 Interface Testing

To test this interface, the queries that were used to evaluate the database were used to get similar results using the written query and using the graphical user-interface. This helped to validate the query construction procedures in the background of the interface and to detect any failure in the system when a certain query is required (Lo, Binnig et al. 2010). Since the interface does not accept any input data, there was no need to test how the interface stores data in the database.

3.3.3.5 Usability Study

To evaluate the validity of the interface and to measure its usability, a controlled user usability testing was conducted. In addition, to interpret the effectiveness and efficiency of this system, and to answer the research question: whether web-based database with user-interface is more effective and efficient to be used for data retrieval and analysis than the original data files, the

effectiveness and efficiency of using original data files for data retrieval were measured and used as benchmarks. In this usability study, a controlled user testing method was used where a small sample of participants, who have some level of technical skills that enable them to use traditional software (Excel, Google Sheets), were asked to perform data retrieval tasks using two methods, the iRDMS, and Excel. Furthermore, the “thinking aloud” technique was followed where the participants were encouraged to say what they are doing, thinking, liking, and disliking about the two systems while performing the tasks (Ericsson & Simon, 1980).

In this usability study, the International Organization for Standardization (ISO) usability model was adopted. ISO FDIS 9241-210 defines usability as:” The extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.” ISO 9241-11 defines effectiveness as: “The accuracy and completeness with which users achieve specified goals” (ISO, 1998). Furthermore, Oxford dictionary defines effectiveness as “The degree to which something is successful in producing a desired result.” Therefore, since effectiveness is purely goal driven measure, the main criteria to use in data management system effectiveness evaluation is whether or not participants can complete a set of specified tasks in a fashion that meets the user requirements and the goals of developing such a system (Harrison et al., 2013). The main goal of developing the iRDMS was to solve challenges that researchers face when dealing with large-scale and complex datasets. One of the main characteristics of complex datasets is the heterogeneity of the original data files. This means that each one of these files needs a different technique to be extracted and analyzed (Karpathiotakis, Alagiannis, & Ailamaki, 2016). This system would be considered effective in solving this issue if the participants in the usability study managed to get data regarding different tasks using the same interface or procedure.

Furthermore, since completeness is an essential part of system effectiveness, and since errors in data is one of the challenges that researchers face when handling complex datasets, the rate of errors which users make when performing the tasks can be used to measure task completeness and system effectiveness (Frøkjær et al., 2000; Gil et al., 2010). Therefore, the systems' effectiveness in reducing the rate of errors was evaluated by measuring errors in the retrieved data for each task by the usability study participants.

Another issue that researchers face with complex datasets is how to format the data to be analysis friendly. Analysis friendly means that the data does not need any further formatting or cleaning to be analyzed (Dipnall et al., 2014). Furthermore, even if researchers managed to format the data to be analysis-friendly, they usually suffer from magnificent time loss to accomplish this task, which is dependent on the efficiency of the method used for the data management and retrieval (Streit, Schulz, Lex, Schmalstieg, & Schumann, 2012). ISO 9241-11 defines efficiency as: “resources expended in relation to the accuracy and completeness with which users achieve goals” (ISO, 1998). Therefore, efficiency can be measured by the time that the participants consume to complete each task in the usability study (Frøkjær et al., 2000; Gil et al., 2010). Furthermore, ISO 9241-11 defines satisfaction as: “The freedom from discomfort, and positive attitudes towards the use of the product” (ISO, 1998). Satisfaction reflects the attitudes of the user towards the software and is usually subjective and varies from one individual to another. Therefore, satisfaction questionnaires and other attitude rating scales are usually used to measure satisfaction. To compare the effectiveness and efficiency of the iRDMS and Excel and to assess the satisfaction of the users with the interface, the participants were asked to perform data retrieval tasks using both methods, the iRDMS and Excel, and then were asked to answer a satisfaction questionnaire that measures their experience with the interface

after performing all the tasks. All participants were introduced to the data structure of each Excel file and were given a short tour of the iRDMS before they started working on the tasks.

To assess the effectiveness of each method from an accuracy perspective, the acquired datasets by the participants for each task were compared with the datasets that were acquired during the interface validation to find whether the participants have made any errors in the data retrieval and the number of errors they have made. Furthermore, to assess the efficiency of each method from a time perspective, the time that each participant took to finish each task was calculated and the time difference between the two methods was investigated. A significant difference between the two methods means that the method with the higher mean score is more effective and efficient to use than the lower one. These evaluation procedures were chosen because they have been proven to detect the usability difference between different systems and because they successfully show the practical advantages of each data management method (Mohammad, Breß, & Schallehn, 2012). Furthermore, suggestions and feedback from the participants were evaluated and considered to improve the design of the database and the interface.

Since this usability study is a designed artifact, an internal validity testing was conducted to detect any issues in the usability study's test procedures and materials prior to conducting the usability testing. In this internal validity study, an initial walk-through was conducted with a member of the data collection project. Based on recommendation from that member, it was decided to make some changes on the usability study materials and on the interface. Therefore, another participant who was part of the data collection project too was included, and our internal validity study was stopped after there were no more changes to be made to the design of the

usability test. The last participant from the internal validity study was included to be the first real participant in the usability study (J. R. Lewis, 2006).

3.3.3.6 Usability Study Participants

Participants in usability testing studies should be sampled from the potential users of the system under evaluation (Kushniruk & Patel, 2004). Therefore, in this study, a sample of the individuals from the data collection project who have access to the web interface and will be using this system in the future were included. However, to test the generalizability of the used data management approach and to ensure the external validity of this solution, convenient sampling was used to recruit participants that were not part of the data collection project. These participants were selected from the students, faculty members and staff of the School of Health and Rehabilitation Sciences and the School of Information Sciences at the University of Pittsburgh. The sample size rule 10 ± 2 was used in this usability study since this sample size have been shown to be sufficient to reach 80% overall discovery rate of usability issues (Hwang & Salvendy, 2010). Furthermore, since paired t-test or its non-parametric equivalent was used in this quantitative analysis, studies have shown that this statistical analysis does not require large sample size to produce results with high statistical power (De Winter, 2013; Sauro & Lewis, 2016). Therefore, a sample size of 10 ± 2 should be sufficient to satisfy the usability testing and the statistical analysis sample size requirements.

3.3.3.7 Usability Study Tasks

Since the goal of this usability study was to compare the original data management method (Excel) with the new one (iRDMS), each one of the tasks in this study was conducted using both

methods. In each task, the participants were asked to acquire a certain sub-dataset from the collected data using the two methods, the iRDMS, and Excel. Tasks that cover most the aspects of the collected dataset were developed to have a comprehensive study. Furthermore, these tasks vary in their difficulties starting from asking the participants to retrieve one data point, to asking them to retrieve multiple data points based on multiple conditions. For example, task 1 and 2, ask participants to retrieve five values in one column, task 3, ask participants to retrieve five values in one column but for two subjects, task 4, 7 and 8 ask participants to create a new column with five values based on the values of one column, two columns, and three columns, respectively, task 5 asks participants to subtract one column from another, task 6 asks participants to map two files, and task 10 and 9 ask participants to get the average of one column based on the values of another column. This progressive difficulty was important to provide some insights into the kind of data retrieval tasks each data management method can handle. Furthermore, since the goal of the proposed data management system was to provide an alternative data management interface and to conduct automated Extract, Transform, Load (ETL) procedures, each task required a different extraction and processing procedure to assess the effect of the data processing and the effect of the interface difference between the website and the original data files.

Although all the participants performed the same tasks, the tasks were presented to the participants in a random order to eliminate any task-order effect. This randomization included both the tasks and the used methods inside each task. The task-order effect can be expected when a user fails to perform easy tasks that were provided at the beginning, and as a result, the user loses interest in performing later harder tasks. Furthermore, this effect might be present if the users were asked to use the iRDMS to retrieve the data subset in each task first and then use Excel since the prior hypothesis indicate that the web interface will be easier and faster (Sauro &

Lewis, 2016). The materials of each one of these tasks were presented to the participants in separate folders. Each folder contained the description of the task, which was also presented as a hard copy, an Excel document for the participants to type the gathered dataset, and the files that the participants used to gather the dataset. The performance of each participant on each task was timed, and the number of data cells they have missed or gotten wrong was counted. At the end of each task, the participants were asked about their overall satisfaction with each one of the used data retrieval methods.

For the purpose of measuring the participants' satisfaction after completing each task or each scenario, the ASQ questionnaire was used, which was developed by J. R. Lewis (1991). This questionnaire is extremely short, and only have three questions measured on a 7-points scale. The three questions measure the: ease of task completion ("Overall, I am satisfied with the ease of completing the tasks in this scenario."), time to complete a task ("Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario."), and adequacy of support information ("Overall, I am satisfied with the support information (on-line help, messages, documentation) when completing tasks.") (J. R. Lewis, 2006). The overall score of this questionnaire is the average of the three questions. However, if one question was not answered by the participant or was not used in the questionnaire, the average of the two answered questions is the overall score of the ASQ (J. R. Lewis, 1995). Therefore, since the question about online help was not related to the investigation and outside the scope of this study, only the two questions about the ease of completing each task and the consumed time were included.

Along with the data retrieval tasks, the participants were provided with two surveys. The first was provided at the beginning of the study, which collected the demographics of the participants and their prior knowledge and experience with Microsoft Excel. Based on the

literature review that was conducted on usability studies, the common two collected demographics are age and gender. Some studies, especially the ones that test healthcare devices, collect other personal information, such as race, ethnicity, and social status (Carroll, Marrero, & Downs, 2007; Carstens & Patterson, 2005). Furthermore, some usability studies that test the adoption of new systems in an organization collect data about the position of the participants in the organization, income, and other information related to what specific aspects of the system the participants are going to use. However, in this usability study, it was not intended to test the difference in system adoption among different groups based on social status, or organization position. Therefore, only age group, gender, race, and education level were collected as it was desired to explain the results of this analysis using these factors. Furthermore, the participants were provided with two questions about their prior experience with Excel. Since Microsoft Excel is a very common used software, there are many questionnaires that measure the skills and experience of individuals on Excel. However, these questionnaires tend to be very detailed and technical, while this survey was not intended to measure the participants' skills and experience on deep levels. Therefore, two general questions that have been used by many researchers to measure the general experience of individuals with different systems and software were adopted. The first question asks the users to rank their skills on Excel on a scale from one to ten (On a Scale of One to Ten, What is Your Skill Level in Microsoft Excel?) and the second asks about how many years they have been using Excel (About how long have you been using Excel?) (Djenno, Insua, Gregory, & Brantley, 2014). The purpose of these questions is to help to explain the participants' performance differences.

The second survey was given to the subjects after finishing all the required tasks. The aim of this survey was to measure the reactions of the users to the system that they used and to

understand what aspects of the system they liked or disliked the most. To accomplish these purposes, the Post-Study System Usability Questionnaire (PSSUQ) was used. PSSUQ was developed originally by J. R. Lewis (1992) with 18 items measured on a 7-point Likert scale to measure four out five system characteristics associated with usability, effective, efficient, engaging, error tolerant, and easy to learn. However, a 19-item version of the PSSUQ was introduced with changes on the order of the items that captures all the five system characteristics associated with usability. J. R. Lewis (1995) has conducted a factor analysis to discover the measured subscales in PSSUQ. The factor analysis has revealed that PSSUQ measures four components of usability, The overall satisfaction score (Overall) with all the 19 items, subscale System Usefulness (SysUse) with 7 items, subscale Information Quality (InfoQual) with 6 items, and subscale Interface Quality (IntQual) with 3 items. The remaining three items were either highly cross loading on more than one factor or not loading on any factor. Therefore, they were part of the overall scale but not part of any of the subscales. Using the scores of the items of each scale and subscale a score for each one of these components was calculated. This survey was chosen because it perfectly fits the usability testing purpose, which was to measure the satisfaction and usability of the participants and to measure the quality of the interface. Also, this survey was used because it showed high reliability and validity when tested and validated by third parties (Fruhling & Lee, 2005).

Although each one of the questions in the PSSUQ has a comment section, it was decided to include four open-ended questions to capture any general comments or suggestions for improvements and to be able to explain the participants' performance on some tasks. The first two open-ended questions (What did you like about the site? and What do you dislike about the site?) were adopted from George (2005), where they conducted a usability testing and design of a

library website. George (2005) indicated that their approach and method could be generalized to other usability studies for any website. The third question (If you could change one thing about this system, what would it be?) was adopted from Hee Kim, H., & Ho Kim, Y. (2008), where they build an evaluation framework of institutional repositories and data management systems (Hee Kim & Ho Kim, 2008). The fourth question (what did you find confusing or a problem on the website?) was adopted from McMullen (2001) where they conducted a usability testing of a library website after a redesign project.

Both surveys, the demographics, and the PSSUQ, and the usability tasks can be reviewed in detail in APPENDIX (A).

3.3.3.8 Usability Study Data Processing

During the sessions where participants performed the tasks, the performance time for each task was written on the task sheet and was entered in the participants' performance table, and the participant demographic information were entered in the demographics table. At the end of each session, the acquired dataset by the participant was reviewed to find any errors or missing data cells. In the analysis dataset, each row represented the performance of a participant on a task using both data retrieval methods. Therefore, each row contained the subject id, task id, the order of the task presentation, performance time, whether they have made any errors or not, the number of missing or wrong data cells, and their answer to the question at the end of each task for both methods. An issue that is worth mentioning here is that the number of errors cannot be used as a measurement of error for both data retrieval methods. If a participant made an error in data retrieval while using the original data files, the error would most likely happen in only one data cell. However, if they made an error using the interface, all the retrieved dataset would be

wrong since the possible errors are either selecting the wrong test or the wrong data field. Therefore, the number of errors is not the same scale for the two methods since in the case of the interface it means whether or not the participant has made an error rather than the number of errors they have made. To account for that, it was decided to add a field for whether the participant has made any errors or not for both methods since they share this scale, and keep the number of errors made by the participant when using Excel only.

3.3.3.9 Usability Study Data Analysis

The main purpose of this data analysis was to find whether there was a significant difference between using the original data files and the web interface for data retrieval in terms of the consumed time and errors.

3.3.3.10 Usability Study Missing Data

Since it was hypothesized that the iRDMS will be faster and easier to use than Excel, it was concerned that many participants would choose not to do the data retrieval tasks using Excel, especially for the harder tasks. In case this phenomenon would have occurred, the decision on how to encounter its effect would have been made based on the pattern of its occurrence. The first pattern would have been if only a few participants show this behavior on few tasks, then these performances would have been simply dropped from the analysis. The second pattern would have been if many participants show this behavior on many tasks, then dropping these performances would not be feasible since it would mean removing the observations on harder tasks that were hypothesized to capture the significant difference between the two data retrieval

methods. Therefore, to remedy that, different data analysis would have been used and another one that fits the data the best would have been chosen.

The first data analysis scenario in the case of the latter pattern would have been to simply give the participant who chose not to perform the task a time of zero and consider all the data cells missing and thus make the value of the field whether they made an error or not as 1. Then, use this error field as a covariate in the time mean difference analysis between the two data retrieval methods. However, this approach might have not been sufficient to capture the true difference in time between the two data retrieval methods. The second data analysis scenario would have been to give the participant who chose not to perform the task a time equal to the maximum time that took the other participants to perform the same task. However, this might have not been feasible if many participants chose not to perform the same task. The third data analysis would have been to use MICE to impute the missing data since this method account for the performance of the participant on other tasks and the difficulty of the task by measuring the performance of the other participants on the same task. The concern here is that the participant who chose not to do the harder tasks would be less fatigue when performing the simpler ones. Thus, their performance would be faster compared to the other participants, which would result in a bias estimation of their performance on the missing ones. However, after collecting the data, data screening and multiple data analysis were performed and it was found that no participants have chosen not to perform any task and that there was no missing data.

3.3.3.11 Usability Study Data Screening

Before performing the statistical analysis, the final table or dataset that contains the participants' information and performance was validated by selecting random records and comparing their

data with the raw materials that have been used during the test sessions. Next, the data were screened to find any missing data as mentioned in the missing data section.

3.3.3.12 Usability Study Time and Error Difference Analysis

To investigate whether there is a significant difference in data retrieval time between the two methods, a mean difference analysis was conducted. The participants' performance times were not normally distributed based on examining the generated density plots and the results of the Shapiro Wilk's normality test. Therefore, Wilcoxon signed-rank test paired samples was used, which is a non-parametric test that is equivalent to the paired Student's t-test, to test the significant of difference (Wilcoxon, 1945). The non-parametric test that is equivalent to the paired t-test was used instead of using the unpaired student t-test because data that were collected from the same sample on two occasions, performance on Excel and on the iRDMS. However, if it had been decided to use the errors as covariates in the analysis, a repeated measure ANCOVA would have been conducted to investigate the mean difference. A significance level of 0.05 was used to assess the p-value of the all the tests. Furthermore, Kruskal-Wallis rank sum test and post hoc Nemenyi test, which are non-parametric test equivalent to one-way ANOVA, were performed to investigate whether the participants have performed significantly different on the ten tasks using each method (Kruskal & Wallis, 1952; Nemenyi, 1963).

Furthermore, the participants were grouped according to their age, gender, education, and race and investigated whether there is a significant performance difference between the groups when using Excel and the DBMS. Again, Kruskal-Wallis rank sum test and post hoc Nemenyi test, which are non-parametric test equivalent to one-way ANOVA, were performed to perform this analysis. Groups that have less than five participants were excluded since four participants

and less were not sufficient to perform the analysis. For example, in education groups, there were only two participants with PhD and only one with less than Bachelor's degree and both were excluded. Therefore, for some comparisons, both, the Kruskal-Wallis rank sum test and the post hoc Nemenyi test gave the same results since only two groups were included in the analysis. Furthermore, the relationship between the participants Excel skills and experience with their performance on the ten tasks using both methods were investigated. In addition, it was investigated whether or not the order in which the tasks were presented to the subjects correlates with the participants' performance on the tasks using the two methods. Also, it was investigated whether participants' answers on the two Excel skills level and experience questions correlate or predict their performance on the tasks using the two methods. Moreover, because multiple comparisons were conducted, Benjamini–Hochberg procedure, which is a False Discovery Rate (FDR) method, was used to adjust p-values to control for the type I error (Benjamini & Hochberg, 1995). All the reported p-values are unadjusted except for significant p-values that are part of multiple comparisons, adjusted p-values will be reported too.

To investigate the difference in the made errors during the data retrieval using the two methods, a Chi-square test was conducted to test for equality of proportions of errors between the two methods. This statistical method was used since two dichotomous data fields were used to conduct this analysis, whether the participant made an error or not and the used data retrieval method. Each task in this usability testing was investigated separately.

3.3.3.13 PSSUQ and ASQ Analysis

Based on the factor analyses by J. R. Lewis (1995) to discover the measured subscales in PSSUQ, they developed the rules for calculating the scale and sub-scale scores for the PSSUQ.

The Overall can be calculated by averaging the responses to items 1 through 19. The sub-scale SysUse can be calculated by averaging the responses to items 1 through 8. The sub-scale InfoQual can be calculated by averaging the responses to items 9 through 15. The sub-scale IntQual can be calculated by averaging the responses to items 16 through 18. The standard deviation of each scale was also calculated to provide insight on the degree of the participants in this usability study agreement up on their satisfaction level. Furthermore, the responses from the participants on the two questions that were acquired from the ASQ were examined to investigate whether participants prefer to perform the tasks on one method or another. Also, the answer of the participants on the two questions and how they correlate with their performance using Excel and DBMS along with the task order, Excel skills, and Excel experience were examined.

3.4 DATA ANALYSIS

3.4.1 Sample size

The sample size is a critical subject in this dissertation for several reasons. First, this data analysis might suffer a shortage in the number of subjects, because the used dataset contains data from a population of limited access. As indicated before, although aphasia is relatively a common condition, sampling a subset of individuals from the population of aphasics who are high likely to complete the tasks of the study is a challenge. This problem is posed by the characteristics of this population, such as age, mobility, and cognitive abilities. Thus, concerns regarding dropouts and incomplete cases are present in this study. Second, sample size is a

critical subject in this dissertation because of the complexity of the proposed data analysis approaches. Although SEM is a practical data analysis choice to test a theory that is complex and nested in nature, it requires relatively large sample size. Therefore, different methods to estimate the needed sample size to conduct SEM were used and alternative analyses that require smaller sample size than SEM were specified in case of failure to acquire the necessary sample size. It is worth mentioning that the small sample size becomes a problem when trying to conduct CFA and SEM but not when conducting EFA. The sample size requirements for the EFA are less strict compared to the ones for CFA and SEM as measurement and standard errors are not estimated (Pearson, 2008). In fact, a systematic review of the used sample size in EFA showed that 40.5% of studies used the ratio 5:1 subjects per variable (SPV) less and 14% used the 2:1 SPV or less ratio (Pearson, 2008). Therefore, the sample size should not be a major concern when conducting EFA.

In the case of CFA and SEM, the first approach of calculating the required sample size is using the rule-of-thumb methods. These methods use only the number of variables in the model or the number of parameters to be estimated. Although these methods do not provide precise sample size estimation due to their disregarding of other attributes, such as the number of factors and cutoff values of fit indices, they still can be used as general guidelines of sample size estimation. These rules-of-thumb vary from 20:1 SPV ratio to 5:1. Per Bryant and Yarnold (1995); Garson (2008); Gorsuch (1983); MacCallum, Widaman, Zhang, and Hong (1999); and Everitt (1975). 5:1 SPV ratio is acceptable if the sample size is 100 or larger. In fact, MacCallum and Austin (2000) conducted a systematic review of the studies that used SEM; they found that 18% of these studies used a sample size less than 100. These results were also reported by

Breckler (1990) where he surveyed 72 studies that used SEM. Breckler (1990) reported that the range of sample sizes was between 40 and 8,650, with 22% less than 100.

When using 5:1 SPV rule-of-thumb, if the EFA suggested to retain 7 factors with 21 variables, then the sample size minimum should be 105. However, MacCallum, Browne, and Sugawara (1996) have computed sample size in SEM framework based on the desired statistical power, the RMSEA levels for exact and close fit, and degree of freedom (*df*). Because RMSEA is the most commonly used fit index in SEM framework, they based their power calculation on how much power is needed to detect a certain RMSEA cutoff value to be able to reject the null hypothesis, which is that the modeled data matrix fits the observed data. First, the *df* in the SEM was calculated based on the assumption that 7 factors will be retained with 21 variables as following:

$$df = m * \frac{(m+1)}{2} - 2 * m - \xi * \frac{(\xi-1)}{2} \quad (1)$$

Where (*m*) is the number of variables in the model, and (ξ) is the number of factors or hypothesized constructs in the model (Rigdon, 1994). That makes the first part of the equation “ $m * (m + 1) / 2$ ” as the number of unique elements in the model, where the second part “ $2 * m$ ” is the number of parameters to be estimated in the model. Therefore, in the case of this dissertation’s hypothesized model, the term (*m*) will be equal to 21 variables and the term (ξ) will be equal to 7 factors. Thus, the *df* was calculated as following:

$$df = 21 * \frac{(21 + 1)}{2} - 2 * 21 - 7 * \frac{(7 - 1)}{2} = 168$$

The used the cutoff value for the statistical power was 0.80, which is the recommended value for reaching an adequate power by Cohen (1992). The value 0.05 was used as the alpha

level of significance. The value 0.05 was also used as the null RMSEA and 0.08 as the alternate RMSEA as suggested by MacCallum et al. (1996). When these values were used to conduct the sample size calculation using MacCallum et al. (1996) method, the results indicated that a sample size of 94 subjects is sufficient to have the statistical power to detect RMSEA levels that can reject the null hypothesis. Therefore, taking both methods into consideration, the sample size 100 was decided to be the desired sample size to be able to conduct the CFA and SEM analyses.

In the case of failing to collect the desired data from 100 subjects, however, a SEM with factor scores or, in other words, a multiple linear regression would have been conducted. However, the required sample size of multiple linear regression is also a subject of debate. As in SEM, multiple linear regression required sample size has many rule-of-thumbs based on the SPV ratio. For example, Austin and colleagues have concluded that “Linear regression models require only two SPV for adequate estimation of regression coefficients, standard errors, and confidence intervals” (Austin & Steyerberg, 2015). Another rule-of-thumb by R. J. Harris (2001) suggests that the number of subjects should exceed the number of variables by at least 50. Schmidt (1971) concluded that sample sizes in the range of 15-20 SPV are adequate to conduct multiple linear regression with adequate power. Green (1991) suggested a minimum SPV number as 50 plus eight times the number of variables. Furthermore, R. J. Harris (2001) suggested that 10 SPV is the minimum sample size for multiple linear regression in many areas of research. Finally, there are other scholars that argue that the sample size calculations for multiple linear regression should be conducted by including the anticipated effect size, desired statistical power, the used alpha probability level, and the number of predictors. Therefore, when using the rules-of-thumb, it can be argued that 14 subjects should be enough to conduct the multiple linear regression depending on Austin and colleagues analysis if 7 factors were retained. Even if more

conservative rule-of-thumb like the 10 SPV was used, only a sample size 70 would be enough to be able to conduct multiple linear regression.

3.4.2 Data Screening

Although data were processed and screened by automated extraction algorithms, the data were screened before the analysis to look for any missing values and to test statistical assumptions. All variables were explored graphically and tested statistically to determine their distributions. Mardia's skewness and kurtosis tests of multivariate normality were used (Mardia, 1970). Furthermore, all variables were plotted and statistically tested to find outliers and data influential points. Generally, in the SEM framework, ML estimation is a robust method against skewed or non-normal data. However, studies have suggested that under severe skewness ML should not be used as the extraction method in SEM. Furthermore, because of having missing data, the EM algorithm with the ML estimation was used to estimate the missing data and to estimate the model parameters. This method was chosen since was not much of a room to drop any observations from the analysis and since SEM was conducted, which uses the ML estimation to fit the model (Dempster et al., 1977). Furthermore, to remedy for the assumption of normality, EM-ML along with Yuan and Bentler (2000) scaled chi-square statistic was used (Allison, 2003). Although this estimation method was implemented to handle data that are MCAR, Yuan and Bentler (2000) have shown that EM-ML produced robust and unbiased results even with data MAR .

If the sample size was not adequate to run SEM an multiple regression was used as an alternative analysis to SEM with latent variables, missing data would have been produced after

the estimation of the CFA solution. Therefore, all missing data would have still been estimated by the EM-ML estimation method prior to calculating the factor scores. However, assumptions of multiple regression, such as linearity, multivariate normality, multicollinearity, and homoscedasticity, would have been tested and remedy for if necessary. First, the multivariate normality would have been tested using Cox-Small test of multivariate normality. If this assumption was violated, then the outcome variable would have been transformed (The sentence comprehension scores) to remedy for this assumption. Box and Cox (1964) procedure would have been used to select the transformation that minimizes the sum of squares of error for a linear regression and, thus, remedy for multivariate normality assumption. Next, the linearity assumption would have been tested by visually plotting the observed versus the predicted values and measuring the slope of the linear relationship. In case this assumption was violated, the dependent variables (the factor scores) would have been transformed. Furthermore, since homoscedasticity has a close relationship to the linearity of the relationship between the predictors and the predicted variables, transforming the dependent variables would high likely resulted in fixing the homoscedasticity assumption (Cohen, Cohen, West, & Aiken, 2013). Multicollinearity was not likely to be an issue since all factor scores would have been calculated by Bartlett factor scores method which minimizes the correlation between the factor scores.

3.4.3 Descriptive Data

Univariate descriptive statistics, such as mean, median, range, and standard deviation, were calculated for all variables. In addition, data regarding the demographics of the participants are reported. Bivariate descriptive statistics are presented in the correlation matrix in the results.

3.4.4 Data Modeling

As mentioned in the introduction of the methods chapter, SEM was used to build the data model that was used to test the hypotheses. It is worth mentioning here that in SEM framework there are two main parts, the factor analysis, and the path analysis. In the factor analysis (EFA and CFA) a measurement model was developed to detect and build the factors or the latent variables. In the path analysis, a path model of the latent variables that were developed in the factor analysis phase was constructed and called structural model. However, in the literature, it is very common to call the path analysis part or the structural model as SEM as it is the unique part that differentiates SEM from regular factor analysis or regression (Engle et al., 1999).

SEM is a statistical technique for testing causality hypotheses by presenting them as paths in a graphical model and express them as regression equations (Aaronsen, Frey, & Boyd, 1988). SEM test these hypotheses using theoretical unobserved variables that are called latent variables, which are measured by observable variables that are called manifest variables or construct indicators. In the SEM framework, a correlation or covariance matrix of the observed variables is used to judge on whether the hypothesized model fits the observed data or not. To explain, once the researcher specifies a model or the hypothesized relationships between the observed variables, SEM will try to reproduce the correlation or covariance matrix based on the specified or hypothesized model. Next, SEM will deploy fit measures that measure how different is the observed and the modeled matrices are, and thus conclude whether the model fits the data or not.

In this analysis, seven latent variables that represent the hypothesized constructs of phonological LP, semantic LP, syntactic LP, phonological STM, semantic STM, syntactic STM,

and CR were proposed. These hypothesized constructs were proposed based on what was found in the literature as possible predictors of the sentence comprehension deficits (STM and CR) along with hypothesized constructs that were used as controlled variables (LP) to observe the unique contribution of STM and CR. However, as discussed in the sample size section, the sample size is a major concern in SEM. Unlike other analysis approaches, a lower power in CFA means that there is not enough power to detect the difference between the observed matrix and the modeled one. Thus, it fails to reject the null hypothesis and show that the model is acceptable (DeCoster, 1998). Another issue that makes sample size important in SEM is that there is some level of uncertainty in estimating the errors and the unobserved variables in the SEM model, an adequate sample size is required to minimize this uncertainty (In'nami & Koizumi, 2013). Because of the nature of the population of aphasics and the limit access to such individuals, the data analysis plan, and the alternative approaches that would have been used in case the sample size did not satisfy the requirements were detailed.

The first step in the data analysis was to perform an EFA to build the measurement model which represents the relationships between the observed variables and their hypothesized constructs or factors. As mentioned in the sample size section, EFA has less strict sample size requirements than the CFA and SEM. Therefore, a relatively small sample size should be sufficient to perform EFA. EFA was an important step in the analysis of this dissertation, not only because it might answer one of the research questions, but because it might determine how many variables and factors to be used in the following analysis. Again, as mentioned in the sample size section, the number of variables and factors has a direct influence on the required sample size. Since the collected sample size was satisfactory to the minimum required sample size to conduct the SEM analyses with the retained factors in from the CFA in this dissertation,

alternative analyses were not needed. However, if the sample size was not adequate to perform the CFA and SEM, one option was to reduce the number of variables that represent the covariate variable LP that is being controlled for in the analysis. In the analysis, nine variables were included to represent LP to capture the three proposed specific domains of LP (phonological, semantic and syntactic). However, since the LP was included to only control for its effect and there was no interest in its relationship to the other variables, only three variables that represent the LP domain in general could be included and carry on with the analysis. Another option was to perform a multiple linear regression using the calculated factor scores to test the hypotheses and answer the questions. Therefore, to answer the research question: Are CR, STM, and LP separable and domain-specific components in PWA? and test the hypotheses, measurement model was developed.

3.4.4.1 Measurement model

In the measurement model, an EFA was deployed to investigate the relationships between the observed variables and their hypothesized constructs. For example, there were three measures that were hypothesized to be affected by semantic STM, in the measurement model, it was desired to find whether a common construct between these three measures that eliminates as much noise or irrelevant variance as possible can be extracted. This irrelevant variance could be related to contributions of other cognitive components or functions that might be necessary to perform some tasks but not the others. The measurement model was built based on the analysis of the covariance matrix between the 21 variables of interest. For this analysis, the functions “fa” in the R “psych” package was used.

The first step in building the measurement model was to determine the number of factors to retain. This step was critical to the whole data analysis since its results might be used to answer a major research question and might also be used in all the subsequent analysis. Therefore, multiple methods were used to make the decision on how many factors to retain. One method that is commonly used for determining the number of factors is K1-Kaiser's eigenvalue-greater-than-one rule proposed by Kaiser (1960). Eigenvalues are the constant values that when multiplied by a vector (eigenvector) give the same results as when multiplying this vector by the correlation or covariance matrix. Both eigenvalues and eigenvectors can provide information on the direction of the correlation or covariance matrix and can be used to calculate the variable loadings or the correlation between each observed variable and the extracted factors. K1-Kaiser's method simply suggests that any factors with eigenvalue that is greater than one should be retained as a major factor. Although this method faces some criticism over its efficiency in determining the number of factors, it still can be used as an extra indicator of how many factors to retain. The most common criticized aspect of this method is that it was proposed to be used in principle component analysis (PCA) not to be used in EFA where a rotation method is used. This can overcome by using PCA as a starting point to obtain the eigenvalues and scree plot of eigenvalues and then apply EFA and CFA to either confirm these findings or reject them. Furthermore, another criticized aspect of K1-Kaiser's method is how strictly should someone use the cutoff value one. For example, should a factor with 0.99 eigenvalue be always ignored where another with 1.01 always be retained? The solution when this case happens (having factors close to the one cutoff value) is to use other methods to judge on whether to retain or reject that factor.

Thus, the second method that was used is the Cattell's Scree test proposed by Cattell (1966). The scree plots are basically used to plot eigenvalues against their associated component

or factor in descending order with a line connecting the plotted points. The Cattell's Scree test suggest using this plot to determine the number of factors by selecting the factors that are above the last big drop in the line connecting the plotted points (Figure 10).

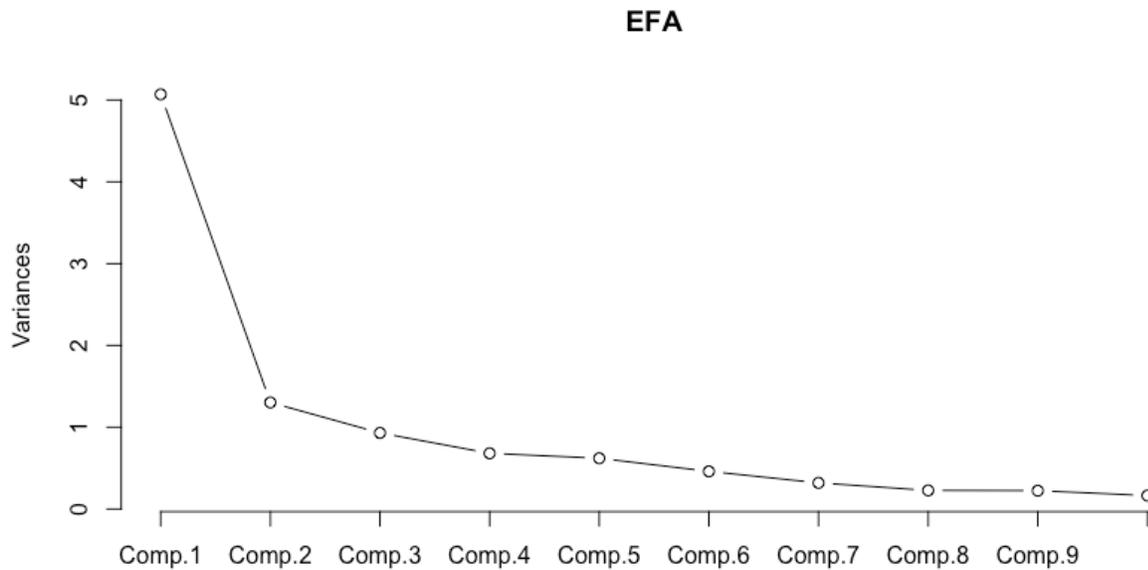


Figure 10. Cattell's Scree test.

This plot generated from randomly generated data. Based on K1Kaiser's rule, two factors should be retained from this analysis.

The third method and arguably the most reliable one that was used to determine the number of factors to retain is parallel analysis (Ledesma & Valero-Mora, 2007). This method, which was proposed by Horn (1965), lunches a Monte Carlo simulation that simulates normal random samples that parallel the observed data in terms of sample size and the number of variables. Then, it uses the eigenvalues from the results of these random samples to set up a cutoff value for what eigenvalues from the original dataset are above the chance. Usually, researchers use the mean of the parallel analysis eigenvalues, which is what was used in the

analysis, or a given percentile, such as the 95th of the distribution of eigenvalues to set up a cutoff value (Figure 11). Finally, after conducting the EFA, the item loadings of each suggested factor were looked at and it was ensured that all factors have at least three variables loading on them.

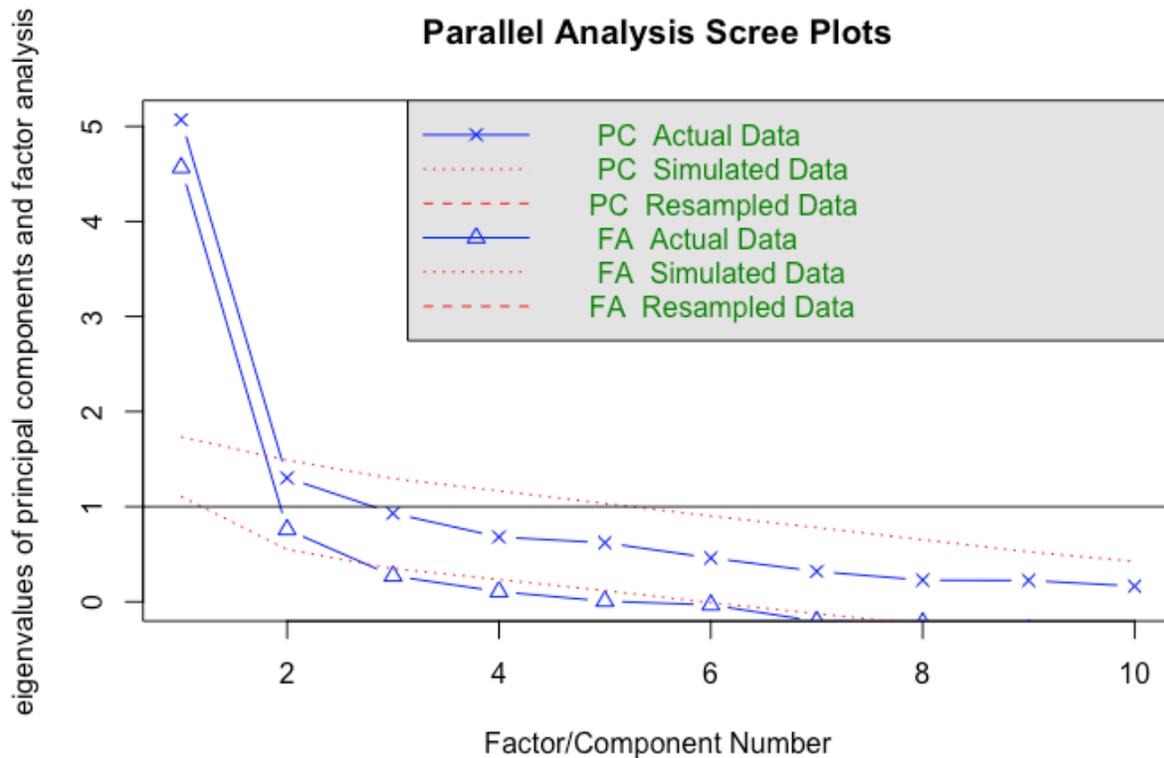


Figure 11. Scree plot by PA.

This plot generated from randomly generated data. Based on PA, two factors should be retained from this analysis as well as the simulated factors have a mean of eigenvalues that is less than the eigenvalues of the first two factors.

As indicated above, after performing the PCA and the parallel analysis multiple EFAs were deployed with the suggested solutions from the PCA and the parallel analysis. In the EFA, rotation was used to rotate or change the direction of each factor to maximize the variable loadings on one factor (closer to one is preferred) while minimizing them on the others (closer to

zero is preferred). The two common types of rotations are orthogonal rotation and oblique rotation (Tabachnick & Fidell, 2007). Orthogonal rotation rotates the factors while keeping 90° between their directions, which basically means that the factors are not allowed to correlate. However, since this is opposing the used theory that suggests that the hypothesized factors do correlate with each other in nature, Varimax rotation was used. Varimax rotation rotates the factors freely and allows them to correlate, which enables having the factors correlation matrix that helps to understand the relationship between the retained factors. To determine what factor each variable loads on, Comrey and Lee (2013) suggestion by ranking the variable loadings as 0.32 (poor), 0.45 (fair), 0.55 (good), 0.63 (very good) or 0.71 (excellent) was used. Therefore, the minimum acceptable variable loading was 0.32, which was also used as the cutoff value to determine the cross-loading variables where a variable loads higher than 0.32 on more than one factor. When cross-loading happened, the variable was assigned to the factor with the highest loading. Complex variables that load on more than one factor were not considered because they are harder to interpret and make the subsequent analysis very complex (Yoo & Donthu, 2001). Each variable loading was used to measure the variable's relationship with the retained factor, the higher the loading, the stronger the relationship.

The next step after conducting EFA was to conduct CFA to confirm the findings from EFA and ensure that the model fits the observed data (Jöreskog, 1967). In CFA, the variables can load on only one factor, which makes the model fit worse since loadings on other factors are eliminated. CFA provides a test of variable loadings significance, where variables with insignificant variable loadings were dropped out of the model. Furthermore, CFA provides a significant test of the correlation between the factors. Therefore, CFA was used to gain a better understanding of the retained model and to refine the model if needed. Even more, the results of

EFA and CFA were used to answer the research question: Are CR, STM, and LP separable and domain-specific components in PWA? For this analysis, the function “cfa” in the R “lavaan” package was used (Rosseel, 2012).

Based on this dissertation’s hypothesis, the outcome of this step would be a model with seven factors that represent the hypothesized constructs of interest (Figure 12). However, one alternative hypothesis states that three hypothesized constructs that represent semantic, phonological, and syntactic LP could be revealed, and hypothesized constructs of STM and CR cannot be separated from these three. The other alternative hypothesis states that three hypothesized constructs of STM, CR, and LP could be revealed but cannot be divided into semantic, phonological, and syntactic domains.

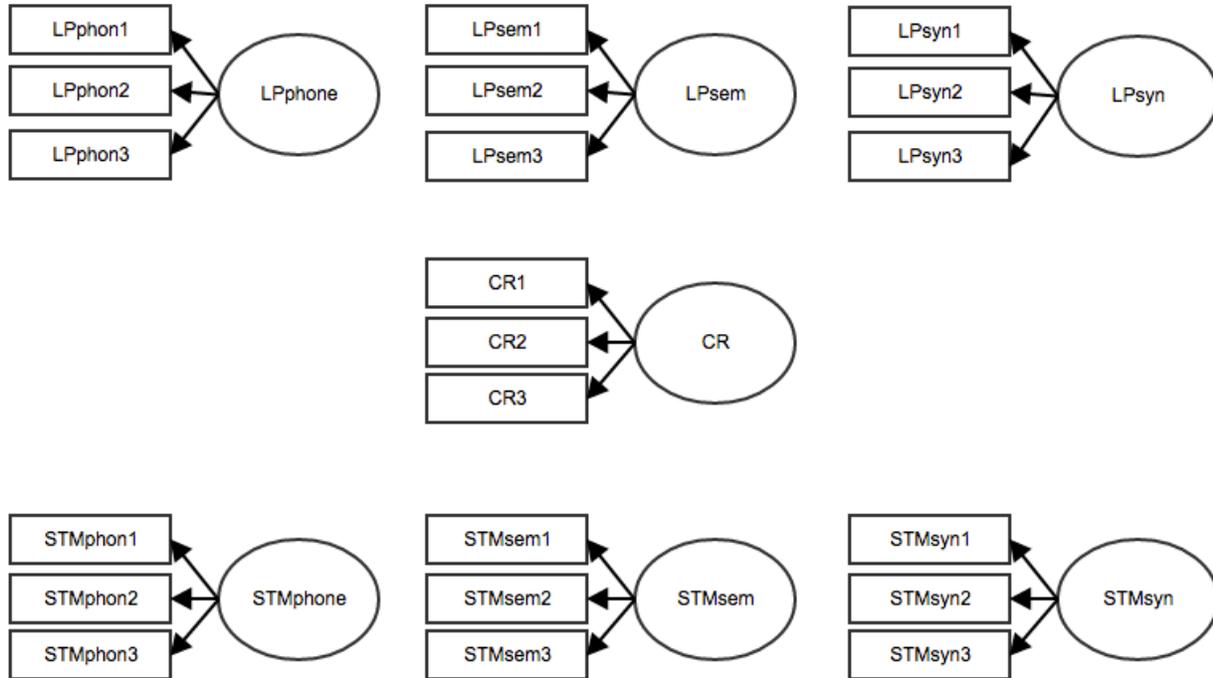


Figure 12. The hypothesized measurement models.

Rhyme Judgments Task (LPphon1), Non-word reading times (LPphon2), Rhyme Judgments Task Pictures (LPphon3), Category judgment Task (LPsem1), Exception word reading and word-picture matching task (LPsem2), Semantic 2-back (LPsem3), Grammaticality Judgment (LPsyn1), List versus sentence reading (LPsyn2), Anagram Test (LPsyn3), Rhyme Probe Span (STMphon1), Rhyme Judgment with a filled interval (STMphon2), Digit Pointing Span (STMphon3), Category Probe Span (STMsem1), Category Judgment with a filled interval (STMsem2), Synonym Judgment with a filled interval (STMsem3), Sentence Probe Span (STMsyn1), Sentence Reading with a filled interval (STMsyn2), Syntactic 2-back (STMsyn3), CRTT-R Stroop (CR1), Picture-Word Interference (CR2), Number Stroop (CR3), Phonological Language Processing (LPphone), Semantic Language Processing (LPsem), Syntactic Language Processing (LPsyn), Phonological Short-Term Memory (STMphone), Semantic Short-Term Memory (STMsem), Syntactic Short-Term Memory (STMsyn), Conflict Resolution (CR), □ = Observed Variable, ○ = Latent Variable.

3.4.4.2 Structural model

In the structural model, a path analysis model with the latent variables that were acquired from the CFA along with the outcome variables (sentence comprehension score or CRTT) were built.

In this structural model, the latent factors that were acquired from the CFA were included as exogenous variables. Exogenous variables are the ones that do not have any variables predicting their variability in the model. The exogenous variables or factors were regressed on the sentence comprehension measure for each sentence of interest. Because there were four types of sentences

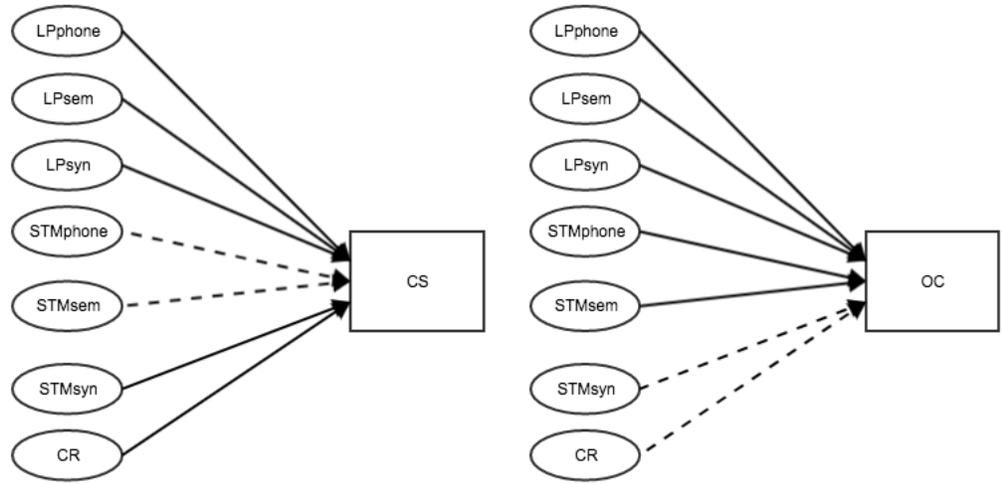
under investigation that are not related to each other, four structural models, each for one sentence type that was presented as an outcome measure, were built. Using these models, the hypothesized relationships between the latent variables and the sentence comprehension measures were tested. A model with a “good fit” meets the following criteria: an RMSEA less than or equal to 0.08 (“poor fit” greater than or equal to 0.10), a CFI greater than 0.9, and an SRMR less than or equal to 0.08 (Kline, 2015). Finally, a significance level of 0.05 was used to assess the model chi-square statistic. Also, the robust fit indices were used because the “Yuan-Banter” statistic was used to adjust for non-normality.

In SEM, when having a set of factors in a model, the relationship between one factor and another is conditional on the existence of the other factors. Therefore, the path from each factor to the outcome variable in the models were used to test the significance of that factor’s prediction of the outcome variable. Each path has an estimate that works similarly to the regression coefficients but describes the relationship between the latent variables and the outcome. The significance of each of these estimates was tested by a z-test of significance with α -level = 0.05. Furthermore, the standard errors of each of these estimates were also reported as indicators of the reflection of the estimates of the population parameters.

In the factor analysis models, the correlations between the factors, the R^2 of each factor and the R^2 of the outcome variables were calculated and reported to indicate to which degree the obtained factors predict the variability in deficits in sentence comprehension and to provide guidance to future research on what can be done to improve the prediction of such deficits. For SEM analysis, the function “sem” in the R “lavaan” package was used (Rosseel, 2012).

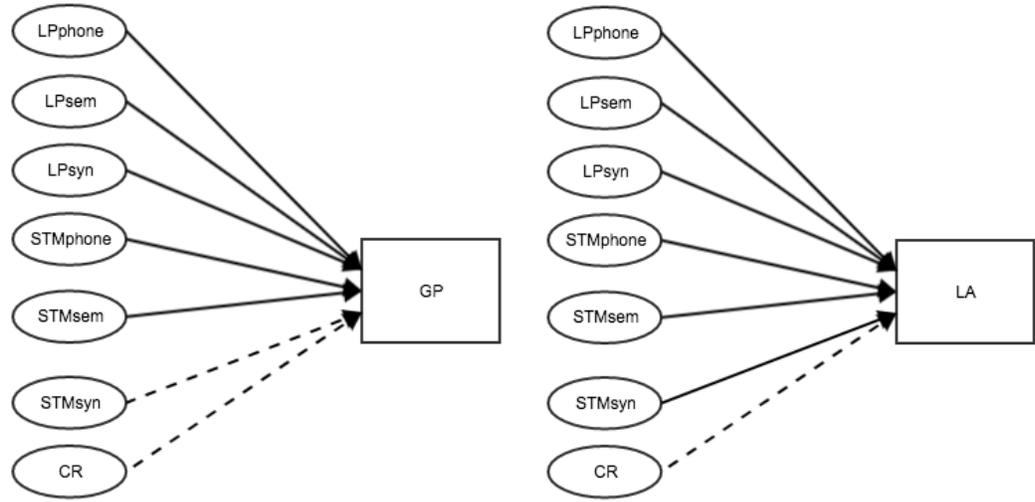
According to the dissertation’s hypotheses, deficits in each one of the sentences would be explained by at least one construct while controlling for the effect of LP. Deficits in CS

sentences will be predicted by phonological and semantic STM (Figure 13a). Deficits in OC sentences will be predicted by syntactic STM and CR (Figure 13b). Deficits in GP sentences will be predicted by syntactic STM and CR (Figure 13c). Deficits in LA sentences will be predicted by CR (Figure 13d). The alternative hypothesis, however, states that no effect will be detected for either STM or CR or both beyond the effect of LP.



(a)

(b)



(c)

(d)

Figure 13. hypothesized structural models.

a) The hypothesized structural model for the compound sentences (CS). b) The hypothesized structural model for the Object Cleft (OC) sentences. c) The hypothesized structural model for the Garden Path (GP) sentences. d) The hypothesized structural model for the Lexical Ambiguity (LA) sentences. LP = Language Processing, Sem= Semantic, Phone= Phonological, Syn= Syntactic, STM= Short-Term Memory, CR= Conflict Resolution, □ = Observed Variable, ○ = Latent Variable, → = Significant Relation, - - -> = Non-Significant Relation.

3.4.4.3 Multiple Linear Regression

As an alternative analysis that would have been conducted instead SEM if the sample size was not adequate to perform SEM, the factor scores would have been calculated after conducting the CFA and would have been used them to perform multiple linear regression. Although by conducting multiple linear regression some power of estimating the errors of each variable might be lost, this approach is still valid to be used for hypothesis significance testing, especially when factor scores are calculated properly (Knofczynski & Mundfrom, 2008). To calculate the factor scores, Bartlett factor scores method would have been used. This method uses maximum likelihood estimation to estimate the factor scores based on the row vector of observed variables, the diagonal matrix and the factor pattern matrix of loadings (Bartlett, 1937). The main advantage of this method is that it produces unbiased factors scores that truly represent the unique and common score of a factor from the set of variables that load on it (Hershberger, 2005). These scores can be calculated after conducting the CFA without fitting the model (so that large sample size is not required), where only the variables that have been indicated to be loading on the factor would be included in the analysis. These factors scores would have been obtained by using the function “Predict” in the R “lavaan” package (Rosseel, 2012).

Another justification for using multiple linear regression as a legitimate alternative to SEM with latent variables is that the SEM model of interest has only one level, where all the variables are exogenous and predict only one endogenous variable. Therefore, the path analysis, which has the advantage of modeling endogenous variables that predict other endogenous variables, in this case would lose its advantage since it would produce the same results as the multiple linear regression. Furthermore, in this dissertation’s analysis, the outcome variables (sentence comprehension scores) for all the four sentences of interest were observed and were

directly analyzed in the model without representing them as latent variables. Therefore, if the multiple linear regression was used, the outcome variables would still be the same as in SEM. In the multiple linear regression models, all the factors from the factors scores analysis in the model would be used as predictors or independent variables (IV) and each sentence comprehension score as a dependent variable (DV).

Similar to SEM, four multiple linear regression models would have been built, each would predict one of the four sentences of interest. In this analysis, the significance testing of the coefficients associated with each of the IVs would have been used to answer the research questions regarding which cognitive domains predict deficits in sentence comprehension. For this significance testing, a t-test would have been used to test whether each one of these coefficients or slopes is significantly different from zero. A zero coefficient means that the corresponding variable has no effect on the DV. Furthermore, the R^2 would have been used to indicate to which degree the factors explain the variability in deficits in sentence comprehension and to provide guidance to future research on what can be done to improve the prediction of such deficits. Furthermore, the semi-partial correlations would have been used to indicate to which degree each one of the factors explains the variability in deficits in sentence comprehension while holding the other factors in the model constant. The prior hypotheses for this analysis would have been the same as the ones for SEM.

4.0 RESULTS

4.1 PRELIMINARY RESULTS

As part of the requirement collection and design validation, two members of the data collection research team were asked to perform usability testing on the web interface. The first member has completed six out of ten tasks due to a shortage of the time availability, and the second member has completed the entire ten tasks. Both members have expressed high satisfaction with the usability of the web interface. Their satisfaction was reflected in their response to the PSSUQ where they averaged an Overall satisfaction score of 1.89 and 1.29 out of 7 (The lower the score the better the satisfaction), respectively. Furthermore, their satisfaction was reflected on each one of the subscales of the PSSUQ where their score on the SysUse subscale was 2.25 and 1.12, on the SysUse subscale was 1.71 and 1.8, and on the IntQual was 1.5 and 1 out of 7, respectively. Detailed scores are viewed in (Table 2).

Table 2. The scores of each item by the two members (1st and 2nd) of the usability testing.

Q#	Question	1 st	2 nd
1	Overall, I am satisfied with how easy it is to use this system.	2	2
2	It was simple to use this system.	2	1
3	I can effectively complete my work using this system	2	1
4	I am able to complete my work quickly using this system.	2	1
5	I am able to efficiently complete my work using this system.	2	1
6	I feel comfortable using this system.	3	1
7	It was easy to learn to use this system.	2	1
8	I believe I became productive quickly using this system.	3	1
9	The system gives error messages that clearly tell me how to fix problems.	2	NA
10	Whenever I make a mistake using the system, I recover easily and quickly.	1	4
11	The information (such as online help, on-screen messages, and other documentation) provided with this system is clear.	2	NA
12	It is easy to find the information I needed.	3	1
13	The information provided for the system is easy to understand.	2	2
14	The information is effective in helping me complete the tasks and scenarios.	1	1
15	The organization of information on the system screens is clear.	1	1
16	The interface of this system is pleasant.	1	1
17	I like using the interface of this system.	2	1
18	This system has all the functions and capabilities I expect it to have.	2	1
19	Overall, I am satisfied with this system.	1	1

Furthermore, they expressed higher satisfaction when using the iRDMS to perform the tasks compared to using Excel, which was reflected in their answers to the questions after performing each task using each method. Both members have a significant time difference between using the iRDMS to retrieve the datasets and using Excel (Table 3).

Table 3. The mean and SD of performance of the two members (1st and 2nd) on the usability testing tasks.

	iRDMS		Excel		P-value
	Mean	SD	Mean	SD	
1 st	87	40	334	160	0.01
2 nd	46	23	285	256	0.01

iRDMS= Internet-Based Research Data Management System.

Both members have committed errors while performing the tasks. The first member made the error while using the interface by selecting the column “ItemType1” instead of the column “ItemNumber1”. When the member was asked about the reason behind this error, he stated that he thought that task was asking about the column “ItemType1”, not the column “ItemNumber1”. Although this error was a subject error, some changes were made to the interface to help the users to avoid such errors. The interface was changed to be viewing the tasks and column names in alphabetical order. Furthermore, a simple algorithm that scans the tasks and column names looking for two or more tasks or columns that have the same name with only one different character was implemented. Next, this algorithm will put this character between parentheses and capitalize it in case it was a letter. For example, when having the tasks “STMsem1a” and “STMsem1b”, they will be viewed as “STMsem1 (A)” and “STMsem1 (B)”. The second member committed her errors while using the original files, where she selected the wrong column once and the wrong subject in another. The second subject did not commit any errors while using the web interface and did not express any concerns regarding confusing one task or column with another.

The first member has expressed some confusion in the tasks wording, where he complained about the order of the requirements of each usability task. For example, each usability task used to ask the users to retrieve a certain column in a certain test for a certain subject. However, when performing the tasks, the users will first select the task then the columns then the subject ID. Therefore, the wordings of the tasks were changed to reflect the order in which the users expect to perform the requirements of each usability task. Furthermore, in the initial usability testing design, one question of the PSSUQ was adapted to give to the participants after completing each task. However, after testing the first subject, and after observing their

confusion when answering this question, the After-Scenario Questionnaire (ASQ) was adapted as mentioned in the usability testing section. Finally, after all these modifications to the interface and the usability testing materials, the second member has expressed her satisfaction with the entire process, and the usability testing team did not find any issues to be fixed before conducting the usability study.

4.2 USABILITY STUDY

4.2.1 Participants

The participants of this study were sampled from the students, faculty members, and staff of the School of Health and Rehabilitation Sciences and the School of Computing and Information Science at the University of Pittsburgh. Participants were recruited using flyers that were distributed in these two schools. Most of the participants were between 18 and 34 years old 75% (18/24), 58% (14/24) were females. Ten participants (41%) reported that they have been using Excel for at least ten years. On a scale from zero to ten (higher is better), ten participants (41%) evaluated themselves to have Excel skills between 7 and 10 (Table 4).

Table 4. *Participant characteristics (N=24).*

Variable	N (%)
Sex	
Female	14 (58%)
Male	10 (42%)
Age (years)	
18-24	9 (37.5%)
25-34	9 (37.5%)
35-60	5 (21%)
61-75	1 (4.1%)
Education	
<College	1 (4.1%)
College graduate	14 (58.3%)
Master's	7 (29.2%)
PhD	2 (8.3%)
Race	
White	11 (45.8%)
African American, Black	2 (8.3%)
Asian	11 (45.8%)
Excel years of experience	
1-3	5 (20.8%)
4-6	4 (16.6%)
7-9	5 (20.8%)
+10	10 (41.6%)
Excel skills level	
1-3	1 (4.1%)
4-6	13 (54.2%)
7-10	10 (41.6%)

4.2.2 Efficiency

The efficiency of each one of the data retrieval and management methods was measured by the time, in seconds, that took participants to perform the data retrieval tasks while using it. As mentioned before, each participant was asked to perform 10 data retrieval tasks using each one of

the data retrieval and management methods. The participants' performance times were not normally distributed based on examining the generated density plots and the results of the Shapiro Wilk's normality test. Therefore, Wilcoxon signed-rank test paired samples was used, which is a non-parametric test that is equivalent to the paired Student's t-test, to test the significance of the difference. As mentioned in the methods section, it is fair to compare the iRDMS and Excel on the performance of the participants on the first three tasks only since they involve pure data retrieval. Other tasks involve data processing when using Excel, which is not possible when using iRDMS since all data processing was already conducted automatically. However, other tasks were used to compare the performance on Excel with and without data processing and to gain insights on how both, data processing and data archiving, have increased the usability of the dataset. The participants' performance time on each of the ten tasks using the iRDMS was significantly faster than using Excel. Using Excel, the task that the participants performed the fastest was task No.1 ($M = 63.46$, $SD = 26.39$, range = 27-147), and the slowest was task No.8 ($M = 228.92$, $SD = 136.25$, range = 72-731). Using the iRDMS, the task that the participants performed the fastest was task No.4 ($M = 23.88$, $SD = 4.82$, range = 17-34), and the slowest was task No.5 ($M = 31.29$, $SD = 7.11$, range = 19-51). Participants' performance on the two methods only correlated on task 1 and 9, others were not significant (Table 5).

Table 5. Participants' performance, in seconds, on each one of the ten tasks on both methods.

ID	Excel			iRDMS			Difference and correlation		
	M (SD)	RSD	Range	M (SD)	RSD	Range	MD	Wilcoxon test	<i>Spearman</i>
1	63.46 (26.39)	0.42	27-147	25.04 (7.10)	0.28	13-47	38.4	300**	0.56*
2	79.75 (43.34)	0.54	35-206	25.29 (6.67)	0.26	14-40	54.5	300**	0.2
3	94.08 (34.40)	0.37	32-183	30 (8.74)	0.29	18-49	64.1	300**	0.27
4	80.67 (24.70)	0.31	50-142	23.88 (4.82)	0.2	17-34	56.8	300**	0.15
5	207.75 (52.83)	0.25	110-315	31.29 (7.11)	0.23	19-51	176.5	300**	0.33
6	151.96 (38.22)	0.25	81-212	26.21 (6.22)	0.24	17-39	125.8	300**	0.4
7	173.08 (133.07)	0.77	86-761	30.29 (18.05)	0.6	16-109	142.8	300**	0.34
8	228.92 (136.25)	0.6	72-731	27.46 (7.51)	0.27	18-53	201.5	300**	0.35
9	148.71 (54.10)	0.36	79-323	26.71 (9.16)	0.34	16-55	122.0	300**	0.52*
10	139.38 (38.25)	0.27	67-202	24.71 (7.65)	0.31	14-45	114.7	300**	0.33

ID = Task ID, V-value = Wilcoxon signed-rank test (Paired samples), SD = Standard Deviation, RSD = Relative Standard Deviation (SD/mean), m = average of participants' performance on all ten tasks for each of the two methods, MD= mean difference, *P<.01, **P<.001.

Furthermore, Kruskal-Wallis rank sum test and post hoc Nemenyi test were performed, which are non-parametric test equivalent to one-way ANOVA, to investigate whether the participants have performed significantly different on the ten tasks using each method (Kruskal & Wallis, 1952; Nemenyi, 1963). Kruskal-Wallis rank sum test suggested that there is a significant interaction between the performance on the ten task when using the iRDMS $X^2= 21$, $df = 9$, $p\text{-value} = 0.01$. Therefore, the post hoc Nemenyi test was performed, which revealed that task 5 was significantly slower than task 4 and 10 with $X^2= 4.88$, $p\text{-value} = .02$ and $X^2= 4.56$, $p\text{-value} = .04$, respectively (Table 6). Furthermore, Kruskal-Wallis rank sum test suggested that there is a significant interaction between the performance on the ten task when using the Excel $X^2= 140$, $df = 9$, $p\text{-value} < 0.001$. Therefore, the post hoc Nemenyi test was performed, which revealed that tasks 5,6,7,8,9 and 10 were significantly slower than tasks 1,2,3 and 4 except for task 9 and 10 compared to task 3 (Table 7).

Table 6. Participants' performance on tasks using iRDMS Pairwise comparisons using Tukey and Kramer (Nemenyi) test with Tukey-distribution approximation.

TaskID	1	2	3	4	5	6	7	8	9
2	0.28	-							
3	2.84	2.55	-						
4	0.63	0.92	3.47	-					
5	4.25	3.97	1.41	4.88*	-				
6	0.89	0.60	1.95	1.52	3.36	-			
7	1.45	1.17	1.39	2.08	2.80	0.56	-		
8	1.58	1.30	1.26	2.21	2.67	0.69	0.13	-	
9	0.54	0.26	2.30	1.17	3.71	0.35	0.91	1.04	-
10	0.31	0.59	3.15	0.32	4.56*	1.20	1.76	1.89	0.85

* $P < .05$

Table 7. Participants' performance on tasks using Excel Pairwise comparisons using Tukey and Kramer (Nemenyi) test with Tukey-distribution approximation.

TaskID	1	2	3	4	5	6	7	8	9
2	1.4	-							
3	2.88	1.49	-						
4	1.56	0.16	1.32	-					
5	10.99**	9.59**	8.1**	9.43**	-				
6	7.96**	6.56**	5.08*	6.4**	3.03	-			
7	7.52**	6.13**	4.64*	5.96**	3.46	0.44	-		
8	10.32**	8.92**	7.44**	8.76**	0.67	2.36	2.8	-	
9	7.35**	5.96**	4.47	5.8**	3.63	0.61	0.17	2.97	-
10	7.04**	5.64**	4.15	5.48**	3.95	0.92	0.49	3.28	0.32

* P<.05 ** P<.01

Moreover, the performance of each participant on the ten tasks was analyzed to investigate the change in performance between using Excel and the iRDMS on individuals' level. When only considering the first three tasks, which did not involve any data processing and only require pure data retrieval, the fastest average time on the three tasks using the iRDMS was (M = 15.67, SD = 3.09, range = 13-20) and the slowest was (M = 41.67, SD = 3.86, range = 38-47). The fastest average time on the ten tasks using the Excel was (M = 37.67, SD = 6.02, range = 32-46) and the slowest was (M = 135.33, SD = 38.39, range = 89-183). It is worth mentioning here that the fastest times on both methods were scored by the same participant and slowest times on both methods were scored by the same participant too. However, since there were only three scores per participant, a mean difference between the two methods on the first three tasks for each participant could not be performed, and this analysis was only performed across participants as mentioned before. Furthermore, the average RSD on the performance of the first three tasks was 0.14 for the iRDMS and 0.32 for Excel (Table 8).

The fastest average time on the ten tasks using the iRDMS was (M = 17.80, SD = 3.89, range = 13-25) and the slowest was (M = 46.10, SD = 22.12, range = 33-109). The fastest

average time on the ten tasks using the Excel was ($M = 88.40$, $SD = 31.04$, range = 44-145) and the slowest was ($M = 46.10$, $SD = 22.12$, range = 33-109). The overall participants' performance time on the ten tasks using the iRDMS was significantly faster than using Excel for each participant. Furthermore, the average Relative Standard Deviation (RSD), which measures how diverse the performance of each participant on different tasks using the same method, was 0.21 for the iRDMS and 0.47 for Excel. Detailed are provided in APPENDIX (B).

Table 8. Participants' performance on the first three tasks using Excel and iRDMS.

Subject ID	Excel			iRDMS		
	M (SD)	RSD	Range	M (SD)	RSD	Range
1	103.67 (33.98)	0.33	64-147	37.67 (5.91)	0.16	33-46
2	120 (37.96)	0.32	69-160	31 (3.74)	0.12	27-36
3	135.33 (38.39)	0.28	89-183	41.67 (3.86)	0.09	38-47
4	89.33 (32.27)	0.36	56-133	38.67 (3.86)	0.1	35-44
5	66.67 (28.08)	0.42	27-88	22 (6.38)	0.29	17-31
6	77.67 (22.29)	0.29	59-109	21 (0.82)	0.04	20-22
7	80.67 (20.27)	0.25	52-95	19 (1.41)	0.07	18-21
8	56.33 (21.75)	0.39	39-87	23 (4.55)	0.2	17-28
9	79.67 (33.07)	0.42	35-114	21.33 (0.94)	0.04	20-22
10	72.67 (19.36)	0.27	53-99	24 (2.16)	0.09	21-26
11	66.33 (22.13)	0.33	38-92	23.33 (3.09)	0.13	19-26
12	37.67 (6.02)	0.16	32-46	15.67 (3.09)	0.2	13-20
13	65.33 (23.70)	0.36	37-95	21 (2.83)	0.13	19-25
14	83.67 (44.50)	0.53	38-144	31 (7.87)	0.25	24-42
15	71 (11.86)	0.17	57-86	22 (2.83)	0.13	18-24
16	63.67 (15.52)	0.24	45-83	26 (0.82)	0.03	25-27
17	72.67 (23.01)	0.32	41-95	35.67 (9.98)	0.28	25-49
18	72 (21.12)	0.29	44-95	29 (2.16)	0.07	27-32
19	66.67 (23.89)	0.36	33-86	23.33 (3.77)	0.16	18-26
20	101.33 (55.78)	0.55	57-180	31 (6.48)	0.21	25-40
21	57.33 (16.86)	0.29	43-81	25.67 (5.91)	0.23	21-34
22	122 (65.22)	0.53	47-206	27 (2.16)	0.08	24-29
23	62.33 (13.20)	0.21	45-77	24.33 (2.87)	0.12	21-28
24	74.33 (15.63)	0.21	53-90	28.33 (6.34)	0.22	22-37

ID = Participants ID, Wilcoxon test = Wilcoxon signed-rank test (Paired samples), SD = Standard Deviation, RSD = Relative Standard Deviation (SD/mean), m = average of participants' performance on all ten tasks for each of the two methods, iRDMS= Internet-Based Research Data Management System.

Furthermore, the participants were grouped according to their age, gender, education, and race and investigated whether there is a significant performance difference between the groups when using Excel and the iRDMS. An ANOVA with a factorial design that investigates the interaction between these variables would have been very informative, but it was not feasible to conduct such analysis due to the small sample size. Therefore, Kruskal-Wallis rank sum test was

performed, and post hoc Nemenyi test, which are non-parametric test equivalent to one-way ANOVA since the data violate the assumptions of normality of residuals, to perform this analysis. Groups that have less than three participants were excluded since two participants and less are not sufficient to perform the analysis. For example, in education groups, there were only two participants with Ph.D. and only one with less than Bachelor's degree, and both were excluded. Therefore, for some comparisons, both, the Kruskal-Wallis rank sum test and the post hoc Nemenyi test gave the same results since only two groups were included in the analysis. Furthermore, the relationship between the participants Excel skills and experience with their performance on the ten tasks using both methods were investigated.

When analyzing age groups 18-24, 25-34, and 35-60 (the group 61-75 was excluded since there was only one participant), Kruskal-Wallis rank sum test suggested that there is a significant difference between the groups performance on task 1 using Excel $X^2= 10.47$, $df = 2$, $p\text{-value} < .01$. Therefore, the post hoc Nemenyi test was performed, which revealed that group 25-34 ($M = 44.78$, $SD = 12.45$, $range = 27-69$) was significantly faster than group 35-60 ($M = 92$, $SD = 29.83$, $range = 59-147$) $X^2= 4.38$, $p\text{-value} < .001$. However, when adjusting for type I error, the $p\text{-value}$ of the Kruskal-Wallis rank sum test becomes 0.44, which suggests that the difference here is by chance. Moreover, the difference between the education groups BS and MS (groups <BS and PhD were excluded for insufficient number of participants) was tested. Kruskal-Wallis rank sum test suggested that the performance on task 3 using Excel of participants who finished at least BS ($M = 85.43$, $SD = 37.08$, $range = 32-183$) was significantly faster than the performance of the group that at least finished their MS ($M = 114.57$, $SD = 25.95$, $range = 81-160$) $X^2= 4.69$, $df = 1$, $p\text{-value} = .03$. However, when adjusting for type I error, the $p\text{-value}$ of the Kruskal-Wallis rank sum test becomes 0.58, which suggests that the difference here

is by chance. Furthermore, difference between the race groups Asian and White (group African American was excluded for insufficient number of participants) was tested. Kruskal-Wallis rank sum test suggested that the performance on task 4 using Excel of White ($M = 72.45$, $SD = 24.34$, range = 50-142) participants was significantly faster than the performance of Asian participants ($M = 90.64$, $SD = 21.97$, range = 58-132) $X^2 = 4.43$, $df = 1$, $p\text{-value} = .04$. However, when adjusting for type I error, the $p\text{-value}$ of the Kruskal-Wallis rank sum test becomes 0.58, which suggests that the difference here is by chance. Even more, the performance on task 8 using iRDMS of Asian participants ($M = 24.73$, $SD = 3.84$, range = 20-31) was significantly faster than the performance of White participants ($M = 31.18$, $SD = 8.87$, range = 19-53) $X^2 = 4.17$, $df = 1$, $p\text{-value} = .04$. However, when adjusting for type I error, the $p\text{-value}$ of the Kruskal-Wallis rank sum test becomes 0.57, which suggests that the difference here is by chance. Additionally, the performance on task 10 using iRDMS of Asian participants ($M = 22.45$, $SD = 8.13$, range = 15-45) was significantly faster than the performance of White participants ($M = 27.45$, $SD = 6.01$, range = 21-43) $X^2 = 5.16$, $df = 1$, $p\text{-value} = .02$. However, when adjusting for type I error, the $p\text{-value}$ of the Kruskal-Wallis rank sum test becomes 0.57, which suggests that the difference here is by chance. There was no significant difference in the performance of males and females on any task using both method. There was no significant difference in the performance of any groups averaged across tasks using both methods. Detailed results are provided in APPENDIX (B).

In addition, it was investigated whether the order in which the tasks were presented to the subjects correlates with the participants' performance on the tasks using the two methods. There was a significant negative correlation between the participants' performance on task 1, 9 and 10 using Excel and the order in which they have received these tasks $r = -.59$, $p\text{-value} < .01$, $r = -$

.51, p -value $< .01$, and $r = -.64$, $p < .001$, respectively. All of these p -values are still significant even after the type I error adjustment. Moreover, there was a significant negative correlation between the participants' performance on task 2, 3, 6, 7 and 9 using iRDMS and the order in which they have received these tasks $r = -.49$, p -value = .01, $r = -.5$, p -value = .01, $r = -.48$, p -value = .01, $r = -.52$, p -value $< .01$, and $r = -.54$, $p < .01$, respectively. All of these p -values are also still significant even after the type I error adjustment. Other correlations besides these significant correlations ranged between $-.01$ and $-.40$, none of which was significant. Even more, the effect of asking the participant to perform the tasks using one method first or the other was investigated. As mentioned before, the tasks for each participant were randomized so that participants use the iRDMS first for five tasks and use Excel for the other five. Only one task was affected by the order of the used method, task 7 using iRDMS $W = 12.5$, p -value = .04. However, the number of participants who performed this task first to the ones who performed it second was 20 to 4, which indicates a large imbalance in the two samples and probable bias results. Furthermore, after the adjustment for type I error, the p -value became .16, which suggests that the order effect here is by chance. Other results show that the participants' performance using Excel or iRDMS was not affected by whether they used Excel first or not (Table 9).

Table 9. Participants' performance, in seconds, on each one of the ten tasks categorized by the order of methods they used on both methods.

Method	TaskID	1st		2nd		Kruskal-Wallis	
		N	M(SD)	N	M(SD)	W-value	P-value
iRDMS	1	8	26.75(9.57)	16	24.19(5.96)	70	0.74
	2	12	27.50(6.96)	12	23.08(6.17)	106	0.05
	3	14	28.29(6.51)	10	32.40(11.47)	57	0.46
	4	11	23.18(5.25)	13	24.46(4.77)	62.5	0.62
	5	7	30.71(6.68)	17	31.53(7.67)	56	0.85
	6	15	27.80(5.72)	9	23.56(6.78)	101	0.05
	7	20	26.15(8.14)	4	51(38.76)	12.5	0.04
	8	8	29.25(11.54)	16	26.56(5.06)	67	0.88
	9	13	27.23(8.64)	11	26.09(10.53)	81.5	0.58
	10	11	25.73(7.18)	13	23.85(8.50)	83	0.52
Excel	1	16	66.88(29.25)	8	56.62(21.78)	79	0.37
	2	12	82(36.10)	12	77.50(52.77)	93	0.24
	3	10	103.20(37.59)	14	87.57(33.12)	96	0.14
	4	13	80.23(29.10)	11	81.18(21.17)	65	0.73
	5	17	207.47(56.18)	7	208.43(52.44)	60.5	0.97
	6	9	156.22(31.81)	15	149.40(43.68)	75	0.68
	7	4	337.50(287.96)	20	140.20(49.45)	65	0.06
	8	16	243.19(164.35)	8	200.38(66.09)	56	0.65
	9	11	147.18(66.15)	13	150(46.94)	63	0.64
	10	13	135.08(42.22)	11	144.45(36.32)	63.5	0.66

1st = Indicates that this method was used first. 2nd = Indicates that this method was used second, iRDMS= Internet-based Research Data Management System.

Also, it was investigated whether participants' answers on the two Excel skills level and experience questions correlate or predict their performance on the tasks using the two methods. There was a significant correlation between participants' number of years using Excel and their performance of tasks 6, 7 and overall average across tasks using the iRDMS $r = .52$, $p\text{-value} < .01$, $r = .56$, $p\text{-value} < .01$, and $r = .44$, $p = .03$, respectively. After adjusting for type I error, the $p\text{-value}$ for the correlation between the overall average across tasks and number of years using Excel became .11, where the $p\text{-values}$ for tasks 6 and 7 remained significant, .047 and .046,

respectively. Other correlations besides these significant correlations ranged between -.22 and .37, none of which was significant.

4.2.3 Effectiveness

The number of tasks that participants did not perform correctly was analyzed to investigate whether there is a significant difference in the effectiveness of the two data retrieval methods. When only considering the first three tasks, participants have made 3 errors when using Excel and only one error when using iRDMS. When considering all the tasks, participants have incorrectly performed 16 tasks out of the 240 tasks they performed using Excel and 3 tasks out of 240 tasks they performed using the iRDMS. Chi-square test of equality of proportions suggested that the participants performed incorrect tasks using Excel (16 tasks out of the 240 tasks) significantly higher than using iRDMS (3 tasks out of the 240 tasks) $X^2= 7.9$, $df = 1$, $p\text{-value} < .01$ (Table 10). Errors were distributed across participants, and there was not any specific group with a higher number of errors, detailed are in APPENDIX (B).

Table 10. *Number of incorrect performances on each of the tasks using Excel and iRDMS.*

Task ID	Excel	iRDMS
1	1	1
2	1	0
3	1	0
4	2	0
5	3	0
6	0	0
7	1	0
8	1	1
9	3	0
10	3	1

iRDMS= Internet-based Research Data Management System.

4.2.4 Satisfaction

As mentioned in the methods section, the PSSUQ was used to measure the participants' satisfaction with the iRDMS. PSSUQ can be used to measure three sub-scales, System Usefulness (SysUse), Information Quality (InfoQual) and Interface Quality (IntQual) along with the overall score (Overall). The sub-scale SysUse, relatively, acquired the highest satisfaction and the InfoQual acquired the lowest. There were no significant correlations between participants' satisfaction and their performance on the iRDMS or Excel nor with their Excel skills or experience (Table 11). Detailed scores on each demographic group response on PSSUQ are provided in APPENDIX (B).

Table 11. *PSSUQ overall scores and each sub-scale with their correlations with different measures.*

Scale	M (SD)	Range	Correlations			
			VS Experience	VS Skills	VS Excel	VS iRDMS
SysUse	1.13 (0.3)	1-2.43	-0.25	-0.35	0.12	-0.20
InfoQual	1.27 (0.34)	1-2.2	0.07	-0.17	0.12	0.09
IntQual	1.25 (0.35)	1-2.33	0.04	0.03	-0.15	-0.14
Overall	1.21 (0.27)	1-2.25	0.07	-0.12	0.05	-0.08

SysUse = System Usefulness, InfoQual = Information Quality, IntQual = Interface Quality, and Overall = overall score, iRDMS= Internet-based Research Data Management System.

Furthermore, the responses from the participants on the two questions that were acquired from the After-Scenario Questionnaire (ASQ) were examined. Participants were significantly more satisfied with using the iRDMS to perform each one of the ten tasks than using Excel, which is reflected in their answers to both questions (Table 12). The task with least participants' satisfaction when using Excel from both, ease of completion and the time it takes to be performed perspective, was task 5 (M = 4.58, SD = 1.78, range = 2-7) and (M = 4.67, SD = 1.72, range = 1-7), respectively. The task with least participants' satisfaction when using iRDMS from

ease of completion perspective was task 2 ($M = 1.38$, $SD = 0.63$, range = 1-3). Tasks 5 and 6 were tied as the least participants' satisfaction when using iRDMS the time it takes to be performed perspective ($M = 1.29$, $SD = 0.45$, range = 1-2) and ($M = 1.29$, $SD = 0.61$, range = 1-3), respectively. The task with highest participants' satisfaction when using Excel from both, ease of completion and the time it takes to be performed perspective, was task 1 ($M = 3$, $SD = 1.58$, range = 1-7) and ($M = 2.83$, $SD = 1.55$, range = 1-7), respectively. The task with highest participants' satisfaction when using iRDMS from ease of completion perspective was task 1 ($M = 1.12$, $SD = 0.33$, range = 1-2). The task with highest participants' satisfaction when using iRDMS from the time it takes to be performed perspective was task 7 ($M = 1.04$, $SD = 0.2$, range = 1-2). Detailed scores on each demographic group response on ASQ are provided in APPENDIX (B).

Table 12. ASQ scores on each task and overall with their test of difference between Excel and the iRDMS.

Task	Q	Excel			iRDMS			Wilcoxon test
		M (SD)	RSD	Range	M (SD)	RSD	Range	W-value
1	1	3 (1.58)	0.53	1-7	1.12 (0.33)	0.29	1-2	171**
	2	2.83 (1.55)	0.55	1-7	1.08 (0.28)	0.26	1-2	171**
2	1	3.42 (1.58)	0.46	1-7	1.38 (0.63)	0.46	1-3	210**
	2	3.38 (1.89)	0.56	1-7	1.25 (0.52)	0.42	1-3	190**
3	1	3.54 (1.66)	0.47	1-7	1.17 (0.37)	0.32	1-2	210**
	2	3.58 (1.82)	0.51	1-7	1.12 (0.33)	0.29	1-2	190**
4	1	3.50 (1.63)	0.47	1-7	1.17 (0.47)	0.4	1-3	249.5**
	2	3.67 (1.75)	0.48	1-7	1.12 (0.44)	0.39	1-3	249**
5	1	4.58 (1.78)	0.39	2-7	1.21 (0.41)	0.34	1-2	300**
	2	4.67 (1.72)	0.37	1-7	1.29 (0.45)	0.35	1-2	276**
6	1	4.08 (1.78)	0.44	1-7	1.25 (0.52)	0.42	1-3	231**
	2	3.96 (1.84)	0.46	1-7	1.29 (0.61)	0.47	1-3	231**
7	1	4.12 (1.96)	0.48	1-7	1.12 (0.33)	0.29	1-2	253**
	2	4.29 (2.21)	0.51	1-7	1.04 (0.20)	0.19	1-2	253**
8	1	4.54 (2.00)	0.44	1-7	1.21 (0.41)	0.34	1-2	231**
	2	4.58 (1.82)	0.4	2-7	1.25 (0.52)	0.42	1-3	276**
9	1	4.08 (1.58)	0.39	1-7	1.21 (0.41)	0.34	1-2	276**
	2	4 (1.87)	0.47	1-7	1.08 (0.28)	0.26	1-2	253**
10	1	4.25 (2.03)	0.48	1-7	1.25 (0.52)	0.42	1-3	231**
	2	4.25 (1.92)	0.45	1-7	1.17 (0.37)	0.32	1-2	231**

** <.001, iRDMS= Internet-based Research Data Management System, RSD= Relative Standard Deviation.

Also, the answers of the participants on the two questions and how they correlate with their performance using Excel and iRDMS along with the task order, Excel skills, and Excel experience were examined. Most of the significant correlation has been between the participants' answers on the ASQ questions after performing tasks using Excel and their performance on these tasks. However, there is still some significant correlations between participants' answers on the ASQ questions and their performance using iRDMS and the other included variables too (Table 13).

Table 13. Significant correlations between ASQ scores with participants' performance on Excel and iRDMS, task order and Excel skills and experience.

Q	Task	M (SD)	ASQ VS				
			Excel	iRDMS	Order	Skills	Experience
Excel-Q1	2	3.42 (1.61)	0.7**	0.13	-0.42*^	-0.12	0.03
Excel-Q1	4	3.50 (1.67)	0.49**	-0.16	-0.35	-0.08	-0.12
Excel-Q1	7	4.12 (2.01)	0.55**	-0.09	0.07	-0.06	-0.07
Excel-Q1	8	4.54 (2.04)	0.27	-0.46*^	-0.11	-0.29	-0.32
Excel-Q1	10	4.25 (2.07)	0.65**	-0.13	-0.31	-0.44*^	-0.35
Excel-Q2	2	3.38 (1.93)	0.71**	0.13	-0.37	-0.11	0.03
Excel-Q2	4	3.67 (1.79)	0.56**	0.05	-0.22	-0.15	-0.17
Excel-Q2	5	4.67 (1.76)	0.5**	-0.03	0.09	-0.29	-0.29
Excel-Q2	7	4.29 (2.26)	0.74**	0.03	-0.06	-0.14	0.02
Excel-Q2	8	4.58 (1.86)	0.5**	-0.35	-0.08	-0.19	-0.19
Excel-Q2	10	4.25 (1.96)	0.77**	0.04	-0.38	-0.49*^	-0.35
iRDMS-Q1	4	1.17 (0.48)	-0.32	-0.44*^	-0.04	-0.21	0.09

** <.01. *<.05. ^=Not significant after adjustment for type I error. iRDMS= internet-based Research Data Management System, M= Mean, Order= Task Order.

Furthermore, the correlation between the two ASQ questions for Excel and iRDMS was investigated. The Spearman correlations results show that the answers to the two questions for each method correlate with each other but not for the questions of the other method on each of the ten tasks across participants (Table 14).

Table 14. Correlations between ASQ questions on Excel and iRDMS for each task.

Task ID	iRDMS Q1 VS			iRDMS Q2 VS		Excel Q1 VS
	iRDMS Q2	Excel Q1	Excel Q2	Excel Q1	Excel Q2	Excel Q2
1	0.8**	0.31	0.23	0.24	0.23	0.94**
2	0.83**	0.2	0.06	0.11	0.03	0.92**
3	0.51*	0.06	-0.08	0.02	-0.01	0.92**
4	0.82**	0.05	-0.22	-0.1	-0.31	0.9**
5	0.57*	0.12	0.08	0.21	0.17	0.9**
6	0.73**	0.23	0.1	0.24	0.16	0.88**
7	0.55*	0.01	-0.1	0.06	-0.22	0.88**
8	0.75**	-0.1	-0.11	-0.03	-0.1	0.91**
9	0.59*	0.11	-0.02	0.09	-0.06	0.88**
10	0.62*	0.29	0.38	0.33	0.22	0.94**

** <.001. *<.01, iRDMS= Internet-based Research Data Management System.

Furthermore, as part of the satisfaction assessment, the users were asked to answer four open-ended questions. For the first question “What did you like about the system?”, participants expressed that the iRDMS was simple, easy to go between data, very easy once you learn how to use, the interface was not busy, easy to understand, learn and use, time utilization was less, organized, and quick. For the second question “What do you dislike about the system?”, only one participant expressed that the iRDMS needs more graphics, others answered with nothing. For the third question “If you could change one thing in the system, what would it be?”, participants expressed that they would add more functions, put the data selection options on the web-page and not in popup windows, and change the color of sample data that is viewed in step 3 in the system. For the fourth question “What did you find confusing or a problem in the system?”, the participants said nothing because the introduction to the system was clear and that the study administrator has answered all their questions, and not confusing but something new that takes time to get used to.

Moreover, since the participants were encouraged to perform the tasks while “thinking aloud” and saying what they like and dislike, participants have provided good feedback of their emotions while performing the tasks. The general theme was that they were frustrated while using Excel as they tend to forget or jump a step, or when Excel acts in an unexplainable way due to the high load of the dataset. Furthermore, participants were mostly happy when they learn that the next task will be performed using the iRDMS by expressing that they like the iRDMS more, they did not have to think much while using it, or that the iRDMS was easier than Excel. Also, some participants have expressed their discomfort on performing the tasks on an Apple MacOS instead of Microsoft Windows, their native OS environment. Furthermore, some participants have suggested some design changes while they were using the iRDMS, such as add search function to narrow down the listed columns or the values to choose from in selection customization and change the color of the provided sample data from red to any other color, because some participants interpreted red as some error message. These suggestions were implemented in the system by adding the search function, changing the color to be green unless there is no sample data available, and made the back/next buttons always visible as it was noticed that some participants were confused what to do next after choosing tasks or columns.

4.3 DATA ANALYSIS (SENTENCE COMPREHENSION)

4.3.1 Participants

One hundred PWA were recruited to participate in the McNeil et al. (2014) study. All participants were native speakers of English, 18 years or older, and have at least 8 years of education. Most participants were males given the gender distribution within the veteran population, but females were recruited and balanced across groups. Race and ethnicity were consistent with population distributions in the greater Pittsburgh, Seattle, Martinez, and Philadelphia areas (Table 15). PWA had clinical characteristics of aphasia consistent with the McNeil and Pratt (2001) definition of aphasia, which includes: a) acquired language processing deficits due to the language-dominant hemisphere brain lesion due to stroke, as confirmed by clinical neurology or brain imaging reports, b) language processing deficits that cross modalities as measured by the language battery from the Comprehensive Aphasia Battery CAT (Swinburn, Porter, & Howard, 2004), c) be without medical record or self-report history of degenerative nervous system illness, dementia, schizophrenia, manic-depression, or schizoaffective disorder. A screening subtest for each CRTT-R condition assesses knowledge of all vocabulary, ability to see and read the lexical items for 3 of the 4 outcome measures, ability to select each word in the self-paced reading tasks, color and shape perception of the response items, the visual-manual ability to select the token on the screen and overall ability to follow the task demands. Participants had to perform at 100% accuracy (CRTT-R scores of 15) on screening items before any test items were administered.

Table 15. *Participants characteristics (N=100).*

Variable	N (%)	Mean (SD)
Sex		
Female	32 (32%)	
Male	68 (68%)	
Age (years)		
26-34	2 (2%)	
35-45	4 (4%)	
46-60	31 (31%)	64 (11)
61-75	54 (54%)	
76-84	9 (9%)	
Education (years)		
10-13	23 (23%)	
14-17	60 (60%)	15 (3)
18-21	14 (14%)	
22-24	3 (3%)	
Race/Ethnicity		
American Indian/Alaskan Native	2 (2%)	
Asian	3 (3%)	
Black/African-American	12 (12%)	
Hispanic/Latino	1 (1%)	
White/Caucasian	77 (77%)	
Mixed	4 (4%)	
Other	1 (1%)	
MPO (Months)		
4-53	37 (37%)	
54-106	26 (26%)	84 (56)
107-170	28 (28%)	
171-216	9 (9%)	

MPO= Months Post Onset (Stroke), M= Mean, SD= Standard Deviation.

4.3.2 Data Descriptions

As mentioned in the methods section, data that describes the participant's performance on 21 tasks that were hypothesized to measure their LP, STM, WM, and CR were collected. Accuracy of responses were used as the dependent measure for all tasks. The tasks that were hypothesized to measure the CR effect were calculated as the difference in accuracy between the items that do not require interference inhibition and the items that do (control tasks). Therefore, the

participants' scores for these tasks could be negative, which means that participant's performance better when they were required to inhibit interference or in the control condition. A positive score means that the participant's performed with fewer correct items than the control condition. A score of zero indicates that there was no difference between the CR (incongruent) condition and the control condition (Table 16). All scores were scaled and centered using the Scale function in R, which does not affect the distribution of the scores and only standardize the scales across tasks.

Table 16. *Task Descriptive Statistics.*

Task Name	N	Mean (SD)	Median	Range	Possible Range
CR2	100	2.57 (8.51)	1	-7-68	(-74,74)
CR3	100	0.83 (3.66)	0	-6-34	(-36,36)
CRTT-R-Stroop	79	11.10 (12.31)	9	-25-45	(-80,80)
LPphon1	100	51.41 (7.93)	54	22-60	(0,60)
LPphon2	99	21.27 (5.22)	22	9-30	(0,30)
LPphon3	100	20.72 (4.51)	21	12-29	(0,32)
LPsem1	100	30.51 (4.22)	32	17-36	(0,36)
LPsem2	99	44.91 (5.02)	46	26-52	(0,52)
LPsem3	100	66.06 (11.21)	71	23-72	(0,72)
LPsyn1	100	33.81 (5.25)	35	20-40	(0,40)
LPsyn2	100	32.19 (4.59)	32	20-40	(0,45)
LPsyn3	99	9.88 (5.11)	12	0-15	(0,15)
STMphon1	98	2.67 (2.09)	2	0-6	(0,6)
STMphon2	99	21.65 (4.86)	22	7-30	(0,30)
ST/WMphon3	98	104.84 (15.24)	108	38-132	(0,141)
STMsem1	100	1.65 (1.64)	1	0-6	(0,6)
STMsem2	99	27.51 (5.27)	28	12-36	(0,36)
ST/WMsem3	97	108.56 (14.39)	107	62-141	(0,141)
STMsyn1	100	0.86 (1.23)	0	0-6	(0,6)
STMsyn2	97	20.45 (3.61)	22	6-26	(0,30)
ST/WMsyn3	96	114.77 (17.43)	120	59-138	(0,141)
LA	100	4.30 (3.55)	4	-14-13	(-15,15)
CS	83	10.58 (2.36)	10.97	3.88-14.3	(0,15)
GP	81	10.00 (0.83)	9.98	7.67-12.16	(0,15)
OC	80	10.39 (1.08)	10.19	4.79-12.88	(0,15)

STMp1= Auditory Phonological Rhyme Span--Words, **LPp1**= Auditory Phonological Rhyme Judgment - Words, **STMp2**= Auditory Phonological Rhyme Judgment Span-Words, **LPsy1**= Auditory Syntactic Grammaticality Judgments- Sentence, **LPp2**= Reading Phonological Rhyme Judgments-Words, **LPsy3**= Reading Syntactic Anagram-Words, **ST/WMsm3**= Auditory Semantic N-Back-Words, **ST/WMsy3**= Auditory Syntactic N-Back-Sentences, **STMp3**= Auditory Phonological N-Back-Words, **LPsm1**= Auditory Semantic Category Judgments-Words, **LPsm3**= Visual Semantic PWI with living/nonliving Judgments-Pictures & Words, **STMsm1**= Auditory Semantic Category Judgment Span-Words, **STMsy1**= Auditory Syntactic Span-Sentence, **LPsy2**= Visual & Written Syntactic Picture Matching-Sentence, **LPp3**= Visual Phonological Pictured Rhyme Judgments-Word, **LPsm2**= Visual Semantic Pyramids and Palm Trees-Words, **STM**= Short-term memory, **WM**= Working memory, and **LP**= Language processing, **CR2**= Picture-Word interference, **CR3**= Number Stroop, **CRTT-R-Stroop**= Computerized Revised Token Test (Stroop), **OC**= Relative Clause, **GP**= Garden Path, **CS**= compound sentences, **LA**= Lexical Ambiguity, **SD**= Standard Deviation, **N**= Count in each column out of 100, **Possible Range**= The possible scores of each task, **Range**= The range of the scores for this particular dataset.

4.3.3 Missing data

As presented in (Table 16), the collected data has suffered from missing data points. The main reason for all the missing data points, except for the CRTT, is due to a confusion in tasks presentation where the investigators presented one task twice instead of two different tasks for the same participant. For the CRTT, the high number of missing data points is due to a technical malfunction that resulted in an entire database being corrupted and deleted from one of the laptops that were used for data collection. Therefore, since all these data points were missing at random, the EM-ML algorithm was used to impute the missing data points.

4.3.4 EFA

Before conducting EFA, the correlation matrix that was obtained by corFIML function in the R psych package, which uses EM-ML method to estimate the missing values, was examined. In general, it was observed that most of the items correlated at least .3 with at least one other item, suggesting reasonable factorability (Table 17). Therefore, an EFA was performed using ML as the estimation method and Varimax as the rotation method using the “fa” function from the R psych package. As mentioned in the methods section, the criteria for choosing the number of factors to retain are: a) eigenvalue higher than one, b) factors that are before the last substantive drop in the eigenvalues, and c) factors with eigenvalues higher than the average of the eigenvalues of the factors that were randomly generated by parallel analysis. The results of the initial EFA model suggested six factors with eigenvalues higher than one. Furthermore, the scree plot was examined to locate the last significant drop in the eigenvalues, which showed a

significant gap between the second and the third factors suggesting that the first two factors should be retained. A parallel analysis was performed to gain insights into the number of factors to retain. The parallel analysis also supported the retention of the first two factors (Figure 14).

Table 17. Tasks Pearson Correlation matrix.

	LPp1	LPp2	LPp3	LPsy3	LPsy1	LPsy2	LPsm1	LPsm3	LPsm2	STMp3	STMsm3	STMsy3
LPp1	1	.31	.31	.25	.36	.28	.35	.29	.25	.19	.20	.03
LPp2	.31	1	.45	.58	.48	.57	.32	.35	.56	.49	.47	.32
LPp3	.31	.45	1	.43	.33	.44	.24	.31	.46	.32	.31	.16
LPsy3	.25	.58	.43	1	.41	.56	.38	.40	.51	.32	.43	.40
LPsy1	.36	.48	.33	.41	1	.38	.34	.36	.39	.34	.28	.30
LPsy2	.28	.57	.44	.56	.38	1	.31	.46	.56	.26	.33	.21
LPsm1	.35	.32	.24	.38	.34	.31	1	.50	.40	.29	.26	.20
LPsm3	.29	.35	.31	.40	.36	.46	.50	1	.51	.18	.28	.19
LPsm2	.25	.56	.46	.51	.39	.56	.40	.51	1	.33	.36	.37
STMp3	.19	.49	.32	.32	.34	.26	.29	.18	.33	1	.63	.52
STMsm3	.20	.47	.31	.43	.28	.33	.26	.28	.36	.63	1	.58
STMsy3	.03	.32	.16	.40	.30	.21	.20	.19	.37	.52	.58	1
STMsm1	.22	.36	.33	.42	.21	.52	.18	.20	.37	.26	.37	.20
STMsy1	.30	.34	.51	.35	.41	.44	.21	.13	.27	.25	.24	.09
STMp1	.46	.50	.41	.40	.44	.46	.22	.18	.34	.40	.30	.26
STMp2	.53	.34	.38	.37	.44	.42	.33	.25	.23	.37	.28	.17
STMsm2	.55	.27	.40	.35	.22	.25	.35	.22	.36	.14	.29	.19
STMsy2	.17	.30	.33	.38	.30	.48	.21	.14	.40	.12	.12	.10
CR2	-.05	-.17	-.11	-.25	-.12	-.23	-.18	-.12	-.43	-.17	-.16	-.23
CR3	.05	.14	-.14	.05	.06	-.02	.06	.04	.04	.08	.01	.14
CRTT-R-Stroop	.14	-.07	-.10	-.18	.06	-.15	.09	-.15	.05	-.08	-.13	-.07
OC	-.06	.28	.17	.22	.10	.33	.10	.01	.08	.33	.37	.25
GP	.36	.22	.36	.22	.24	.24	.05	.10	.21	.27	.44	.17
CS	.10	.30	.22	.29	.20	.25	.21	.02	.23	.22	.39	.12
LA	-.13	-.25	-.18	-.20	-.17	-.30	-.10	-.11	-.20	-.14	-.24	-.14

STMp1= Auditory Phonological Rhyme Span--Words, **LPp1**= Auditory Phonological Rhyme Judgment - Words, **STMp2**= Auditory Phonological Rhyme Judgment Span-Words, **LPsy1**= Auditory Syntactic Grammaticality Judgments- Sentence, **LPp2**= Reading Phonological Rhyme Judgments-Words, **LPsy3**= Reading Syntactic Anagram-Words, **STMsm3**= Auditory Semantic N-Back-Words, **STMsy3**= Auditory Syntactic N-Back-Sentences, **LPsm1**= Auditory Semantic Category Judgments-Words, **LPsm3**= Visual Semantic PWI with living/nonliving Judgments-Pictures & Words, **STMsm1**= Auditory Semantic Category Judgment Span-Words, **STMsy1**= Auditory Syntactic Span-Sentence, **LPsy2**= Visual & Written Syntactic Picture Matching-Sentence, **LPp3**= Visual Phonological Pictured Rhyme Judgments-Word, **LPsm2**= Visual Semantic Pyramids and Palm Trees-Words, **STM**= Short-term memory, **WM**= Working memory, and **LP**= Language processing, **CR2**= Picture-Word interference, **CR3**= Number Stroop, **CRTT-R-Stroop** = Computerized Revised Token Test (Stroop), **OC**= Relative Clause, **GP**= Garden Path, **CS**= compound sentences, **LA**= Lexical Ambiguity.

Table 17 (continued). Tasks Pearson Correlation matrix.

	STMsm1	STMsy1	STMp1	STMp2	STMsm2	STMsy2	CR2	CR3	CRTT-R-Stroop	OC	GP	CS	LA
LPp1	.22	.30	.46	.53	.55	.17	-.05	.05	.14	-.06	.36	.10	-.13
LPp2	.36	.34	.50	.34	.27	.30	-.17	.14	-.07	.28	.22	.30	-.25
LPp3	.33	.51	.41	.38	.40	.33	-.11	-.14	-.10	.17	.36	.22	-.18
LPsy3	.42	.35	.40	.37	.35	.38	-.25	.05	-.18	.22	.22	.29	-.20
LPsy1	.21	.41	.44	.44	.22	.30	-.12	.06	.06	.10	.24	.20	-.17
LPsy2	.52	.44	.46	.42	.25	.48	-.23	-.02	-.15	.33	.24	.25	-.30
LPsm1	.18	.21	.22	.33	.35	.21	-.18	.06	.09	.10	.05	.21	-.10
LPsm3	.20	.13	.18	.25	.22	.14	-.12	.04	-.15	.01	.10	.02	-.11
LPsm2	.37	.27	.34	.23	.36	.40	-.43	.04	.05	.08	.21	.23	-.20
STMp3	.26	.25	.40	.37	.14	.12	-.17	.08	-.08	.33	.27	.22	-.14
STMsm3	.37	.24	.30	.28	.29	.12	-.16	.01	-.13	.37	.44	.39	-.24
STMsy3	.20	.09	.26	.17	.19	.10	-.23	.14	-.07	.25	.17	.12	-.14
STMsm1	1	.46	.38	.31	.31	.34	-.11	.04	-.02	.20	.20	.32	-.16
STMsy1	.46	1	.51	.39	.34	.39	-.15	-.07	.04	.27	.34	.26	-.15
STMp1	.38	.51	1	.58	.38	.40	-.29	.14	-.05	.17	.30	.15	-.25
STMp2	.31	.39	.58	1	.34	.31	-.06	.02	.02	.16	.33	.34	-.20
STMsm2	.31	.34	.38	.34	1	.34	-.09	.04	.02	.05	.32	.24	-.09
STMsy2	.34	.39	.40	.31	.34	1	-.23	.11	.06	.31	.21	.21	-.10
CR2	-.11	-.15	-.29	-.06	-.09	-.23	1	-.04	.01	-.06	-.03	-.06	-.01
CR3	.04	-.07	.14	.02	.04	.11	-.04	1	-.01	.22	-.12	-.02	-.10
CRTT-R-Stroop	-.02	.04	-.05	.02	.02	.06	.01	-.01	1	-.05	-.03	.03	.13
OC	.20	.27	.17	.16	.05	.31	-.06	.22	-.05	1	.21	.35	-.17
GP	.20	.34	.30	.33	.32	.21	-.03	-.12	-.03	.21	1	.33	-.04
CS	.32	.26	.15	.34	.24	.21	-.06	-.02	.03	.35	.33	1	-.19
LA	-.16	-.15	-.25	-.20	-.09	-.10	-.01	-.10	.13	-.17	-.04	-.19	1

STMp1= Auditory Phonological Rhyme Span--Words, **LPp1**= Auditory Phonological Rhyme Judgment - Words, **STMp2**= Auditory Phonological Rhyme Judgment Span- Words, **LPsy1**= Auditory Syntactic Grammaticality Judgments- Sentence, **LPp2**= Reading Phonological Rhyme Judgments-Words, **LPsy3**= Reading Syntactic Anagram-Words, **STMsm3**= Auditory Semantic N-Back-Words, **STMsy3**= Auditory Syntactic N-Back-Sentences, **LPsm1**= Auditory Semantic Category Judgments-Words, **LPsm3**= Visual Semantic PWI with living/nonliving Judgments-Pictures & Words, **STMsm1**= Auditory Semantic Category Judgment Span-Words, **STMsy1**= Auditory Syntactic Span-Sentence, **LPsy2**= Visual & Written Syntactic Picture Matching-Sentence, **LPp3**= Visual Phonological Pictured Rhyme Judgments-Word, **LPsm2**= Visual Semantic Pyramids and Palm Trees-Words, **STM**= Short-term memory, **WM**= Working memory, and **LP**= Language processing, **CR2**= Picture-Word interference, **CR3**= Number Stroop, **CRTT-R-Stroop** = Computerized Revised Token Test (Stroop), **OC**= Relative Clause, **GP**= Garden Path, **CS**= compound sentences, **LA**= Lexical Ambiguity.

Parallel Analysis

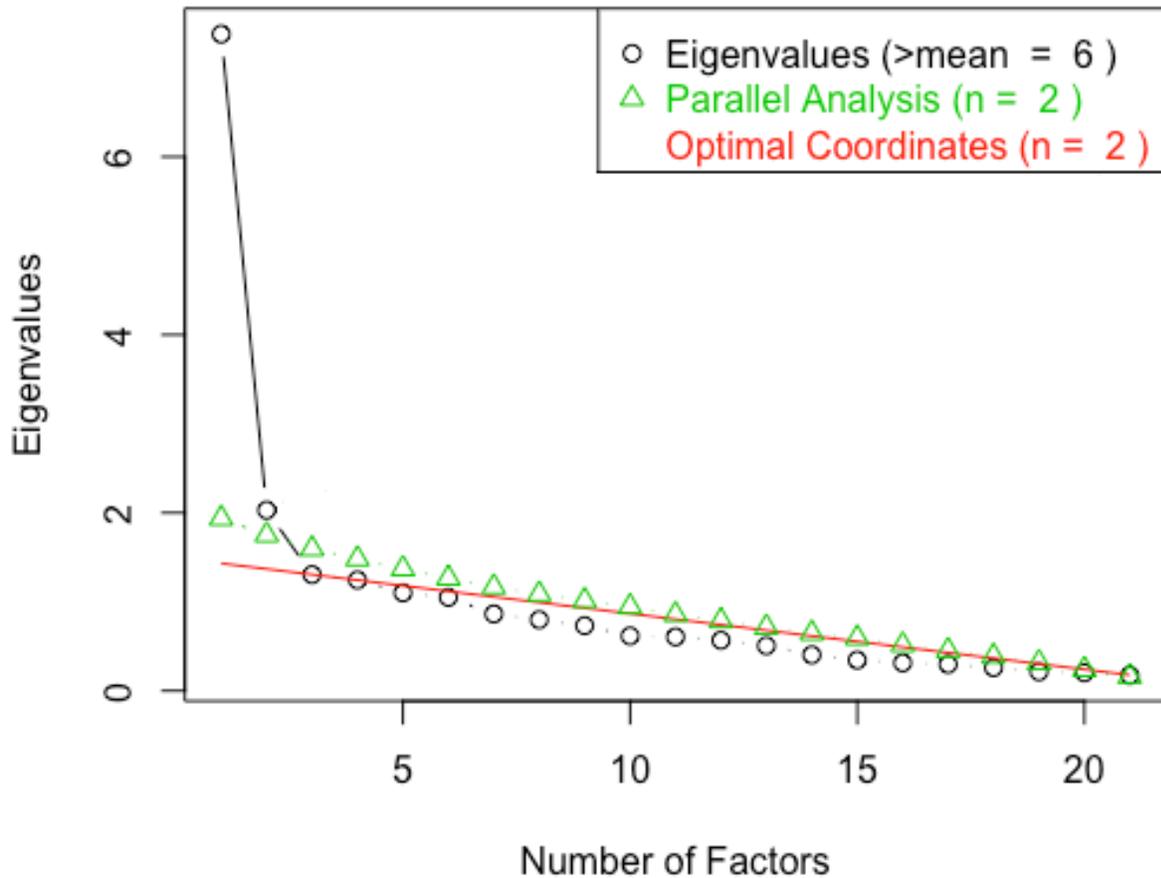


Figure 14. Parallel analysis.

Parallel analysis along with other tests of how many factors to retain. Acceleration factor = where the elbow of the scree plot appears. Optimal coordinates = the extrapolated coordinates of the previous eigenvalue that allow the observed eigenvalue to go beyond this extrapolation. This graph was built and obtained from the `eigenComputes`, `parallel`, `nScree`, and `plotnScree` function from the `nFactors` package in R.

Therefore, two EFA models were generated, one with six factors and one with two factors, to observe the item loadings matrix and model fit. Furthermore, a CFA was performed on each of these two models along with the further adjustments (dropping insignificant item loadings or items with loadings less than .32 and implementing modification indices) to confirm the model fit for both solutions. The initial eigenvalues for each of the six factors account for

11%, 10%, 10%, 9%, 9%, and 4% of the variance respectively. However, when the item loadings matrix as examined, it was found that the task “STMsem2” loads on the fourth factor by itself, which led to run another model with five factors. The initial eigenvalues from the 5-factor model indicated that the five factors accounted for 13%, 11%, 10%, 9%, and 6% of the variance respectively. Based on the item loadings from the 5-factor model, a CFA model was built with five factors where each item loads on the factor on which it received the highest loading (Table 18). Since the data were not multivariate normal (Mardia’s skewness = 246.71, $p < .001$; Mardia’s kurtosis = 542.3754, $p < .001$), the Maximum likelihood with robust adjustments (Yuan & Bentler, 2000) was used for the entire analysis. The CFA of this model yielded a significant difference between the observed and modeled covariance matrix, Yuan-Bentler $X^2(142, N = 100) = 223.016$, $p < .001$; CFI = .895, RMSEA = .072 (%90 CI: .053, .090), SRMR= .067, which suggests that the model is on the edge of a poor model fit because of the disagreement between the fit indices. Next, the initial eigenvalues from the 2-factor model indicated that the two factors explained 23% and 13% of the variance respectively. Based on the item loadings from the 2-factor model, a CFA model was built with two factors where each item loads on the factors that it had the highest item loading on (Table 18). Task “CR2” was dropped after the initial CFA model was performed since its loading on the second factor became -.27, which is lower than the cutoff value of item loadings (.32). The CFA of this model suggested that there was a significant difference between the observed and modeled covariance matrix, Yuan-Bentler $X^2(128, N = 100) = 205.469$, $p < .001$; CFI = .891, RMSEA = .076 (%90 CI: .056, .095), SRMR= .069, which also suggests that the model had a moderate to poor model fit because of the disagreement between the fit indices.

Table 18. Item loadings from the 5 and 2-factor EFA solutions (item loading $<|0.32|$ were suppressed).

	5-factor solution					2-factor solution	
	F1	F2	F3	F4	F5	F1	F2
LPsy2	0.67			0.33		0.69	
STMsy1	0.58		0.32			0.63	
STMsy2	0.52					0.57	
STMsm1	0.51					0.50	
STMp1	0.50		0.47			0.63	
LPsy3	0.48	0.34				0.57	0.42
LPp3	0.44					0.60	
LPp2	0.42	0.41		0.35		0.56	0.47
STMp2	0.39		0.56			0.60	
STMsm3		0.75					0.77
STMp3		0.75					0.70
STMsy3		0.69					0.74
LPp1			0.81			0.56	
STMsm2			0.54			0.52	
LPsy1			0.34			0.53	
LPsm1			0.33		0.42	0.42	
LPsm2				0.92		0.56	0.38
CR2				-0.42			
LPsm3					0.85	0.42	
CR3							
CRTT-R-Stroop							

STMp1= Auditory Phonological Rhyme Span--Words, **LPp1**= Auditory Phonological Rhyme Judgment - Words, **STMp2**= Auditory Phonological Rhyme Judgment Span-Words, **LPsy1**= Auditory Syntactic Grammaticality Judgments- Sentence, **LPp2**= Reading Phonological Rhyme Judgments-Words, **LPsy3**= Reading Syntactic Anagram-Words, **ST/WMsm3**= Auditory Semantic N-Back-Words, **ST/WMsy3**= Auditory Syntactic N-Back-Sentences, **STMp3**= Auditory Phonological N-Back-Words, **LPsm1**= Auditory Semantic Category Judgments- Words, **LPsm3**= Visual Semantic PWI with living/nonliving Judgments-Pictures & Words, **STMsm1**= Auditory Semantic Category Judgment Span-Words, **STMsy1**= Auditory Syntactic Span-Sentence, **LPsy2**= Visual & Written Syntactic Picture Matching-Sentence, **LPp3**= Visual Phonological Pictured Rhyme Judgments-Word, **LPsm2**= Visual Semantic Pyramids and Palm Trees-Words, **STM**= Short-term memory, **WM**= Working memory, and **LP**= Language processing, **CR2**= Picture-Word interference, **CR3**= Number Stroop, **CRTT-R-Stroop** = Computerized Revised Token Test (Stroop).

After carefully examining each one of these models, it was concluded that the 5-factor model is not interpretable since there is no coherent theory that can explain the obtained factors. For the 2-factor model, one interpretation is that the second factor represents WM since the three tasks that load on this factor were hypothesized to engage WM more than any of the other tasks.

However, since the model fit of this model is mediocre to poor according to many fit indices, multiple CFA models were computed where each represents a theory of the constructs that these tasks share.

4.3.5 CFA

A confirmatory factor analysis was performed, using the “lavaan” package in R, to answer the experimental question: Are CR, STM, and LP separable and domain-specific components in PWA? Therefore, several models were built, each corresponding to a theory regarding the structure and the nature of the LP, STM, and CR cognitive systems. This technique was followed since a model that tests a specific theory could have a good fit but still not as good as another model that tests an alternative theory. Comparing CFA models is helpful to select the best theory that explains the greatest variance in the data from among the competing theories. To accomplish such comparison, the nested CFA models were tested using the (Satorra & Bentler, 2001) Scaled Chi-Square Difference Test, which tests whether or not the difference in X^2 in the compared models is significant. If this test suggests insignificant difference or the two models were not nested, and there is no clear advantage of one model fit over the other, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which are comparative measures of fit while penalizing models for extra complexity, were used. All of the presented models have been modified based on the modification indices that were suggested by the CFA. However, during these modifications, only the suggested correlation of errors of the observed variables were added while being extremely cautious not to add any correlation unless it is supported by the theory behind the tasks.

To answer whether LP, STM, and CR are separate, the goodness of fit of two CFA models were tested. The first model was a single factor model where all tasks load on one factor. The second model was where each task loads on its hypothesized cognitive system despite its hypothesized language domain. After dropping the insignificant tasks (the three CR tasks, which resulted in having only LP and STM factors in the second model), there was a significant difference between the observed and modeled covariance matrix for the single factor solution, Yuan-Bentler $\chi^2(128, N = 100) = 267.119, p < .001$; CFI = .805, RMSEA = .10 (%90 CI: .084, .118), SRMR= .081, which suggests a poor model fit. For the 2-factor model, there was also a significant difference between the observed and modeled covariance matrix, Yuan-Bentler $\chi^2(127, N = 100) = 263.899, p < .001$; CFI = .811, RMSEA = .10 (%90 CI: .083, .117), SRMR= .080, which also suggests a poor model fit. From the fit indices, the two suggested models are equally poor, which is confirmed by the (Satorra & Bentler, 2001) Scaled Chi-Square Difference Test, which suggests that these two models are not significantly different from each other. However, when comparing the two models using AIC and BIC, the two-factor model was better than the single factor model, which suggest, that LP and STM are separable (Table 19).

Table 19. *Scaled Chi Square Difference Test.*

	df	AIC	BIC	χ^2	χ^2 diff	df diff	P-value
2-Factor	127	4577.0	4738.5	246.21			
1-Factor	128	4581.1	4740.1	252.39	2.51	1	0.11

df= degrees of freedom, AIC= Akaike Information Criterion, BIC= Bayesian Information Criterion.

Furthermore, to answer the question whether LP and STM are domain-specific (phonology, syntax and semantics), a model with six factors was built, where each task loads on its hypothesized cognitive system and language domain. Therefore, a model with 18 tasks and 6 factors that were hypothesized to correspond to phonological LP, semantic LP, syntactic LP,

phonological STM, semantic STM, and syntactic STM was tested. For this model (6- factors), there was also a significant difference between the observed and modeled covariance matrix, Yuan-Bentler $\chi^2(114, N = 100) = 244.055, p < .001$; CFI = .824, RMSEA = .10 (%90 CI: .084, .12), SRMR= .79, which suggests a poor model fit. Again, both models, the 2-factor and the 6-factor, have similar model fit indices, which is also confirmed by the (Satorra & Bentler, 2001) Scaled Chi-Square Difference Test. According to this test, there is no significant difference between these two models. Thus, the AIC and BIC were examined, which both suggest that the 2-factor model is better than the 6-factor, which suggests that LP and STM are domain-general (Table 20).

Table 20. *Scaled Chi Square Difference Test.*

	df	AIC	BIC	χ^2	χ^2 diff	df diff	P-value
6-Factor	114	4579.4	4774.8	222.62			
2-Factor	127	4581.1	4738.5	246.21	21.15	13	0.07

df= degrees of freedom, AIC= Akaike Information Criterion, BIC= Bayesian Information Criterion.

Since the 2- factors model is still not a good fit, and since the EFA suggested a model with a WM factor, the next step in the analysis was to test an alternative hypothesis that suggests a 3- factors model, a domain-general LP factor, a domain-general WM factor, and a domain-general STM factor. Therefore, a CFA model with 15 items that were hypothesized to load on 3 factors was built. For this model, there was a significant difference between the observed and modeled covariance matrix, Yuan-Bentler $\chi^2(84, N = 100) = 113.793, p = .017$; CFI = .949, RMSEA = .057 (%90 CI: .026, .083), SRMR= .059, which all, except for the χ^2 , suggest a good model fit. Since the 2- factors model and this 3- factors model are not nested, the AIC and BIC of both models were examined to test whether one is better than the other. Similar to the model fit indices, the AIC and BIC show an advantage of the 3- factors model over the 2- factors

model, AIC = 3767.29 and BIC= 3900.15 for the 3- factors model, and AIC = 4581.1 and BIC= 4738.5 for the 2- factors model. Therefore, this model also confirms that LP and STM along with WM are separable constructs.

Furthermore, since the models that were used to test whether these constructs are domain-general or domain-specific were very close and poor in terms of the model fit, a model that tests whether the domain-general LP factor can be broken down to 3 domain-specific factors in the 3-factor model was built. For this 5-factor model, there was also a slightly significant difference between the observed and modeled covariance matrix, Yuan-Bentler $\chi^2(77, N = 100) = 102.26, p = .029$; CFI = .957, RMSEA = .055 (%90 CI: .019, .081), SRMR= .055, which is, again, a very similar model fit to the model where LP was represented by only one domain-general factor. Furthermore, the (Satorra & Bentler, 2001) Scaled Chi-Square Difference Test suggests that the 5-factor model did not improve the fit of the 3-factor model. In fact, although the fit indices of the 5-factor model have a slightly better fit than the 3-factor model, the AIC and BIC of both models suggest that this improvement is more likely to be due to the decrease in df rather than an actual better fit and that the 3-factor model is still better than the 5-factor model (Table 21).

Table 21. *Scaled Chi Square Difference Test.*

	df	AIC	BIC	χ^2	χ^2 diff	df diff	P-value
5-Factor	77	3769.4	3920.5	93.74			
3-Factor	84	3767.3	3900.2	105.65	11.25	7	0.12

df= degrees of freedom, AIC= Akaike Information Criterion, BIC= Bayesian Information Criterion.

Furthermore, after examining the correlations of the three LP factors, which ranged from .89 to .99, another CFA model with a second order factor that combines the three LP factors was built. For this model, there was also a significant difference between the observed and modeled covariance matrix, Yuan-Bentler $\chi^2(81, N = 100) = 112.921, p = .011$; CFI = .945, RMSEA =

.060 (%90 CI: .030, .085), SRMR= .058, which is, again, a very similar model fit to the model where LP was represented by only one domain-general factor. The AIC and BIC of this model are 3771.468 and 3912.147, respectively. From the AIC and the BIC, the 5-factor model and the model with a second-order factors are similar to each other, but both are still not better than the 3-factor model. Therefore, the 3-factor model was retained as the final CFA model that will be used to answer the research questions and to conduct the subsequent analyses (Figure 15).

The Cronbach's alpha for LP = .85, WM = .81, and STM = .71, which all considered to be acceptable. The factor loadings of the LP items were all significant. The factor loadings ranged from .45 to .77 (communalities, ranged from .19 to .58). The factor loadings of the STM items were also significant. The factor loadings ranged from .61 to .72 (communalities, ranged from .37 to .51). The factor loadings of the WM items were all significant. The factor loadings ranged from .69 to .83 (communalities, ranged from .47 to .69; see Table 22 & Figure 15). The correlations between the factors were moderate to moderate-high (Table 23).

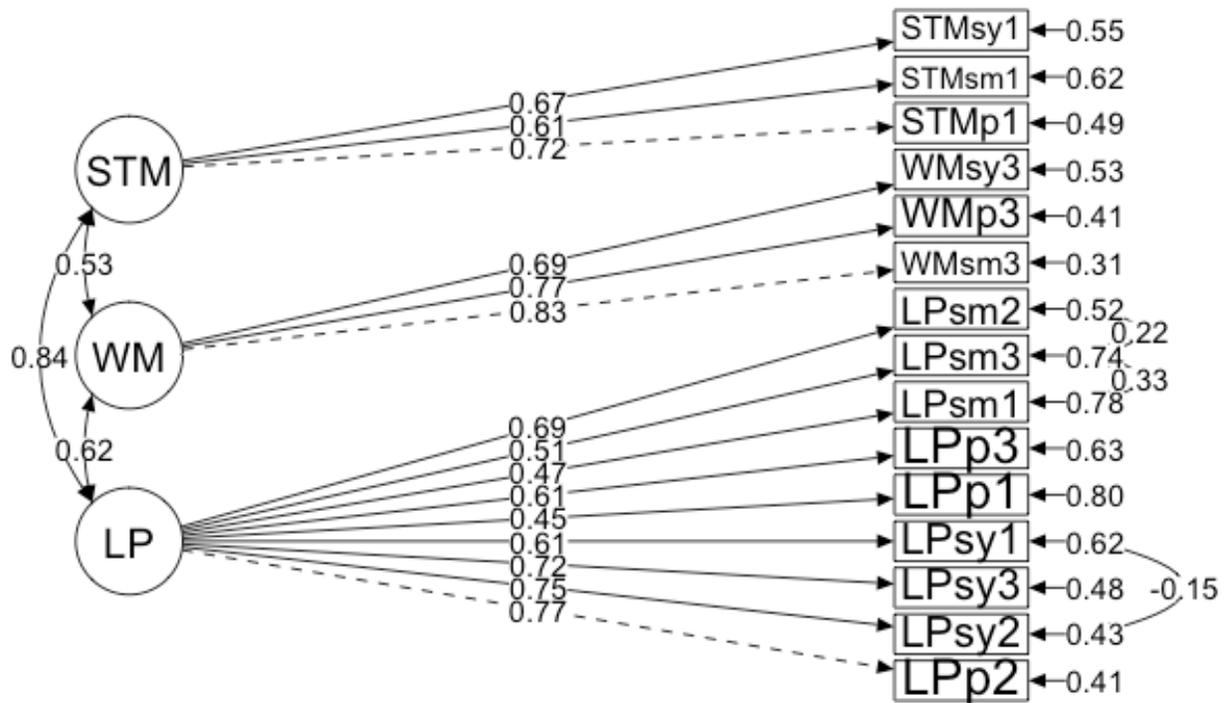


Figure 15. The 3-factor CFA model.

STMp1= Auditory Phonological Rhyme Span--Words, LPp1= Auditory Phonological Rhyme Judgment - Words, STMp2= Auditory Phonological Rhyme Judgment Span-Words, LPsy1= Auditory Syntactic Grammaticality Judgments- Sentence, LPp2= Reading Phonological Rhyme Judgments-Words, LPsy3= Reading Syntactic Anagram-Words, STMsm3= Auditory Semantic N-Back-Words, STMsy3= Auditory Syntactic N-Back-Sentences, LPsm1= Auditory Semantic Category Judgments-Words, LPsm3= Visual Semantic PWI with living/nonliving Judgments-Pictures & Words, STMsm1= Auditory Semantic Category Judgment Span-Words, STMsy1= Auditory Syntactic Span-Sentence, LPsy2= Visual & Written Syntactic Picture Matching-Sentence, LPp3= Visual Phonological Pictured Rhyme Judgments-Word, LPsm2= Visual Semantic Pyramids and Palm Trees-Words, STM= Short-term memory, WM= Working memory, and LP= Language processing.

Table 22. Factor loadings and communalities of the 3-factor CFA model.

Path (From -> To)	B	β	SE	z-value	P-value	R ²
LP -> LPp2	1	0.767				0.588
LP -> LPsy2	0.978	0.753	0.132	7.419	< .001	0.568
LP -> LPsy3	0.938	0.724	0.129	7.262	< .001	0.524
LP -> LPsm2	0.899	0.694	0.13	6.902	< .001	0.481
LP -> LPsy1	0.798	0.615	0.134	5.936	< .001	0.378
LP -> LPp3	0.795	0.612	0.133	5.971	< .001	0.374
LP -> LPsm3	0.659	0.511	0.135	4.872	< .001	0.261
LP -> LPsm1	0.605	0.466	0.135	4.466	< .001	0.217
LP -> LPp1	0.578	0.445	0.136	4.251	< .001	0.198
STM -> STMp1	1	0.717				0.514
STM -> STMsy1	0.942	0.673	0.167	5.657	< .001	0.453
STM -> STMsm1	0.858	0.613	0.171	5.03	< .001	0.376
WM -> WMsm3	1	0.833				0.694
WM -> WMp3	0.919	0.767	0.133	6.933	< .001	0.588
WM -> WMsy3	0.824	0.687	0.126	6.562	< .001	0.472

STMp1= Auditory Phonological Rhyme Span--Words, LPp1= Auditory Phonological Rhyme Judgment - Words, STMp2= Auditory Phonological Rhyme Judgment Span-Words, LPsy1= Auditory Syntactic Grammaticality Judgments- Sentence, LPp2= Reading Phonological Rhyme Judgments-Words, LPsy3= Reading Syntactic Anagram-Words, STMsm3= Auditory Semantic N-Back-Words, STMsy3= Auditory Syntactic N-Back-Sentences, LPsm1= Auditory Semantic Category Judgments-Words, LPsm3= Visual Semantic PWI with living/nonliving Judgments-Pictures & Words, STMsm1= Auditory Semantic Category Judgment Span-Words, STMsy1= Auditory Syntactic Span-Sentence, LPsy2= Visual & Written Syntactic Picture Matching-Sentence, LPp3= Visual Phonological Pictured Rhyme Judgments-Word, LPsm2= Visual Semantic Pyramids and Palm Trees-Words, STM= Short-term memory, WM= Working memory, and LP= Language processing, SE= Standard error.

Table 23. Factor Correlations of the 3-factor CFA model.

LP	1.000		
STM	.84	1.000	
WM	.62	.52	1.000

STM= Short-term memory, WM= Working memory, and LP= Language processing.

4.3.6 SEM

SEM was conducted to answer the question: For PWA, do STM and CR predict comprehension success beyond the contribution of LP, on sentence structures that have been hypothesized to rely on these functions? As was shown in the CFA section, the CR function could not be measured separately with the set of tasks that were hypothesized to measure it. Therefore, this

analysis has investigated whether STM and WM predict comprehension success beyond the contribution of LP on the four sentence structures that were measured (compound sentences, Object Cleft, Garden Path, and Lexical Ambiguity). The data were tested for multivariate outliers using plots, Mahalanobis distance, and Bonferroni p-values. Two outliers were identified with OC and one with LA, none of them had any impact when removed. Therefore, these outliers were kept to retain an adequate sample size. To answer this research question, four SEM models were built where the LP, STM, and WM factors were regressed on a measure of comprehension success of one of the four sentences for each model. For the first model, the LP, STM, and WM factors were regressed on the comprehension success of the Object Cleft sentences (OC) measured by the CRTTrp task. For this model, there was a significant difference between the observed and modeled covariance matrix, Yuan-Bentler $X^2(96, N = 100) = 132.417, p = .008$; CFI = .941, RMSEA = .058 (%90 CI: .031, .081), SRMR= .060, which all, except for the X^2 , suggest a good model fit. The correlations between the OC and the LP, WM and STM were 0.26, 0.38, and 0.28, respectively. The results indicated that only WM has a significant effect on the comprehension success of the Object Cleft sentences while controlling for the effects of LP and STM (Table 24 and Figure 16).

Table 24. *Test of significance of the regression coefficients for the Object Cleft sentences model.*

Path (To <- From)	β	SE	z-value	P-value	R^2
OC <- LP	-0.408	0.468	-1.132	0.258	
OC <- WM	0.442	0.194	2.713	0.007	0.23
OC <- STM	0.437	0.488	1.258	0.208	

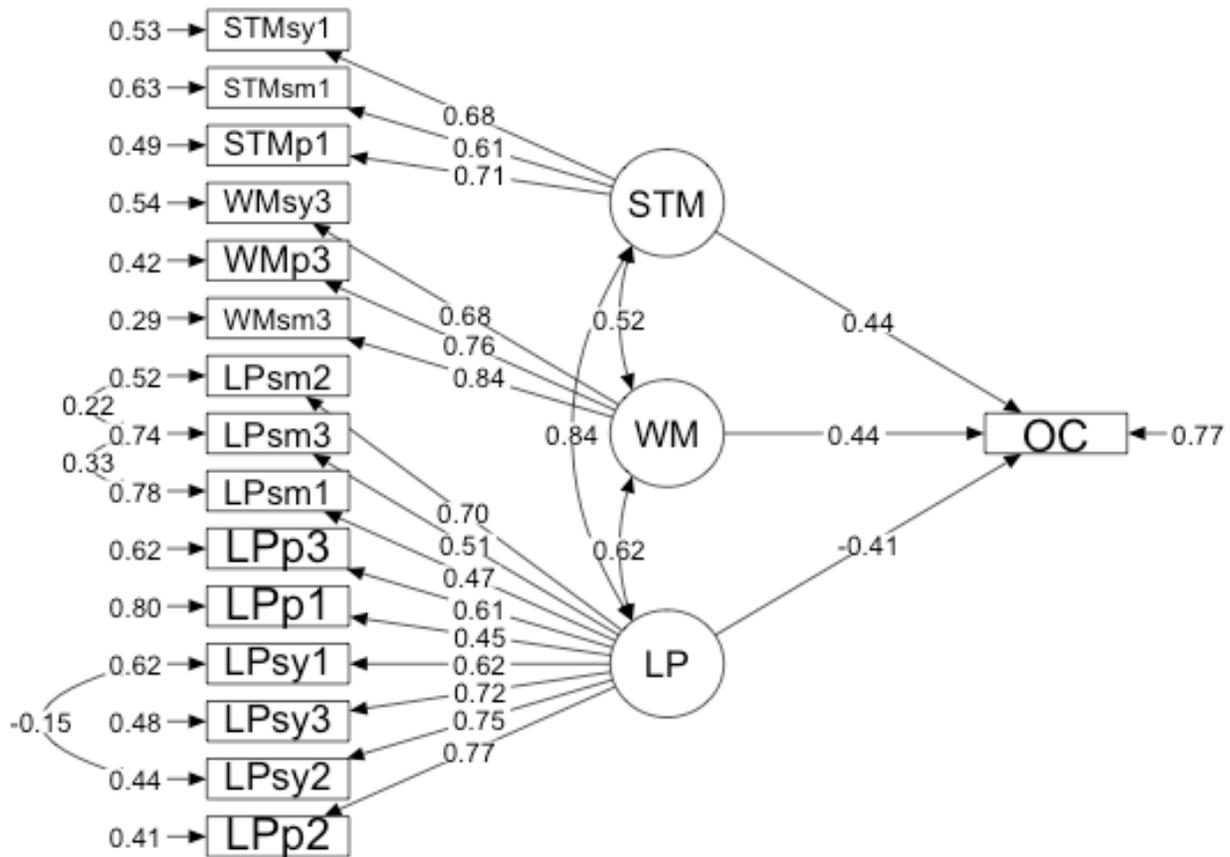


Figure 16. The Object Cleft sentences model.

STMp1= Auditory Phonological Rhyme Span--Words, LPp1= Auditory Phonological Rhyme Judgment - Words, STMp2= Auditory Phonological Rhyme Judgment Span- Words, LPsy1= Auditory Syntactic Grammaticality Judgments- Sentence, LPp2= Reading Phonological Rhyme Judgments- Words, LPsy3= Reading Syntactic Anagram- Words, STMsm3= Auditory Semantic N-Back- Words, STMsy3= Auditory Syntactic N-Back- Sentences, LPsm1= Auditory Semantic Category Judgments- Words, LPsm3= Visual Semantic PWI with living/nonliving Judgments- Pictures & Words, STMsm1= Auditory Semantic Category Judgment Span- Words, STMsy1= Auditory Syntactic Span- Sentence, LPsy2= Visual & Written Syntactic Picture Matching- Sentence, LPp3= Visual Phonological Pictured Rhyme Judgments- Word, LPsm2= Visual Semantic Pyramids and Palm Trees- Words, OC= Object Cleft, STM= Short-term memory, WM= Working memory, and LP= Language processing.

For the second model, the LP, STM, and WM factors were regressed on the comprehension success of the Garden Path sentences (GP) measured by the CRTTgp task. For this model, there was a significant difference between the observed and modeled covariance matrix, Yuan-Bentler $X^2(96, N = 100) = 131.030, p = .01$; CFI = .942, RMSEA = .058 (%90 CI: .029, .081), SRMR= .064, which all, except for the X^2 , suggest a good model fit. The correlations between the GP and the LP, WM and STM were 0.37, 0.4, and 0.38, respectively. The results

indicated that only WM has a significant effect on the comprehension success of the Object Cleft sentences while controlling for the effects of LP and STM (Table 25 and Figure 17).

Table 25. *Test of significance of the regression coefficients for the Garden Path sentences model.*

Path (To <- From)	β	SE	z-value	P-value	R ²
GP <- LP	-0.18	0.447	-0.521	0.602	
GP <- WM	0.362	0.181	2.329	0.02	0.26
GP <- STM	0.382	0.495	1.081	0.28	

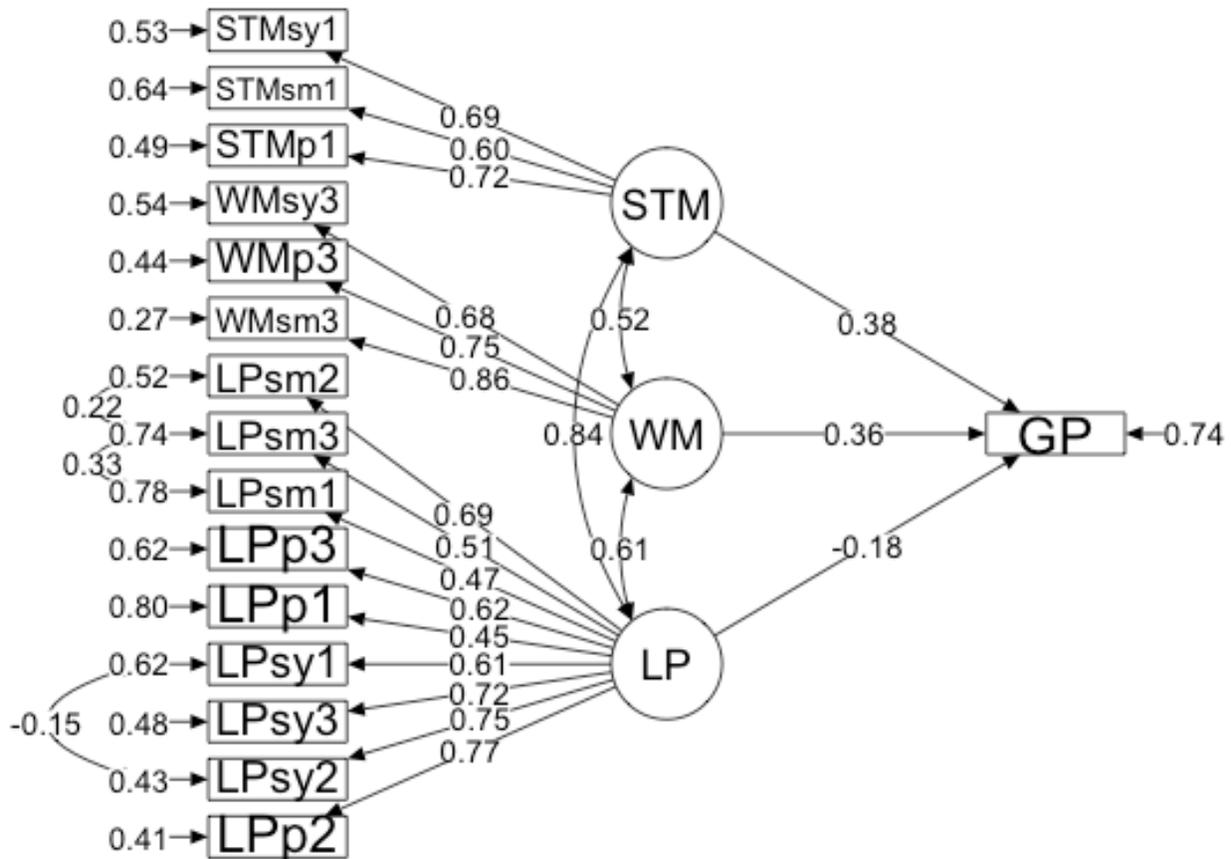


Figure 17. The Garden Path sentences model.

STMp1= Auditory Phonological Rhyme Span--Words, LPp1= Auditory Phonological Rhyme Judgment - Words, STMp2= Auditory Phonological Rhyme Judgment Span-Words, LPsy1= Auditory Syntactic Grammaticality Judgments- Sentence, LPp2= Reading Phonological Rhyme Judgments-Words, LPsy3= Reading Syntactic Anagram-Words, STMsm3= Auditory Semantic N-Back-Words, STMsy3= Auditory Syntactic N-Back-Sentences, LPsm1= Auditory Semantic Category Judgments-Words, LPsm3= Visual Semantic PWI with living/nonliving Judgments-Pictures & Words, STMsm1= Auditory Semantic Category Judgment Span-Words, STMsy1= Auditory Syntactic Span-Sentence, LPsy2= Visual & Written Syntactic Picture Matching-Sentence, LPp3= Visual Phonological Pictured Rhyme Judgments-Word, LPsm2= Visual Semantic Pyramids and Palm Trees-Words, GP= Garden Path, STM= Short-term memory, WM= Working memory, and LP= Language processing.

For the third model, the LP, STM, and WM factors were regressed on the comprehension success of the Compound sentences (CS) measured by the CRTT-IV task. For this model, there was a significant difference between the observed and modeled covariance matrix, Yuan-Bentler $\chi^2(96, N = 100) = 129.592, p = .013$; CFI = .943, RMSEA = .057 (%90 CI: .028, .081), SRMR= .060, which all, except for the χ^2 , suggest a good model fit. Unlike the OC and GP models, there

were no outliers found for this model. The correlations between the CS and the LP, WM and STM were 0.36, 0.35, and 0.34, respectively. The results indicated that STM and WM have no significant effect on the comprehension success of the Compound sentences while controlling for the effect of LP (Table 26 and Figure 18).

Table 26. *Test of significance of the regression coefficients for the Compound sentences model.*

Path (To <- From)	β	SE	z-value	P-value	R ²
CS <- LP	0.189	0.431	0.568	0.57	
CS <- WM	0.22	0.173	1.481	0.139	0.17
CS <- STM	0.058	0.462	0.178	0.859	

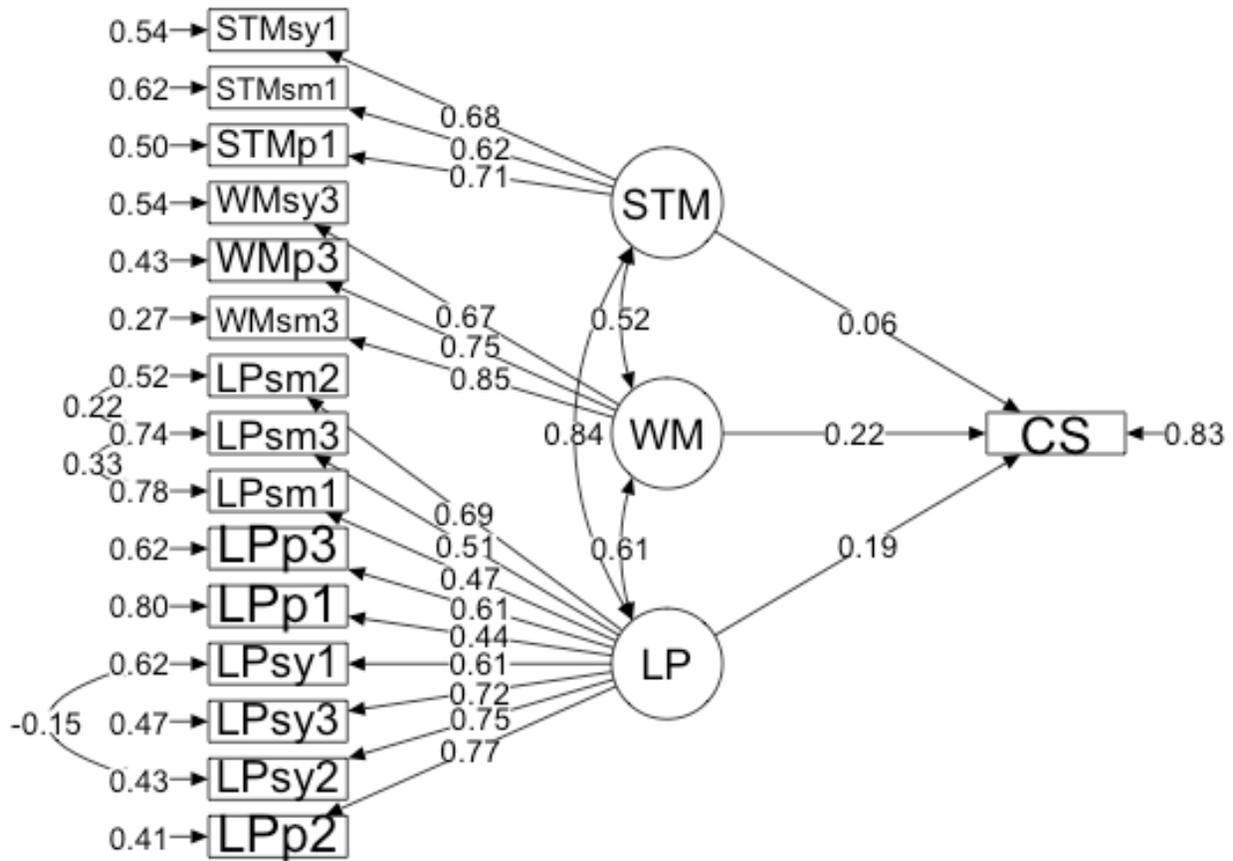


Figure 18. The Compound sentences model.

STMp1= Auditory Phonological Rhyme Span--Words, LPp1= Auditory Phonological Rhyme Judgment - Words, STMp2= Auditory Phonological Rhyme Judgment Span-Words, LPsy1= Auditory Syntactic Grammaticality Judgments- Sentence, LPp2= Reading Phonological Rhyme Judgments-Words, LPsy3= Reading Syntactic Anagram-Words, STMsm3= Auditory Semantic N-Back-Words, STMsy3= Auditory Syntactic N-Back-Sentences, LPsm1= Auditory Semantic Category Judgments-Words, LPsm3= Visual Semantic PWI with living/nonliving Judgments-Pictures & Words, STMsm1= Auditory Semantic Category Judgment Span-Words, STMsy1= Auditory Syntactic Span-Sentence, LPsy2= Visual & Written Syntactic Picture Matching-Sentence, LPp3= Visual Phonological Pictured Rhyme Judgments-Word, LPsm2= Visual Semantic Pyramids and Palm Trees-Words, CS= Compound Sentences, STM= Short-term memory, WM= Working memory, and LP= Language processing.

For the fourth model, the LP, STM, and WM factors were regressed on the comprehension success of the Lexical Ambiguity (LA) sentences measured by the Sentence Picture Matching task. For this model, there was a slightly significant difference between the observed and modeled covariance matrix, Yuan-Bentler $X^2(96, N = 100) = 120.769, p = .045$; CFI = .958, RMSEA = .051 (%90 CI: .00, .074), SRMR= .057, which all, except for the X^2 ,

suggest a good model fit. Again, unlike the OC and GP models, there were no outliers found for this model. The correlations between the LA and the LP, WM and STM were -0.30, -0.23, and -0.28, respectively. The results indicated that STM and WM have no significant effect on the comprehension success of the Compound sentences while controlling for the effect of LP (Table 27 and Figure 19).

Table 27. *Test of significance of the regression coefficients for the Lexical Ambiguity sentences model.*

Path (To <- From)	β	SE	z-value	P-value	R ²
LA <- LP	-0.215	0.391	-0.715	0.475	
LA <- WM	-0.058	0.176	-0.394	0.693	0.10
LA <- STM	-0.073	0.408	-0.25	0.803	

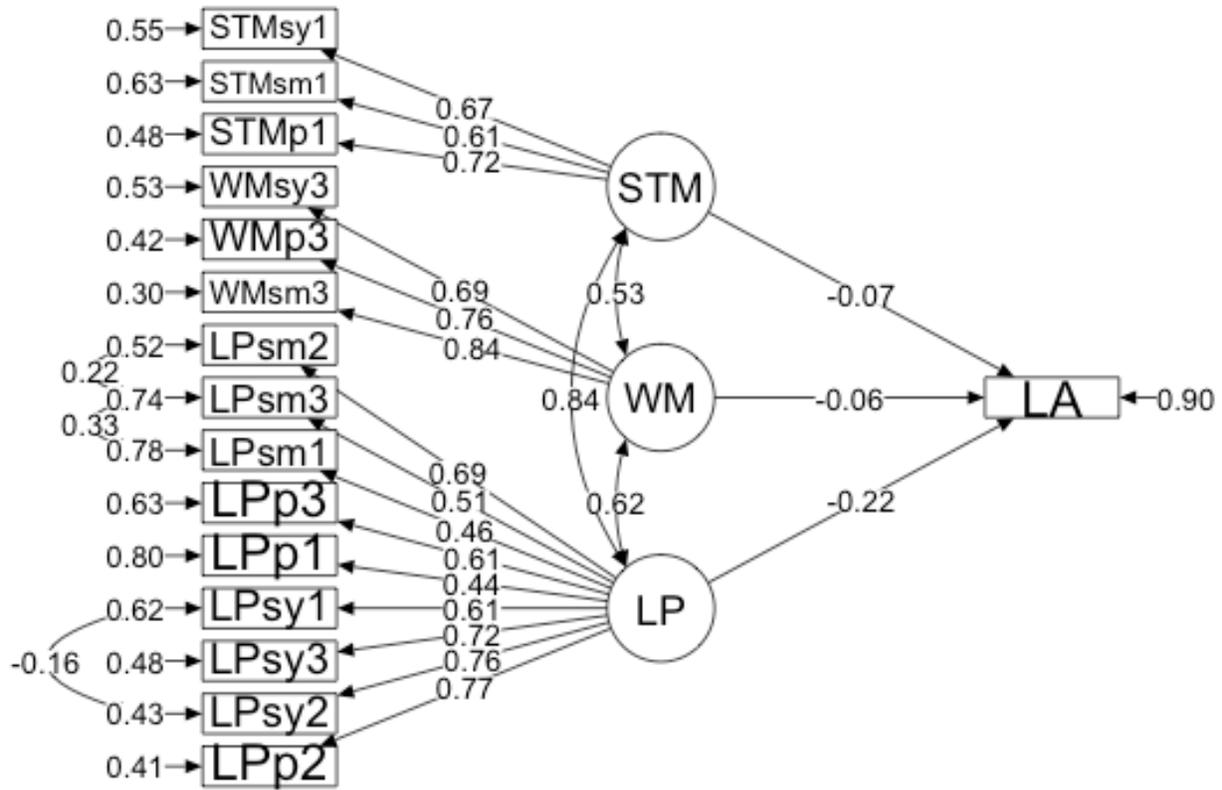


Figure 19. The Lexical Ambiguity sentences model.

STMp1= Auditory Phonological Rhyme Span--Words, LPp1= Auditory Phonological Rhyme Judgment - Words, STMp2= Auditory Phonological Rhyme Judgment Span-Words, LPsy1= Auditory Syntactic Grammaticality Judgments- Sentence, LPp2= Reading Phonological Rhyme Judgments-Words, LPsy3= Reading Syntactic Anagram-Words, STMsm3= Auditory Semantic N-Back-Words, STMsy3= Auditory Syntactic N-Back-Sentences, LPsm1= Auditory Semantic Category Judgments-Words, LPsm3= Visual Semantic PWI with living/nonliving Judgments-Pictures & Words, STMsm1= Auditory Semantic Category Judgment Span-Words, STMsy1= Auditory Syntactic Span-Sentence, LPsy2= Visual & Written Syntactic Picture Matching-Sentence, LPp3= Visual Phonological Pictured Rhyme Judgments-Word, LPsm2= Visual Semantic Pyramids and Palm Trees-Words, AM= Lexical Ambiguity, STM= Short-term memory, WM= Working memory, and LP= Language processing.

5.0 DISCUSSION

This dissertation extends prior work on research data management to create models and techniques that maximize and ensure optimal use of research datasets. Findings in this field suggest that the outcome of the data analysis process should be a product that can be used to serve purposes beyond the one that is related to the primary research project. This universality nature cannot be achieved without having a plan for the lifecycle of the product starting from the data collection throughout the data processing and ending with data archiving (Hey et al., 2009). The Data Documentation Initiative (DDI) data management and documentation model “Data Lifecycle” outlines the proper phases that researchers should follow from the early stages of the research to the conclusion of their studies. As mentioned in the methods section, a collaboration with McNeil et al. (2014) was established on their research project, that was sponsored by the VA, where they collected data from PWA on many tasks. The scope of this dissertation is to conduct the data management for such a huge research project starting from data collection and ending with data exploration and analysis.

5.1 DATA PROCESSING

The data processing for this project was started in the early stages during the data collection. As outlined in the methods section, data processing involved intensive communications with the

clinicians who were collecting the data to provide consultations on the proper use of the data collection methods and to fix any issues that might occur. These consultations were provided based on the anticipated final data structure that the collected data will be stored in. This anticipation was built after conducting multiple requirement collection sessions with the clinicians and the researchers who are going to be using the collected dataset. The main advantage of having the data management team involved early in the data collection process was to detect any potential issues with the collected data and to fix them in early stages. For example, as will be discussed in detail later, data validation procedures were built in the early stages of the project to evaluate the accuracy and correctness of the data extraction procedures and, more importantly, the data collection process. This was important since the project was dealing with a population with limited access, which meant that there is not a comfortable level of freedom to drop any data from the analysis and, thus, there was a minimal tolerance for any issues that might lead to that.

Moreover, data processing involved a de-identification procedure of the dataset before it can be transferred to the online system. Basically, the de-identification involved removing the dates from the data collection sessions since they can be traced back to the individuals who participated in the study. However, although dates will not be included in the data analysis of this dissertation, future research might investigate the interaction between time and the performance of the subjects. Therefore, removing the dates would have resulted in a huge loss of important data, which led to the implementation of a de-identification algorithm that replaces the dates with the number of days since the first session for each subject. The main challenge in this technique is that it cannot be conducted manually due to the massive amount of data. However, this

challenge was solved by implementing automated and customized techniques that use regular expression to recognize dates and then do the calculations and replacement.

As mentioned before, a critical part of the data processing was conducting the data extraction from the original files. This part was critical since all the subsequent data processing and analysis procedures will be directly affected by the quality of the data extraction. The main challenges in this step was dealing with different types of files that must be merged before conducting the analysis, each file had a different format, massive files containing data for multiple tasks, and each task only uses a subset of columns. The challenge here was how to extract the right data for each task from the appropriate columns and structure a new dataset that contains the processed data. However, this challenge was solved by implementing the extraction map and the extraction algorithms that were discussed in detail in the methods section. The extraction map was constructed by the clinicians who were the expert in the data in terms of where each information should come from and where it should go to. The map was an excellent tool to create a direct communication line between these clinicians and the extraction algorithm. Therefore, during the data processing, the extraction algorithm had to be developed only at the beginning of the project, but multiple versions of data extraction procedures were conducted. In other words, the data extraction algorithm did not need to be changed each time a new data extraction rule emerges. Clinicians were able to change the extraction rules as needed and run the extraction map that will automatically detect and apply the new rules without the involvement of anyone from the data management team. This had a huge impact on the speed of conducting the data extraction since there was no need for meetings between the clinicians and the technicians to explain the new rules, and no need for the technicians to spend some time re-customizing the extraction algorithm to fit the new rules.

In addition, the same technique was used for data validation where clinicians specify what they expect from each column for each task in a validation map. Then, a validation algorithm accesses this map and evaluates each column accordingly. If any column violates the rules, the algorithm generates a report that explains what was expected and what was actually found to the clinicians. Again, this algorithm was developed once and was used on a regular basis by the clinicians to test the newly collected data without involving any technicians in the process. Another advantage of this technique is that it gave the clinicians a strong control over the quality of this massive dataset since they can view any abnormal data points in an understandable and interpretable format. Furthermore, the results of this validation process were used to trace back and fix issues that might have occurred in the data collection, de-identification, or extraction procedures. Therefore, this technique had a positive impact on the quality of the processed data and, thus, on the analyses that were and will be conducted on this dataset.

To continue the automation of the data processing, an algorithm that reads the data files and builds the SQL create-statements for the database tables, automatically, was implemented and executed. Basically, the algorithm can recognize whether each field is empty, integer, decimal, short text, or long text and then creates the tables with the proper columns and data field types. This procedure is highly important since it builds upon the results from the data extraction procedure, which changes automatically and rapidly with the change of the extraction map by the clinicians. Therefore, when a clinician adds, deletes, or changes the content of a column, it would be inefficient to manually change the database schema every time, which was solved by implementing this automated algorithm. Another major advantage is that the automation gives a chance to easily change the database engine if needed. For example, in this project, MySQL was chosen to be the database engine for some reasons that were mentioned earlier. However, if the

SQL statements were written manually and for some reason it was decided to switch to a new engine, then all the statements would have had to be rewritten in the new SQL syntax. Using the automated SQL building procedure, however, it is only required to let the procedure use the new syntax, and it will automatically apply it to all the automatically generated SQL statements.

5.2 DATA ARCHIVING

The next phase in the data lifecycle was the data archiving phase, which included transforming data to be stored in a format that can be easily shared with all the involved parties. Databases are highly recommended in this situation since they use a Structured Query Language that makes data insertion, deletion, updating, and retrieval faster and easier. Furthermore, databases are widely used and can be hosted on any web-server, which provides a more convenient and secure way of sharing data. In the database, the normalization forms were followed in the design to eliminate any redundancy and make the insertion, selection, and update easier and faster. Another benefit of the normalization is ensuring that there is no inconsistency in the database, especially in the case of updating. For example, if a specific task description needed to be updated and all the data were in one table, then the task description for each subject for each item would have had to be updated. However, since the tasks descriptions were separate from the items description from the subjects' scores, only a single cell in the tasks description table needs to be updated, and other tables will automatically be pointing at the new data row since they are connected using the task ID. Another advantage of normalization was that there was not any data loss when the dataset was transformed to the database. For example, if there was a specific task item where no one has replied to, and all the data were in one table with subject ID as the

primary key, the description of that item would have been lost since it cannot be linked to any subject. Therefore, the database was divided to have separate tables for subjects' description, tasks description, items description, and subjects scores on each item for each task. Thus, in the database, there was high flexibility to update, add, and drop any specific data entry without having to update, add, or delete unrelated or have duplicate data points.

5.3 DATA DISTRIBUTION

The next phase in the data lifecycle was the data distribution phase. Data distribution is a critical phase since data in this phase are exposed to loss and corruption. The main goal of this phase was to securely deliver the full collected dataset to all the entities of the research project in a useable and interpretable format. As discussed before, having the data in flat files (text or sheets) stored in a shared location or in one computer might be efficient for small research projects for a short-term period. However, due to the low-security measures that protect flat files, and due to the existence of many copies of the collected dataset in multiple locations over time, tracking back the originally collected dataset becomes challenging. Therefore, a web-based database that stores the original dataset, along with some processed and calculated data points, that is accessible to all the entities of the research project via a usable and around the day available interface was proposed and implemented. Using this interface, it is easier to track who has gained access to the collected dataset and it is easier to provide users with different access privileges. For example, although not implemented in the current version of the system, the admin can block some users from viewing parts of the dataset if they do not have the privilege to do so. To explain, if some researchers have been given the permission to upload a dataset with

dates to their server and were asked to make the dates available to only the approved users and not to statisticians or data administrators who were not essentially part of the data collection project. Then, one table in the database that links users to the columns they can view in each table would have been enough to accomplish such a requirement. This technique, however, would not have been feasible without the integration of the database with an online system that interprets these privileges and gives access to users accordingly. However, other data management tools, such as Excel or text files do not support such a feature, which would have made implementing this requirement very complicated or even impossible. Furthermore, in the current version of the iRDMS, users can view the history of the subsets of data they have requested from the iRDMS, which can help them to track what data points they requested and what selection conditions they applied.

In addition to the security and confidentiality issues that the online system solves, it facilitates the distribution of new data updates, especially in long-term projects that involve continuous data extraction and processing. In other words, the data administrators can always update the online dataset and guarantee that researchers will be using the latest version instantly. This can be extremely helpful in case an error was found in the dataset, or an urgent data extraction or processing requirement has emerged. Furthermore, managing and retrieving data using an online system that reduces the cognitive load of digging for data in huge and poorly structured data files can help researchers to save time and effort. In the next section, a review and interpretation the findings of the usability study, that compares using Excel and iRDMS for data management and retrieval, is presented.

5.3.1 Usability Study

The usability study was conducted with two investigations in mind, the benefits of conducting automated data processing and the benefits of implementing an Internet-based research data management system. A group of participants that represent the academic field were sampled and their demographic characteristics were balanced. To investigate the benefits of implementing iRDMS, the three domains of usability according to ISO, efficiency, effectiveness, and satisfaction were measured.

The analysis of the participants' performance on the first three tasks clearly shows an advantage of using the iRDMS for data retrieval over using Excel. Only The first three tasks were used here since they did not involve any data processing and only required participants to perform some data retrieval tasks. Furthermore, when looking at the RSD of these three tasks, the spread of participants' performance on iRDMS was almost half the spread of their performance on Excel. This, to some extent, proves that the iRDMS has standardized procedures for data retrieval that can be followed by different participants on an equal level. Also, this conclusion can be drawn when looking at the performance of individual participants on the first three tasks using both methods. The variance of the performance of most participants on the first three tasks using iRDMS was only half, and for some individuals was only third, the variance of the performance of the same tasks using Excel. Therefore, these results basically suggest that participants were more stable in their performance and that they did not have to deal with different challenges when performing different data retrieval tasks.

Furthermore, when looking at the analysis of the difference between participants' performance on the ten tasks using Excel, it was observed that data processing requirements slow down the participants' data retrieval performance. For example, tasks that involve processing of

more than one column are significantly slower than tasks that only involve the retrieval of one column for one subject. Task 9 and 10, which involve calculating a single final score of five items were significantly slower than tasks that involve the retrieval or processing of only one column for one subject but not tasks that involve the retrieval or processing of more than one column or more than one subject. These results basically show the benefit of the automated data processing procedures even without having an internet-based data management system since all the compared tasks were performed using Excel. Furthermore, since dealing with data from more subjects slows down the participants' performance, which was suggested by the performance on task 3 (requires the retrieval of two subjects' data) and that it was not significantly faster than tasks 9 and 10, unlike tasks 1 and 2 (require the retrieval of only one subject's data) which were significantly faster than tasks 9 and 10, the iRDMS was implemented to solve this issue by making the selection of a subset of data an easier task. Again, this argument can be made based on the fact the performance of participants on task 3 (94.08 seconds on Excel and 30 seconds on iRDMS) was 31% slower than the average of the performance of participants on tasks 1 and 2 (71.60 seconds on Excel and 25.2 seconds on iRDMS) when using Excel and was reduced to be only 20% slower when using iRDMS.

Using the iRDMS, there was no significant difference between the tasks except between task 4 and 5, and 10 and 5 for multiple reasons. First, task 4 has the fastest time of all the tasks, and task 5 has the slowest. This happened because task 5 asks for the column ISI6_OnSetToOnSet, which needs very careful selection as there are columns with lots of common characters in different parts of the column name. For example, there is an identical column to "ISI6_OnSetToOnSet" with the letter T instead of S "ITI6_OnSetToOnSet". Furthermore, there are other identical columns with a different number instead of the number 6.

Even more, there are other columns that have “ISI6_” but with a different last part of the name, such as “ISI6_OnSetTime” or “ISI6_OnSetDelay” instead of “ISI6_OnSetToOnSet”. In task 4, however, participants are required to select column “ItemType3”, which might only be confused with “ItemType1” or “ItemType2”. It is worth mentioning here that an algorithm that detects any identical column names with only one different character at the end of the names and puts this character between parentheses and makes an upper case if it is a letter was implemented. All these factors have contributed to slower times on task 5 and faster times on the other tasks, especially task 4. This leads to the emphasis on the effect of the database content on the usability of the system. For example, if the naming of the variables was changed to some unique names that are not easily confused, it is hypothesized that the difference between task 5 and the other tasks would disappear.

Additionally, different group’s performance on Excel and on the iRDMS was investigated. Basically, a significant difference between any two groups on only one task or two tasks cannot be used to draw a conclusion that one group differ from another. Therefore, the analysis of groups performance on each task and their overall performance on all tasks was performed. However, a true difference is only concluded if it occurred on the overall performance or on more than few tasks. The results did not show any significant difference between any two groups on the overall performance. The only difference that occurred twice was the difference between Asian and White participants on tasks 8 and 10 when using iRDMS. However, this difference has mostly occurred by chance since it only occurred in two tasks and since the p-values 0.04 and 0.02 are very close to the alpha level and became insignificant after controlling for the type I error due to multiple comparisons.

Moreover, it was investigated whether prior knowledge and experience on Excel would influence the performance of participants on both, Excel and iRDMS. Both aspects, knowledge, and experience with Excel, had no significant correlation with the participants' performance on Excel, which raises questions about the validity of these two questions when used to measure knowledge and experience on Excel. However, the question that asks "About how long have you been using Excel? (in years)" showed a significant positive correlation with the overall participant's performance using iRDMS, which means that participants who have been using Excel longer were slower on iRDMS. This could be a sign that individuals number of years using an old system could predict their adoption to the new one. Furthermore, although the order of the tasks presentation was randomized for each participant, the analysis indicates that the performance of the participants was influenced by the "order effect." The analysis showed that the later in the session the participants receive that tasks, the better they perform, which was the case for three tasks using Excel (tasks 1,9, and 10) and five tasks using iRDMS (tasks 2,3,6,7 and 9). This effect has been reported by many usability testing studies and was explained by the probability that participants' knowledge about the general theme of the tasks was increasing as they go by discovering and learning new techniques to perform the tasks (Page, 2013; Strack, 1992). Moreover, when analyzing the effect of asking the participants to first use one method or another, Excel and iRDMS showed no effect of being performed first or second.

In addition to the benefit of data retrieval time reduction, the analysis shows that participants have committed fewer errors when using the iRDMS (1 error) compared to using Excel (3 errors) when performing the first three tasks. Furthermore, when considering the overall picture that includes the data processing procedures and the iRDMS by analyzing errors made on all ten tasks, the iRDMS (3 errors) significantly lowers the error rate compared to using Excel

(16 errors). From these results, it can be observed that most the committed errors when using Excel came from performing tasks that involve data processing (13 out of 16 errors). Moreover, it is worth mentioning here that all these errors, based on their nature and the feedback from the participants when asked why they committed them, were slips rather than mistakes. For example, participants selected the data of the wrong subject, the wrong tasks or miscalculated the processed data, not because they did not know how to perform these tasks, but because they lost their attention for a moment or they looked at the wrong data cell. As discussed before, one advantage of automating the data processing procedures is the stability of the computerized algorithms. For example, when the extraction algorithm and the extraction map were built to process the e-prime files, the algorithm and the map were validated by random samples of data, because once the algorithm starts to perform correctly, there is no more room for random errors in the extraction. Humans, on the other hand, could make different and random errors for each record they process, and selecting random samples to validate the manually extracted data is not valid. Therefore, this effectiveness analysis is further evidence of the importance of conducting the extraction tasks as automatically as possible. In addition, this effectiveness analysis is evidence that the iRDMS reduces the error rate by preventing the users of making errors by preventing the users from entering any values and by viewing a summary of the users' selection before they download the data.

In consequence of the better efficiency and effectiveness that the iRDMS produced in comparison to Excel, participants have expressed high satisfaction and comfort with iRDMS. This was reflective in the satisfaction scores that were measured by PSSUQ and its sub-scales. The overall scale and the three subscales ranged from 1.13 to 1.27 on average on a 7-point Likert scale. This tight range of values suggests that the iRDMS provides a satisfactory replacement of

using Excel to manage this particular research dataset, and probably most of the research datasets, on all four scales of PSSUQ (SysUse, InfoQual, IntQual, and overall). Furthermore, the absence of any significant correlation between the scales of PSSUQ and participants' experience with Excel, skills on Excel, performance on Excel, and performance on iRDMS in evidence that participants were not biased by their performance or prior knowledge.

Even more, the response of the participants to the ASQ questions that measure the satisfaction with time spent and ease of completing each task using Excel and the iRDMS show a significantly higher satisfaction with using iRDMS instead of Excel in every single task, including the first three tasks. Furthermore, an issue worth mentioning here is that the participants' satisfaction scores on the ASQ on the iRDMS were suffering the ceiling effect. This is observable by looking at the average of the two ASQ questions of each of the ten tasks when using the iRDMS, which ranged from 1.04 to 1.38 with no SD that exceeds 0.63. Therefore, performing correlation analysis on these scores, although was reported for descriptive purposes, was too sensitive, thus uninterpretable, due to lack of variability. However, when looking at the answers on the two ASQ questions on tasks using Excel, scores have ranged from 1-7 or 2-7 in all the tasks. Therefore, the performance of correlations between the answers to the two ASQ questions on tasks using Excel and the participants' performance on Excel, skills on Excel, experience on Excel, performance on iRDMS, and task order was feasible and interpretable. Again, if only one question or one task was significantly correlated with any of these variables then it is not logical to draw any conclusion based on that, and it would be safer to conclude that it occurred by chance alone. Therefore, the only significant correlations that are worth mentioning are the correlations between the participants' answers to the two ASQ questions on multiple tasks with the participants' performance on these tasks using Excel. All these

correlations were positive, which means that the slower the participants perform on Excel the less they were satisfied with its time consumption and ease of use.

In addition, the results of the two questionnaires were also supported by the feedback that the participants gave during the think-aloud process and to the four open-ended questions in the PSSUQ. The think-aloud was helpful in capturing the participants' emotions and feelings toward the two systems as they were living them, which reduces the effect of time on these feelings as they might fade by the time they answer the PSSUQ. Also, feedback like "I like it when I know that the next task will be on iRDMS" and "I like iRDMS because I do not have to think while performing the tasks" capture some important qualitative measures that other quantitative measures do not. For example, Cooper (1998) refers to cognitive load as "the total amount of mental activity on working memory at an instance in time". Therefore, the qualitative feedback from the participants suggest that the they have experienced less cognitive load and frustration when they were using iRDMS. The four open-ended questions, on the other hand, gave the participants the room to express their overall thoughts after experiencing the whole testing session. The feedback in these two questions is also important since all participants, at that point, have gone through the same experience, unlike the think-aloud feedback, which could be biased by the order of the tasks since a participant might be frustrated because they just had all the difficult tasks one after another. This can be seen where participants suggested some design changes during the think-aloud process and did not mention these suggestions in the survey at the end of the study. Therefore, based on the findings of this study, it is highly recommended that investigators try to capture the participants' emotion during and after conducting the tasks.

As mentioned in the results section, the first suggestion was provided based on the issue that participants know how to filter rows based on the values in one column in Excel but did not

know how to filter the columns based on their names. Therefore, the search function was added to both, columns names and values in selection customization. As a result, it is anticipated that the gap in the users' performance between Excel and iRDMS would be even wider after this change. Moreover, the suggestion to change the red color shows the importance of mimicking the real world of the users where red color in street signs or papers is usually associated with warnings and safety information (Young, 1991).

The satisfaction results from the PSSUQ, ASQ and the think-aloud were supported by the fact the participants were more effective and efficient in task performance when they were using iRDMS. Even though the following might not be true in other situations, the ISO three domains of usability, in this study, had the same conclusion, which is that the iRDMS was more usable than Excel. However, the fact that a significant correlation between most of the satisfaction measures and the efficiency of the participants' performance was not observed, emphasizes that these domains are measuring different aspects of usability and one does not predict the other. Also, looking at the PSSUQ and ASQ, where they include some questions that measure learnability, cognitive load, ease of use, and satisfaction with system design, it can be concluded that even though a system is not efficient and effective, it can still be satisfactory based on the mentioned domains (learnability, cognitive load, ease of use, and good system design) and vice versa. However, in the case of iRDMS, such a conclusion must be made with extreme caution as the PSSUQ might be suffering from the ceiling effect. This basically means that if iRDMS was not efficient and effective, participants' satisfaction would have varied and would have shown some correlation with their performance on iRDMS. Anyhow, even though the general and final results of each domain, in this study, points to the same conclusion, any usability testing should

include each one of these domains as they might vary in their levels and, thus, show some usability issues that would not have been discovered otherwise.

5.4 DATA ANALYSIS (SENTENCE COMPREHENSION)

One strength of the dataset that has been used to conduct the data management and analysis is the diversity of the participants that were included in the study. McNeil et al. (2014) did an excellent job in including participants from different locations who were balanced by their gender, education, race, age, and time post-onset based on the population of PWA. Furthermore, although the participants of this study were included based on a specific definition of aphasia and some other inclusion criteria that ensure the existence of the condition in each individual, the sample was representative of multiple types and severity levels of aphasia. This characteristic of the sample helps the researchers who use this dataset to gain more generalizability power of the results of their analyses. However, this also might make it challenging to find a model that fits all these diverse participants in this dataset. Therefore, one thing to keep in mind while the results of the factor analyses are discussed, is the existence of sub-groups in the sample where each sub-group confirms a different theory. Another strength in the used dataset is that it measures the performance of the participants on different cognitive systems and functions. Therefore, this makes it suitable for performing the analyses that aim to capture unobserved constructs through the common variance in the tasks. However, this might also introduce the challenge of having tasks that are affected by many factors and probably some noise other than the hypothesized construct to be measured by each task. The tasks in this dataset are more likely to be affected by this issue since they are aiming to measure a very complicated, overlapping, and, at some level,

occult system. Thus, conclusion from the results of any analysis on these tasks might be different when the same analysis is done on a different set of tasks.

Exploratory Factor Analysis is a great data discovery technique that might show some “hidden” relationships that researchers would not have discovered otherwise. However, as mentioned before, the results of the EFA are very dangerous since they might show some good models that are not true in reality and based only on noise or measurement error. Therefore, any EFA results should be validated by theory to ensure the legitimacy of the suggested relations between the modeled variables. In the factor analysis, 21 tasks that have been hypothesized to measure 7 cognitive factors were used. EFA showed that these tasks load on five factors as suggested by the eigenvalues. However, when these five factors and the tasks that each factor is measured by were examined, there was no valid interpretation for most of these factors and it was concluded that they are either measuring some factors that were not yet discovered or they are basically measuring noise. The second model suggested by EFA was a 2-factor model that introduced a new WM factor that was not part of the 7 factors that were hypothesized. This WM was interpreted to be WM because it was measured by three tasks that were hypothesized to measure some aspects of STM but also well known to be measuring WM. However, since the model fit was poor for this model, and since a question that needs testing multiple theories to find the correct answer was being asked, CFA was performed. The difference between CFA and EFA is that in CFA, the researcher specifies the model structure that they want to test by restricting each task to load on the factors that they specified. In the EFA, on the other hand, the model is built by letting the variables freely load on all the factors that the researchers can only specify how many of them to retain.

CFA is a great technique to test multiple theories and compare how the suggested structure fits the observed data in each model. For the question: Whether CR, STM, and LP separable and domain-specific components in PWA, two models were needed to test the possibility of each answer to each part of the question. For the first part, whether they are separable, a model where all the tasks load on one factor, which means that CR, LP, and STM are not separable, to a model where tasks load on three factors, CR, LP, and STM were compared. Both, the EFA and CFA, suggested that the three CR tasks did not belong to any factor and they have nothing in common among themselves nor with the other tasks in the dataset. This might be as a result of the nature of the CR function and that it is very specific and very prone to noise, which makes it very difficult to detect, or because there were only three tasks to measure this function while there were nine tasks to measure the different domains of the LP and STM systems. Therefore, due to the fact that these CR tasks did not load on any factor, not even on the WM factor which is the closest factor to CR in theory, these tasks that have been hypothesized to load on CR were excluded. For the second part of the question, whether they are domain-specific or not, the model with the LP and STM factors was used to represent domain-general systems theory, and another model with six factors, each represents a specific language domain of each system was used to represent the domain-specific systems theory. Although the models for each part of the question were very competitive and close in model fit, results suggest that LP and STM are separable but domain-general systems. However, due to the poor model fit of all models, and based on the insights from EFA, a model with LP, STM, and WM factors was built, which turned to be a good fit model. This model was the best model to fit the data, even better than models that separate the LP into domain-specific factors.

The final CFA model, domain-general LP, STM, and WM cognitive systems, confirms many theories that suggested that LP, STM, and WM are different systems that perform different cognitive tasks. It also suggests that these systems are domain-general in terms of the resources they use and the nature of the impairments they suffer. However, due to the tight competition between the models that represent domain-general factors and the ones that suggest domain-specific factors the possibility that LP, STM, and WM might have some level of domain-specific structure or nature was not ruled out. For example, although the final model was better than the model with three domain-specific LP factors according to AIC and BIC, which compare models while including the sample size and model complexity in the comparison, the model fit of the latter was better than the final model. This suggests that the theory of domain-specific systems could be true. This was also observed by the model with a secondary factor, where the three domain-specific LP factors load on one factor that represents LP in general. According to AIC and BIC, this model suggested that having a secondary factor is better than having three independent domain-specific LP factors. However, since both models are still not better than the final model, this was interpreted as that the effect of the domain-specific nature of these systems are so little that they are not worth the added complexity to have them separated. Although the collected dataset contains tasks that were hypothesized to measure the language domains, the results of these analyses might not be the same when they are repeated with a different set of tasks that might measure the language domains from a different angle.

The first factor in the final model was interpreted as an LP factor that represents LTM along with storages that keep the language representations and rules that are used to build or understand different pieces of language. In other words, this factor was interpreted as the box on the left on Engle's WM model plus LTM (Figure 20). Furthermore, the WM factor was

interpreted to be the Central Executive box in Engle’s WM model (Figure 20), which is responsible for keeping attention activated, including goal maintenance and inhibition of distractions, and acts as the engine that processes the representations and rules of language that are provided by LP. The third factor is the STM factor which was interpreted as a domain-general temporary storage that acts as the interface between the environment, WM, and LTM.

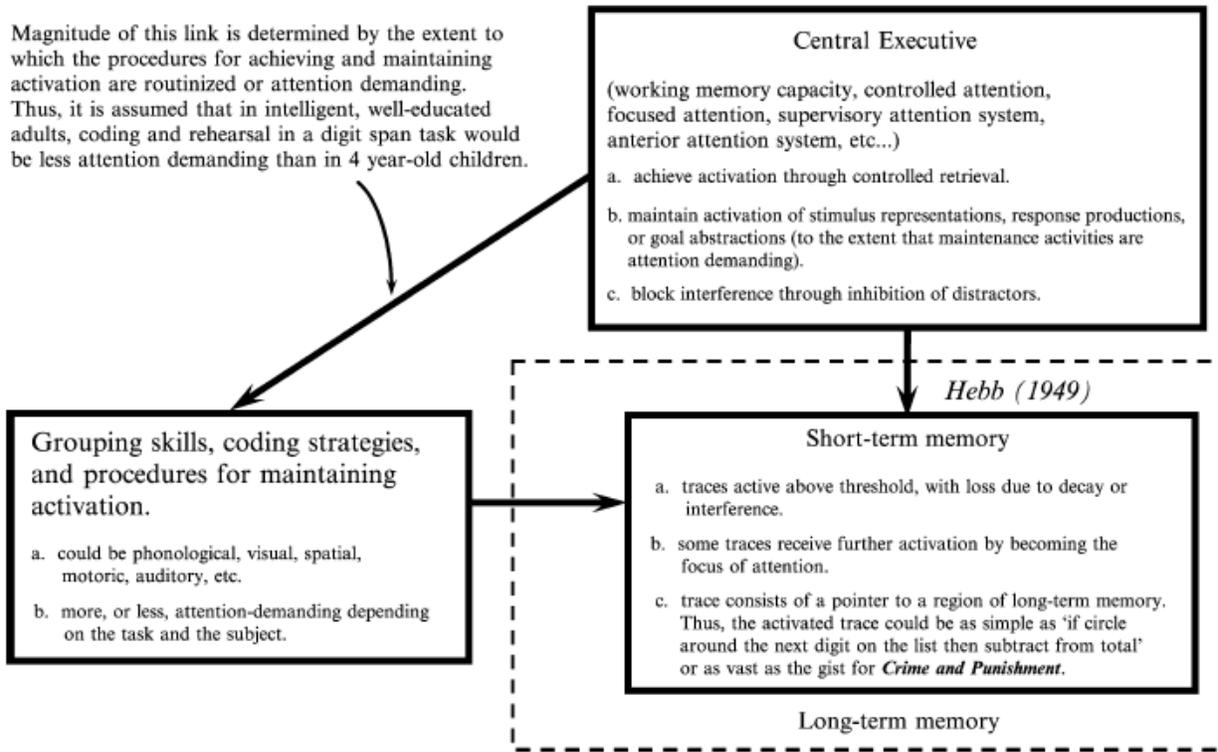


Figure 20. Engle, 2004 WM model.

Relationships of components of the working memory system as proposed by Engle et al. This diagram shows the three components of WM, attention, rehearsal procedures, and STM, which is an activated part of the LTM. (from Engle, 2004, p. 148).

Furthermore, the moderate to high correlations between these three factors were interpreted as a confirmation that these systems do overlap. For example, it is believed that STM is the part of WM that is responsible for holding the components of language while they are being processed and integrated by the Central Executive system. Also, it is believed that STM has a very similar structure to LTM and that it is the activated part of LTM. This overlapping is a

good example of the complexity of the constructs that are measured in this analysis, since after all the purifications that each factor has been through, such as selecting tasks that highly measure one function but not the others and running CFA which only retains the common variance in each factor, there still moderate to high correlations between these factors.

The overlapping nature of the factors was also confirmed by the SEM models where the relations between LP, STM, and WM and four sentence types was investigated. When tested individually, these relations between each of these factors and each of the sentence comprehension measures show moderate correlations and significant regression coefficients. However, when the SEM models, where the effects of all the factors were controlled for in each model, were built, the effects of LP and STM became much smaller, insignificant, and in some cases, changed their directions. The same phenomenon was true for WM except in its relation to OC and GP sentences where it maintained a significant effect even after controlling for the effects of LP and STM. Furthermore, the fact that only WM was significantly predicting OC and GP sentences gives more confidence that the WM factor is truly measuring the WM system. GP, such as “The blue circle touched by the green square is above the green circle on this one” and OC, such as “It was the blue circle that the green square touched on this one”, are known to be dependent on WM since they require some level of manipulation and choosing between candidate techniques of sentence processing and comprehension (E. Chen et al., 2005; R. C. Martin, 1990; Vuong & Martin, 2011). However, this is not the case for CS, where minimum manipulation is required and it is hypothesized that CS require more STM than WM (R. C. Martin, 1987; Miyake et al., 1994). In general, the analysis shows that WM has different relations with the sentence comprehension measures than the relations between LP and STM with the same measures. Even more, a general pattern in the analysis shows that LP and STM

cancel each other's effect. Again, when modeled individually, STM and LP show significant prediction of the targeted sentence. When modeled together, however, the variance explained does not change much compared to the variance explained by each one of them individually, and the relation with the targeted sentence becomes insignificant.

The interpretation of these observations is that WM is the soul of language comprehension. Furthermore, it is believed that WM is the most important factor since when a specific sentence requires WM, WM does not get shadowed or affected by other factors, which also suggests its independence from the other factors. On the other hand, LP and STM require WM to be properly functioning, since none of them was able to significantly predict any sentence while controlling for the effect of the other factors, especially STM with CS. In fact, this observation emphasizes on an important advantage of the methodology that was used to answer the research questions. To explain, there are many studies that have investigated the relations between cognitive systems or functions and language comprehension. However, most of these studies take one function and study its effect on language comprehension without controlling for the effects of any other cofounders. Consequently, this leads the researchers to believe that there is a significant effect of the cognitive function they are investigating on language comprehension, which would have been the case in this analysis if the effect of the other factors was not controlled for. Moreover, the fact that the three factors were able to only predict 10% of the LA sentences is interpreted by that LA sentences rely heavily on the CR function, which the factor analysis could not detect or separate from the other factors (E. Chen et al., 2005; R. C. Martin & He, 2004). Therefore, it is believed that it would have been possible to explain a significant variance of LA sentences if a conflict resolution or interference inhibition factor was detected. Furthermore, although WM significantly predicted the interference in the

GP and OC sentences and did not predict the interference in the LA sentences, the interference in the LA sentences is on the semantic level that requires CR, which is a very specific function of WM. On the other hand, the interference in GP and OC sentences is on the level of what rule of sentence comprehension to be used and requires many functions of WM including, but not only, the CR function, which led to the easier detection and prediction by WM.

5.5 LIMITATIONS AND FUTURE WORK

Although the usability study has increased the understanding of how users' performance significantly changes when using one system or another, it is recognized that the sample size was a limitation of this study, not because it was not sufficient to detect usability issues but because it was not sufficient to investigate the difference between different groups and how they react to both systems. Another limitation in the usability study that can be solved in future work is to focus on tasks that have the same requirements when performed on the two systems so that the difference between the two systems can be analyzed on the individuals' level. One challenge that might be investigated in depth in the future is when to start timing the performance of the participants on the two methods. For example, in this study, the timer was started after the participants open the Excel file and are ready to start searching for the required information on Excel, and when they open the mode-choose page on the iRDMS. However, someone might disagree with this method and start the timer when participants log into the operating system or click to start the Excel program and the internet browser. Furthermore, the fact that the number of years of using Excel predicted the performance on the iRDMS and that the performance on task 1 using Excel had a significant correlation with the performance of the same task on iRDMS

suggest that the adoption of the new system might be predicted by the users' performance and experience on the old one. Therefore, future work could implement a model that includes multiple factors that measure the users' relationship with the old system to predict their adoption to the new one. Moreover, future work could investigate the influence of the database content on the usability of the system.

Even more, the results of the usability study indicate that there was a higher variability in the participants' performance on Excel compared to the variability of their performance on iRDMS. This can be explained by many factors. One, is that participants have different levels of experience and skills on Excel since they have been using it for some time, which was not the case on iRDMS since they all have the same level of knowledge about the system. A future usability study could investigate the variability of the iRDMS users' performance after they use it for a couple of months or years. The conclusion of such study would truly answer the question whether the iRDMS completely standardize the process of data retrieval across users. Furthermore, future work could involve implementing data extraction functions into the system. In other words, future work could merge and integrate the data extraction techniques that have been used in this dissertation with the iRDMS. For example, with such technology, clinicians could upload their original data files and could use a control panel that enables them of creating new columns and specifying the extraction rules of these columns similar to what has been done in the extraction map. This technique has a great potential to help clinicians who do not have access to technicians that can help them to build customized extraction algorithms.

Although the SEM analysis have increased the understanding of the nature of the interactions between critical cognitive systems, it is recognized that this analysis is far from modeling the entire language comprehension process and all the underlying supporting cognitive

functions. Therefore, it is recognized that one major challenge in this study is the number and type of tasks were collected for each hypothesized construct. A future research opportunity might be to collect data on different tasks that might measure the CR effect more precisely. Also, future research might be able to investigate the nature of the interaction between LP and STM by using tasks that detect deeper levels of these systems from different aspects, such as different domains of language, or different measures of memory capacity.

Another challenge that is recognized in SEM analysis is the sample size. Although 100 was the magical number that was needed to be able to conduct CFA and SEM, the sample size in this dissertation was still on the edge of being not sufficient to be used to draw any generalizable conclusions. Furthermore, with this small sample size, there was not a comfortable level of freedom to drop any outliers or missing values, even though they had no effect on the results of the analysis. Moreover, a larger sample size would have made it possible to conduct cluster analysis to detect any sub-groups in the used dataset and conduct separate CFA and SEM for each group to investigate whether different groups have different cognitive interactions. Similarly, a larger sample size with control or normal participants would have helped to investigate whether the findings are limited to PWA or they can be generalized to a none impaired cognitive-linguistic system.

This analysis opens a window for developing health information systems that use insights from the results to improve the healthcare of PWA. One example of a health information system that might be developed based on the results of this analysis, is a predictive software that enables clinicians to predict the improvement in sentence comprehension of their patients if they improved one of the cognitive systems that were modeled. Furthermore, the results of this analysis can be used to develop health mobile applications that bring some of the tasks that were

used in this study to the hands of the patients and their families with some scales to compare their performance with the performance of the participants from this dissertation. Such mobile apps can also be used to collect the performance of the users to conduct future analysis with some control measures for the uncontrolled and unsupervised administration.

6.0 CONCLUSION

This dissertation has demonstrated a comprehensive data management and analysis of a large research dataset. This dissertation involved the adoption of a large dataset from its early data collection stages and managing it through its full data lifecycle. Furthermore, since each research project has its unique challenges and barriers, one takeaway from this dissertation is the importance of involving a data management team that have the resources and ability to design and implement unique and customized solutions to solve the challenges of each project. The main advantage of having the data management team involved early in the data collection process was to detect any potential issues with the collected data and to fix them in early stages. The data extraction map was an example of how data management challenges can be solved by implementing tools that exactly fit the purpose of the current task and can be reused in the future. The extraction map was an excellent tool to create a direct communication line between the clinicians and the extraction algorithm. This had a huge impact on the speed of conducting the data extraction since there was no need for meetings between the clinicians and the technicians to communicate the new rules, and no need for the technicians to spend time re-customizing the extraction algorithm to fit the new rules. Even more, this technique facilitated the conduction of the data validation in a fast and systematic fashion. Therefore, this technique had a positive impact on the quality of the processed data and, thus, on the analyses that were and will be conducted on this dataset.

Moreover, another example of the importance of involving a data management team is the implementation of the iRDMS that helped the research project to ensure high data quality and security. These important features would have been jeopardized if public or general tools were used to conduct the data management of this project. In addition to the security and confidentiality issues that the online system solves, it facilitates the distribution of new data updates, especially in long-term projects that involve continuous data extraction and processing. Furthermore, one of the important investigations in this dissertation is the usability study that was used to test the efficiency, effectiveness, and satisfaction of the users of the iRDMS.

As mentioned earlier, the methodology that was followed in the SEM analysis had the advantage of controlling multiple factors to gain a more comprehensive picture of the nature of the relation between the cognitive systems and sentence comprehension. This analysis has revealed that LP and STM are highly correlated factors, which resulted in a high interaction and overlapping between the two factors. The overlapping nature of the factors was also confirmed by the SEM models where the relations between LP, STM, and WM and four sentence types was investigated. Furthermore, this analysis has confirmed theories that viewed attention and WM in general as the soul of language comprehension. Moreover, this analysis has showed that WM is the most important factor since when a specific sentence requires WM, WM does not get shadowed or affected by other factors, which also suggests its independence from the other factors.

APPENDIX A

GLOSSARY

International Organization for Standardization (ISO): The International Organization for Standardization is an international standard-setting body composed of representatives from various national standards organizations

Compound Sentences (CS): For example, “Touch the little red square and the big blue circle on this one”

Object Cleft (OC): For example, “It was the blue circle that the green square touched on this one”

Garden Path (GP): For example, “The blue circle touched by the green square is above the green circle on this one”

Lexical Ambiguity (LA): For example, “He drank the port quickly”

Data lifecycle: A the process of managing data from its point of creation or collection to its final destination.

The Post-Study System Usability Questionnaire (PSSUQ): A free to use questionnaire that was originally developed by a group of human factors engineers and usability specialists as an internal project at International Business Machines Corporation (IBM)

Think-aloud: A usability technique where the participants, while performing tasks, are encouraged to say aloud what they are doing, thinking, liking, and disliking about the two systems while performing the tasks

Computerized Revised Token Test (CRTT): A sentence comprehension test that comes in different versions based on the administration method and the used sentence type.

Explanatory factor analysis (EFA): A statistical technique that discovers common variance between variables.

Confirmatory factor analysis (CFA): A statistical technique that tests theories of the common variance between variables.

Structural Equation Modeling (SEM): A family of statistical techniques, including EFA and CFA, that deals with unobserved measures. However, it is very common to use the term SEM to refer to the path analysis part of SEM.

Maximum Likelihood (ML): A method that is used by factor analysis to estimate the modeled correlation matrix.

Internet-based research data management system (iRDMS): A system that was developed as part of this dissertation to manage the collected dataset.

E-Prime: A software that was developed by Psychology Software Tools, Inc. to be used in computerized behavioral experiments and research.

Normalization forms: Set of rules that should be followed in database design to ensure having a database with no redundant data across its tables.

DEMOGRAPHICS QUESTIONNAIRE

Age Group:

- 18-24
- 25-34
- 34-60
- 60-75
- 75+

Gender:

- M
- F

What is the highest level of education you have completed:

- 2-year degree
- Bachelor's Degree
- Master's Degree
- Doctoral Degree
- Professional Degree

About how long have you been using Excel (in years)?

0 1 2 3 4 5 6 7 8 9 10+

On a Scale of One to Ten, What is Your Skill Level in Microsoft Excel?

Low 1 2 3 4 5 6 7 8 9 10 High

SATISFACTION QUESTIONNAIRE

1. Overall, I am satisfied with how easy it is to use this system.

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

COMMENTS:

2. It was simple to use this system.

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

COMMENTS:

3. I can effectively complete my work using this system.

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

COMMENTS:

4. I am able to complete my work quickly using this system.

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

COMMENTS:

5. I am able to efficiently complete my work using this system.

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

COMMENTS:

6. I feel comfortable using this system.

STRONGLY									STRONGLY
AGREE	1	2	3	4	5	6	7		DISAGREE

COMMENTS:

7. It was easy to learn to use this system.

STRONGLY									STRONGLY
AGREE	1	2	3	4	5	6	7		DISAGREE

COMMENTS:

8. I believe I became productive quickly using this system.

STRONGLY									STRONGLY
AGREE	1	2	3	4	5	6	7		DISAGREE

COMMENTS:

9. The system gives error messages that clearly tell me how to fix problems.

STRONGLY									STRONGLY
AGREE	1	2	3	4	5	6	7		DISAGREE

COMMENTS:

10. Whenever I make a mistake using the system, I recover easily and quickly.

STRONGLY									STRONGLY
AGREE	1	2	3	4	5	6	7		DISAGREE

COMMENTS:

11. The information (such as online help, on-screen messages, and other documentation) provided with this system is clear.

STRONGLY
AGREE 1 2 3 4 5 6 7 STRONGLY
DISAGREE

COMMENTS:

12. It is easy to find the information I needed.

STRONGLY
AGREE 1 2 3 4 5 6 7 STRONGLY
DISAGREE

COMMENTS:

13. The information provided for the system is easy to understand.

STRONGLY
AGREE 1 2 3 4 5 6 7 STRONGLY
DISAGREE

COMMENTS:

14. The information is effective in helping me complete the tasks and scenarios.

STRONGLY
AGREE 1 2 3 4 5 6 7 STRONGLY
DISAGREE

COMMENTS:

15. The organization of information on the system screens is clear.

STRONGLY
AGREE 1 2 3 4 5 6 7 STRONGLY
DISAGREE

COMMENTS:

16. The interface of this system is pleasant.

STRONGLY
AGREE 1 2 3 4 5 6 7 STRONGLY
DISAGREE

COMMENTS:

17. I like using the interface of this system.

STRONGLY
AGREE 1 2 3 4 5 6 7 STRONGLY
DISAGREE

COMMENTS:

18. This system has all the functions and capabilities I expect it to have.

STRONGLY
AGREE 1 2 3 4 5 6 7 STRONGLY
DISAGREE

COMMENTS:

19. Overall, I am satisfied with this system.

STRONGLY
AGREE 1 2 3 4 5 6 7 STRONGLY
DISAGREE

COMMENTS:

20. What did you like about the site?

21. What do you dislike about the site?

22. If you could change one thing about this system, what would it be?

23. what did you find confusing or a problem in the website?

TASKS

Task 1 (Files)

In this task, you will be using the file “**E-prime**” to retrieve the requested dataset. The data in this file are raw, unprocessed and non-extracted. Therefore, some data processing might be needed to collect the final data points.

In this task, you are asked to retrieve data points for the test “**OUTsem1b**” for one subject “**5555**”.

You are required to retrieve data points for the variable “**Probe_RT**”. However, this variable is presented under the column “**Question.RT**” in “**E-Prime**”.

- Please open the Excel file called “Task 1” and type in the generated data under the column “**Time_StartIns**”.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 1 (Interface)

In this task, you will be using the web interface to retrieve the requested data.

In this task, you are asked to retrieve data points for the test “**OUTsem1b**” for one subject “**5555**”.

You are required to retrieve data points for the variable “**Probe_RT**”.

- Please paste the downloaded Excel file in the task’s folder.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 2 (Files)

In this task, you will be using the file “**E-prime**” to retrieve the requested dataset. The data in this file are raw, unprocessed and non-extracted. Therefore, some data processing might be needed to collect the final data points.

In this task, you are asked to retrieve data points for the test “**STMsem3**” for one subject “**2222**”.

You are required to retrieve data points for the variable “**ITI_ACC**”. However, this variable is presented under the column “**ITI.ACC[Block]**” in “**E-Prime**”.

- Please open the Excel file called “Task 2” and type in the generated data under the column “**ITI_ACC**”.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 2 (Interface)

In this task, you will be using the web interface to retrieve the requested data.

In this task, you are asked to retrieve data points for the test “**STMsem3**” for one subject “**2222**”.

You are required to retrieve data points for the variable “**ITI_ACC**”.

- Please paste the downloaded Excel file in the task’s folder.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 3 (Files)

In this task, you will be using the file “**E-prime**” to retrieve the requested dataset. The data in this file are raw, unprocessed and non-extracted. Therefore, some data processing might be needed to collect the final data points.

In this task, you are asked to retrieve data points for the test “**OUTsem1b**” for the subjects: “**3333**” AND “**6666**”.

You are required to retrieve data points for the variable “**ItemNumber1**”. However, this variable is presented under the column “**GlobalNum**” in “**E-Prime**”.

- Please open the Excel file called “Task 3” and type in the generated data under the column “**ItemNumber1**”.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 3 (Interface)

In this task, you will be using the web interface to retrieve the requested data.

In this task, you are asked to retrieve data points for the test “**OUTsem1b**” for the subjects: “**3333**” AND “**6666**”.

You are required to retrieve data points for the variable “**ItemNumber1**”.

- Please paste the downloaded Excel file in the task’s folder.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 4 (Files)

In this task, you will be using the file “E-prime” to retrieve the requested dataset. The data in this file are raw, unprocessed and non-extracted. Therefore, some data processing might be needed to collect the final data points.

In this task, you are asked to retrieve data points for the test “STMsyn3” for one subject “1111”.

You are required to retrieve data points for the variable “ItemType3”. However, this variable is not presented directly in “E-Prime”. The values of “ItemType3” depend on the values of the column “Sequence” in “E-Prime” as following:

- If “Sequence” = "Filler", the value of “ItemType3” should be = “0”.
- If “Sequence” = “Target”, the value of “ItemType3” should be = “1”.
- If “Sequence” = "Prime", the value of “ItemType3” should be = “2”.
- Please open the Excel file called “Task 4” and type in the generated data under the column “ItemType3”.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 4 (Interface)

In this task, you will be using the web interface to retrieve the requested data.

In this task, you are asked to retrieve data points for the test “**STMsyn3**” for one subject “**1111**”.

You are required to retrieve data points for the variable “**ItemType3**”.

- Please paste the downloaded Excel file in the task’s folder.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 5 (Files)

In this task, you will be using the file “**E-prime**” to retrieve the requested dataset. The data in this file are raw, unprocessed and non-extracted. Therefore, some data processing might be needed to collect the final data points.

In this task, you are asked to retrieve data points for the test “**STMsem1**” for one subject “**9999**”.

You are required to retrieve data points for the variable “**ISI6_OnsetToOnset**”. However, this variable is not presented directly in “**E-Prime**”. You need to calculate the values of “**ISI6_OnsetToOnset**” using two columns (“**Stimg.OnsetTime**” and “**Waitf.OnsetTime**”) in “**E-Prime**” as following:

$$\text{“ISI6_OnsetToOnset”} = \text{“Stimg.OnsetTime”} - \text{“Waitf.OnsetTime”}$$

- Please open the Excel file called “Task 5” and type in the generated data under the column “**ISI6_OnsetToOnset**”.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 5 (Interface)

In this task, you will be using the web interface to retrieve the requested data.

In this task, you are asked to retrieve data points for the test “STMsem1” for one subject “9999”.

You are required to retrieve data points for the variable “ISI6_OnsetToOnset”.

- Please paste the downloaded Excel file in the task’s folder.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 6 (Files)

In this task, you will be using the file “**E-prime**” to retrieve the requested dataset. The data in this file are raw, unprocessed and non-extracted. Therefore, some data processing might be needed to collect the final data points.

In this task, you are asked to retrieve data points for the test “**LPsyn2a**” for one subject “**5555**”.

You are required to retrieve data points for the variable “**ReadTm2**”. However, this variable is not presented in “**E-Prime**”. You need to collect the value of “**ReadTm2**” from the column “**ReadT**” in a file called: “**Subject5555_Session6_LPsyn2a**”.

To find the score of “**ReadT**” that corresponds to each item in task “**LPsyn2a**”, you need to find the row in “**Subject5555_Session6_LPsyn2a**” where:

- 1- The value of column “**ItemOrderTxt**” = the value of column “**Experimental.Sample**” in the “**E-Prime**”.

AND

- 2- The value of column “**PhraseLock**” = -2

- Please open the Excel file called “**Task 6**” and type in the generated data under the column “**ReadTm2**”.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 6 (Interface)

In this task, you will be using the web interface to retrieve the requested data.

In this task, you are asked to retrieve data points for the test “**LPsyn2a**” for one subject “**5555**”.

You are required to retrieve data points for the variable “**ReadTm2**”.

- Please paste the downloaded Excel file in the task’s folder.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 7 (Files)

In this task, you will be using the file “**E-prime**” to retrieve the requested dataset. The data in this file are raw, unprocessed and non-extracted. Therefore, some data processing might be needed to collect the final data points.

In this task, you are asked to retrieve data points for the test “**LPsem1a**” for one subject “**6666**”.

You are required to retrieve data points for the variable “**ItemType1**”. However, this variable is not directly listed in “**E-prime**”.

To generate the values of the “**ItemType1**”, you will need to look at the columns “**Running[Block]**” and “**CorrectAnswer[Block]**”. There are two possible values combination of these two columns for each row, and value of “**ItemType1**” will be different based on each one of them. Only one combination applies:

A- If “Running[Block]” = "Experiment"	B- If “Running[Block]” = "Experiment"
AND	AND
“CorrectAnswer[Block]” = "5"	“CorrectAnswer[Block]” = "3"
Then the value of “ItemType1” = “REL”.	Then the value of “ItemType1” = “UN”.

Please open the Excel file called “Task 7” and type in the generated data under the column “**ItemType1**”.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 7 (Interface)

In this task, you will be using the web interface to retrieve the requested data.

In this task, you are asked to retrieve data points for the test “**LPsem1a**” for one subject “**6666**”.

You are required to retrieve data points for the variable “**ItemType1**”.

- Please paste the downloaded Excel file in the task’s folder.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 8 (Files)

In this task, you will be using the file “**E-prime**” to retrieve the requested dataset. The data in this file are raw, unprocessed and non-extracted. Therefore, some data processing might be needed to collect the final data points.

In this task, you are asked to retrieve data points for the test “**STMsem2a**” for one subject “**7777**”.

You are required to retrieve data points for the variable “**ItemType4**”. However, this variable is not directly listed in “**E-prime**”.

To generate the values of the “**ItemType4**” you will need to look at the columns “**Running[Block]**”, “**PrimeType**” and “**TargetType**”. There are four possible values combination of these three columns for each row, and value of “**ItemType4**” will be different based on each one of them. Only one combination applies:

3- If “Running[Block]” = “Experiment”
AND
PrimeType = “t”
AND
TargetType = “t”

Then the value of “**ItemType4**” = “1”.

4- If “Running[Block]” = “Experiment”
AND
PrimeType = “a”
AND
TargetType = “t”

Then the value of “**ItemType4**” = “2”.

1- If “Running[Block]” = “Experiment”
AND
PrimeType = “t”
AND
TargetType = “a”

Then the value of “**ItemType4**” = “2”.

2- If “Running[Block]” = “Experiment”
AND
PrimeType = “a”
AND
TargetType = “a”

Then the value of “**ItemType4**” = “3”.

Please open the Excel file called “Task 8” and type in the generated data under the column “**ItemType4**”.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 8 (Interface)

In this task, you will be using the web interface to retrieve the requested data.

In this task, you are asked to retrieve data points for the test “STMsem2a” for one subject “7777”.

You are required to retrieve data points for the variable “ItemType4”.

- Please paste the downloaded Excel file in the task’s folder.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 9 (Files)

In this task, you will be using the file “**E-prime**” to retrieve the requested dataset. The data in this file are raw, unprocessed and non-extracted. Therefore, some data processing might be needed to collect the final data points.

In this task, you are asked to retrieve the final score for the test “**LPsem1b**” for subject “**5555**”. This means that the outcome of this task is one number that represents the performance of an individual on test “**LPsem1b**”. This score is the average of the scores in column “**Resp_RT**” for ONLY the items where column “**Itemtype2**” equals to “**REL**”.

However, since you are dealing with non-extracted data, you need to do some data extraction first:

- 1- “**Itemtype2**” is represented in “**E-Prime**” as column “**CorrectAnswer[Block]**”. If the value of “**CorrectAnswer[Block]**” equals to “**5**” in “**E-Prime**” then the value of “**Itemtype2**” equals to “**REL**”.
- 2- “**Resp_RT**” is represented in “**E-Prime**” as column “**Response.RT[Block]**”.

Please open the Excel file called “**Task 9**” and type in the generated data under the column “**FinalScore**”.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 9 (Interface)

In this task, you will be using the web interface to retrieve the requested data.

In this task, you are asked to retrieve the final score for the test “**LPsem1b**” for subject “**5555**”. This means that the outcome of this task is one number that represent the performance of an individual on test “**LPsem1b**”. This final score is called “**RT_Negative**”.

Please paste the downloaded Excel file in the task’s folder.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 10 (Files)

In this task, you will be using the file “E-prime” to retrieve the requested dataset. The data in this file are raw, unprocessed and non-extracted. Therefore, some data processing might be needed to collect the final data points.

In this task, you are asked to retrieve the final score for the test “CR2a” for subject “6666”. This means that the outcome of this task is one number that represents the performance of an individual on test “CR2a”. This score is the average of the scores in column “Probe_RT” for ONLY the items where column “CorrectAnswer” equals to “5”.

However, since you are dealing with non-extracted data, you need to do some data extraction first:

- 1- “CorrectAnswer” is represented in “E-Prime” as column “CorrectAnswer[Block]”.
- 2- “Probe_RT” is represented in “E-Prime” as column “Probe.RT[Block]”.

Please open the Excel file called “Task 10” and type in the generated data under the column “FinalScore”.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Task 10 (Interface)

In this task, you will be using the web interface to retrieve the requested data.

In this task, you are asked to retrieve the final score for the test “**CR2a**” for subject “**6666**”. This means that the outcome of this task is one number that represent the performance of an individual on test “**CR2a**”. This final score is called “**RT_Positive**”.

Please paste the downloaded Excel file in the task’s folder.

Overall, I am satisfied with the ease of completing this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

Overall, I am satisfied with the amount of time it took to complete this task:

STRONGLY AGREE	1	2	3	4	5	6	7	STRONGLY DISAGREE
-------------------	---	---	---	---	---	---	---	----------------------

APPENDIX B

PSSUQ groups.

Grouping Variable	Group	N	Scale	M (SD)	Range
Age	18-24	9	SysUse	1.07 (0.14)	1-1.38
			InfoQual	1.14 (0.31)	1-2
			IntQual	1.22 (0.42)	1-2.33
			Overall	1.12 (0.16)	1-1.53
	25-34	9	SysUse	1.27 (0.44)	1-2.43
			InfoQual	1.37 (0.39)	1-2.20
			IntQual	1.41 (0.31)	1-2
			Overall	1.34 (0.37)	1-2.25
	35-60	5	SysUse	1.02 (0.05)	1-1.12
			InfoQual	1.36 (0.22)	1-1.67
			IntQual	1.07 (0.13)	1-1.33
			Overall	1.16 (0.10)	1-1.29
	61-75	1	SysUse	1 (0)	1-1
			InfoQual	1.14 (0)	1.14-1.14
			IntQual	1 (0)	1-1
			Overall	1.05 (0)	1.05-1.05
Education	<BS	1	SysUse	1 (0)	1-1
			InfoQual	1.14 (0)	1.14-1.14
			IntQual	1.33 (0)	1.33-1.33
			Overall	1.11 (0)	1.11-1.11
	BS	14	SysUse	1.19 (0.37)	1-2.43
			InfoQual	1.38 (0.38)	1-2.20
			IntQual	1.38 (0.40)	1-2.33
			Overall	1.30 (0.32)	1-2.25
	MS	7	SysUse	1.07 (0.13)	1-1.38
			InfoQual	1.07 (0.12)	1-1.33
			IntQual	1.05 (0.12)	1-1.33
			Overall	1.06 (0.07)	1-1.16

	PhD	2	SysUse	1 (0)	1-1
			InfoQual	1.33 (0.33)	1-1.67
			IntQual	1 (0)	1-1
			Overall	1.15 (0.15)	1-1.29
Gender	Female	14	SysUse	1.07 (0.14)	1-1.50
			InfoQual	1.22 (0.29)	1-2
			IntQual	1.10 (0.15)	1-1.33
			Overall	1.13 (0.16)	1-1.53
	Male	10	SysUse	1.22 (0.42)	1-2.43
			InfoQual	1.34 (0.39)	1-2.20
			IntQual	1.47 (0.43)	1-2.33
			Overall	1.32 (0.35)	1-2.25
Race	African American	2	SysUse	1 (0)	1-1
			InfoQual	1.07 (0.07)	1-1.14
			IntQual	1.17 (0.17)	1-1.33
			Overall	1.06 (0.06)	1-1.11
	Asian	11	SysUse	1.06 (0.11)	1-1.38
			InfoQual	1.15 (0.25)	1-1.86
			IntQual	1.24 (0.40)	1-2.33
			Overall	1.12 (0.15)	1-1.53
	White	11	SysUse	1.24 (0.41)	1-2.43
			InfoQual	1.44 (0.37)	1-2.20
			IntQual	1.27 (0.31)	1-2
			Overall	1.32 (0.34)	1-2.25

ASQ groups.

Grouping Variable	Group	N	Variable	M (SD)	Range
Age	18-24	9	Files_Q1	4.28 (1.85)	1-7
			Files_Q2	4.14 (1.97)	1-7
			Interface_Q1	1.12 (0.33)	1-2
			Interface_Q2	1.12 (0.33)	1-2
	25-34	9	Files_Q1	4 (1.90)	1-7
			Files_Q2	4.09 (1.93)	1-7
			Interface_Q1	1.30 (0.57)	1-3
			Interface_Q2	1.24 (0.54)	1-3
	35-60	5	Files_Q1	3.26 (1.44)	1-6
			Files_Q2	3.28 (1.63)	1-7
			Interface_Q1	1.24 (0.43)	1-2
			Interface_Q2	1.16 (0.37)	1-2

	61-75	1	Files_Q1	3.10 (1.76)	1-7
			Files_Q2	3.60 (2.06)	1-7
			Interface_Q1	1 (0)	1-1
			Interface_Q2	1 (0)	1-1
Education	<BS	1	Files_Q1	4.90 (1.14)	3-7
			Files_Q2	5 (1.48)	2-7
			Interface_Q1	1 (0)	1-1
			Interface_Q2	1 (0)	1-1
	BS	14	Files_Q1	3.67 (1.85)	1-7
			Files_Q2	3.64 (1.91)	1-7
			Interface_Q1	1.23 (0.48)	1-3
			Interface_Q2	1.18 (0.44)	1-3
	MS	7	Files_Q1	4.50 (1.65)	1-7
			Files_Q2	4.54 (1.69)	1-7
			Interface_Q1	1.21 (0.44)	1-3
			Interface_Q2	1.23 (0.48)	1-3
	PhD	2	Files_Q1	3.05 (1.83)	1-6
			Files_Q2	3.15 (2.15)	1-7
			Interface_Q1	1.15 (0.36)	1-2
			Interface_Q2	1 (0)	1-1
Gender	Female	14	Files_Q1	4.10 (1.83)	1-7
			Files_Q2	4.13 (1.92)	1-7
			Interface_Q1	1.20 (0.45)	1-3
			Interface_Q2	1.17 (0.45)	1-3
	Male	10	Files_Q1	3.65 (1.80)	1-7
			Files_Q2	3.63 (1.89)	1-7
			Interface_Q1	1.22 (0.46)	1-3
			Interface_Q2	1.17 (0.40)	1-3
Race	African American	2	Files_Q1	4.25 (1.95)	1-7
			Files_Q2	4.20 (2.16)	1-7
			Interface_Q1	1.25 (0.62)	1-3
			Interface_Q2	1.25 (0.62)	1-3
	Asian	11	Files_Q1	4.32 (1.78)	1-7
			Files_Q2	4.26 (1.91)	1-7
			Interface_Q1	1.15 (0.35)	1-2
			Interface_Q2	1.20 (0.42)	1-3
	White	11	Files_Q1	3.45 (1.75)	1-7
			Files_Q2	3.53 (1.82)	1-7
			Interface_Q1	1.26 (0.50)	1-3
			Interface_Q2	1.13 (0.38)	1-3

Subjects with first three tasks only.

ID	Excel			iRDMS			Wilcoxon test	
	M (SD)	RSD	Range	M (SD)	RSD	Range	W-value	P-value
1	285 (243.52)	0.85	64-761	46.10 (22.12)	0.48	33-109	55	<0.01
2	183 (62.28)	0.34	69-278	31.90 (5.17)	0.16	23-40	55	<0.01
3	160.90 (62.23)	0.39	89-315	32.90 (7.94)	0.24	22-47	55	<0.01
4	156.30 (67.52)	0.43	56-268	33.40 (5.90)	0.18	23-44	55	<0.01
5	140.60 (62.37)	0.44	27-240	23.60 (5.02)	0.21	17-31	55	<0.01
6	117.20 (41.11)	0.35	59-177	22.20 (2.99)	0.13	18-29	55	<0.01
7	155.30 (85.24)	0.55	52-372	19.30 (2.15)	0.11	16-24	55	<0.01
8	150.20 (88.31)	0.59	39-284	22.20 (4.28)	0.19	17-31	55	<0.01
9	125.90 (56.65)	0.45	35-253	23.80 (6.01)	0.25	18-37	55	<0.01
10	133.40 (55.61)	0.42	53-229	24.10 (3.30)	0.14	18-29	55	<0.01
11	92.30 (45.59)	0.49	38-216	23.40 (3.83)	0.16	15-29	55	<0.01
12	107 (58.95)	0.55	32-178	17.80 (3.89)	0.22	13-25	55	<0.01
13	149 (126.71)	0.85	37-500	22.80 (4.53)	0.2	19-34	55	<0.01
14	125.80 (55.36)	0.44	38-239	24.90 (7.57)	0.3	15-42	55	<0.01
15	98 (33.04)	0.34	57-170	20.60 (2.62)	0.13	16-24	55	<0.01
16	111.10 (39.03)	0.35	45-168	27.20 (5.34)	0.2	22-42	55	<0.01
17	112.10 (50.74)	0.45	41-190	32.90 (10.52)	0.32	18-53	55	<0.01
18	88.40 (31.04)	0.35	44-145	29.10 (3.86)	0.13	20-34	55	<0.01
19	125.60 (61.87)	0.49	33-231	25.80 (6.10)	0.24	18-38	55	<0.01
20	129 (54.64)	0.42	57-219	28.90 (4.81)	0.17	24-40	55	<0.01
21	131.90 (74.87)	0.57	43-271	27.70 (8.22)	0.3	19-45	55	<0.01
22	160.30 (60.95)	0.38	47-259	30.10 (3.65)	0.12	24-37	55	<0.01
23	116.20 (59.26)	0.51	45-253	26.40 (4.36)	0.17	21-34	55	<0.01
24	128.10 (52.83)	0.41	53-214	33 (9.39)	0.28	21-51	55	<0.01

Number of incorrect performances of each participant using Excel and iRDMS.

Method	Subject ID	Age	Education	Gender	Race	# Errors
Excel	1	35-60	PhD	Female	White	2
	4	61-75	BS	Female	White	1
	5	25-34	BS	Male	Asian	2
	7	18-24	MS	Male	Asian	1
	8	25-34	BS	Male	White	1
	9	18-24	MS	Female	Asian	1
	12	25-34	BS	Female	African American	1
	13	18-24	BS	Male	Asian	1
	14	25-34	MS	Female	Asian	1
	15	18-24	BS	Female	White	1
	18	35-60	PhD	Female	White	1
	19	25-34	MS	Female	White	2
	23	35-60	BS	Male	White	1
DBSM	11	18-24	BS	Male	Asian	1
	17	35-60	BS	Female	White	1
	24	25-34	BS	Male	White	1

BIBLIOGRAPHY

- Aaronson, L. S., Frey, M. A., & Boyd, C. J. (1988). Structural equation models and nursing research: Part II. *Nursing research*, 37(5), 315-317.
- Abran, A., Khelifi, A., Suryn, W., & Seffah, A. (2003). *Consolidating the ISO usability models*. Paper presented at the Proceedings of 11th international software quality management conference, Sofia, Bulgaria.
- Adani, F., & Fritzsche, T. (2015). *On the relation between implicit and explicit measures of child language development: evidence from relative clause processing in 4-year-olds and adults*. Paper presented at the Proceedings of the 39th Annual Boston University Conference on Language Development, Boston, Ma.
- Alanazi, A. (2015). *A COMPREHENSIVE MODEL TO EXPLAINING USERS' ACCEPTANCE AND INTENTION TO USE ELECTRONIC HEALTH RECORDS (EHR) IN REHABILITATION FACILITIES IN SAUDI ARABIA*. (PhD), University of Pittsburgh,
- Allen, C. M., Martin, R. C., & Martin, N. (2012). Relations between short-term memory deficits, semantic processing, and executive function. *Aphasiology*, 26(3-4), 428-461.
- Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of abnormal psychology*, 112(4), 545.
- Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *The psychology of learning and motivation*, 2, 89-195.
- Austin, P. C., & Steyerberg, E. W. (2015). The number of subjects per variable required in linear regression analyses. *Journal of clinical epidemiology*, 68(6), 627-636.
- Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4(11), 417-423.

- Baddeley, A. (2003). Working memory and language: An overview. *Journal of communication disorders*, 36(3), 189-208.
- Baddeley, A., Chincotta, D., Stafford, L., & Turk, D. (2002). Is the word length effect in STM entirely attributable to output delay? Evidence from serial recognition. *The Quarterly Journal of Experimental Psychology: Section A*, 55(2), 353-369.
- Baddeley, A., & Hitch, G. (1974). Working memory. *The psychology of learning and motivation*, 8, 47-89.
- Baddeley, A., & Warrington, E. K. (1970). Amnesia and the distinction between long-and short-term memory. *Journal of verbal learning and verbal behavior*, 9(2), 176-189.
- Bagozzi, R. P., & Yi, Y. (2012). Specification, evaluation, and interpretation of structural equation models. *Journal of the academy of marketing science*, 40(1), 8-34.
- Bajpai, N. (2015). Metrics for leveraging more in Clinical Data Management: proof of concept in the context of vaccine trials in an Indian pharmaceutical company. *Asian Journal of Pharmaceutical and Clinical Research*, 8(3), 350-357.
- Ball, A. (2012). *Review of data management lifecycle models*. Retrieved from <http://opus.bath.ac.uk/28587/>
- Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of school psychology*, 48(1), 5-37.
- Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology. General Section*, 28(1), 97-104.
- Batini, C., Ceri, S., & Navathe, S. B. (1992). *Conceptual database design: an Entity-relationship approach* (Vol. 116): Benjamin/Cummings Redwood City, CA.
- Batini, C., Lenzerini, M., & Navathe, S. B. (1986). A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys (CSUR)*, 18(4), 323-364.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the royal statistical society. Series B (methodological)*, 57(1), 289-300.
- Bentler, P. M., & Weeks, D. G. (1980). Linear structural equations with latent variables. *Psychometrika*, 45(3), 289-308.

- Berry, D. C., & Broadbent, D. E. (1990). The role of instruction and verbalization in improving performance on complex search tasks. *Behaviour & Information Technology*, 9(3), 175-190.
- Bowers, V. A., & Snyder, H. L. (1990). Concurrent versus retrospective verbal protocol for comparing window usability. *Proceedings of the Human Factors Society Annual Meeting*, 34(17), 1270-1274.
- Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the royal statistical society. Series B (methodological)*, 211-252.
- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5), 662-679.
- Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological bulletin*, 107(2), 260.
- Broca, P. (1861). Nouvelle observation d'aphemie produite par une lesion de la troisieme circonvolution frontale. *Bulletins de la Societe d'anatomie (Paris)*, 398-407.
- Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194), 4-7.
- Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. J. Grimm & P. R. Yarnold (Eds.), *Reading and Understanding Multivariate Statistics* (pp. 99-136). Washington, DC: American Psychological Association.
- Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3).
- Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming*: Sage.
- Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and brain Sciences*, 22(01), 77-94.
- Carroll, A. E., Marrero, D. G., & Downs, S. M. (2007). The HealthPia GlucoPack™ Diabetes phone: a usability study. *Diabetes technology & therapeutics*, 9(2), 158-164.
- Carstens, D. S., & Patterson, P. (2005). Usability study of travel websites. *Journal of Usability Studies*, 1(1), 47-61.

- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, 1(2), 245-276.
- Chen, E., Gibson, E., & Wolf, F. (2005). Online syntactic storage costs in sentence comprehension. *Journal of Memory and Language*, 52(1), 144-169.
- Chen, P. P.-S. (1976). The entity-relationship model—toward a unified view of data. *ACM Transactions on Database Systems (TODS)*, 1(1), 9-36.
- Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). *Development of an instrument measuring user satisfaction of the human-computer interface*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Washington, D.C., USA.
- Clark, D. G. (2011). Sentence comprehension in aphasia. *Language and Linguistics Compass*, 5(10), 718-730.
- Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, 13(6), 377-387.
- Cohen, J. (1992). A power primer. *Psychological bulletin*, 112(1), 155.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*: Routledge.
- Comrey, A. L., & Lee, H. B. (2013). *A first course in factor analysis*: Psychology Press.
- Connolly, T. M., & Begg, C. E. (2005). *Database systems: a practical approach to design, implementation, and management*: Pearson Education.
- Cooper, G. (1998). *Research into cognitive load theory and instructional design at UNSW*. Citeseer. Retrieved from <http://dwb4.unl.edu/Diss/Cooper/UNSW.htm>
- Council, N. R. (2011). *The prevention and treatment of missing data in clinical trials*: National Academies Press.
- Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (Vol. 20, pp. 506): Cambridge University Press.
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, 19(4), 450-466.

- Dawber, T. R., Meadors, G. F., & Moore Jr, F. E. (1951). Epidemiological Approaches to Heart Disease: The Framingham Study*. *American Journal of Public Health and the Nations Health*, 41(3), 279-286.
- Dawson, B., & Trapp, R. (2001). *Basic and clinical biostatistics*. New York: McGraw-Hill.
- De Winter, J. C. (2013). Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation*, 18(10), 1-12.
- DeCoster, J. (1998). *Overview of factor analysis*. Retrieved from <http://www.stat-help.com/notes.html>
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.
- Dennis, M., Agostino, A., Roncadin, C., & Levin, H. (2009). Theory of mind depends on domain-general executive functions of working memory and cognitive inhibition in children with traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology*, 31(7), 835-847.
- Dipnall, J. F., Berk, M., Jacka, F. N., Williams, L. J., Dodd, S., & Pasco, J. A. (2014). Data Integration Protocol In Ten-steps (DIPIT): A new standard for medical researchers. *Methods*, 69(3), 237-246.
- DIS, I. (2009). 9241-210: 2010. Ergonomics of human system interaction-Part 210: Human-centred design for interactive systems. *International Standardization Organization (ISO)*. Switzerland.
- DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, 14(20), 1-11.
- Djenno, M., Insua, G., Gregory, G. M., & Brantley, J. S. (2014). Discovering Usability: Comparing two discovery systems at one academic library. *Journal of Web Librarianship*, 8(3), 263-285.
- Dogac, A., Yuruten, B., & Spaccapietra, S. (1989). A generalized expert system for database design. *IEEE Transactions on Software Engineering*, 15(4), 479-491.
- Dollaghan, C. A. (2004). Evidence-based practice in communication disorders: What do we know, and when do we know it? *Journal of communication disorders*, 37(5), 391-400.

- Donner, A. (1982). The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. *The American Statistician*, 36(4), 378-381.
- Eason, K. (1988). *Information Technology and Organizational Change*: Taylor & Francis, Inc.
- Eberwein, C. A., Pratt, S. R., McNeil, M., Fossett, T. R., Szuminsky, N., & Doyle, P. J. (2007). Auditory performance characteristics of the computerized Revised Token Test (CRTT). *Journal of Speech, Language, and Hearing Research*, 50(4), 865-877.
- Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 44, pp. 145-200).
- Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of experimental psychology: General*, 128(3), 309.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological review*, 87(3), 215.
- Everitt, B. (1975). Multivariate analysis: The need for data, and other problems. *The British Journal of Psychiatry*, 126(3), 237-240.
- Few, S. (2004). *Show me the numbers: Designing tables and graphs to enlighten* (Vol. 1): Analytics Press Oakland, CA.
- Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). *Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?* Paper presented at the Proceedings of the SIGCHI conference on Human Factors in Computing Systems, New York, NY.
- Fruhling, A., & Lee, S. (2005). *Assessing the reliability, validity and adaptability of PSSUQ*. Paper presented at the AMCIS 2005 Proceedings, Omaha, USA.
- Fu, L., Ding, S., & Chen, T. (2010). *Clinical Data Management System*. Paper presented at the 2010 International Conference on Biomedical Engineering and Computer Science, Wuhan, China.
- Garson, D. G. (2008). Factor analysis. Retrieved from <http://tx.liberal.ntu.edu.tw/~PurpleWoo/Literature/DataAnalysis/Factor%20Analysis-types.htm>

- George, C. A. (2005). Usability testing and design of a library website: an iterative approach. *OCLC Systems & Services: International digital library perspectives*, 21(3), 167-180.
- Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic bulletin & review*, 19(5), 847-857.
- Gerritsen, M., Sartorius, O., vd Veen, F., & Meester, G. (1993). *Data management in multi-center clinical trials and the role of a nation-wide computer network. A 5 year evaluation*. Paper presented at the Proceedings of the Annual Symposium on Computer Application in Medical Care.
- Geschwind, N. (1979). Specializations of the human brain. *Scientific American*, 241(3), 180-201.
- Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1), 1-76.
- Gil, Y., Ratnakar, V., & Fritz, C. (2010). *Assisting Scientists with Complex Data Analysis Tasks through Semantic Workflows*. Paper presented at the AAAI Fall Symposium: Proactive Assistant Agents, Arlington, VA.
- Gorsuch, R. (1983). Factor analysis. 2nd. *Hillsdale, NJ: LEA*.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60, 549-576.
- Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate behavioral research*, 26(3), 499-510.
- Greenes, R., Pappalardo, A. N., Marble, C., & Barnett, G. O. (1969). Design and implementation of a clinical data management system. *Computers and biomedical research*, 2(5), 469-485.
- Grimes, D. A. (2010). Epidemiologic research using administrative databases: garbage in, garbage out. *Obstetrics & Gynecology*, 116(5), 1018-1019.
- Güting, R. H., Almeida, V., Ansorge, D., Behr, T., Ding, Z., Höse, T., . . . Telle, U. (2005). *Secondo: An extensible DBMS platform for research prototyping and teaching*. Paper presented at the ICDE 2005., Tokyo, Japan.
- Gvion, A., & Friedmann, N. (2012). Does phonological working memory impairment affect sentence comprehension? A study of conduction aphasia. *Aphasiology*, 26(3-4), 494-535.

- Haarmann, H. J., Davelaar, E. J., & Usher, M. (2003). Individual differences in semantic short-term memory capacity and reading comprehension. *Journal of Memory and Language*, 48(2), 320-345.
- Hakuta, K. (1981). Grammatical description versus configurational arrangement in language acquisition: The case of relative clauses in Japanese. *Cognition*, 9(3), 197-236.
- Hanten, G., & Martin, R. C. (2000). Contributions of phonological and semantic short-term memory to sentence processing: Evidence from two cases of closed head injury in children. *Journal of Memory and Language*, 43(2), 335-361.
- Haraty, R. A., Mansour, N. a., & Daou, B. (2001). *Regression testing of database applications*. Paper presented at the Proceedings of the 2001 ACM symposium on Applied computing, Las Vegas, NV, USA.
- Harris, L., Olson, A., & Humphreys, G. (2014). The link between STM and sentence comprehension: A neuropsychological rehabilitation study. *Neuropsychological rehabilitation*, 24(5), 678-720.
- Harris, R. J. (2001). *A primer of multivariate statistics*: Psychology Press.
- Harrison, R., Flood, D., & Duce, D. (2013). Usability of mobile applications: literature review and rationale for a new usability model. *Journal of Interaction Science*, 1(1), 1.
- Hebb, D. (1949). *The Organization of Behavior*. New York: John Weley & Sons Inc.
- Hee Kim, H., & Ho Kim, Y. (2008). Usability study of digital institutional repositories. *The electronic library*, 26(6), 863-881.
- Hellerstein, J. M. (2008). *Quantitative data cleaning for large databases*. Retrieved from <http://db.cs.berkeley.edu/jmh>
- Hernández, M. A., & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data mining and knowledge discovery*, 2(1), 9-37.
- Hershberger, S. L. (2005). Factor score estimation. In B. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science*. London, UK.: John Wiley and Sons, Inc.
- Hey, A. J., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1): Microsoft research Redmond, WA.
- Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation*, 3(1), 134-140.

- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179-185.
- Hovenga, E. J. (2010). *Health informatics: an overview* (Vol. 151): Ios Press.
- Hox, J., & Bechger, T. (1998). An introduction to structural equation modelling. *Family Science Review*, 11(354-373).
- Hwang, W., & Salvendy, G. (2010). Number of people required for usability evaluation: the 10±2 rule. *Communications of the ACM*, 53(5), 130-133.
- IEC, I. (2001). 9126-1 (2001). Software Engineering Product Quality-Part 1: Quality Model. *International Organization for Standardization*.
- In'nami, Y., & Koizumi, R. (2013). Review of sample size for structural equation models in second language testing and learning research: A Monte Carlo approach. *International Journal of Testing*, 13(4), 329-353.
- ISO, S. (1998). 9241-11 (1998). *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)—Part II Guidance on Usability*.
- Ivory, M. Y., & Hearst, M. A. (2001). The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys (CSUR)*, 33(4), 470-516.
- January, D., Trueswell, J. C., & Thompson-Schill, S. L. (2009). Co-localization of Stroop and syntactic ambiguity resolution in Broca's area: Implications for the neural basis of sentence processing. *Journal of Cognitive Neuroscience*, 21(12), 2434-2444.
- Joliffe, I., & Morgan, B. (1992). Principal component analysis and exploratory factor analysis. *Statistical methods in medical research*, 1(1), 69-95.
- Jöreskog, K. G. (1967). A general approach to confirmatory maximum likelihood factor analysis. *ETS Research Report Series*, 1967(2), 183-202.
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological review*, 99(1), 122.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1), 141-151.
- Kalinyak-Fliszar, M., Kohen, F., & Martin, N. (2011). Remediation of language processing in aphasia: Improving activation and maintenance of linguistic representations in (verbal) short-term memory. *Aphasiology*, 25(10), 1095-1131.

- Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: the contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of experimental psychology: General*, 132(1), 47.
- Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5), 402-406.
- Karpathiotakis, M., Alagiannis, I., & Ailamaki, A. (2016). Fast queries over heterogeneous data through engine customization. *Proceedings of the VLDB Endowment*, 9(12), 972-983.
- Kemmler, G., Hummer, M., Widschwendter, C., & Fleischhacker, W. W. (2005). Dropout rates in placebo-controlled and active-control clinical trials of antipsychotic drugs: a meta-analysis. *Archives of General Psychiatry*, 62(12), 1305-1312.
- Kent, W. (1983). A simple guide to five normal forms in relational database theory. *Communications of the ACM*, 26(2), 120-125.
- Kim, J.-O., & Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods & Research*, 6(2), 215-240.
- Kline, R. B. (2015). *Principles and practice of structural equation modeling*: Guilford publications.
- Knofczynski, G. T., & Mundfrom, D. (2008). Sample sizes when using multiple linear regression for prediction. *Educational and Psychological Measurement*, 68(3), 431-442.
- Krishnankutty, B., Bellary, S., Kumar, B. N., & Moodahadu, L. S. (2012). Data management in clinical research: an overview. *Indian journal of pharmacology*, 44(2), 168.
- Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260), 583-621.
- Kuchinke, W., Ohmann, C., Yang, Q., Salas, N., Lauritsen, J., Gueyffier, F., . . . Voko, Z. (2010). Heterogeneity prevails: the state of clinical trial data management in Europe—results of a survey of ECRIN centres. *Trials*, 11(1), 1.
- Kumar, R. S., & Arasu, G. T. (2014). Data Management and Data Analysis Interoperability In Medical Research. *IJAICT*, 1(6), 528-533.
- Kushniruk, A. (2002). Evaluation in the design of health information systems: application of approaches emerging from usability engineering. *Computers in biology and medicine*, 32(3), 141-149.

- Kushniruk, A., & Patel, V. L. (2004). Cognitive and usability engineering methods for the evaluation of clinical information systems. *Journal of biomedical informatics*, 37(1), 56-76.
- Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, 5(12), 2032-2033.
- Lagakos, S. W. (2006). Time-to-event analyses for long-term treatments-the APPROVe trial. *New England Journal of Medicine*, 355(2), 113.
- Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation*, 12(2), 1-11.
- Lee, H. (1995). Justifying database normalization: a cost/benefit model. *Information processing & management*, 31(1), 59-67.
- Leung, H. K., & Wong, P. W. (1997). A study of user acceptance tests. *Software quality journal*, 6(2), 137-149.
- Lewis, J. R. (1991). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the ASQ. *ACM SIGCHI Bulletin*, 23(1), 78-81.
- Lewis, J. R. (1992). *Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Santa Monica, CA.
- Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human - Computer Interaction*, 7(1), 57-78.
- Lewis, J. R. (2006). Usability testing. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (Vol. 12, pp. 1275-1316). New York, NY: John Wiley.
- Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, 25(1), 93-115.
- Lim, K., McNeil, M., Dickey, M., Doyle, P., & Hula, W. (2012). Conflict resolution and goal maintenance components of executive attention are impaired in persons with aphasia: evidence from the picture-word interference task. *Procedia-Social and Behavioral Sciences*, 61, 181-182.

- Little, R. J., D'agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., . . .
Murphy, S. A. (2012). The prevention and treatment of missing data in clinical trials.
New England Journal of Medicine, 367(14), 1355-1360.
- Lord, P., & Macdonald, A. (2003). *Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision*. Retrieved from
http://www.jisc.ac.uk/uploaded_documents/e-scienceReportFinal.pdf
- Love, T., & Oster, E. (2002). On the categorization of aphasic typologies: The SOAP (a test of syntactic complexity). *Journal of Psycholinguistic Research*, 31(5), 503-529.
- Lu, Z., & Su, J. (2010). Clinical data management: Current status, challenges, and future directions from industry perspectives. *Open Access J Clin Trials*, 2, 93-105.
- MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual review of psychology*, 51(1), 201-226.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods*, 1(2), 130.
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological methods*, 4(1), 84.
- MacWhinney, B., & Fromm, D. (2016). AphasiaBank as BigData. In *Seminars in speech and language* (Vol. 37, pp. 010-022). New York, NY: Thieme Medical Publishers.
- Majerus, S., Van der Linden, M., Poncelet, M., & Metz - Lutz, M. N. (2004). Can phonological and semantic short - term memory be dissociated? Further evidence from landau - kleffner syndrome. *Cognitive Neuropsychology*, 21(5), 491-512.
- Malhotra, N. K. (1987). Analyzing marketing research data with incomplete information on the dependent variable. *Journal of Marketing Research*, 24(1), 74-84.
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3), 519-530.
- Martin, N., Kohen, F., & Kalinyak-Fliszar, M. (2010). *A processing approach to the assessment of language and verbal short-term memory abilities in aphasia*. Paper presented at the Clinical Aphasiology Conference, Isle of Palms, SC.

- Martin, N., Kohen, F., Kalinyak-Fliszar, M., Soveri, A., & Laine, M. (2012). Effects of working memory load on processing of sounds and meanings of words in aphasia. *Aphasiology*, 26(3-4), 462-493.
- Martin, N., Saffran, E. M., & Dell, G. S. (2013). Recovery in Deep Dysphasia: Evidence for a Relation between Auditory—Verbal STM Capacity and Lexical Errors in Repetition. In G. Cohen, R. A. Johnston, & K. Plunkett (Eds.), *Exploring Cognition: Damaged Brains and Neural Networks: Readings in Cognitive Neuropsychology and Connectionist Modelling* (Vol. 52, pp. 329-355). New York, NY: Psychology Press.
- Martin, R. C. (1987). Articulatory and phonological deficits in short-term memory and their relation to syntactic processing. *Brain and Language*, 32(1), 159-192.
- Martin, R. C. (1990). Neuropsychological evidence on the role of short-term memory in sentence processing. In G. Vallar & T. Shallice (Eds.), *Neuropsychological impairments of short-term memory* (pp. 390-427). New York, NY: Cambridge University Press.
- Martin, R. C., & Feher, E. (1990). The consequences of reduced memory span for the comprehension of semantic versus syntactic information. *Brain and Language*, 38(1), 1-20.
- Martin, R. C., & Freedman, M. L. (2001). Short-term retention of lexical-semantic representations: Implications for speech production. *Memory*, 9(4-6), 261-280.
- Martin, R. C., & He, T. (2004). Semantic short-term memory and its role in sentence processing: A replication. *Brain and Language*, 89(1), 76-82.
- Martin, R. C., & Romani, C. (1994). Verbal working memory and sentence comprehension: A multiple-components view. *Neuropsychology*, 8(4), 506.
- Martin, R. C., Vuong, L. C., & Crowther, J. E. (2007). Sentence-level deficits in aphasia. In M. G. Gaskell & G. Altmann (Eds.), *The Oxford Handbook of Psycholinguistics* (pp. 425). USA: Oxford University Press.
- Maurer, G., Fromkin, V. A., & Cornell, T. L. (1993). Comprehension and acceptability judgments in agrammatism: Disruptions in the syntax of referential dependency. *Brain and Language*, 45(3), 340-370.
- McMullen, S. (2001). Usability testing in a library Web site redesign project. *Reference services review*, 29(1), 7-22.

- McNeil, M., Fassbinder, W., Aldhoayan, M. D., Zhou, L., Qi, M., & Pratt, S. R. (2014). Aphasic Comprehension: Conflict Resolution and Short-Term Memory. <http://cpath.him.pitt.edu/aphasia/>
- McNeil, M., Kim, A., Lim, K., Pratt, S. R., Kendall, D., Pompon, R. H., . . . Kim, H. S. (2010). *Automatic activation, interference and facilitation effects in persons with aphasia and normal adult controls on experimental CRTT-R-Stroop tasks*. Paper presented at the Clinical Aphasiology Conference, Isle of palms, SC.
- McNeil, M., Lim, K., Fassbinder, W., Dickey, M., Kendall, D., Pratt, S. R., & Kim, H. (2012). Self-Paced Reading Stroop-Interference Effects in Persons with Aphasia. *Procedia-Social and Behavioral Sciences*, 61, 64-65.
- McNeil, M., Odell, K., & Tseng, C.-H. (1991). Toward the Integration of Resource Allocation into a General Theory of Aphasia. In T. E. Prescott (Ed.), *Clinical aphasiology* (Vol. 20, pp. 21-39). Austin, TX: Pro-Ed.
- McNeil, M., & Pratt, S. R. (2001). A standard definition of aphasia: Toward a general theory of aphasia. *Aphasiology*, 15(10/11), 901-911.
- McNeil, M., & Prescott, T. E. (1978). *Revised token test*: Pro-ed.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*: Sage.
- Milner, B. (1966). Amnesia following operation on the temporal lobes. In C. W. M. Whitty & O. L. Zangwill (Eds.), *Amnesia* (pp. 109-133). London: Butterworth-Heinemann.
- Mirman, D., Strauss, T. J., Brecher, A., Walker, G. M., Sobel, P., Dell, G. S., & Schwartz, M. F. (2010). A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. *Cognitive Neuropsychology*, 27(6), 495-504.
- Mishra, C., Koudas, N., & Zuzarte, C. (2008). *Generating targeted queries for database testing*. Paper presented at the Proceedings of the 2008 ACM SIGMOD international conference on Management of data, Vancouver, Canada.
- Miyake, A., Just, M. A., & Carpenter, P. A. (1994). Working memory constraints on the resolution of lexical ambiguity: Maintaining multiple interpretations in neutral contexts. *Journal of Memory and Language*, 33(2), 175-202.

- Mohammad, S., Breß, S., & Schallehn, E. (2012). *Cloud Data Management: A Short Overview and Comparison of Current Approaches*. Paper presented at the Grundlagen von Datenbanken, Lübbenau, Germany.
- Molich, R., & Nielsen, J. (1990). Improving a human-computer dialogue. *Communications of the ACM*, 33(3), 338-348.
- Namey, E., Guest, G., Thairu, L., & Johnson, L. (2008). Data reduction techniques for large qualitative data sets. In G. Guest & K. M. MacQueen (Eds.), *Handbook for team-based qualitative research* (pp. 137-161). Lanham, MD: Rowman & Littlefield.
- Nemenyi, P. (1963). *Distribution-free multiple comparisons*. (PhD), Princeton University,
- Nielsen, J., & Molich, R. (1990). *Heuristic evaluation of user interfaces*. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, Seattle, WA, USA.
- Olmsted-Hawala, E. L., Murphy, E. D., Hawala, S., & Ashenfelter, K. T. (2010). *Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability*. Paper presented at the Proceedings of the sigchi conference on human factors in computing systems, Atlanta, GA, USA.
- Page, T. (2013). Usability of text input interfaces in smartphones. *Journal of Design Research*, 11(1), 39-56.
- Parsons, M. A., Godøy, Ø., LeDrew, E., De Bruin, T. F., Danis, B., Tomlinson, S., & Carlson, D. (2011). A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science*, 37(6), 555-569.
- Pearson, R. H. (2008). *Recommended sample size for conducting exploratory factor analysis on dichotomous data*: ProQuest.
- Peute, L. W., Spithoven, R., & WM, P. J. B. M. (2008). Usability studies on interactive health information systems; where do we stand? In S. K. Andersen (Ed.), *EHealth Beyond the Horizon: Get IT There: Proceedings of MIE2008, the XXIst International Congress of the European Federation for Medical Informatics* (pp. 327): IOS Press.
- Redman, T. C. (2008). *Data driven: profiting from your most important business asset*: Harvard Business Press.
- Riel, A. J. (1996). *Object-oriented design heuristics* (Vol. 335): Addison-Wesley Reading.

- Rigdon, E. E. (1994). Calculating degrees of freedom for a structural equation model. *Structural Equation Modeling: A Multidisciplinary Journal*, 1(3), 274-278.
- Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., & Brecher, A. (1996). The Philadelphia naming test: scoring and rationale. *Clinical aphasiology*, 24, 121-134.
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5-12 (BETA). *Journal of Statistical Software*, 48(2), 1-36.
- Royce, W. W. (1970). *Managing the development of large software systems*. Paper presented at the proceedings of IEEE WESCON, Monterey, California, USA.
- Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81): John Wiley & Sons.
- Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., & Lorenzen, W. E. (1991). *Object-oriented modeling and design* (Vol. 199): Prentice-hall Englewood Cliffs, NJ.
- Salthouse, T. A., & Meinz, E. J. (1995). Aging, inhibition, working memory, and speed. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 50(6), P297-P306.
- Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66(4), 507-514.
- Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research*: Morgan Kaufmann.
- Schmidt, F. L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement*, 31(3), 699-714.
- Simberloff, D., Barish, B., Droegemeier, K., Etter, D., Fedoroff, N., Ford, K., . . . Rossmann, M. (2005). *Long-lived digital data collections: enabling research and education in the 21st century*. Retrieved from <https://www.nsf.gov/pubs/2005/nsb0540/>
- Stabler, E. P. (1994). The finite connectivity of linguistic structure. In C. Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing* (pp. 303-336). New York, NY.: Psychology Press.
- Stangl, D. K. (2005). Bridging the gap between statistical analysis and decision making in public health research. *Statistics in medicine*, 24(4), 503-511.

- Storey, V. C. (1991). Relational database design based on the Entity-Relationship model. *Data & knowledge engineering*, 7(1), 47-83.
- Strack, F. (1992). "Order effects" in survey research: Activation and information functions of preceding questions. In N. Schwarz, S. Sudman, H. Schuman, F. Strack, R. Tourangeau, J. M. Feldman, B. A. Bickart, L. L. Martin, T. F. Harlow, D. D. L. Daamen, S. E. de Bie, J. Tarnai, D. A. Dillman, J. B. Billiet, L. Waterplas, G. Loosveldt, G. F. Bishop, T. W. Smith, H.-J. Hippler, E. Noelle-Neumann, J. A. Krosnick, E. S. Knowles, M. C. Coker, D. A. Cook, S. R. Diercks, M. E. Irwin, E. J. Lundeen, J. W. Neville, M. E. Sibicky, G. R. Salancik, J. F. Brand, A. T. Panter, J. S. Tanaka, T. R. Wellens, G. V. Bodenhausen, L. M. Moxey, A. J. Sanford, T. M. Ostrom, A. L. Betz, J. J. Skowronski, & N. M. Bradburn (Eds.), *Context effects in social and psychological research* (pp. 23-34). New York, NY: Springer.
- Streit, M., Schulz, H.-J., Lex, A., Schmalstieg, D., & Schumann, H. (2012). Model-driven design for the visual analysis of heterogeneous data. *IEEE Transactions on Visualization and Computer Graphics*, 18(6), 998-1010.
- Sung, J. E., McNeil, M., Pratt, S. R., Dickey, M. W., Fassbinder, W., Szuminsky, N., . . . Doyle, P. J. (2011). Real-time processing in reading sentence comprehension for normal adult individuals and persons with aphasia. *Aphasiology*, 25(1), 57-70.
- Sung, J. E., McNeil, M., Pratt, S. R., Dickey, M. W., Hula, W. D., Szuminsky, N., & Doyle, P. J. (2009). Verbal working memory and its relationship to sentence - level reading and listening comprehension in persons with aphasia. *Aphasiology*, 23(7-8), 1040-1052.
- Surkis, A., & Read, K. (2015). Research data management. *Journal of the Medical Library Association: JMLA*, 103(3), 154.
- Swanson, H. L., & Luxenberg, D. (2009). Short-term memory and working memory in children with blindness: Support for a domain general or domain specific system? *Child Neuropsychology*, 15(3), 280-294.
- Swets, B., Desmet, T., Hambrick, D. Z., & Ferreira, F. (2007). The role of working memory in syntactic ambiguity resolution: a psychometric approach. *Journal of experimental psychology: General*, 136(1), 64.
- Swinburn, K., Porter, G., & Howard, D. (2004). *Comprehensive aphasia test*: Psychology Press.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Allyn & Bacon/Pearson Education.

- Teorey, T. J., Yang, D., & Fry, J. P. (1986). A logical design methodology for relational databases using the extended entity-relationship model. *ACM Computing Surveys (CSUR)*, 18(2), 197-222.
- Thomasian, A., & Ryu, I. K. (1983). *A decomposition solution to the queueing network model of the centralized DBMS with static locking*. Paper presented at the Proceedings of the 1983 ACM SIGMETRICS conference on Measurement and modeling of computer systems, Minneapolis, Minnesota, USA.
- Thurstone, L. L. (1934). The vectors of mind. *Psychological review*, 41(1), 1.
- Tsichritzis, D., & Klug, A. (1978). The ANSI/X3/SPARC DBMS framework report of the study group on database management systems. *Information systems*, 3(3), 173-191.
- Vardigan, M., Heus, P., & Thomas, W. (2008). Data documentation initiative: Toward a standard for the social sciences. *International Journal of Digital Curation*, 3(1), 107-113.
- Vuong, L. C., & Martin, R. C. (2011). LIFG-based attentional control and the resolution of lexical ambiguities in sentence context. *Brain and Language*, 116(1), 22-32.
- Ware, J. H. (2003). Interpreting incomplete data in studies of diet and weight loss. *New England Journal of Medicine*, 348(21), 2136-2137. doi:10.1056/NEJMe030054
- Waters, G., Caplan, D., & Hildebrandt, N. (1991). On the structure of verbal short-term memory and its functional role in sentence comprehension: Evidence from neuropsychology. *Cognitive Neuropsychology*, 8(2), 81-126.
- Weitzel, J. R., & Kerschberg, L. (1989). Developing knowledge-based systems: reorganizing the system development life cycle. *Communications of the ACM*, 32(4), 482-488.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6), 80-83.
- Wilson, B. A., & Baddeley, A. (1993). Spontaneous recovery of impaired memory span: Does comprehension recover? *Cortex*, 29(1), 153-159.
- Wilson, W. (1990). Management of data quality in a long-term epidemiologic study: The Framingham heart study. In G. E. Liepins & V. R. R. Uppuluri (Eds.), *Data quality control: Theory and pragmatics* (Vol. 112, pp. 57). New York, NY: CRC Press.
- Wright, R. B., & Converse, S. A. (1992). Method bias and concurrent verbal protocol in software usability testing. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 36(16), 1220-1224. doi:10.1177/154193129203601608

- Yin, R. K. (2015). *Qualitative research from start to finish*: Guilford Publications.
- Yoo, B., & Donthu, N. (2001). Developing a scale to measure the perceived quality of an Internet shopping site (SITEQUAL). *Quarterly journal of electronic commerce*, 2(1), 31-45.
- Young, S. L. (1991). Increasing the Noticeability of Warnings: Effects of Pictorial, Color, Signal Icon and Border. *Proceedings of the Human Factors Society Annual Meeting*, 35(9), 580-584. doi:10.1518/107118191786754662
- Yuan, K. H., & Bentler, P. M. (2000). Three likelihood - based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological methodology*, 30(1), 165-200.
- Yuan, K. H., & Bentler, P. M. (2006). Structural Equation Modeling1. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 297): Elsevier B.V.