2021

Optimizing the Use of Normative Data in Dementia Diagnostic Evaluations

Laurie Holler

Walden University

College of Education

This is to certify that the doctoral dissertation by

Laurie Holler

has been found to be complete and satisfactory in all respects,
and that any and all revisions required by
the review committee have been made.

Review Committee
Dr. Sandra Street, Committee Chairperson, Education Faculty
Dr. Jeremy Grabbe, Committee Member, Education Faculty
Dr. Stephen Rice, University Reviewer, Education Faculty

Chief Academic Officer and Provost
Sue Subocz, Ph.D.

Walden University
2021

Abstract

Optimizing the Use of Normative Data in Dementia Diagnostic Evaluations

by

Laurie Holler

MA, Walden University, 2015

BS, Cedar Crest College, 2011

Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Psychology

Walden University

May 2021

Abstract

Cognitive tests are typically scored and interpreted using an appropriate normative reference group, often similar age individuals with similar levels of education. Psychometric testing theory presupposes that demographic correction is always beneficial, supporting the ubiquitous use of age and education correction in clinical practice. In the context of dementia, however, there is some evidence suggesting that demographic correction (specifically age correction) may reduce the sensitivity of cognitive tests to age related cognitive decline. It was hypothesized that age correction would reduce the utility of cognitive tests for detecting cognitive change in individuals with mild cognitive impairment and mild dementia due to Alzheimer's disease. This hypothesis was investigated using the National Alzheimer's Coordinating Center (NACC) database. NACC data are contributed by the NIA-funded Alzheimer's Disease Centers (ADCs). A series of hierarchical multiple linear regressions predicted the CDR® Dementia Staging Instrument Sum of Boxes Score (CDR-SB) from domain specific composite scores derived using different types of demographic correction (i.e., no correction, age correction, education correction, and both age and education correction). When looking at memory scores alone, raw scores captured more variation in the CDR-SB. However, when using a typical neuropsychological (NP) battery approach, correcting for education only produced a superior model. Findings may be used by clinicians for positive social change by recognizing that a diagnosis between normal cognitive aging and dementia is never determined by a single cut off score in clinical practice, correcting for education is an essential component when processing standardized test scores.

Optimizing the Use of Normative Data in Dementia Diagnostic Evaluations

by

Laurie Holler

MA, Walden University, 2015

BS, Cedar Crest College, 2011

Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Psychology

Walden University

May 2021

Dedication

This work is dedicated to all those fighting the battle with dementia and all those yet to be diagnosed. It is for the caregivers, both family and non-family…the true laborers of love. Also, for those on the front lines in the race for the cure, may you be blessed with the insight and discoveries that will change the trajectory of the course of neurodegenerative diseases and improve the lives of those afflicted. I would like to take the time to thank my tribe, who supported me through this journey. My children, Andrew Julian and Bryna Jade, who are the motor driving my passion to never stop evolving. Life would lack meaning and direction without you both. My partner, Tom, one of the most positive, tenacious, and driven humans I know. Thank you for never letting me give up, always feeding me, and for making sure the little details in life were taken care of so I could focus on my work. My circle of sisters, amazingly strong women, who bless me with their friendship daily. You will never know how much our conversations became the ballast when I started to capsize and healed me when I felt broken. Finally, Dr. Peter Stewart, this would not have been possible without the time and attention you gave to this project. My dissertation committee, Dr. Sandra Street and Dr. Jeremy Grabbe, along with Dr. Stephen Rice, University Reviewer, thank you for guiding my journey.

Acknowledgments

The NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by the NIA-funded ADCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P50 AG008702 (PI Scott Small, MD), P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P50 AG005134 (PI Bradley Hyman, MD, PhD), P50 AG016574 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI Robert Vassar, PhD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50 AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P50 AG005131 (PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30 AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P50 AG033514 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), P50 AG047270 (PI Stephen Strittmatter, MD, PhD).

Table of Contents

List of Tables

Chapter 1: Introduction to the Study

The overarching purpose of this study was to investigate the impact of demographic correction on the diagnostic validity of cognitive tests when differentiating between normal cognitive aging and dementia, because it was not clear how to best use normative data in dementia diagnostic evaluations. This goal was achieved by constructing a series of cognitive composite scores from tests subject to different types of demographic correction (i.e., uncorrected test scores, age corrected test scores, education corrected test scores, as well as age and education corrected test scores) and examining their relationship with a gold-standard measure used to determine the presence or absence of dementia, the CDR-SB.

The fields of education and psychology recognize standardized testing and norm-referenced scoring as a significant method of collecting meaningful information about individuals and groups. The American Educational Research Association (AERA) in a joint committee with the American Psychological Association (APA) and the National Council on Measurement in Education (NCME) asserted, “Educational and psychological testing and assessment are among the most important contributions of cognitive and behavioral sciences to our society” arguing that better decisions are made with their proper use (AERA, APA & NCME, 2014, p.1). Psychological testing plays a big part in the diagnosis of dementia, a progressive cognitive decline that is serious enough to take away a person’s independence. Since more people are surviving into old age, the period of highest risk for cognitive deficits related to dementia, the number of people that can be helped using a NP measurement perspective is increasing.

It is imperative to diagnose the condition early and with accuracy, as by the time clinical symptoms are clear enough to make a definitive diagnosis, too many neurons have already been destroyed and the damage is irreversible. There is a need to gain empirical characterizations of normal cognitive aging, mild cognitive impairment (MCI), and dementia using comprehensive NP methods (Bondi et al., 2014). Standardized testing using normative methods provides a comprehensive assessment of general cognitive functioning that identifies strengths and weaknesses in examinees with a variety of neurological conditions. While brief global cognitive assessment screening measures like the Montreal Cognitive Assessment (MoCA) and the Mini-Mental State Examination (MMSE) are used to delineate between healthy aging and dementia using cut off scores, they are limited in their sensitivity to cognitive impairment and do not reliably differentiate normal from diseased individuals until late in the course of the illness (Malek-Ahmadi et al., 2015; Roalf et al., 2013; Stephan et al., 2017).

Psychometric theory has long recognized that the individual's performance on standardized tests is strongly related to demographic variables such as age and education. For example, an 80-year-old man with 8 years of education cannot be expected to achieve the same memory performance as a 30-year-old with 20 years of education. Clearly, comparing the elderly individual's performance to that of the 20-year-old in this scenario would be inappropriate. Rather, cognitive testing is typically interpreted using demographically corrected scores that allow an "apples to apples" comparison. In the example above, the 80-year-old man's performance would be demographically corrected by comparing his performance to other similar aged individuals with similar levels of

education. Therefore, it is standard practice to demographically correct test scores for age and education to increase their sensitivity to detecting impairments and facilitate test score comparison to an appropriate normative cohort (Heaton et al., 2004; Malek-Ahmadi et al., 2015; Petersen et al., 2014; Quaranta et al., 2016; Smith & Bondi, 2013). In certain cases, demographic correction may enhance sensitivity to cognitive changes. For example, if the 20-year-old in the example above is compared to other young, highly educated individuals, then the stringent expectations placed on his performance are more likely to reveal a change has occurred (Heaton et al., 2004; Malek-Ahmadi et al., 2015; Quaranta et al., 2016; Schoenberg & Scott, 2011).

The current debate is whether the use of a normative system universally improves diagnostic accuracy in an aging population. Researchers argue that current norms may underestimate the presence of cognitive impairments in the older population because the norms are contaminated with undiagnosed cognitively impaired individuals (Hassenstab et al., 2016; Hessler et al., 2014; O'Connell & Tuokko, 2010). If individuals with undiagnosed impairments are present in "normal samples" of healthy elderly, then this decreases the performance expectations for the group as a whole and increases variability, thus obscuring the detection of change (Hassenstab et al., 2016; Hessler et al., 2014; O'Connell & Tuokko, 2010). Considering that the base rate of dementia becomes incredibly high, around 40% over age 80 (Alzheimer's Association [AA], 2018) this is a serious problem for NP research. A limited number of studies suggested that raw test scores or education-only test scores may be more sensitive to impairment (Hassenstab et al., 2016; Hessler et al., 2014; O'Connell & Tuokko, 2010). The use of partially corrected

and uncorrected test scores has not been rigorously empirically evaluated, and further research on this topic is justified (Smith & Bondi, 2013).

It is imperative to continually monitor changes and revise key documents in this rapidly evolving field to develop the utility of this body of knowledge (AERA, APA & NCME, 2014; Smith & Bondi, 2013). When a rival hypothesis challenges the status quo, the validation process must continue to obtain empirical evidence by conducting a logical analysis that evaluates the new proposition. Outcome studies using comprehensive NP testing to reveal the patterns and profiles of cognitive dysfunction are critically needed to move the field significantly forward (Bondi et al., 2014; Malek-Ahmadi et al., 2015; Smith & Bondi, 2013). Important contributions can be made to the field if the findings of this investigation support that one set of scores (a) raw scores, (b) scores that are only age corrected, (c) scores that are only education corrected, or (d) a combination of age and education correction demonstrate better predictive power of the patient's dementia severity as measured by the CDR-SB. This study addressed the specific need to differentiate normal aging and dementia using cognitive testing and various combinations of demographic corrections (normative data) to examine the strength of the test scores' relationship to the patient's clinical status.

Problem Statement

The rapid growth of our oldest population, referred to as the "silver tsunami", will cause an unprecedented challenge to our health care industry, namely our Medicare system (AA, 2018; Gill, 2015; He et al., 2016). Because age is the number one risk factor for cognitive impairment due to Alzheimer's disease (AD) and other types of dementia,

the natural increase in dementia cases will create historically high financial demands for medical expenses and ongoing personal care needs (AA, 2018; Gill, 2015; Smith & Bondi, 2013). It is estimated that costs will exceed $47 trillion for medical and long-term care expenses for all the individuals in the United States alive today that will develop dementia (AA, 2018, Zissimopoulos et al., 2014). Researchers are in search of a set of predictors to differentiate those who are experiencing normal cognitive decline due to aging from individuals who are in the process of developing dementia, an irreversible neurodegenerative process. NP methods may be ideal for that purpose.

Significant medical, emotional, and social benefits for both the individual and their families will result from our ability to differentiate between normal aging, MCI, and dementia earlier and more accurately. The presumed benefit to identifying cognitive impairment earlier is that cognitively impaired individuals are at higher-than-average risk of transitioning into dementia (AA, 2018; Dubois et al., 2016; Hessler et al., 2014; Langa & Levine, 2014; Petersen et al., 1999; Smith & Bondi, 2013). Diagnostic accuracy of early dementia is critical because secondary prevention trials, disease-modifying treatments, need to be administered early in the disease process before too much damage has been done (AA, 2018; Dubois et al., 2016; Rockwood et al., 2014; Smith & Bondi, 2013; Sperling et al., 2011; Ward et al., 2013; Zissmopoulos et al., 2014).

There is no single universally accepted test for dementia. A differential diagnosis requires a thorough workup that includes laboratory, structural neuroimaging, neurologic and clinical information. In alignment with psychometric theory, it is widely recognized that an individual’s performance on standardized tests is strongly related to demographic

variables such as age, education, sex, and ethnicity, therefore raw scores are “corrected” to increase the test’s sensitivity to impairment (Heaton et al.; 2004; Malek-Ahmadi et al., 2015; Quaranta et al., 2016; Schoenberg & Scott, 2011;). However, there is a lack of empirical evidence that these demographic corrections improve cognitive testing’s accuracy in late adulthood.

Researchers now question whether the demographically corrected normative system universally improves diagnostic accuracy (Hassenstab et al., 2016; Hessler et al., 2014; Holtzer et al, 2008; O’Connell & Tuokko, 2010). The current study was an outcome study that examined the relationship between the tests, normed different ways, and the dementia diagnostic status, with the goal of finding which way of scoring best captured the real-world changes associated with dementia. Because demographic corrections had not been rigorously empirically evaluated in late adulthood, further research into this area was warranted to determine if raw test scores may demonstrate superior sensitivity for differentiating between normal aging and dementia.

Purpose

The purpose of this quantitative study was to identify optimal normative methods for detecting pathologic cognitive change in elderly individuals. This was achieved by examining the manner in which demographic corrections for age alone, education alone, and age and education together affected the relationship between the NP test scores and a patient’s clinical dementia severity as measured by the CDR-SB. Individuals with normal cognition, mild cognitive impairment (i.e., predementia, and mild dementia due to AD served as the study population. Analyses proceeded in three steps. First, tests were scored

using four different methods of demographic correction (i.e., no correction, age correction, education correction, age and education correction). This yielded four different sets of test scores. Second, test scores were aggregated into cognitive composite scores (i.e., one set of composite scores per norming method). Finally, the relationship of each set of cognitive composite scores to dementia severity status was analyzed using four different hierarchical linear regressions to determine which normative method captured the most variance associated with the cognitive changes that accompany dementia. The R^2 values from these different regressions were compared across the different norming methods to identify which method of demographic correction was most strongly associated with the dementia severity status. Akaike's Information Criterion (AIC) was calculated and evaluated using published cutoffs frequently used to compare non-nested models (Burnham & Anderson, 2004).

Multiple regression was ideally suited for this task given that it is explicitly designed for examining the magnitude of association between a set of continuous predictor variables and a continuous outcome variable (Field, 2013). In this study, composite scores derived from tests scored using four different methods of normative correction served as the continuous predictor variables in four different hierarchical linear regressions (i.e., one for each norming method) modeling the CDR-SB, a gold-standard measure of the extent to which cognitive loss interferes with an individual's real-world functioning ([dementia status]; Burke et al., 1988). As above, the R^2 values and AIC values from the different regressions were compared across norming methods, to

determine which method best captured functional decline due to cognitive impairment in MCI and AD dementia.

Because the order of entry into a hierarchical linear regression model greatly impacts the results, the order of entry was defined *a priori*, based on prior work. This facilitated comparability across models. It was hypothesized that memory and executive function would be most strongly related to dementia severity status, followed by language and attention function. Impairments in memory and executive functioning have been established as prototypical early changes in presentations of AD (Karantzoulis & Galvin, 2011). It was also predicted that age-adjustment would decrease the magnitude of the association between cognitive test scores and the CDR-SB while education-adjustment would increase the magnitude of association between cognitive test scores and the CDR-SB. The findings of this study may aid in determining how to best use normative data in dementia diagnostic evaluations, identify specific and increase the understanding of cognitive biomarkers in a specific neurodegenerative disease.

Significance

Malek-Ahmadi et al., (2015) asserted that classification of impairment using normative data that corresponds with a specific clinical diagnosis is needed to further the field. Mortamais et al. (2017) argued that secondary prevention trials are hindered by a lack of proximal cognitive outcome markers. The results of this study may advance the understanding of the issues raised by these researchers. This project was unique because it is assumed that we should use age- and education-corrected scores in standardized testing, but it has not been rigorously empirically evaluated in our aging population

(Quaranta et al., 2016). Previous researchers who have examined the impact of age and education corrected scoring focused either on a single test score (Sliwinski et al.,1996; Sliwinski et al., 1997), or used heterogeneous composite scores that represented multiple cognitive domains (Hessler et al., 2014). One novel feature of this study was the creation of domain-specific cognitive scores for (a) memory, (b) executive function, (c) language, and (d) attention. This allowed for the precise investigation of the extent to which age and education correction affected the predictive validity of these individual cognitive domains. This is important because it is unlikely that cognitive domains are affected by demographic variables in a uniform manner. For example, we know that processing speed declines with age (Eckert, 2010) but crystallized intelligence such as vocabulary and knowledge remain relatively stable and may even improve during senescence (Harada et al., 2013). It stood to reason that age correction might enhance the accuracy with which changes in processing speed could be detected across the lifespan and is less important when measuring crystallized knowledge. No studies could be located that examined the relationship between dementia severity and demographically corrected cognitive test scores at an individual domain level.

Positive social change results from improvements that promote earlier detection of neurodegenerative diseases allowing for better treatment planning and prediction of progression into dementia. The ability to delay the progression of dementia, even just for 1 year, is shown to have significant medical, emotional, and social benefits for the individual, as well as financial benefits for our nation. Opening this window to earlier interventions gives the individual more time to seek treatment, learn compensatory

strategies, and participate in their own care and estate planning. This has the potential to save money on the historically-costly long course of this disease (AA, 2018; Dubois et al., 2016; Langa & Levine, 2014; Zissimopoulos et al., 2014).

Background

Francis Galton was credited with launching the modern psychological testing movement using systematic data he collected on different psychological processes in 1884; not only did he pioneer tests of sensory discrimination, but he also developed the use of self-report measures (rating scales) along with the statistical methods necessary for analysis of the data (Anastasi & Urbina, 1997). After a chance meeting with an American psychologist, James McKean Cattell, a former student of Wilhelm Wundt, Cattell merged Galton's new testing movement with what he learned in Wundt's experimental laboratory in Leipzig, Germany. Cattell continued his work during his tenure at the University of Pennsylvania and furthered psychological testing when he proposed a series of 10 different tests and measurements to explore the "constancy", "interdependence", and "variations of mental processes" (Cattell, 1890). Cattell also sketched out rudimentary methodology for standardized administration in an effort to gain the uniform results necessary to enable comparisons across different times and places. Cattell's work helped spread the interest in quantifying mental abilities to further psychology as a science.

The next hundred years in psychological testing research saw huge leaps forward with an understanding of measurement error, validation studies, and the development of norms for different populations (Cortina et al., 2017). Alfred Binet created the first comprehensive standardized test in 1904 to determine how students would achieve in a

classroom, and along with his advance came the development of norm-referenced scoring to estimate the position of the individual within the context of a larger population. In other words, was the person's performance "normal" for students of the same age, and by comparing students to each other he determined if a particular student was ahead-of or behind the "norm" (Binet & Simon, 1980). Researchers eagerly adopted these statistical significance and prediction models that emphasized psychometric properties and classification, and an explosion of new measures and data-analysis innovations were developed (Cortina et al., 2017).

The use of testing in the field of neuropsychology began around WWII when the assessment and recovery of brain injured soldiers created a new need in the field of testing beyond the sensory, vocational, and intelligence testing that was currently available. Ralph Reitan, a recent college graduate, was given the task of evaluating brain-injured soldiers and found a lack of publications available for reference (Grant & Heaton, 2015). The profiles and patterns that were revealed by NP testing became a key component in making a differential diagnosis, predicting the progression of, and planning the treatment for neurodegenerative diseases. This research consistently demonstrated that actuarial methods were a necessary component for comprehensive assessment (Heaton et al., 2004; Quaranta et al., 2016; Ritchie, et al., 2015; Smith & Bondi, 2013; Sutphen et al., 2015). The use of a normative system in NP testing, with its roots in the work of pioneer Alfred Binet, has since been assumed to provide more refined estimates of cognitive performance and better detection of cognitive impairments (Heaton et al., 2004; Malek-Ahmadi et al., 2015).

Normative scores are derived from the performance of a large, diverse normative sample that demographically represents the U.S. population. Yet some researchers have argued that despite their attempt at representativeness, it may be advantageous to develop norms based on specific subgroups and subpopulations to improve the utility of testing (Brown & Bryant, 1984; Chew et al., 1984; Hassenstab et al., 2016; Holtzer et al., 2008; Oosterhuis et al., 2016; Svinicki & Tombari, 1981). The aging population may be one specific subgroup that warrants the use of an alternative method because of the presence of undiagnosed cognitive impairments in the normative reference population. Researchers forwarded a proposal that demographic corrections may not improve diagnostic accuracy in the service of diagnosing cognitive impairment in an older population because the norms are tainted with individuals who may be in early stages of a degenerative cognitive decline, which compromises the mean performance and increases the variability in the normative sample (Hassenstab et al., 2016; Hessler et al., 2014; Holtzer et al., 2008; O'Connell & Tuokko, 2010; Wyman-Chick et al., 2018). Yet other researchers like Quaranta et al. (2016) failed to support the hypothesis that raw scores were superior to age-corrected scores and normative scoring remains standard practice. Wyman-Chick et al. (2018) argued that the selection of the normative comparison group greatly impacts both research and clinical interpretations of cognitive data. Yet no studies could provide rigorously validated impact of demographic corrections on the diagnostic accuracy of cognitive testing in individual cognitive domains when employed in dementia diagnostic evaluations. The current study focused on the accuracy of raw scores versus

demographically corrected scores using an outcome study with a large data set gathered through the NACC database.

Framework

This study was grounded in testing and measurement theory that governs how psychological constructs are measured and compared between individuals and groups. Professor James McKean Cattell (1890) at the University of Pennsylvania wrote about the benefits of standardized psychological testing, arguing that efforts:

> Would be of considerable scientific value in discovering the constancy of mental processes, their interdependence, and their variation under different circumstances…the scientific and practical value of such tests would be much increased should a uniform system be adopted, so that determination made at different times and places could be compared and combined. (p. 347)

Frenchman Alfred Binet furthered the utility of testing and measurement in an educational setting when he developed the first comprehensive standardized test in 1904 as a method of classifying which students could or could not achieve in the classroom. Binet never claimed that his scale could measure intelligence like a "ruler can measure a linear surface", but instead he claimed to provide "…a classification, a hierarchy among diverse intelligences; and for the necessities of practice this classification is equivalent to a measure" (Binet & Simon, 1980, p. 41). Binet developed what is now known as norm-referenced scoring to allow an estimation of the individual's position within the context

of a larger population. Thus, norms became a fact of life in educational and psychological assessment.

It was quickly realized that much of variance between test scores could be accounted for by a single demographic variable, and mounting evidence over the evolution of standardized testing showed that even more variances could be accounted for with a combination of multiple demographic variables (Barona et al., 1984; Karzmark et al., 1985; Wilson et al., 1978). The addition of demographic corrections became almost mandatory in psychological testing because these factors are relevant in an individual's diagnosis (Quaranta et al., 2016). In 2004, Heaton et al. published a widely adopted comprehensive set of demographically adjusted NP norms for more than 50 commonly used measure for adults ages 20- to 85-years old which helped solidify the use of a normative system in the field of NP testing at all ages.

Using the established framework, the predictor variables, the demographically corrected scores, should better predict the outcome variable (the dementia severity rating as measured by the CDR-SB) because demographically corrected scores are believed to improve diagnostic accuracy. But researchers argue different reasons that this may not be true. Manuals for standardized testing give national norms, but the utility of norms is questioned in an aging demographic. First and foremost, individuals in the normative sample population may already be transitioning into dementia and contaminating the norms by lowering the mean performance and increasing the variability in cross-sectional normative samples (Hassenstab et al., 2016; Hessler et al., 2014; O'Connell & Tuokko, 2010). Similarly, many individuals in an aging normative sample may be prescribed brain

impairing maintenance medications that also affect their performance and increase the variability within the sample. Second, is the Flynn effect; norms become less accurate when too much time passes since publication due to changes in demographics, socioeconomics, and cultural factors that modify the reference population. Third, norms become more forgiving and tolerant of errors as age increases, possibly underestimating the risk of dementia in our oldest population. The current study hypothesized that age-corrected scores would have the lowest magnitude of association with dementia severity level as measured by the CDR-SB, while education-corrected scores would have a higher magnitude of association with dementia severity level as measured by the CDR-SB.

Research Questions

The overarching research question was: how does demographic correction for age and education affect the relationship between cognitive tests and functional deterioration due to cognitive impairment? Some researchers proposed that raw test scores may have superior sensitivity for detecting cognitive changes accompanying dementia in an aging population, over the standard practice of demographically correcting the scores for age and education level (Hessler et al, 2014; Holtzer et al., 2008; O'Connell & Tuokko, 2010). It is not clear how to best use normative data in dementia diagnostic evaluations using cognitive testing and there is no consensus in the literature making a rigorous empirical investigation using an outcome study warranted. Tests will be scored 4 different ways (a) corrected for gender, age, and education (GEA); (b) corrected for gender and age (GA); (c) corrected for gender and education level (GE); and (d) raw test scores corrected for gender only (G). Cognitive composite scores representing memory

and learning, executive functioning, language, and attention were built for each norming method. Then one multiple linear regression equation per norming method (four total) was built. These models were then compared by examining the R^2 and AIC values to determine which model "best" captured functional decline due to cognitive loss. The multiple linear regression equations outlined below were used to determine which set of scores had the strongest relationship with the patient's clinically determined dementia severity level as measured by the CDR-SB.

1. The relationship between the CDR-SB and cognitive composites built from GEA corrected data.
2. The relationship between the CDR-SB and cognitive composites built from GA corrected data.
3. The relationship between the CDR-SB and cognitive composites built from GE corrected data.
4. The relationship between the CDR-SB and cognitive composites built from "raw" or G corrected test data.

The novel feature of the current study was the creation of domain-specific cognitive composite scores. There were 4 steps to each regression equation because each composite score, memory, executive function, language, and attention, was entered into the equation in a hierarchical fashion allowing for the precise investigation of the extent to which the demographic corrections affected the predictive validity of each individual cognitive domain score.

RQ1: How do age and education correction affect the relationship between cognitive tests and functional deterioration due to cognitive impairment (i.e., dementia severity)?

H_1: Age and education correction increase the extent to which NP tests capture functional decline due to cognitive loss in dementia. This will be tested by comparing R^2 and AIC values from regression equation 1 with regression equation 4. For example, if the R^2 value in equation 1 is moderate ($R^2 > .25$) and the R^2 value in equation 4 is strong ($R^2 > .4$) as per conventions from Cohen (1988), then the null hypothesis can be rejected. Alternatively, if the AIC of model 4 is less than the AIC value of model 1 by 4 or more (Burnham & Anderson, 2016), then the null hypothesis can be rejected.

$H_0$1: Age and education correction decrease or have no effect on the extent to which NP tests are able to capture functional decline due to cognitive loss in dementia.

Sub RQ2: How does age correction affect the relationship between cognitive tests and functional deterioration due to cognitive impairment?

H_2: Age correction decreases the extent to which NP tests are able to capture functional decline due to cognitive loss in dementia. This will be tested by comparing R^2 and AIC values from regression equation 4 with regression equation 2. For example, if the R^2 value in equation 2 is moderate and the R^2 value in equation 4 is strong, then the null hypothesis can be rejected. Alternatively, if the AIC of model 4 is less than the AIC value of model 2 by 4 or more, then the null hypothesis can be rejected.

$H_0$2: Age correction increases or has no effect on the extent to which NP tests are able to capture functional decline due to cognitive loss in dementia.

Sub RQ3: How does education correction affect the relationship between cognitive tests and functional deterioration due to cognitive impairment?

H_3: Education correction increases the extent to which NP tests are able to capture functional decline due to cognitive loss in dementia. This will be tested by comparing R^2 and AIC values from regression equation 4 with regression equation 3. For example, if the R^2 value in equation 4 is moderate and the R^2 value in equation 3 is strong, then the null hypothesis can be rejected. Alternately, if the AIC of model 3 is less than the AIC value of model 4 by 4 or more, then the null hypothesis can be rejected.

$H_0$3: Education correction decreases or has no effect on the extent to which NP tests are able to capture functional decline due to cognitive loss in dementia.

Optimal demographic correction (as determined via the analyses above) will result in better diagnostic performance when differentiating normal controls from individuals with dementia. Comparison of the R^2 values of the various regression equations presented above will clarify the extent to which various types of demographic correction (e.g., age, education, age & education) influence the ability of cognitive tests to detect meaningful variation in dementia severity. While the above tests are not associated with significance levels, R^2 values and AIC values are used commonly to compare non-nested models and

allow the important research questions above to be answered quantitatively albeit not with a given significance level as in a traditional null hypothesis testing approach.

Nature of the Study

This was a quantitative study of concurrent case-referent design for evaluating test-criterion relationships. This study looked at the relationship between the cognitive tests, normed different ways, and the clinical dementia severity rating (CDR-SB) using hierarchical multiple linear regression. Prior to modeling, the data was described by calculating descriptive statistics including the mean, standard deviation, and a range of all study variables. The intercorrelations of the cognitive test scores, age, education, and the CDR-SB was calculated for each of the 4 different norming methods using bivariate Pearson correlations (one correlation matrix per norming method). The correlations provided a direct measure of the strength of the relationship between all quantitative study variables and aided in the interpretation of the hierarchical multiple linear regression analyses for testing the primary study hypothesis, help detect suppression, and served as a measure of variable importance (Nathans et al., 2012). Following descriptive statistics and analysis of intercorrelations between tests, the primary study hypotheses was tested using least squares hierarchical multiple linear regression. Each set of cognitive composites, one per norming method, was regressed on the CDR-SB. Because there are no universally accepted statistical tests by which to compare these non-nested models, these differences were evaluated quantitatively using R^2 values and AIC values, but were not be associated with a significance level. This told us the extent to which these

different methods of scoring captured clinically meaningful variations in dementia severity status and the extent to which they differed from one another.

Hierarchical multiple regression was chosen for this analysis because it allowed us to fit a model to the data that enabled the prediction of the outcome variable, the CDR-SB, from a number of different independent variables, the cognitive tests for each of the 4 cognitive domains scored 4 different ways. This technique is appropriate when examining the magnitude of association between a set of continuous predictor variables and a continuous outcome variable (Field, 2013). The comparison of the R^2 and AIC values from the resulting models facilitated an exploration of the extent to which different scoring methods differed in their ability to capture meaningful variability in dementia severity status. Identifying optimal norming methods facilitates greater diagnostic accuracy in the context of dementia.

Sources of Data

The University of Washington's NACC is funded by the National Institute on Aging and maintains a valuable resource, a cumulative database with which researchers can collaborate (NACC, 2010). The NACC shares all data, providing an excellent resource for investigating cognitive aging and dementia in a well-defined cohort. This data, the Universal Data Set (UDS), will be obtained from the NACC who took the first steps to standardize data collection across the ADCs in 1999 in an effort to advance better research hypotheses; by the end of 2016 data had been collected from 34,748 participants (Weintraub et al., 2018a). Participants with normal cognition, MCI, and various etiologies of dementia are recruited and followed annually. Data collection using Version

3 began in March 2015 as part of an ongoing effort to produce a uniform data set with an updated NP battery allowing research institutions to collaborate using freely available standardized instruments (Besser et al., 2018). The current study used the updated 3rd version of the NACC's UDS Neuropsychological Assessment Battery (UDS-3 NAB) which included measures of dementia severity, verbal and nonverbal memory, executive function, language, and attention (Besser et al., 2018). The instruments produced both raw test scores and age- and education-corrected scores and were modeled as a function of the patient's dementia severity level as measured by the CDR-SB. All identifying information of the subjects was scrubbed prior to the dissemination of the data to assure complete confidentiality.

Summary

Since the first standardized psychological testing was used to quantify an individual's performance and compare it across groups, norm-referenced scoring was used to make raw scores more relevant and useful. Researchers draw conclusions about an individual's performance by comparing scores to national norms that allow them to find if the person's performance is "typical" and determine whether their functioning is at, ahead, or behind the norm. The presence of appropriate national norms is necessary for these conclusions. A recent proposal questions the utility of age- and education-corrected scores in late adulthood, postulating that these norms may be contaminated with undiagnosed individuals in preclinical or prodromal stages of neurodegenerative diseases, making the tests less sensitive to the detection of cognitive impairments in our aging population.

AD and other dementias are some of the costliest conditions in our society and predicted to reach crisis levels as our oldest population grows faster than any other demographic and threatens to place great financial strain on our health care system. Secondary prevention trials are handicapped by a lack of cognitive outcome markers and the ability to classify impairment using standard normative data. The current search for a set of predictors, including cognitive biomarkers, that distinguish individuals who are experiencing the effects of normal aging from those in the process of developing neurodegenerative diseases gives rise to the question concerning the utility of raw scores versus demographically-corrected scores in norm-referenced cognitive testing of an older population. The current study aimed to explore the strength of the relationship between tests scored different ways and dementia severity with the ultimate goal of finding which way of scoring was most diagnostically accurate.

Chapter 2: Literature Review

The literature review in this chapter justifies the need for additional research to examine whether the current standard practice of age- and education-correcting scores in NP testing are best practice when attempting to differentiate between normal cognitive aging and dementia in an aging population. Some researchers championed the proposal that national norms are polluted with undiagnosed individuals in the early stages of dementia, decreasing the tests' ability to detect cognitive impairment (Hessler et al., 2014; Holtzer et al, 2008; O'Connell & Tuokko, 2010; Wyman-Chick et al., 2018). But others failed to support this hypothesis and current practice still adheres to demographic correction (Malek-Ahmadi et al., 2015; Quaranta et al., 2016). This study addressed a current debate in the literature over which set of data, normative or raw scores, best predict the patient's clinical level of dementia using data obtained from NACC's ADCs. The current study may further the utility of NP testing in the service of earlier and more accurate detection of neurodegenerative diseases because optimal demographic correction will result in better diagnostic performance when differentiating normal controls from individuals with dementia.

This chapter starts by addressing the major social problem, the public health burden of dementia. It then covers an overview of dementia including the concept of MCI, and the criteria for dementia diagnosis. From there, dementia is discussed starting with the most common etiology, Alzheimer's disease (AD), followed by how the testing profiles of the next common causes of dementia compare and contrast, including vascular dementia, Lewy body disease, frontotemporal dementia, and Parkinson's disease. A

history of psychological and educational testing is explored and connected to the role of NP testing in a clinical setting with an emphasis on norm referenced interpretation in NP assessment, methods of norming, and norm referenced interpretation specifically for dementia diagnostic evaluations.

The literature search included articles electronically accessed through Walden University library's databases; Academic Search Complete; Google Scholar; Mental Measurements Yearbook with Tests in Print; Proquest; psycTESTS; PubMed; SAGE; Taylor and Francis Online; Thoreau; World Health Organization (WHO), and open access articles from PubMed.gov and NIH.gov. Other resources such as UpToDate and Elsevier were accessed through an alternative institution's library databases. Search terms included; Alzheimer's disease; dementia; dementia assessment; diagnostic accuracy and dementia assessment; diagnostic accuracy and memory impairment; diagnostic accuracy and mild cognitive impairment; NP assessment and dementia; demographic correction and neuropsychology and dementia; and demographic correction of NP test scores. Multiple books, both in print and electronically, were also accessed and reviewed for relevant information. The search had a rough scope over the last decade with an emphasis on the last 5 years of research, even though it was necessary to look into the annals of history when tracing the origins of psychological and educational testing theory.

The Public Health Burden of Dementia

Our healthcare system is about to face an unprecedented challenge as the growth of our oldest population, the baby boom generation, reaches the age of high risk for the development of neurodegenerative diseases like AD and other types of dementia (AA,

2018; Gill, 2015; Livingston et al., 2017; Smith & Bondi, 2013; WHO, 2018;). As the number of older Americans increases rapidly due to medical advances and better environmental conditions, so will the number of new cases of dementia. Significant financial, medical, and emotional benefits will result from earlier intervention because disease-modifying and psychosocial interventions are most effective early in the disease process; delaying the onset of the disease, even just for one year, has significant benefits due to the long duration of the illness prior to death that carries a heavy emotional and financial burden. (AA, 2018; Dubois et al., 2016; Hurd et al., 2013; Langa & Levine, 2014; Rockwood et al., 2014; Smith & Bondi, 2013; Sperling et al., 2011; Ward et al., 2013; Wei-Hong et al., 2017; Zissmopoulos et al., 2014).

Lifetime expenditures for an individual with dementia are roughly $341,840, three times more than the cost of care for people without dementia for the same age group (AA, 2018). Medicare and Medicaid cover 67% or $186 billion of the total $277 billion in these costs, deeming this issue a major social problem (AA, 2018). Payments in every category; primary care physicians; specialists; lab services; medication; emergency room visits; inpatient hospital stays; skilled nursing facilities; and hospice care are higher for those with dementia, 23 times greater than those who remain dementia free (AA, 2018). Even with the financial assistance from Medicare and private insurance, out-of-pocket expenses are an additional burden to an already emotionally stressed family dealing with the dementia diagnosis of a loved one.

In addition to medical expenses and lost economic value for unpaid care, dementia caregivers reported more physical and mental health issues than the general

population (AA, 2018; Ma et al., 2018; Roth et al., 2015; Solway, 2017). Caregivers spend an average of 21.9 unpaid hours a week caring for a loved one with dementia (AA, 2018). This often exacerbates their own health issues; increases emotional stress and depression; and depletes income and finances due to disruption in employment and additional personal health care expenses (AA, 2018; Ma et al., 2018; Roth et al., 2015; Solway, 2017). This caregiver strain was even shown to increase the caregiver's risk of death (Roth et al., 2015). While most caregivers reported caring for their loved one was rewarding, they also acknowledged the role is highly stressful (AA, 2018; Solway, 2017). This is such a prominent issue that the National Academies of Sciences, Engineering, and Medicine (2016) released a report entitled *Families Caring for an Aging America* which focused on national health care reform efforts that recognize the role of family members and encourage health care providers to deliver evidence-based services to both care recipients and their caregivers.

Other benefits for the individual include early intervention programs such as cognitive rehabilitation that maximizes reserved cognitive resources by teaching compensatory strategies, behavioral interventions like diet and exercise that may increase quality of life and prolong independence, and estate and care planning while the person can still participate (AA, 2018; Livingston et al., 2017). Changing the trajectory of AD and other neurodegenerative diseases has the potential to improve the lives of patients, their families, and society as a whole.

The WHO (2018) called to prioritize dementia as a global health issue and reported the prevalence and financial burdens of people living with syndromes of

cognitive impairment worldwide, proposing the need for public policies to address the impending crisis:

> Almost 9.9 million people develop dementia each year, the majority (63%) of whom reside in low- and middle- income countries. Dementia currently affects approximately 50 million people worldwide; a number that is projected to grow to 82 million by 2030 and 152 million by 2050. It is the second largest cause of disability for individuals aged 70 years and older, and the seventh leading cause of death. Dementia imposes an estimated economic cost of approximately US $818 billion per year globally – or 1.1% of global gross domestic product. Left unaddressed, dementia could represent a significant barrier to social and economic development. (p. 6)

Research focused on earlier and more accurate diagnoses is part of the formula leading to improvements in biomedical, psychological, and social interventions that have the potential to reduce the number of new cases by 10-20% because of their potential to ease the physical, psychosocial, and financial hardships for individuals, their families, and developing nations (AA, 2018; WHO, 2018).

Overview of Dementia

Normal or Abnormal Cognitive Aging?

The life-span perspective classifies human development from conception to death, encompassing all stages and phases of growth, maturity, and aging. Viewing aging through this lens allows for a model that avoids pejorative or abnormal terms when psychological processes such as cognition change during the maturation process. It is

clear that adults peak cognitively between the ages of 20 and 40, but it is also established that fluid intelligence, efficient functioning of the central nervous system, declines steadily over adulthood beginning at age 35 or 40 (Boyd & Bee, 2019). Given ample time, older adults will still come up with the adequate answers, just not as quickly as younger adults. Using a life-span perspective, cognitive changes due to aging are not seen as an abnormal condition, but rather a normal developmental stage of life. The difference between normal cognitive aging and abnormal cognitive aging is an impairment that is distinct from normal aging, not typical of age-matched peers, and objectively measurable using NP testing measures.

Cognitive changes are measured by NP testing, tests designed to detect quantitative or qualitative changes in the main cognitive domains of memory, executive function, language, attention, processing speed, and visuospatial skills. When an individual scores > 1 SD below the age corrected normative mean on a testing measure in a single domain with no interference in Instrumental Activities of Daily Living (IADLs; e.g., driving, managing one's finances, self-managing medications, using the community, etc.) a diagnosis of MCI may be made. When the individual scores > 1 SD below the age corrected normative mean in multiple domains leading to difficulty with IADLs then a diagnosis of dementia may be considered. There is a general acknowledgment that preclinical dementia-related neuropathology is present in normally aging individuals prior to any measurable cognitive decline (Mortamais et al., 2017; Ritchie et al., 2015; Rockwood et al., 2014; Smith & Bondi, 2013; Sperling et al., 2011; Sutphen et al., 2015 Ward et al., 2013; Wei-Hong et al., 2017). This grey area, the phase between normal

cognitive functioning and clinical dementia, is seen as the most promising period for disease-modifying interventions that have the potential to alter the trajectory of the disease and has come to be widely accepted as the concept of MCI.

Our ability to predict which patients with MCI will remain stable, typical of normal aging, versus which will convert to dementia and continue to decline, is a major goal in current research. NP testing plays an important role in the multidisciplinary search for answers (Bondi et al., 2014; Smith & Bondi, 2013). NP testing becomes a key front-line component in the detection of preclinical dementia because, unlike a lumber puncture, it is non-invasive, does not require expensive medical equipment like brain imaging, and is easy to administer in a variety of clinical settings.

The Evolution of the Concept of MCI

Kral's (1962) seminal work delineated the difference between "benign" (normal) and "malignant" (pathological) aging, many terms have been proposed to describe the concept of the "not-normal but not-demented clinical state", and MCI has clearly gained widest acceptance (Smith & Bondi, 2013, pp. 72-73). Petersen et al., (1995) adopted the term MCI as a diagnostic entity to reflect the earliest objectively measurable deficits in cognition when it is no longer normal relative to expectations for age, but the individuals can still live and function independently. The first guidelines for MCI proposed by Petersen et al. (1999) recommended that general criteria include a non-demented individual (does not meet the DSM criteria for a dementia syndrome) with generally intact cognition, preserved Activities of Daily Living (ADLs; e.g., personal hygiene, continence management, dressing, feeding, and ambulating) and minimal impairment of

IADLs, but experiencing subjective memory complaints that can be objectively measured. Smith and Bondi (2013) explained the concept of MCI and its meaning for the clinician:

> MCI constitutes that level of cognitive function wherein low-functioning normal older persons and high functioning dementia patients cannot be reliably distinguished. If all persons labeled as MCI are conceived of as belonging in either a normal population, not destined to develop dementia, or from a population that is developing dementia, then MCI can be thought of not as a condition present in the patient, but rather as a state of uncertainty in the clinician. (p. 73)

MCI was incorporated into the DSM-5 as a mild neurocognitive disorder and is central in the field because it is considered a significant risk factor for the subsequent development of dementia. While a percentage of people with MCI remain stable, and a smaller percentage may recover completely, estimates varied from 43% to 83% conversion rate to dementia depending on the methodology used; but it is agreed that these patients are at higher risk for developing dementia (Bondi et al., 2014; Mazaheri et al, 2018; Mitchell & Shiri-Feshki, 2009; Petersen et al., 1999; Petersen et al., 2013; Petersen et al., 2015; Smith & Bondi, 2013; Ward et al. 2013; Wimblad et al., 2004). MCI is also important because it is the window when the least damage has occurred making it an ideal target for interventions. Thus, the concept defined as MCI has become a primary focus for research in neurodegenerative diseases.

Early in the MCI research, total learning score, also referred to as immediate recall, emerged as the single most sensitive and specific measure for distinguishing MCI from normal aging; researchers found the addition of a delayed recall measure enhanced classification accuracy and improved prediction of progression to AD dementia (Bondi et al., 2014; Smith & Bondi, 2013). As the conception of MCI evolved, it became recognized as a pathologically heterogeneous disorder and the concept was broadened to include deficits in other domains besides just memory (Smith & Bondi, 2013; Wimblad et al., 2004).

A multidisciplinary consensus conference in 2003 expanded MCI into three subtypes, amnestic, multiple domain, and single nonmemory domain (e.g., language or visuospatial) and listed multiple possible etiologies; degenerative; vascular; metabolic; traumatic; psychiatric; and "others" (Wimblad et al., 2004). This was an important revision in the concept as subcortical dementias like Parkinson's and Huntington's diseases may leave memory relatively intact in the early stages while the first measurable deficits appear as compromised attention and processing speed (Smith & Bondi, 2013; Wimblad et al, 2004). Comprehensive NP testing is the best way to classify the specific subtype of MCI (Bondi et al., 2014; Wimblad et al., 2004).

In 2013, the DSM-5 (APA, 2013) classified MCI as Mild Neurocognitive Disorder and established the following criteria for diagnosis:

A. Evidence of modest cognitive decline from a previous level of performance in one or more cognitive domains (complex attention,

executive function, learning and memory, language, perceptual-motor, or social cognition) based on:

1. Concern of the individual, a knowledgeable informant, or the clinician that there has been a mild decline in cognitive function: and
2. A modest impairment in cognitive performance, preferably documented by standardized neuropsychological testing or, in its absence, another quantified clinical assessment.

B. The cognitive deficits do not interfere with capacity for independence in everyday activities (i.e., complex instrumental activities of daily living such as paying bills or managing medications are preserved, but greater effort, compensatory strategies, or accommodation may be required).

C. The cognitive deficits do not occur exclusively in the context of a delirium.

D. The cognitive deficits are not better explained by another mental disorder (e.g., major depressive disorder, schizophrenia). (p. 605)

When testing outcome shows that an individual's memory is significantly lower than age expectations, but other domains (attention, language, visuospatial skills, and executive functions) remain intact, amnestic MCI is the preferred classification. If mild deficits are found in a number of different domains, multidomain MCI (with or without a memory component) is more appropriate. When one nonmemory domain is impaired,

such as visuospatial skills, then single nonmemory domain MCI is the most applicable. After a diagnosis of MCI is made and classified, then the clinician must attempt to determine the etiology of the impairment and plan for monitoring or treatment. The differential diagnosis of a cognitive disorder requires an extensive workup given the serious consequences of progressive degeneration and impending disability. NP testing plays an important role in the push for earlier and more accurate diagnoses that will allow for disease-modifying treatments to be developed, tested, and utilized successfully (Bondi et al., 2014; Mortamais et al., 2017; Smith & Bondi, 2013).

The difference between MCI and dementia is the severity and prognosis. Once the criteria are met for a dementia diagnosis, there is progression over time. MCI, on the other hand, while considered a significant risk for future dementia does not always progress. Current research estimates the majority of patients with MCI transition to dementia within 5 years of the MCI diagnosis (Mazaheri et al., 2018). Differential diagnosis demands a comprehensive clinical assessment that includes a full neurological exam, brain imaging studies, and NP testing (Bondi et al., 2014; Mazaheri et al., 2018; Mortamais et al., 2017; Petersen et al., 2014; Smith & Bondi, 2013). The early and accurate diagnosis of MCI is increasingly important as patients are presenting concerns to their primary care physicians earlier, and secondary prevention trials seek to intervene sooner in the disease process to limit permanent damage to the brain.

Dementia Criteria

The term dementia is customary in most settings, but the DSM-V reclassified dementia as major neurocognitive disorder (APA, 2013). Some of the earliest research on

aging defined two types of changes related to cognition and behavior, “benign” and “malignant” (Kral, 1962). Benign changes were typical, developmental changes associated with aging and unrelated to diseased brain tissue, in other words, normal aging. Malignant changes were histopathological brain changes that were progressive in nature. As research continued, the focus became the ability to distinguish a normally aging individual, who was worried about their memory function enough to complain to their primary care provider, from a malignant or neurodegenerative process (Smith & Bondi, 2013). Currently, the DSM-V sets the consensus diagnostic criteria for a major neurocognitive disorder as:

A. Evidence of significant cognitive decline from a previous level of performance in one or more cognitive domains (complex attention, executive function, learning and memory, language, perceptual-motor, or social cognition) based on:
    1. Concern of the individual, a knowledgeable informant, or the clinician that there has been a significant decline in cognitive function; and
    2. A substantial impairment in cognitive impairment in cognitive performance, preferably documented by standardized neuropsychological testing or, in its absence, another quantified clinical assessment.
B. The cognitive deficits interfere with independence in everyday activities (i.e., at a minimum, requiring assistance with complex

instrumental activities of daily living such as paying bills or managing medications).

C. The cognitive deficits do not occur exclusively in the context of a delirium.

D. The cognitive deficits are not better explained by another mental disorder. (APA, 2013, pp. 602-603)

The DSM-V requires the clinician to specify the etiology of the major neurocognitive disorder, a discussion of all the causes is beyond the scope of this paper, but the most prevalent causes are addressed in the following subsections. The most common cause of dementia is AD. The second most common cause is vascular disease, followed by Lewy body disease, and frontotemporal lobar degeneration (Ramirez-Gomez et al., 2017; Smith & Bondi, 2013). While AD, Lewy body disease, and frontotemporal dementia are all classified as neurodegenerative diseases, vascular disease is more diverse and does not always conform to the same standards. NP profiles and patterns aid in distinguishing these underlying pathologies (Ramirez-Gomez et al., 2017; Smith & Bondi, 2013; Stephan et al., 2017). This is why the number of people that can be helped using a NP perspective is increasing.

Alzheimer's Disease

Over a century ago Alois Alzheimer, considered the father of neuropathology, was the first to describe a patient with the progressive form of dementia that now bears his name (Möller & Graeber, 1998). AD is the most common cause of dementia accounting for up to 80% of all cases (AA, 2018; Sutphen et al., 2015). In the United

States, one person develops the disease every 65 seconds (AA, 2018). Current estimates support that 10% of people age 65 have AD, with the prevalence of the disease increasing exponentially with age: three percent of people age 65-74, 17% of people age 75-84, and over 40% of people age 85 and older (AA, 2018; Hebert et al., 2013).

The first individuals to be born into the baby boom generation turned 72 in 2018, placing them at high risk for neurodegenerative disorders (AA, 2018). The rapid increase of our oldest population over the coming decades will stress our health care system as demand for medical care and long-term services will increase, placing a huge burden on Medicare that will cause a major economic ripple on our Nation's budget (Barnett, et al., 2014; Dubois et al., 2016; Hurd et al., 2013; Rockwood et al., 2014; Zissmopoulos et al., 2014). Estimates claim a $935 billion in savings that can be realized over the 10-year period from 2026-2035 with an overall $7.9 trillion savings for the current U. S. population (AA, 2015). Our Nation will benefit from earlier and more accurate detection of neurodegenerative diseases like Alzheimer's because of the potential to improve the lives of the millions of individuals yet to be diagnosed.

Prior to updated guidelines in 2011, a formal diagnosis of AD required that an individual already exhibit *significant* problems with learning, thinking, or memory. The seminal work of Braak and Braak (1991) changed how AD was viewed based on the discovery of neurofibrillary tangles (a known biomarker of AD) in people as young as 30. This sparked a surge of research that led to Jack et al.'s (2010) continuum model of AD that begins with a preclinical period, decades before symptoms appear but when biological changes are already taking place in the central nervous system, moving into

MCI where cognitive decline can be objectively measured, and finally full-blown Alzheimer's dementia. This revised model of the Alzheimer's trajectory continues to guide most of the current direction in research and practice with the hopes that disease-modifying interventions will be developed and utilized during the earliest stages to change the course of the disease prior to total dementia setting in (Jack et al., 2013).

It is certain that a preclinical stage of AD begins decades before symptoms appear, when biological changes take place, but the individual remains asymptomatic (AA, 2018; Jack et al., 2013; Jack et al., 2015; Mazaheri, 2018; Mortamais et al., 2017; Ritchie et al., 2015; Sutphen et al., 2015). This has led some researchers to categorize AD as a disease of midlife rather than of old age (Ritchie et al., 2015; Sutphen et al., 2015). Jack et al. (2013) proposed the main AD biomarkers change in a temporally ordered manner; starting with an abnormal accumulation of amyloid β protein (Aβ) as plaques, and hyper phosphorylated tau protein as neurofibrillary tangles that can be assessed by measures of cerebral spinal fluid (CSF) Aβ and tau. This is followed by the biomarkers of neurodegeneration indicated by brain imaging, hypo metabolism on fluorodeoxyglucose (FDG) PET, and structural MRI, that finally lead into the clinical symptoms and measurable cognitive decline (cognitive biomarkers). This insidious onset is included in the diagnostic criteria for AD. The current model assumes "the maximum rate of change moves sequentially from one biomarker class to the next, and as the disease progresses all biomarkers become progressively more abnormal simultaneously…at rates that change over time in an ordered manner" (Jack et al., 2013, p. 207). The rate of progressive cognitive impairment is "loosely coupled" with the amount CSF Aβ, but "closely

coupled" with imaging neurodegenerative biomarkers (Jack et al., 2013). It was also noted that the time needed to travel the course of the disease varies among individuals because it is mediated by baseline differences in brain plasticity and cognitive reserve, as well as the presence of other pathophysiology like cerebrovascular disease or Lewy bodies, which often co-occur with AD and contribute to an individual's variation and presentation (Jack et al., 2013; Karantozoulis & Galvin, 2011; Smith & Bondi, 2013).

In the Alzheimer's continuum, MCI is thought of as the period that reveals the first objectively measurable cognitive deficits due to neuropathological brain changes that occur in the course of the disease, and also where therapeutic trials designed to prevent cognitive decline are most useful (AA, 2018; Livingston et al., 2017; Mortamais et al., 2017; Sutphen et al., 2015). Once brain changes are so significant that cognition and physical functioning decline, risk reduction and medical interventions are of little value (AA, 2018; Peall & Robertson, 2015). The problem remains in the obvious detail that by definition of being asymptomatic, the preclinical phase avoids detection using current cognitive measures. Empirically validated innovations for detection using a NP testing perspective are critically needed.

Our ability to diagnose individuals earlier and more accurately for the purpose of testing and utilizing disease-modifying treatments are the key to changing the trajectory of neurodegenerative diseases. Mortamais et al. (2017) asserted:

> The design of secondary prevention trials targeting the preclinical period has thus been handicapped up to this point by that lack of proximal cognitive outcome markers. The cognitive tests currently used to describe

> AD, having been largely derived from comparisons of persons with and without dementia, are by definition inappropriate for preclinical studies. Such early cognitive changes, if they exist, are likely to be subtle, requiring highly sensitive tests that target specific brain regions affected early in the disease process. (p. 469)

AD is a prototypical cortical dementia with the most salient cognitive biomarkers markers being episodic memory impairments (learning and retention measures), that when coupled with the presence of a molecular biomarkers, like CSF Aβ and tau or FDG PET, clinicians can be fairly certain the individual will progress into AD dementia (Mortamais et al., 2017; Smith & Bondi, 2013). Episodic memory impairment is the most prominent predictor of dementia, but aphasia and apraxia are also common features and can be measured by lower performance in verbal fluency, processing speed, and fluid reasoning (Karantzoulis & Galvin, 2011; Mortamais et al., 2017; Smith & Bondi, 2013). During recognition memory testing, patients with AD do not benefit from cueing and tend to show greater false-positive errors (Karantzoulis & Galvin, 2011). As the disease progresses, language skills continue to deteriorate, and global aphasia and muteness are common. Traditionally, cognitive dysfunction was viewed as the outcome of AD and other dementias. Mortamais et al. (2017) argued that increasing evidence supports that cognitive changes can be detected in preclinical stages of dementia rather than waiting for a clinical diagnosis and there is a need for “comprehensive evidence-based guidelines for preclinical cognitive assessment”. The current study attempted to address these needs.

Vascular Dementia

Vascular dementia (VaD) is an umbrella term that refers to any dementia caused by impaired cerebral blood flow due to cerebrovascular disease or brain injury. It was included in the DSM-5 as a major cognitive disorder. VaD is the second most common form of dementia but it is not classified as a neurodegenerative disease, only AD leads in incidence. Pure VaD is relatively uncommon, but it is a contributor in an estimated half of clinical- and population-based studies, most often in combination with AD and diagnosed as mixed dementia ([MD]; Blom et al., 2014; Kang et al., 2016; Smith, 2017; Smith & Bondi, 2013). While rare in its pure form, it can cause dementia (Ramirez-Gomez et al., 2017; Smith, 2017; Smith & Bondi, 2013; Stephan et al., 2017). VaD generally has a more abrupt onset than AD (Karantzoulis & Galvin, 2011; Smith & Bondi, 2013). Typically, it is identified in one of two ways, either a stroke is diagnosed which is then followed by the onset of dementia, or a patient with no history of stroke complains of cognitive decline and neuroimaging or neuropathology reveals the vascular brain injury (Smith, 2017). Cognitive impairments due to vascular issues are also diagnosed on a spectrum, with the severity of the vascular disease correlated with the extent of the cognitive impairment ranging from MCI with a vascular etiology, also known as vascular cognitive impairment to all-out VaD when the criteria for dementia is met (Ramirez-Gomez et al., 2017; Stephan et al., 2017). Just like with AD the prevalence of vascular dementia increases after 65 years of age and factors related to the brains ability to compensate for the level of pathology makes it difficult to use neuroimaging

alone to diagnose the severity of VaD. NP testing becomes an important factor in the comprehensive clinical assessment used make the final diagnosis.

In addition to age, there are other risk factors associated with vascular dementia; hypertension; diabetes; high cholesterol; sedentary lifestyle; low or high body mass index; smoking; coronary artery disease; and atrial fibrillation (Livingston et al., 2017; Smith & Bondi, 2013). Cerebral blood flow becomes impaired through slow cumulative processes that lead into cerebral small vessel disease or may have a sudden onset from a single major event such as a hemorrhagic stroke (Blom et al., 2014; Ramirez-Gomez et al., 2017; Smith, 2017; Smith & Bondi, 2013; Stephan et al., 2017). Cortical signs of stroke may include aphasia and apraxia, but the NP profile of an individual post stroke varies because it is directly related to the stroke location and severity (Smith & Bondi, 2013; Stephan et al., 2017). The mere presence of a cerebrovascular brain injury does not necessarily signal dementia or indicate impending dementia.

There are many different cardiovascular and cerebrovascular incidents that lead to cognitive dysfunction and different vascular disorders have different patterns of cognitive impairment (Kang et al., 2016; Ramirez-Gomez et al., 2017; Stephan et al., 2017). People with vascular dementia tend to experience motor issues more often than those with AD, such as a slowing of gait, and neuropsychiatric signs like depression, apathy, psychosis, or sudden and inappropriate laughing or crying known as pseudobulbar affect (Smith & Bondi, 2013). When vascular dementia is suspected, the cognitive profile is examined along with a complete health history, risk factors, brain imaging, and the presence or

absence of biomarkers of other neurodegenerative diseases that may be contribute to cognitive decline such as the presence of CSF Aβ (Smith & Bondi, 2013).

The NP profile for VaD is typically characterized by poor executive function that includes decreased inhibition and processing speed; poor planning and problem solving; and difficulties with task changing, working memory, and attention, but the variety of incidents that lead to vascular cognitive dysfunction makes it difficult to typify a pattern across all vascular conditions (Karantzoulis & Galvin, 2011; Ramirez-Gomez et al., 2017; Smith & Bondi, 2013; Stephan et al., 2017). While episodic memory impairment is a hallmark of the AD diagnosis, those with VaD typically respond better to recognition and cueing of learned information (Karantzoulis & Galvin, 2011; Ramirez-Gomez et al., 2017). Measures of verbal fluency showed greater impairment of phonemic (letter) fluency in VaD versus greater impairment of semantic (category) fluency in AD (Ramirez-Gomez et al., 2017). Ramirez-Gomez et al. (2017) found phonemic and semantic differences alone did not distinguish AD from VaD, but when they generated a formula that incorporated the first learning trial of a word memory test, they were able to classify AD from VaD in an autopsy confirmed cohort with moderate sensitivity and specificity, but asserted additional independent studies were necessary to confirm their hypothesis.

Vascular dementia may mimic AD depending on the location of the infarct, and the fact that it is often found in combination with AD as MD makes it even more difficult when testing profiles have significant overlap (Karantzoulis & Galvin; 2011; Kang et al., 2016). When comparing a pure AD etiology to MD, the frontal lobe deficit patterns of

VaD overlap with the medial and lateral temporal lobe deficit patterns of AD. Kang et al. (2016) found significant differences in the milder stage of dementia where the NP profiles of MD demonstrated lower performance on executive function and semantic fluency but maintained a memory advantage over AD into a moderate stage of dementia. As the severity of the dementia increased, testing patterns and profiles were harder to distinguish from one another as no significant differences were found in attention, language, visuospatial, or memory scores (Kang et al., 2016). Another conclusion from Kang et al.'s study was that AD patients appeared to maintain better ADLs making the rating of functional performance an important piece of the differential diagnostic puzzle. The current study includes a clinical dementia staging instrument, CDR, considered a gold standard for capturing daily functional performance
related to dementia. Clearly the variations in cognitive profiles across VaD make it challenging to come to a consensus on which cognitive tests best capture the information needed to support an MCI due to vascular conditions or a VaD diagnosis (Stephan et al., 2017). Comprehensive testing across all domains allows clinicians to identify cognitive strengths and weaknesses to rule out alternative explanations for the impairments, and the use of the CDR as a measure of functional performance helps support the final diagnosis.

Lewy Body Disease

Dementia with Lewy bodies (DBL) is considered one of the three most common forms of dementia and is the second leading neurodegenerative cause (McKeith et al., 2017; Smith & Bondi, 2013). It has distinct clinical features including cognitive fluctuations; hallucinations; rapid eye movement (REM) sleep behavior disorder (RBD);

and parkinsonism; which typically appear early and persist throughout the course of the disease (McKeith et al., 2017). Unlike AD and VaD, the incidence of DLB does not appear to increase with age (Smith & Bondi, 2013). The First International Workshop of the Consortium on DLB convened in 1996 and established the classification of dementia with Lewy bodies (Rizzo et al., 2012). The pathologic hallmark of DLB is the Lewy body, an intracytoplasmic inclusion in deep cortical layers in the brain, especially the frontal and temporal lobes (McKeith et al., 2017; Rizzo et al., 2017; Smith & Bondi; 2013). While short-term memory is typically the earliest deficit of an AD patient, impaired visuospatial function, attention, and executive function appear to be the most prominent deficits in early DLB (McKeith et al., 2017; Rizzo et al., 2017; Smith & Bondi, 2013). In fact, an absence of visuospatial impairment helps clinicians exclude DLB. Attention deficits vary from seconds to days and are interspersed with periods of near normal function; they may take the form of a brief cognitive fluctuation that interrupts the flow of an ADL or be severe enough for the individual to appear catatonic for a length of time (Smith & Bondi, 2013; McKeith et al., 2017). The CDR helps capture these fluctuations in consciousness.

Visual hallucinations are rare in AD but occur in up to 80% of individuals with DLB (McKeith et al., 2017). They are an early sign of DLB and are often under reported because the patient has a lack insight regarding the nature of the hallucinations. They may come in the form of people, animals, or inanimate objects that move or shape shift; or even more complex visual interactions like ongoing conversations with the dearly departed. Auditory hallucinations, hearing music, a TV, or voices nearby; olfactory

hallucinations, both pleasant and foul; and tactile hallucinations like feathers or fur brushing up against an arm or leg, or even insects crawling on their skin are also present in patients with DLB (McKeith et al., 2017). Another early sign that occurs in 85% of individuals and may precede clinical diagnosis by up to 20 years is RBD, characterized by recurrent dream enactment and vocalizations (Donaghy et al., 2018; McKeith et al., 2017; Smith & Bondi, 2013). It can be mild or severe, and injuries can happen from striking a bed partner or suddenly jumping out of bed. RBD is not exclusive to DLB and can also occur in patients with Parkinson disease dementia. Gait disorders, limb rigidity, or a combination of the two, termed parkinsonism, is also present in 70%-90% of patients with DLB but usually in a milder degree than someone with Parkinson's disease.

These types of overlapping clinical features make differential diagnosis challenging which is why NP testing plays an important role in a comprehensive clinical assessment to avoid serious negative side effects of certain treatment protocols. DLB continues to be under-recognized, and misdiagnosed as AD or Parkinson's disease, so there is a need to refine diagnostic criteria to improve sensitivity and specificity as treatment efficacy is highly specific to DLB (McKeith et al., 2017; Rizzo et al., 2017). There are serious consequences when DLB is treated with the wrong types of medications; 30% to 50% of patients with DLB have severe antipsychotic sensitivity with reactions that may include irreversible parkinsonism, impaired consciousness, and even death (McKeith et al., 2017; Smith & Bondi, 2013; Rizzo, et al, 2017). These reactions are less common in Parkinson's disease and have not been observed in AD. Donaghy et al. (2018) concluded the addition of neuropsychiatric symptoms other than hallucinations

(e.g., delusions, anxiety, depression, and apathy) to the core diagnostic features supported the differential diagnosis between DLB and AD.

The NP profile of DLB is a mixture of cortical and subcortical symptoms characterized by disproportional impairment in visuospatial, attention, and executive functions early on (Karantzoulis & Galvin, 2011; McKeith et al., 2017; Smith & Bondi, 2013). When memory is impaired, usually later in the disease, encoding is typically less affected than retrieval; object naming is also typically preserved (Karantzoulis & Galvin, 2011; McKeith et al., 2017; Smith & Bondi, 2013). No specific testing battery has been developed (Donaghy et al., 2018; McKeith et al., 2017), but comprehensive measures that includes spatial and perceptual tasks like complex figure copy and line orientation, and executive and processing speed measures like trail making tests and coding, are especially helpful in the differential diagnosis process when used in tandem with word memory lists, and object naming tasks.

Frontotemporal Dementia

Frontotemporal dementia (FTD) has become an umbrella term for a group of clinically heterogeneous degenerative disorders affecting the frontal lobe alone, an isolated temporal lobe, or a degeneration of both the frontal lobes and temporal lobes. The most common subtypes are the behavioral variant (bvFTD) and two forms of Primary Progressive Aphasia (PPA), nonfluent and semantic variants (Lee, 2019; Ravskoski et al., 2011; Smith & Bondi, 2013). The main symptoms of bvFTD are persistent and significant changes in behavior and personality, while the main changes in PPA are a progressive deterioration of language skills. Because the pathology of each

variant is different, a consensus on the neuropsychology of FTD remains elusive. When contrasted with AD, memory is spared in the early course of the disease with better recall and recognition across all FTD syndromes (Smith & Bondi, 2013). Wittenberg et al., (2008) asserted that the difficulties to finding a consensus stems from still evolving diagnostic criteria, inconsistent findings in research, and the rare prevalence of FTD. Smith and Bondi (2013) claimed that FTD accounted for only about 5% of all dementias in an unselected autopsy series.

The most common subtype bvFTD accounts for nearly 50% of all FTD cases with an onset most common in the 6th decade of life but uncovered as early as the 2nd decade and as late as the 9th decade of life with only a 0.02% incidence rate in the general population (Lee, 2019). The primary characteristics are the pervasive behavioral changes that are often ignored or misdiagnosed for several years causing significant impact on the caregiver stress levels before a formal clinical diagnosis (Lee, 2019; Smith & Bondi, 2013). Rascovsky et al. (2011) included the following symptoms in their outline for diagnostic criteria:

A. Disinhibition – inappropriate and embarrassing public behavior
B. Apathy, inertia, loss of sympathy, empathy, or changes in humor – indifference to others' needs and feelings, less warmth and affection
C. Hyperorality – changes in food preferences or decline in table manners
D. Compulsive behaviors – obsessions with new hobbies or interests; smoking, alcohol use; or new religious and spiritual pursuits

E. A neuropsychological profile that shows a relative sparing of memory and visuospatial functions with deficits in executive functions

It is challenging to interpret NP test results in many patients with bvFTD due to substantial overlapping profiles with other neurodegenerative diseases. First, other medical illnesses (infarction, tumors, abscess, or trauma), substance abuse, psychiatric disorders, and other dementias such as AD or LBD must be ruled out (Lee, 2019; Smith & Bondi, 2013). Because individuals with bvFTD rarely have insight into their behavioral changes, a proper diagnosis is heavily dependent on the testimony of a knowledgeable informant. This is obtained via informant interviews during the clinical interview and through the CDR, a structured interview and informant testimony. Clearly, the significant personality and behavioral changes in bvFTD dwarf any behavioral disturbances present in AD, but in the temporal variants that are discussed below language and semantic knowledge are the most pronounced deficits.

The other syndrome of FTD is PPA characterized by an insidious onset of progressive language impairment that is evident in the early stages of the disease; prevalent deficits in word finding, word comprehension and usage, and sentence construction are present while other cognitive domains and ADLs are relatively spared (Gorno-Tempini et al., 2011; Mesulam, 2013; Smith & Bondi, 2013). Two variants of PPA have been delineated based on the type of language impairment; nonfluent or agrammatic, and semantic (Gorno-Tempini et al., 2011; Smith & Bondi, 2013). Word finding is the common feature across both subtypes of PPA, but the nonfluent variant has

more pronounced articulation problems while the semantic variant has more pronounced comprehension difficulties.

The main characteristic of nonfluent PPA is apraxia of speech as demonstrated by effortful, halting speech with speech-sound errors or distortions and agrammatism in language production; comprehension is spared for single words and simple sentences, but complex syntax poses problems (Smith & Bondi, 2013; Gorno-Tempini et al., 2011). The semantic variant PPA is marked by preserved fluent output, but simple comprehension becomes impaired through a loss of single word or object meaning and as the disease progresses comprehension becomes more globally impaired, episodic memory may decline, and behavioral symptoms such as rigidity of personality and loss of empathy may emerge (Mesulam, 2013; Smith & Bondi, 2013). A differential diagnosis is made by first ruling out other medical issues like cerebrovascular disease or tumors, then testing is used to discern the pattern of language deficits; the patient must also initially present with no impairments of episodic or visual memory, no visuospatial impairment, and no prominent behavioral disturbances (Lee, 2019). There is inconclusive evidence for the utility of NP testing in the diagnosis of FTD as the pattern of executive dysfunctions has not been distinguished from AD (Smith & Bondi, 2013). The challenges remain to improve the utility of NP testing and to clear the confusion behind the diagnostic criteria of FTD.

Parkinson's Disease

Parkinson's disease (PD), once considered a disorder of only the motor system, is now widely recognized as a clinically diverse disease with three major subtypes; two of

which have more neuropsychiatric and nonmotor manifestations in addition to motor symptoms (Chou, 2019). The traditional tremor-dominant subtype has slower progression and less cognitive impairment than the akinetic-rigid subtype and the postural instability and gait difficulty subtype (Chou, 2019). The common clinical motor manifestations include tremor, bradykinesia, rigidity, and postural instability, while 97% of patients also present with nonmotor symptoms, some which manifest before motor symptom onset:

- Cognitive dysfunction and dementia
- Psychosis and hallucinations
- Mood disorders – depression, anxiety, and apathy
- Sleep disturbance
- Fatigue
- Autonomic dysfunction
- Olfactory dysfunction
- Gastrointestinal dysfunction
- Pain and sensory disturbances
- Dermatologic findings - seborrhea
- Rhinorrhea

PD and DLB share many overlapping clinical symptoms and pathological similarities such as parkinsonian features, psychosis, visual hallucinations, and fluctuating cognition making the differential diagnosis even more challenging (Chou, 2019). Clinicians and researchers use the convention of the “one-year rule”; if motor symptoms begin more than a year prior to the onset of dementia, then PD is diagnosed. When motor symptoms

present concurrently, or they start during the same year, then the diagnosis of DLB is given.

Cognitive dysfunction and dementia are common in PD with an estimated 78% incidence rate of dementia uncovered in longitudinal studies (Chou, 2019; Rodnitzky, 2018). Subcortical dementias like PD typically spare memory in the early stages with the first NP manifestations appearing as psychomotor retardation, compromised executive function, and impaired attention and processing speed (Chou, 2019; Smith & Bondi, 2013). As the disease progresses memory recall and visuospatial skills are more notably impaired with full dementia typically setting in later during the course of the disease. Cognitive testing plays a large role in determining the etiology of AD and all other types of dementia.

The History of Educational and Psychological Testing

Anastasi and Urbina (1997) wrote the classic comprehensive text on psychological testing that is still widely used in graduate programs. The authors wrote, "The roots of testing are lost in antiquity" (p. 32) but sketched an outline of the ground-breaking pioneers that laid the foundation for educational and psychological testing as a valid and reliable method for collecting meaningful data about individuals. English biologist Francis Galton launched the modern testing movement when he established the need to measure characteristics of related and unrelated persons in the effort to further his research interests in heredity, leading to the first large systematic collection of data from his anthropometric laboratory (Anastasi & Urbina, 1997). In addition to developing rating-scales and self-report questionnaires, Galton also advanced the statistical methods

necessary for data analysis. Around the same time American psychologist James McKeen Cattell has just finished his dissertation on reaction time under the tutelage of Wilhelm Wundt in the first experimental psychology laboratory in Leipzig, Germany. Cattell's chance encounter with Galton in 1888 while lecturing at Cambridge University, inspired Cattell to merge Wundt's newly established science of experimental psychology with Galton's even newer testing movement. The result of this early work was an upsurge of interest in testing and measures designed to quantify and classify human behavior and cognition.

Frenchman Alfred Binet built on that foundation and constructed the first comprehensive test of intelligence in 1905 at the request of the Minister of Public Instruction in an effort to create proper procedures to educate children with mental retardation (Anastasi & Urbina, 1997). The 1905 scale, as it was known, originally consisted of 30 problems that required comprehension, reasoning, and judgment and was arranged from least to most difficult. The scale was administered to 50 "normal" children aged 3 to 11, and other children and adults with mental retardation. This preliminary scale had no objective method for arriving at a total score, but it caught the attention of psychologists around the world and was translated and adapted in many countries, including the United States. In the revised version, the 1908 scale, the researchers dropped the unsatisfactory tests and added others that had more promise. Simon and Binet then collected data on 300 normal children and grouped the results by age level. Any tests passed by 80 to 90% of normal 3-year-olds were grouped into a 3-year level; all tests passed by 4-year-olds were grouped into a 4-year level; and the same was done with

each age group up to 13. Thus, the foundation for a normative scoring system as a uniform frame of interpretation and reference was based on research with children, but Binet's ill-timed death in 1911 left much work to be done.

L. M. Terman and associates from Stanford University adopted Binet's work in 1916 and used it to build the broader and more psychometrically sound Stanford-Binet coining the use of a ratio between mental age and chronological age as an intelligence quotient (IQ). The work of these early psychometricians quickly diffused throughout the world and standardized psychological and educational testing forged ahead as an explosion of new measures and methods for analyzing data quickly followed. The *Journal of Applied Psychology* (JAP) chronicles the history that led to our understanding of measurement error and validation, as well as many other methodological areas of psychological testing (Cortina et al., 2017). From 1917 to 1925 the journal published the early work on the development and norming of cognitive ability testing for different populations and the beginning of statistical significance and prediction models with emphasis on psychometric properties and classification. Over the next 40 years, work on test scoring methods and cross validation dominated the publication while the following 43 years witnessed an explosion of new measures and the data-analysis innovations that came along with them (Cortina et al., 2017). The current study builds upon the work of these early psychometricians and examines subgroup differences to validate optimal norms for an aging population.

The use of testing in the field of neuropsychology began around WWII when the evaluation and recovery of brain injured soldiers created a new need in the field of testing

beyond sensory, vocational, and intelligence testing. Ralph Reitan, a recent college graduate, was given the task of assessing these brain-injured soldiers and found a lack of publications available for reference (Grant & Heaton, 2015). With the help of the hospital's chief neurologist, John Anita, they published a series of four articles on the psychological consequences of brain injury. Anita encouraged Reitan to consult with psychologist Ward Halstead who he had seen lecture on the effects of brain injury. Through his new relationship with Halstead, he met Louis Thurston a mathematical psychologist from the University of Chicago Medical School and together the men encouraged Reitan to enter a graduate program in psychology (Grant & Heaton, 2015). "Through a combination of mishaps and serendipity", Reitan ended up splitting his studies between medical school and psychology; and as a graduate student Reitan tested patients in Halstead's laboratory using the instruments Halstead developed (Grant & Heaton, 2015). The brain-behavior relationship began to unfurl as data was gathered through testing, medical, surgical, and autopsy channels and the field of neuropsychology was born. Reitan's work:

> …refined and standardized what most neuropsychologists now take for granted as they write their reports: the approach to inference in individual cases that takes into account such information as levels of performance, patterns of test results, right-left comparisons, and pathognomonic signs. (Grant & Heaton, 2015, para. 3)

The Halstead-Reitan battery is a collection of NP tests that assess the functioning of the brain and is still used both in its complete form and as individual test components today.

The Role of Neuropsychological Testing in Clinical Settings

NP testing and measurement play an important role in diagnosis, prediction of progression, and treatment planning for neurodegenerative diseases because research demonstrates that these actuarial methods are superior to clinical judgment alone (Smith & Bondi, 2013). The FDA established cognitive measures in this critical role when they required NP measures be included as a co-primary outcome in research studies seeking to demonstrate efficacy in dementia treatments (Leber, 1990). A *test* is defined as any set of tasks, procedures, or stimuli designed to elicit responses that sample an examinee's performance or behavior in a specified domain, while *assessment* is the broader term referring to the process that integrates the gathered data with other sources of evidence such as interviews about a participant's social, educational, employment history, health history, and psychological history (AERA, APA & NCME, 2014). NP testing batteries gather quantified and meaningful data about an individual's various cognitive and behavioral domains for diagnostic or predictive value, yet the final diagnosis should always include a full assessment including physical and neurological examinations, the patient's medical and family history, and blood tests or brain imaging to rule out other potential causes of cognitive impairment.

As there is no single test for dementia, a variety of different measures are used during an assessment to take an inventory of the strengths and weaknesses of the major cognitive domains including verbal and nonverbal IQ, memory domains that encompass both encoding (learning) and retention (delayed recall and recognition), executive function processes, language production, attention, visuospatial skills, and processing speed. The

patterns and profiles attained from testing aid in a differential diagnosis between the underlying pathologies (Karantzoulis & Galvin, 2011; Ramirez-Gomez et al., 2017; Ramlall et al., 2014; Smith & Bondi, 2013; Stephan et al., 2017).

Normal aging individuals have no deficits on psychometric test performance relative to their age-matched peers, whereas individuals "at risk" for dementia may have borderline or impaired cognitive function in one or more areas of cognition when compared to age-matched peers. Ideally, more than one measure is used in each domain so that evidence converges to illuminate the relationship between tests intended to assess similar constructs. This also helps assure that discriminant evidence between measures intended to measure different constructs are also accurate. Educational and psychological testing methods are some of the most significant and vital contributions of the behavioral sciences to society (AERA, APA & NCME, 2014). The procedures for neuro-psychological testing and data collection are highly operationalized and demonstrate strong reliability (Ramirez-Gomez et al., 2017; Smith & Bondi, 2013).

> Tests that are valid for their intended purposes provide substantial benefits for test takers and test users…proper use can result in better decisions about individuals and programs than would result without their use... The improper use of tests, on the other hand, can cause considerable harm to test takers and other parties affected by test-based decisions. (AERA, APA & NCME, 2014, p.1)

A primary consideration in developing and evaluating tests is validity, an accumulation of evidence that scientifically supports that the test measures the construct

it is intended to measure. The validity and reliability of testing methods are the foundation for accurate assessment (AERA, APA & NCME, 2014; Heaton et al., 2004). The first empirical paper published in the Journal of Applied Psychology in 1917 was a validation study by Terman and his colleagues (Cortina et al., 2017). The process of validation deliberates arguments both for and against the intended interpretation of the test scores relevant to their proposed use (AERA, APA & NCME, 2014). The validation process is continual, constantly evolving as new data is gathered, and often necessitates revisions to accommodate the latest articulated evidence (Wilkenson & Robertson, 2006). The proposed study is a concurrent study, particularly useful for psychodiagnostic tests (AERA, APA & NCME, 2014). It also continues the validation process of the measures being used and may provide evidence to refine and reevaluate the utility of the tests and their interpretations for use in an aging population.

Smith and Bondi (2013) asserted that NP measures are essential in clinical and research efforts focused on neurodegenerative disease and defined five roles for such measurements in preclinical and clinical dementia populations. First, NP measures serve as biomarkers because they are highly operationalized and help distinguish between underlying pathologies (Smith & Bondi, 2013). The FDA's requirement that cognitive measures must be included as a co-primary outcome in secondary prevention studies solidified NP measures in this key role (Smith & Bondi, 2013). Second, they serve as predictors for the development of AD and other dementias; they detect the clinical manifestations of neurodegenerative disorders, so they should also predict their future development (Smith & Bondi, 2013).

Third, measures can dynamically capture countervailing influences on disease trajectory, studies suggested that memory function does not decline at an even rate but goes through periods of stabilization that may reflect biological and psychological compensatory mechanisms such as the mediating and moderating factors of compensatory strategies or cognitive reserve (Smith & Bondi, 2013). Fourth, NP measures are proxies for important functional deficits; patients may be unreliable reporters of their own functional , so the measures serve to estimate functional impairment which helps family members determine what matters most, because it identifies when their loved one is no longer safe living independently (Smith & Bondi, 2013). Finally, NP measure can provide insights into interventional targets (Smith & Bondi, 2013). Just as important as identifying impaired cognitive domains, measurements also have the ability to identify cognitive domains with preserved strengths. Cognitive rehabilitation services can capitalize on an individual's residual strengths in order to compensate for weaker areas. These five functions demand that the NP measures contain the optimal sensitivity and specificity for their intended purpose.

A test that is sensitive to detecting a neurodegenerative disease like AD must have a high probability that patients with AD score in the abnormal range, while higher specificity assures that patients without dementia will score within normal range. Smith and Bondi (2013) argued that positive predictive value rather than sensitivity, is statistically more relevant to a diagnostic situation, and enhancements to specificity are more important for this purpose. The use of norms is assumed to enhance specificity. Researchers argue that aging is the major risk factor for dementia, therefor it undermines

the sensitivity of NP measures to control for age (Hessler et al., 2014; Holtzer et al., 2008; Malek-Ahmadi et al., 2015; O'Connell & Tuokko, 2010; Quaranta et al., 2016). The current study will obtain empirical evidence and conduct logical analyses to evaluate the proposition that age- and education-corrected norms may not universally improve utility of testing measures in an aging population.

Norm-Referenced Interpretation in Neuropsychological Assessment

Educational and psychological testing theory assumes the concept of an "ideal" or "normal" level of functioning against which the test taker's performance can be compared. Therefore, the most fundamental level of interpretation is the participant's performance in relation to the general population as established by a standardized sample to derive normative scores (AERA, APA & NCME, 2014; Heaton et al., 2004). Raw scores, simply the number of items correct on any given test, are thought to be of little use because tests vary in difficulty and the number of items they contain, making it difficult to make meaningful comparisons to other scores (AERA, APA & NCME, 2014; Bryant & Brown, 1984; Schoenberg & Scott, 2011). Therefore, *derived* scores, a statistical concept illustrating the participant's exact position relative to individuals in the normative group, has become a far more meaningful and significant metric. Norms provide a point of reference that make raw scores valuable by allowing interpretations that indicate if the individual's performance is typical for the normative group. They also allow for comparison of performance across various tests, track change or progress across time, and diagnose strengths and weaknesses (Bryant & Brown, 1984; AERA, APA & NCME, 2014; Schoenberg & Scott, 2011). Thus, norm-referenced interpretation has

become gospel in testing theory (Quaranta et al., 2016). One of the fundamental principles relevant to norming is the selection of the appropriate comparison group for the tests being used.

At a most basic level, a normative sample is taken from the population that is thought to be large enough to represent the current U.S. population, and sufficient enough to be proportional across certain demographic characteristics such as sex, age, and level of education, geographic region, and race or ethnicity. The mean score becomes the average and expected level of performance for age and education. High scores are classified in terms of being "high average", "superior", and "very superior", while low scores are described in terms of "low average", "weak/mildly to moderately impaired", and "exceptionally weak/severely impaired". The use of demographically corrected normative scores is recommended for most diagnostic purposes. Yet, despite the quest to achieve a fair representation in a normative sample, researchers established that there are times when it may be advantageous to develop norms based on the performance of individuals in a specific subpopulations; especially if the mean of the subgroup is 1 to 1½ SD away from the mean of the normative group, or when test performance is tied to a specific therapy or treatment (Bryant & Brown, 1984; Hessler et al., 2014; Holtzer et al., 2008; Malek-Ahmadi et al., 2015; O'Connell & Tuokko, 2010; Svinicki & Tombari, 1981). This study examines whether late adulthood, a time of exponentially increased risk for cognitive impairment, may be one of the exceptions to the standard practice of using an age- and education-corrected normative system.

Methods of Norming

Both educational and psychological testing assume a "normal" level of functioning, which is traditionally established by the mean test performance in the standardization sample. The most fundamental interpretation of an individual's performance is their standing in relation to the general population. When the individual's raw score is compared to the distribution of scores across the sample population, it becomes a snapshot of where they fall in that distribution. This gives clinicians and researches a uniform frame of reference to determine the individual's relative position within the context of the larger population. Understanding the principles relative to developing norms is imperative for the test user as "psychological test norms are in no sense absolute, universal, or permanent" (Anastasi & Urbina, 1997, p. 68).

The main consideration in traditional norming methods is representativeness of the standardization sample to the general population, but it is equally important that the sample represent the population using the test (Anastasi & Urbina, 1997). Data should be collected at multiple sites that represent different geographic regions of the U.S. The advantage of traditional norming is its simplicity, but the greatest disadvantage is that separate norm groups must be defined arbitrarily for continuous covariates like age and as a result can change an interpretation of an individual's test performance; a corrective measure would be to define more categories, but the smaller sample size produces less precise norms (Oosterhuis, et al., 2016). Zachary and Gorsuch (1985) introduced linear regression to avoid categorizing continuous covariates. According to Oosterhuis et al.

(2016) regression-based norming, which requires a smaller sample size but claims equally precise norms is gaining in popularity.

It has been established that participants should be screened and excluded for characteristics that might interfere with their performance such as sensory impairment, brain impairing medications, or a medical history that includes stoke, epilepsy, or any other neurological issues that affect cognition (White & Stern, 2003), but in an aging population this is a difficult task. Robust norms, norms that follow the normative cohort for a length of time removing anyone who develops dementia and keeping only those who remain dementia free, may have more clinical utility than conventional norms when dealing with an aging population (Holtzer et al., 2008). Hassenstab et al. (2016) found that removing preclinical participants from normative samples yielded higher means and less variability on episodic memory, visuospatial ability, and executive function measures reducing age-affects, but provided no substantive benefit for diagnostic classification. However, the considerable investment of resources needed to establish robust norms leaves researchers looking for alternative methods to estimate the prevalence of preclinical cases and consequentially adjust interpretation guidelines for cognitive testing.

Norm-Referenced Interpretation in MCI and Dementia

Norms are assumed to enhance specificity, the probability that a person without dementia will have normal test scores. But because age is the number one risk factor for dementia, researchers currently debate the use of norms in an aging population arguing that norms reduce the sensitivity of the test scores to abnormal cognitive impairment (Hessler et al., 2014; O'Connell & Tuokko, 2010; Quaranta et al., 2016). There are

several reasons stated why this may be true. It is noted in the literature that norms become less accurate as time between publication and use increases due to changes in demographics and socioeconomic factors that modify the composition of the reference population (Quaranta et al., 2016). It is also argued that individuals in the normative sample population for age corrected norms may already be transitioning into dementia and contaminating the norms by lowering the mean performance and increasing the variability in cross-sectional samples (Holtzer et al., 2008). Late adulthood is also a time when brain impairing maintenance medications are routinely prescribed to manage chronic health conditions. Age correction is also thought to decrease the sensitivity of measures because norms become more forgiving and tolerant of errors as age increases, possibly decreasing the sensitivity of the tests to cognitive impairment and underestimating the risk of dementia in our aging population (Smith & Bondi, 2013). Optimal demographic correction as determined via the analyses in this study could result in better diagnostic performance when differentiating normal controls from individuals with dementia when employing cognitive testing.

Summary

It is recognized that symptoms of dementia may not appear for 20 years or more after brain changes start to occur (Jack et al., 2015; Ritchie et al., 2015; Rockwood et al., 2014; Smith & Bondi, 2013; Sperling et al., 2011; Sutphen et al., 2015; Ward et al., 2013). Research is only beginning to address how many people may be in preclinical stages of dementia or have MCI due to neurodegenerative diseases (AA, 2018) and these individuals are included in the standard norms which influence a diagnostic outcome

when employing a NP testing perspective. Smith and Bondi (2013) asserted the number of patients that can be helped through neuropsychology is only increasing and the ability to differentiate between normal aging, dementia, and the phase between the two, accepted as MCI, has emerged as a primary focus of research.

Psychometric measures are the necessary tools that provide the data needed to distinguish between normal aging and a neurodegenerative process (Smith & Bondi, 2013). As there is no single measure for dementia, the continued validation of current measures to optimize their ability to differentiate between normal cognitive aging and a neurodegenerative disease process is paramount. There is no cure for AD and the race against the clock continues to inspire researchers to search for new ways to diagnose the condition earlier and more accurately. Treatments and interventions must be administered as early as possible if there is any hope of changing the course of the disease. Researchers argue that outcome studies using comprehensive actuarial methods to examine the patterns and profiles of NP dysfunctions are needed to move the field forward (Bondi et al., 2014; Malek-Ahmadi et al., 2015, Smith & Bondi, 2013). The current study aims to clarify the current debate over which normative method is most diagnostically accurate.

Chapter 3: Methodology

The following methodology was review and approved by the Walden University IRB, approval number 01-21-20-0529160. The overarching purpose of this study was to investigate the impact of demographic corrections on the diagnostic validity of cognitive tests in differentiating between normal cognitive aging and dementia. The standard practice of demographically correcting raw test scores for age and education is widely believed to universally improve the scores' ability to detect cognitive impairment for all age groups. However, there is a debate in the literature about the use of demographically corrected scores in our aging population because the norms for older individuals may be tainted by individuals with preclinical AD and thus underestimate the presence of cognitive impairments. Consequently, it is not clear how to best use normative data in dementia evaluations. This section is an exposition of the data and the analytic strategies that will be used for this investigation.

First, the methodology and rationale is introduced along with a brief review of how clinicians determine the presence or absence of dementia. Next, participant selection procedures, recruiting strategies, and data collection techniques designed to minimize threats to internal validity are addressed. The primary variables involved in the analysis are presented, and the manner in which demographic corrections were applied and the creation of aggregate cognitive composite scores from individual tests follows. The data sources are thoroughly explained, including a review of the reliability and validity of the tests taken from the most recent version of the UDS-3 NAB, with special attention paid to the CDR-SB the primary criterion measure in the study. Subsequently the detailed data

analysis plan that was used to investigate primary study hypotheses (including assumption testing and regression diagnostics that protect against threats to internal and external validity) are presented. Finally, *a priori* power analysis for tests of the overall models was generated, and the regression coefficients are disclosed.

Hierarchical multiple linear regression was used to investigate the strength of the relationship between the patient's test scores (scored 4 different ways) and the patient's clinical dementia severity rating. Because the order of entry into the hierarchical multiple linear regression model can have great impact on the results, the order of variable entry was defined *a priori*, which avoided pitfalls inherent in methods such as stepwise regression and will promoted better generalization as opposed to overfitting of sample data (Harrell, 2001; Roa, 2003). Comparison of the R^2 and AIC values of the various regression equations clarified the extent to which the raw scores, or the various types of demographic corrections (e.g., age, education, age and education) influenced the ability of cognitive tests to detect meaningful variation functional changes due to cognitive loss in dementia and MCI. While this cannot be explicitly tested for significance because there are no universally accepted means for quantitatively comparing non-nested models, it was one of the more important features of this study. Differences in R^2 and AIC values of the various regression models allowed for quantitative analysis of model differences, albeit not with a specific statistical test.

Determining the Presence or Absence of Dementia

The primary difference between dementia and MCI is the extent to which the cognitive decline influences the individual's day-to-day functioning (APA, 2013;

McKhann et al, 2011). MCI requires an objectively determined decline in cognitive function as evidenced by mental status screening or formal NP testing in the setting of relatively well-preserved day-to-day functioning (APA, 2013; Albert et al., 2011). Dementia by definition requires significant functional impairment that represents a decline from the individual's previously higher level of functioning (APA, 2013; McKhann et al, 2011). Heuristically and in clinical practice, this is often defined as the loss of ability to independently complete IADLs (e.g., driving, managing one's finances, self-managing medications, using the community, etc.) with an adequate performance level. More formally, global staging instruments such as the CDR measure the extent to which cognitive loss interferes with an individual's ability to perform day-to-day activities (Morris, 1997). The CDR measure represents the ultimate quantitative standard for the presence or absence of dementia and thus serves as an optimal criterion for the purposes of the present study.

Study Variables

The primary variables involved in the analyses presented below included the CDR as an outcome measure, a global measure of dementia severity that has been neuropathologically validated and is considered an international "gold-standard" for ascertaining the presence or absence of dementia (Olde-Rikkert et al., 2011). The CDR also allows clinicians to derive the CDR-SB by summing clinician ratings in the 6 different domains, which provides a more fine-grained, pseudo-continuous measure of dementia severity. Published criteria derived from large clinical samples exist for determining the presence or absence of dementia using either the global CDR score

(Morris, 1993) or the CDR-SB (O'Bryant et al., 2008, 2010). Individuals with a CDR-SB of 0 are considered "normal," those with a CDR-SB of 0.5 - 2.5 are considered to have "questionable impairment," those with a CDR-SB of 3 - 4 are considered to have "very mild dementia," whereas individuals with a CDR-SB of 4.5 - 9.0 are considered to have "mild dementia." The predictor variables were derived from a subset of the tests from the UDS-3 NAB. The memory domain was represented by Craft Story immediate and delayed recall and Benson complex figure delayed recall. The executive function domain was defined by verbal fluency (F & L) and the Trail Making Test B. The language domain was defined by the Multilingual Naming Test (MINT) and semantic fluency (Animals). And Digit Span Forward (DGF) and Digit Span Backward (DGB) backward were used to define the attention domain. Additional clinical and demographic variables needed to process the data, such as etiologic diagnosis, age, education, gender, and visit number were also utilized

Participant Selection and Stratification

Most individuals with Alzheimer's disease develop the condition later in life, but Alzheimer's can develop in a subset of individuals any time after the second decade of life (Rossor et al., 2010). Classically, Alzheimer's dementia is considered "early onset" or "young onset" when it develops prior to age 65. While this threshold is somewhat arbitrary, there is a large body of evidence suggesting that individuals with early-onset AD often present with atypical forms of the illness and different cognitive profiles characterized by more visuospatial disturbance, executive dysfunction, and higher rates of behavioral or neuropsychiatric disturbance (Ossenkoppele et al., 2015). The primary

goal of this study was to inform the clinical use of NP tests in the types of situations most commonly encountered by professionals in clinical practice. Accordingly, all subjects younger than 60 years of age were excluded from analysis in order to increase the likelihood that the study sample was most reflective of typical presentations of AD.

Similarly, different dementia subtypes present with different degrees and types of cognitive difficulties. For example, dementia due to Lewy body disease is classically thought to present with early visuospatial and constructional impairments while memory may remain preserved until well into the disease course (Karantzoulis & Galvin, 2011). Inclusion of non-Alzheimer's dementia may thus introduce phenotypic-related variability that would obscure detection of subtle differences due to demographic factors. Accordingly, only individuals with a primary etiologic diagnosis of AD were selected from the data set for analysis. Dementia due to AD is a progressive condition and individuals traverse several stages during the course of the illness. NP tests may be most helpful in the earlier phases of disease (e.g., MCI, Mild Dementia), as patients become too cognitively impaired to participate meaningfully in assessment as they transition from mild to later stages of dementia. There is also a high likelihood that cognitive test data from moderate to severely demented patients may be less reliable and thus less meaningful than in individuals with milder forms of the disease (Weintraub et al., 2018b). For example, attention, language, memory, and executive function test results may be substantially influenced by general confusion, language comprehension problems, difficulty appropriately engaging in the task, or other impairments that render cognitive tests unreliable indicators of the processes they are purported to measure. Accordingly,

individuals with a CDR-SB > 9 (i.e., moderate to severe dementia) were removed from all analyses.

Creation of Cognitive Composites

Most studies addressing the effects of age and education correction on the predictive validity of cognitive tests in dementia have focused on single test scores (Sliwinski et al.,1996; Sliwinski et al., 1997) or heterogenous composite scores representing multiple cognitive domains (Hessler et al., 2014). It is likely that age and education correction influence different cognitive domains differently. For example, processing speed invariably declines with increasing age (Eckert, 2010) whereas crystallized abilities such as vocabulary or fund of knowledge remain stable or even improve throughout senescence (Harada et al., 2013). A novel feature of the present study was the creation of domain-specific cognitive composite scores for memory, executive function, language, and attention that allowed for precise investigation of the extent to which age and education affected the predictive validity of individual cognitive domains using a hierarchical multiple linear regression analysis.

Creation of cognitive composites was also advantageous statistically. Composite measures may be more sensitive to longitudinal cognitive changes in preclinical dementia and more reliable than the measures from which they are derived (Riordan, 2017). Cognitive composites also offer a better and more complete sampling of participant cognitive abilities than do single cognitive test scores by virtue of their broader item content. Models with fewer explanatory variables are also more desirable than complex models due to enhanced interpretability (James et al., 2017). Such models also tend to

demonstrate less variance across samples, which increases the likelihood that they will generalize to new samples. Composite scores can generally be categorized as empirically derived, theoretically derived, or some combination of these two methods (Weintraub et al., 2018b). This project made use of a theoretically driven strategy that is grounded in well-established principles of NP function and localization. The method used for aggregation of tests within a domain has been widely used as exemplified by Donohue et al. (2014) and involves first grouping tests into different cognitive domains, transforming scores to the same metric (i.e., Z-scores), and then summing them to create aggregate cognitive composite measures with similar psychometric properties and scales of measurement. Theoretical groupings of the particular tests used in this study into different cognitive domains (memory, executive function, language, and attention) are the same as in Weintraub et al. (2018a).

Impairments in memory and executive functioning are highly characteristic of the cognitive phenotype associated with typical presentations of AD (Karantzoulis & Galvin, 2011). Memory measures in particular have been shown to be some of the earliest indicators decline, even in minimally symptomatic individuals (Weintraub et al., 2018b). Tests in this domain included Craft Story immediate and delayed recall and the Benson complex figure delayed recall. The executive function domain included verbal fluency (F & L) and Trail Making Test B. Impairments in language including deficits in semantic verbal fluency and visual object confrontation naming are also characteristic of AD (Salmon & Bondi, 2009). This domain was constructed using the Multilingual Naming Test (MINT) and semantic fluency (Animals). Individuals with AD may additionally

show difficulties with attention and working memory, though these abilities are often better preserved until later into the illness (Cherry et al., 2002). Digit Span Forward (DGF) and Digit Span Backward (DGB) backward were used to create the attention domain. Following transformation of raw scores into Z-scores using the various norming methods described in the following section, Z-scores were summed across tests within a given cognitive domain to create a composite score for each domain. These composite scores served as the predictor variables in the series of hierarchical regression analyses.

Cognitive Test Scoring

Subjects' raw test scores were analyzed before and after a series of demographic corrections. Cognitive tests are scored on different metrics (i.e., seconds to completion as opposed to number of words remembered) and thus must be transformed to a common scale in order to facilitate comparison. Weintraub and colleagues (2018b) developed a normative calculator for the UDS-NAB through fitting linear regression models to the cognitive test data of 3602 cognitively normal participants over the age of 60. Specifically, cognitive test scores were predicted using age, gender, education, and the combination of these variables. These regression models can then be used to standardize observed test scores, adjusting for the demographic variables of choice.

The intercepts of the regression models represent the mean performance of the overall study sample holding age and education constant. The root mean squared error (RMSE) represents the average squared difference between the observed scores and the predicted scores, which can be used as a surrogate measure of the population standard deviation (Weintraub et al., 2018b). To facilitate comparison between the models, the intercept,

slope, and demographic variable regression weights from the full model incorporating age, sex, and education will be utilized and thus, all analyses were corrected for gender.

To generate "raw scores," subjects' scores were first transformed into Z-scores by generating predicted scores using each subjects' actual gender, the average education and age level of the overall sample, subtracting the predicted score for a participant from their observed score, and then dividing by the RMSE of the model. There were no systematic adjustments for age or education, allowing the variability of those factors to remain in the model. To generate age-corrected scores, participant scores were transformed as above but with an adjustment for age by multiplying the participant's actual age by the coefficient for age when generating the predicted scores. To generate education-corrected scores, the same procedure was applied but with an adjustment for each participant's actual education level. To generate age- and education-corrected scores, subjects' actual age and education was used to generate predicted scores.

The National Alzheimer's Coordinating Center Uniform Data Set 3

NACC was established by a division of National Institutes of Health. NACC's ultimate goal is to provide a comprehensive approach to research on AD (Besser, 2018). To date, there are 39 present and past ADCs. In 2005, the ADCs began longitudinally collecting demographic, clinical, NP, and diagnostic data on the original version of the UDS (Morris et al., 2006). Version 2 was implemented in 2008, which represented a minor update to data collection elements including several new forms, restructuring the form logic, and adding a few NP test elements. The 3rd and most recent revision of the Uniform Data Set (UDS-3) represents a major advance, including a fully updated NP test

battery, additional supplemental data, and updated diagnostic criteria to reflect changes in how dementia syndromes are classified in current clinical practice. Data collection with the UDS-3 was implemented in March 2015. As of the most recent data freeze in March 2019 approximately 6, 266 individuals in the UDS 3 had completed the UDS-3 NAB. Further details regarding the study sample are available at: www.alz.washington.edu/WEB/UDS_NEUROonepage.pdf.

Each ADC enrolls subjects according to its own protocol. Subjects may come via clinician referral, self-referral by the patient or family members, active recruitment through community organizations, or by volunteering. In addition to patients with dementia and mild cognitive impairment, most centers also enroll normal control participants. As such, the NACC subjects are not a statistically representative sample of the U.S. Population. They are best described as a referral-based or volunteer case series. This renders UDS data inappropriate for studies of the prevalence or incidence of dementia. UDS data are collected via standardized evaluation of subjects enrolled in the ADCs. Data are generated using a standard order of administration for the NP tests, collected by trained clinician and clinic personnel, and diagnosis is made by either a consensus team of multiple practitioners or a single physician dependent upon the individual ADCs protocol. Subjects are seen for an initial visit and followed longitudinally with approximately annual visits until they can no longer participate or are lost to follow up. The complete UDS data-collection protocol is available at https://www.alz.washington.edu/WEB/qaqc_protocol.html.

Clinical Dementia Rating

The CDR is a dementia severity rating that is clinician administered, structured interview of a patient and a knowledgeable informant (often a family member) that can be conducted by a physician, nurse, social worker, or other trained staff member (Morris, 1997). Following the interview, clinicians rate a patient's functioning in the areas of memory, orientation, judgment and problem solving, function at home and hobbies, function in the community, and personal care. A global score can be calculated based on published scoring rules and used as a gross summary measure of dementia severity, but the CDR also allows clinicians to derive the CDR-SB which provides a more fine-grained, pseudo-continuous measure of dementia severity. Published criteria derived from large clinical samples exist for determining the presence or absence of dementia using either the global CDR score (Morris, 1993) or the CDR-SB (O'Bryant et al., 2008, 2010). The CDR was developed by John Morris and colleagues at the Washington University in St. Louis ADC in the 1980s (Hughes et al., 1982). It has since become the dominant global staging measure used clinically, in research, and as a primary end point in clinical trials. It has been translated into 14 different languages and was described as the "best-evidenced" measure in a recent review on global dementia severity staging measures (Olde-Rikkert et al., 2011). The CDR has been neuropathologically validated and demonstrates predictive accuracy of 92% for the presence of Alzheimer's pathology in symptomatic individuals with AD (Storandt et al., 2006). Studies indicate high interrater reliability for physicians and non-physicians applying the CDR (Williams et al.,

2013). Accordingly, the CDR was an ideal criterion measure for the purposes of the present study.

Craft Story 21

The trademark characteristic of AD is memory loss, so list learning and story memory tasks are frequently used in the episodic memory evaluation for dementia. Craft and colleagues designed a story recall test with multiple forms that achieved similar psychometric properties to the Weschler Logical Memory test, immediate and delayed recall conditions (Weintraub et al., 2018a). The complete set of stories consisted of 22 narratives that were originally tested on 13 healthy adults and 22 patients with Alzheimer's dementia and rated on the CDR as very mild, mild, moderate, and severe (Craft et al., 1996). Participants listened to a brief story with 25 bits of information and were asked to recall both immediately and after a 10-minute delay, receiving credit for each bit of data that was recalled verbatim or accurately paraphrased. Validity was determined by correlation with The Wechsler Memory Scale normative scores:

> Pearson *r*'s between Logical Memory and paragraph recall scores were 0.73, $p < 0.02$ (immediate recall) and 0.84, $p < 0.004$ (delayed recall) for normal adults and 0.76, $p < 0.0002$ (immediate recall) and 0.88, $p < 0.0001$ (delayed recall) for Alzheimer patients. Group mean and SD were nearly identical. (Craft et al., 1996, p. 126)

When the Neuropsychology Work Group committee convened to make recommendations for the UDS-NAB 3, their pilot study to determine the equivalence of the stories in middle-aged and older adults determined that 3 of the stories offered the greatest

relationship to the Logical Memory subtest of the Weschler Memory Scale and to each other, with a single story chosen because of its applicability to a culturally diverse population and "Craft Story 21" was adopted as the episodic memory measure for UDS-NAB 3 (Weintraub et al., 2018a). In this study, the composite score for the memory domain was composed of the scores from Craft Story 21 and the Benson Complex Figure recall task.

Benson Complex Figure

Asking a patient to copy a figure is the most common method of assessing visuospatial ability in dementia evaluations and having the patient recall the figure after a delay is considered a measure of nonverbal memory. These tasks are new addition to the UDS-NAB 3. Impairments in the visuospatial domain commonly appear in AD, bvFTD, and DLB. Complex figure copy tasks like the Rey-Osterrieth Complex Figure are influenced by visual spatial perception and attention, and also frontally mediated executive skills like organization, strategic planning, and working memory (Possin, Laluz, Alcantar, Miller & Kramer, 2011). The Rey Complex Figure Test was developed by Rey in 1941 and has a long history in neuropsychology (Strauss et al., 2006). Internal reliability was evaluated by split-half and alpha coefficients and achieved greater than .60 for copy trial, and greater than .80 for recall trial. Test-retest reliability examined (r = .76; r = .89) for immediate copy and delayed recall respectively (Strauss et al., 2006). Validity is also supported through independent correlational and factor analytic studies (Meyers & Meyers, 1995). The Benson Figure is a simplified variation of the Rey-Osterrieth figure

developed by Frank Benson (Possin et al., 2011) and was adopted as a measure of visuospatial ability and memory recall in the UDS-3 NAB.

Phonemic Fluency

Verbal fluency tests assess the spontaneous production of individual words under constrained conditions. Phonemic fluency tasks ("F" and "L") are commonly used in many NP batteries (Strauss et al., 2006). Originally developed as a measure of primal mental abilities by Thurstone in 1938, his variant of the word fluency test showed performance improved throughout childhood, peaked about age 30-39, and mildly declined into old age in normal cognition. This pattern was confirmed by subsequent research, and accumulating evidence showed that the test was highly useful for the detection of dementia because it is heavily dependent on the integrity of executive function (Strauss et al., 2006).

The letters F, A, and S are most commonly used, but C, F, and L are also used. The examinee is given the specified letter, in this case F and L, and orally produces as many words as possible in 1 minute. The total score is the sum of all correct words for both letters. Alternate form reliability observed correlations among phonemic fluency tasks high (.85 to .94) with differences between letter sets small (Strauss et al., 2006). For a detailed discussion of letter equivalence across different versions see Borkowski, Benton and Spreen (1967). Studies indicated test-retest correlations typically .70 or higher for letter and fluency at both short (2 week) and long (5 year) intervals (Strauss et al., 2006). Age and education corrected norms are published in manuals (Heaton et al., 2004), and can be statistically computed (Mitrushina et al., 2005; Tombaugh et al., 1999).

For the purpose of this study, verbal fluency F & L and TMT B were chosen as measures of the executive function domain.

Trail Making Tests B

The adult version of the Trail Making Tests (TMT) measures processing speed in people aged 15 to 89. Originally constructed in 1938 as a divided attention test, they were part of the Army Individual Test Battery and adapted by Reitan in 1955 (Strauss et al., 2006). Scoring is expressed in seconds required to complete the test with a maximum time on TMT B set at 300 seconds. Performance is affected by age, education, and IQ; with education becoming progressively more important with increasing age. Test-retest reliability was high in healthy controls (r = .89) for TMT B; but not uniformly reliable in clinical groups (r = .67 to r = .86); practice effects noted in healthy controls leveled off after 5 administrations (Strauss et al., 2006). Alternate form reliability reported a reliability coefficient of .92. Validity was demonstrated through correlations with other measures of executive processing speed including the Category Test; Wisconsin Card Sorting Test; Visual Search and Attention Test; Symbol Digit Modality Test; Paced Serial Addition Test; and Letter Cancellation that were moderate to strong (.36 to .93) with TMT B emerging as more sensitive to executive control (Strauss et al., 2006). Thus, TMT B was chosen to be included in the composite measure of executive control function.

The Multilingual Naming Test

Individuals with AD show deficits in naming speed and accuracy (dysnomia). The Boston Naming Test is one of the most common measures of confrontation naming

(Kaplan et al., 1983) and has been shown to discriminate well between cognitively normal participants and those with dementia (Katsumata et al., 2015). The MINT was specifically designed to be a culturally sensitive measure of picture naming for English, Spanish, Mandarin Chinese, and Hebrew. Sixty-eight black-and-white line drawings, selected from a variety of sources with translation equivalents in each different language, are presented in increasing difficulty of order. Ivanova et al. (2013) found the MINT was highly correlated with the BNT, ranging from $r = .855$ to $r = .893$, $p < 0.001$, and suggested that it had more utility for diagnostic purposes as the BNT was biased in favor of English. Ivanova et al. (2013) established that a 32-item subset of the MINT had adequate sensitivity and provided superior clinical utility because of its contextual diversity to detect naming impairments in AD and controls.

The NP Work Group replaced the Boston Naming Test (BNT) with the MINT for the UDS-NAB 3 (Weintraub et al., 2018a). This study used the MINT and Animal Fluency as the measures for the Language domain.

Animal Fluency

The most common semantic fluency test requires an individual to name as many animals as possible in 1-minute, other categories such as fruits and vegetables or "things to wear" are also used (Strauss et al., 2006). Norms were derived from large samples of participants that ranged in age from 20 to 101 years of age depending on the study; (n = 1148) in Heaton et al. (2004); (n = 2843) in Mitrushina et al. (2005) and (n = 735) in Tombaugh et al. (1999) to name just a few. Tombaugh et al. (1999) found the degree of internal consistency high (r = .83) and test-retest reliability coefficient of .74 after a 5-

year interval in elderly individuals. Studies showed test-retest correlations were typically .70 or higher for semantic fluency at both short (2 week) and long (5 year) intervals (Strauss et al., 2006). Like phonemic fluency tasks, age and education corrected norms are located in published manuals (e.g., Heaton et al., 2004) and can be statistically computed (Mitrushina et al., 2005; Tombaugh et al., 1999).

Digits Forward and Digits Backward

DGF assess attentional capacity using auditory digit repetition, a common method used in most existing tests for this purpose. DGB requires the examinee to reverse orally presented digits as a measure for both attentional capacity and working memory. The number strings are administered up to failure of two trials at the same length with points earned for each completed sequence and, in some cases, a note for the longest digit span completed. Digit span tasks are modeled after the Weschler Memory Scale III (WMS), which was originally designed to assess auditory attention and working memory (Strauss, 2006). The number spans for the UDS-NAB 3 were randomly generated with the restriction that no digit would be adjacent to the next higher or lower digit, and efforts made to avoid recognizable sequences such as common area codes (Weintraub et al., 2018a). Digit span generalizability coefficients to the WMS were high (.80 to .89), and while ‘clinical lore’ espouses that DGB is more demanding of working memory than DGF and more sensitive to advancing age and neurodegenerative conditions, the most recent findings suggest that both DGF and DGB are affected equally as one ages, although large discrepancies between the two tasks may point to a deficiency in working

memory (Strauss et al., 2006). DGF and DGB served as our measures of the attention domain.

Data Analysis Plan

The following section describes the analytic strategies that will be used to answer the primary study hypothesis, data cleaning (e.g., handling of outliers and missing values), construction of hierarchical linear regressions to test primary study hypotheses, and an *a priori* power analysis. Prior to that discussion, it may be helpful to restate the series of regression analyses that are planned, and the primary study hypothesis and sub-hypotheses. This study involved predicting the CDR-SB score by a combination of cognitive composite scores derived from cognitive tests scored using 4 different methods of demographic correction (i.e., no correction, age correction, education correction, age and education correction). The following prototype model was built 4 different times (once for each of the norming methods) and then those models were compared using R^2 and AIC values: CDR-SB scores =

β_1Memory and Learning + β_2 Executive Function + β_3 Language + β_4 Attention

Four sets of regressions were used to examine:

1. The relationship between CDR-SB and the “raw” (i.e., corrected for gender only) cognitive composite scores.
2. The relationship between the CDR-SB and cognitive composite scores derived from age corrected test data.
3. The relationship between the CDR-SB and cognitive composite scores derived from education corrected test data.

4. The relationship between the CDR-SB and cognitive composite scores derived from age and education corrected test data.

There was one model for each norming method as specified above, so determining which model was "best" required comparing the merits of the various models. This represented a statistical conundrum, because the models were built using different data (data subject to different norming methods) and as such are "non-nested" models. In the context of least squares regression, nested models are easily testable for differences at a given significance level but there are no universally agreed upon methods for comparing non-nested models. In practice, various parameters such as a model's R^2 value and AIC are typically used for model selection purposes in lieu of significance tests. For consideration, some authors have convincingly argued that null hypothesis testing itself is undesirable and gives a false sense of precision when none is warranted and as such, the lack of significance testing in this project is not viewed as a particular shortcoming (Harrel, 2001).

R^2 values represent the total amount of variability in the criterion variable, in this study the CDR-SB, accounted for by the predictor variables, in this case the cognitive composite scores (Field, 2013). Thus, models with a higher R^2 value are desirable and indicate that the predictor variables are capturing more variation in the criterion of interest than models with a lower R^2 value. Cohen (1988) has set conventions for interpreting R^2 values that are used widely in the social sciences, with $R^2 = .02$ considered a small effect, $R^2 = .25$ considered a medium effect, and $R^2 = .40$ or greater considered to be a large effect. Thus, there would be clear and meaningful differences between the

models if the R^2 values associated with the different models fell into different effect size categorizations. For example, if the R^2 value in equation 2 was moderate ($R^2 > .25$) and the R^2 value in equation 1 was strong ($R^2 > .4$), then model 1 is clearly superior to model 2. Differences that are smaller in magnitude may suggest the superiority of one model versus another model, but there are no universally accepted criteria for determining the incremental difference required to make a clinically significant contribution to clinical practice. In the context of dementia evaluation, it can be argued that any increase in R^2 value that might lead to greater diagnostic accuracy is desirable and may be of practical significance at a population level.

AIC was developed by Hirotogu Akaike in 1974 (Akaike, 1974) and is typically employed in logistic regression but can also be computed for least squares regression. This parameter is used commonly for model selection and considers both a model's overall fit and its parsimony. While AIC is often used to compare nested models, Akaike's work makes no statement that models must be nested, and thus this statistic is often used to compare non-nested models. One calculates the lowest AIC value of all models being considered and then evaluates the change in AIC between different models (the delta AIC). According to Burnham and Anderson (2004), models having a delta AIC less than or equal to 2 have substantial support and are comparable to the model with the minimum AIC value. Those in which $4 \leq$ delta AIC ≤ 7 can be said to have considerably less support, and those where delta AIC > 10 have essentially no support. Therefore, the models above can be compared semi-quantitatively using their AIC values with the model having the lowest AIC value being the most desirable.

Primary Study Hypothesis

H_1: Age and education correction increase the extent to which NP tests are able to capture functional decline due to cognitive loss in dementia. This will be tested by comparing R^2 and AIC values from regression equation 1 with regression equation 4.

$H_0$1: Age and education correction decrease or have no effect on the extent to which NP tests are able to capture functional decline due to cognitive loss in dementia.

H_2: Age correction decreases the extent to which NP tests are able to capture functional decline due to cognitive loss in dementia. This will be tested by comparing R^2 and AIC values from regression equation 1 with regression equation 2.

$H_0$2: Age correction increases or has no effect on the extent to which NP tests are able to capture functional decline due to cognitive loss in dementia.

H_3: Education correction increases the extent to which NP tests are able to capture functional decline due to cognitive loss in dementia. This will be tested by comparing R^2 and AIC values from regression equation 1 with regression equation 3.

$H_0$3: Education correction decreases or has no effect on the extent to which NP tests are able to capture functional decline due to cognitive loss in dementia.

Optimal demographic correction as determined by the analyses above may result in better diagnostic performance when differentiating normal controls from individuals with dementia. Comparison of the R^2 and AIC values of the various regression equations presented above will clarify the extent to which various types of demographic correction influence the ability of cognitive tests to detect meaningful variation in dementia severity. While the above hypotheses are not associated with significance levels, R^2 values and

AIC values are commonly used to compare non-nested models and allow the important research questions above to be answered quantitatively.

Data Preparation

The plan was to identify missing values and either exclude them on a case-wise basis during analyses or replace them with the group-wise mean value (i.e., a missing value for a patient with mild dementia would be replaced using the mean of that variable for individuals with mild dementia). The strategy used depended on the number of missing data points and other specific data characteristics that were not known prior to receiving the data set. Primary tests of study hypotheses may further benefit from exclusion of outliers and transformation of poorly behaved variables that deviate markedly from normality. It should be noted that linear regression does not explicitly require multivariate normality (Allison, 1999). As with most parametric models however, relative normality prior to model fitting may aid in model stability, generalization, and avoiding violations of other assumptions downstream (Tabachnick & Fidell, 2012).

To assess normality, histograms and Q-Q plots were constructed for each of the cognitive tests for visual inspection (see appendix). Skew and kurtosis values were calculated. In general, skew and kurtosis values < 1 are considered acceptable, particularly given that multiple linear regression does not require normally distributed variables (George & Mallery, 2016). Parametric methods such as the Komolgorov-Smirnoff test were not appropriate for the present study because with the large sample size, they would be overpowered and likely detect tiny departures from normality. Similarly, it was also inadvisable to perform statistical tests of skew and kurtosis values

by dividing them by their standard errors and then comparing them to a Z-score distribution (Field, 2013). Data transformations were considered and applied as needed with the caveat that model interpretability was a primary concern and thus variable transformation was avoided if at all possible, particularly for the CDR-SB. Univariate outliers were identified through the use of Boxplots, with values plus or minus 3 times the interquartile range of a variable screened as potential outliers (Field, 2013). Multivariate outliers were identified within the regression analyses through several different methods described below.

Statistical Analytic Strategies

Prior to modeling, the data was described by calculating descriptive statistics including the mean, standard deviation, and range of all study variables. The intercorrelations of the cognitive test scores, age, education, and the CDR-SB were calculated for each different norming method using bivariate Pearson correlations. These correlations provided a direct measure of the strength of relationship between a given cognitive test and the criterion measure of interest. These bivariate correlations aided in the interpretation of the hierarchical linear regression analyses for testing primary study hypotheses, helped to detect suppression, and served as a measure of variable importance (Nathans et al., 2012). Following descriptive statistics and analysis of intercorrelations between tests, the primary study hypotheses was tested using least squares hierarchical multiple linear regression. This analytic technique is appropriate when one wishes to examine the magnitude of association between a set of continuous predictor variables and a continuous outcome variable (Field, 2013).

The order of variable entry was defined *a priori*, which avoided the pitfalls inherent in methods such as stepwise regression and promoted better generalization as opposed to overfitting of sample data (Harrell, 2001). The cognitive composite scores were entered in the following order in a block wise fashion using an alpha of 0.5 to enter the model: memory, executive functioning, language, and attention. It was hypothesized that memory and executive function would be the most important predictors in the model irrespective of norming method followed by language, then attention. The models were evaluated based on R^2 values and adjusted R^2. Models were compared by examining R^2 differences between models and AIC differences. The importance of individual cognitive domains were evaluated by examining R^2 change, standardized regression weights associated with each cognitive domain, and significance of terms in the final models. The model construction strategy allowed for a precise determination of the different proportions of variance accounted for by each cognitive domain and how the magnitude of those relationships was affected by demographic correction.

Regression requires several assumptions: linearity of the relationship between predictor and criterion variables, absence of multicollinearity, absence of outliers amongst the independent and dependent variables, independence of errors, and homoscedasticity (Tabachnick & Fiddell, 2012). Linearity was examined visually using scatterplots and bivariate Pearson correlations described above. Multicollinearity was addressed *a priori* by design, through creating homogeneous cognitive composites that were relatively distinct from one another by virtue of their item content. Bivariate Pearson correlations between the test variables allowed for quantitative evaluation of any

multicollinearity between variables. In general, variables with correlations above 0.7 – 0.8 should not be used together in a regression equation (Allison, 1999). Variance Inflation Factor (VIF) and tolerance values calculated as a result of the regressions complemented these *a priori* methods of detecting multicollinearity. In general, VIF above 10 and Tolerance values below 0.2 are considered as representing possible problems (Field, 2013). Outliers and influential data points were evaluated using regression diagnostics.

In a multiple regression context, outliers can be defined as points that differ substantially from the main trend of the data (Field, 2013). Examination of standardized and studentized residuals following model fitting was used to identify such points which were further inspected for possible removal. Single data points with standardized residuals above or below 3.29 were considered for removal, as values this high are unlikely to occur based on chance (Field, 2013). The proportion of cases with standardized residuals greater or less than 1.96 was also examined, because 95% of data points should fall within these values in a well-fitting model. Multivariate outliers were also evaluated by calculating their distance from the group centroid via Mahalanobis distance, which is distributed as a Chi-square and can be evaluated at $p < .001$ to detect multivariate outliers (Meyers et al., 2016).

Outlying data points may not be a large concern if they are not overly influential on the overall model. To assess influence, leverage values were examined. Leverage gauges the influence of the observed value of a case on a particular variable over the predicted values of a regression solution. The average expected leverage is defined as

(k+1/n) with cutoff values around 3 times the average leverage value typically considered as indicative of points with undue influence (Pituch & Stevens, 2015). This statistic was complemented by calculation of studentized deleted residuals, which represented the difference between the prediction of an observed value when it is and when it is not included in the model divided by its standard deviation (Field, 2013). Influence on the overall model was evaluated using Cook's distance, with values greater than 1 potentially indicating cause for concern (Cook & Weisberg, 1982).

Independence of errors or lack of autocorrelation amongst residuals was assessed using the Durbin-Watson statistic which tested for serial correlations between adjacent residuals. This statistic ranges from 0 to 4, with values greater than 2 implying negative correlation and below 2 implying positive correlation. Rule of thumb suggested by Field (2013) for evaluating this statistic are that values less than 1 or greater than 3 are indicative of a possible problem. Homoscedasticity was evaluated by plotting standardized residuals vs. standardized predicted values, which should ideally assume a random pattern. Funnel shaped plots may suggest heteroscedasticity and a curved appearance may indicate departures from normality (Field, 2013). Histograms of residual values and normal p-p plots were also calculated to examine normality of residuals.

Power Analysis

Power for multiple regression includes tests of the overall model being significantly different than 0, tests of R^2 increase at each step in variable entry, and tests for the significance of individual regression coefficients. G*Power 3.1.9.4 was used to examine power for each of these tests assuming a moderate effect size ($f^2 = 0.15$) and $\alpha =$

.05 at the desired power level of 0.8, as recommended by Cohen (1988) (Faul et al., 2007). A total sample size of $n = 85$ was required to achieve power of 0.803 for tests of the R^2 deviation from 0 and the same sample size was required for tests of R^2 increase. A total sample of 55 was required for tests of model coefficients.

Summary

The current investigation examined the impact of demographic correction on the diagnostic validity of cognitive tests in the service of differentiating between normal cognitive aging, MCI, and dementia. The proposed study challenges the standard practice of demographic correction thought to universally improve cognitive testing instruments' sensitivity to impairment. The NACC appointed a specific task force to choose the set of measures that are freely available and have adequate discriminatory powers to encourage uniform data collection strategies and collaboration between researchers. The resulting data set, the UDS Version 3, was the focus of the current study because of its size and diversity. It contains healthy control participants as well as those meeting the diagnostic criteria for MCI and Alzheimer's dementia. The results of this study may not only tip the debate towards an optimum scoring method to detect the earliest stages of degenerative cognitive impairment, but it may also advise the illusive search for a set of cognitive biomarkers that distinguish individuals who are experiencing the effects of normal aging from those in the process of developing a neurodegenerative disease. The earlier and more accurately we can diagnose Alzheimer's, the greater chance we have of altering the disease trajectory, potentially improving the future for the many individuals that are yet to be diagnosed and other stakeholders.

Chapter 4: Results

The purpose of the present study was to investigate the impact of demographic corrections on the diagnostic validity of cognitive tests when differentiating between normal cognitive aging and AD dementia. The standard practice of demographically correcting raw test scores for age and education is widely believed to universally improve the scores' sensitivity to detecting cognitive impairment for all age groups but has been challenged by researchers who believe that uncorrected scores may be more sensitive for detecting impairment in the aging population (Hassenstab et al., 2016; Hessler et al., 2014; Holtzer et al., 2008; O'Connell & Tuokko, 2010; Wyman-Chick et al., 2018) . Hierarchical multiple linear regression was used to predict the outcome variable, the CDR-SB scores, by a combination of cognitive composite scores derived from cognitive tests scored using 4 different methods of demographic correction.

The cognitive domain composite scores for memory, executive function, language, and attention were the predictor variables. The composite construction strategy was based on the same theoretical grouping of tests used in the work of Weintraub et al. (2018a). This theoretically driven strategy is well grounded in established principles of NP function and localization. Weintraub and colleagues (2018b) developed a normative calculator for the UDS-NAB using test data of 3602 cognitively normal participants over the age of 60. This model was used to standardize the test scores and adjust for the various demographic corrections that were compared in the regression analysis. The composite score for the memory domain (MEMO) was composed of the scores from 2 tests, Craft Story 21 and the Benson Complex Figure recall task. The executive function

domain (EXEC) was constructed from verbal fluency F & L and TMT B tests. The language domain (LANG) was derived from the MINT and Animal Fluency measures. And the attention domain composite score (ATTN) was composed of DGF and DGB tests. For a full discussion of the reliability and validity of each individual NP test, see chapter 3. Cognitive composites offer a more complete sampling of cognitive abilities than a single cognitive test by virtue of broader item content and are advantageous statistically as they can be more sensitive to cognitive changes in preclinical dementia therefore more reliable than the single measures from which they were derived. These 4 domains were chosen because models with fewer explanatory variables are more desirable than complex models due to enhanced interpretability (James et al., 2017). For more details on how each of the demographic scores were computed see the full discussion in chapter 3 under the subheadings “Cognitive Test Scoring” and “Creation of Cognitive Composites”.

The order of entry into hierarchical regression models was defined *a priori*. Memory measures have been shown to be the earliest indicators of decline in AD even in minimally symptomatic individuals, and executive function impairment usually follows (Karantzoulis & Galvin, 2011). Deficits in language functioning, including semantic verbal fluency and visual object confrontation are also characteristic of AD (Salmon & Bondi, 2009). Individuals with AD may also exhibit impairment with attention and working memory though these abilities may be well preserved until later into the illness (Cherry et al., 2002). The following prototype model was built 4 different times, once for each norming method:

$$\text{CDR-SB scores} = \beta_1 \text{MEMO} + \beta_2 \text{EXEC} + \beta_3 \text{LANG} + \beta_4 \text{ATTN}$$

The first model was corrected for gender only (G) and considered the raw scores for the purpose of the analysis. The second model was age corrected only (GA). The third model was education corrected only (GE). The fourth model was corrected with a combination of both age and education (GEA). Therefore, the abbreviated combination of MEMO_G, is the score for the memory composite score corrected only for gender (raw score), while MEMO_GEA is the memory composite score corrected for gender, education, and age. Four sets of regressions will be used to examine:

1. The relationship between CDR-SB and the G cognitive composite scores.
2. The relationship between the CDR-SB and GA cognitive composite scores.
3. The relationship between the CDR-SB and GE cognitive composite scores.
4. The relationship between the CDR-SB and GEA cognitive composite scores.

Models were compared by examining R^2 differences between the models and AIC values. This strategy clarified the extent to which the raw scores and various demographic corrections influenced the ability of the cognitive tests to detect meaningful variation in functional changes due to cognitive loss from AD dementia. The importance of individual cognitive domains was evaluated by examining R^2 change, standardized regression weights associated with each cognitive domain, and significance of terms in the final models. It also allowed for the precise determination of the different proportions of variance accounted for by each cognitive domain with each norming method. Determining which model was “best” required comparing the merits of the various models because there was one model for each norming method as specified above.

The overarching research question was how demographic corrections affected the strength of the relationship between cognitive test scores and CDR-SB. The primary study hypothesis was that age and education correction would increase the extent to which NP test scores were able to capture functional decline due to cognitive loss in dementia. There were also 2 subhypotheses 1) age correction alone would decrease the extent to which NP tests were able to capture functional decline due to cognitive loss in dementia, and 2) education correction alone would increase the extent to which NP tests were able to capture functional decline due to cognitive loss.

Chapter 4 opens with data collection and participant selection information that includes a review of the inclusion and exclusion parameters that determined the final sample size. Descriptive and demographic characteristics of the final sample extracted from the entire data set provided by NACC. A discussion of the statistical assumptions necessary for the analyses, and a complete report of the findings organized by the research questions, including tables that best illustrated the results of the analyses and effect sizes. The chapter closes with a precise summary of hypotheses testing.

Data Collection

The data set was obtained from the University of Washington's NACC by submitting the abstract from the present study and signing a data use agreement. To date, the NACC coordinated the collection of longitudinal data on 967 different variables for more than 100,000 participants. This study was limited to only data gathered on the most recent version of the UDS-3 implemented in March 2015 because it reflected changes in how dementia syndromes are classified in current clinical practice with an updated

version of the UDS-3 NAB that allowed researchers to collaborate using freely-available standardized testing instruments. For a complete discussion on how each ADC enrolls subjects and gathers data, see chapter 3 “The National Alzheimer’s Coordinating Center Uniform Data Set 3”. NACC subjects are best described as a referral-based or volunteer case series and thus are not a statistically representative sample of the population. The complete UDS-3 NAB data included a robust group of individuals with normal cognition as well as those with various etiologies of neurodegenerative diseases. The primary goal of this study was to inform the use of NP tests to differentiate between normal cognitive aging and dementia in the most commonly encountered situations, so the subjects were filtered to include only a primary etiologic diagnosis of AD in people 60 years and older with a CDR-SB score < 9.5. This cutoff excluded individuals with moderate to severe dementia for reasons fully justified in chapter 3 under “Participant Selection and Stratification”.

Sample Descriptives

The case processing summary showed no missing data. From the more than 100,000 participants in the full data set, 8724 subjects met all inclusion and exclusion criteria (n = 8724), 5192 females and 3532 males. They ranged in age from 60 to 101 with an average age of 74. The majority earned a bachelor’s degree with a range of formal education from 9 years to 21 years. Table 1 contains the descriptive statistics for all study variables. The frequencies tables were visually inspected for anomalies, these values all fell within expected ranges.. Table 2 used the outcome measure, to classify the subjects according to their CDR-SB scores into categories of Normal Cognition, MCI, or

Demented. The majority of subjects, 6237, fell into the normal cognition range, 1505 had a diagnosis of MCI, and 982 subjects met the criteria for dementia.

Table 1

Descriptive Statistics for All Variables

Variable	Mean	SD	Min	Max	Range	IQR	Skew	Kurtosis
AGE	74.12	7.87	60	101	41	12	.32	- .46
SEX	.60	.49	0	1	1	1	-.39	-1.85
EDU	16.24	2.51	9	21	12	4	-.32	.20
CDR-SB	.83	1.68	0	9	9	.50	2.53	6.30
MEMO_GEA	-.4595	1.13	-3.83	2.78	6.62	1.47	-.56	-.34
MEMO_GA	-.4635	1.14	-3.69	2.52	6.21	1.48	-.59	-.20
MEMO_GE	-.4592	1.16	-3.56	2.69	6.25	1.51	-.53	-.25
MEMO_G	-.4632	1.17	-3.37	2.43	5.80	1.52	-.52	-.29
EXEC_GEA	-.3411	1.08	-4.59	2.48	7.07	1.12	-1.07	1.36
EXEC_GA	-.3482	1.12	-4.40	2.46	6.86	1.20	-1.07	1.24
EXEC_GE	-.3408	1.11	-4.45	2.59	7.04	1.17	-1.06	1.22
EXEC_G	-.3479	1.15	-4.28	2.56	6.83	1.23	-1.06	1.10
LANG_GEA	-.3459	1.14	-9.09	2.95	12.04	1.24	-1.70	6.16
LANG_GA	-.3528	1.17	-8.91	3.03	11.94	1.29	-1.58	5.36
LANG_GE	-.3457	1.17	-8.88	3.00	11.88	1.27	-1.63	5.66
LANG_G	-.3525	1.20	-8.80	3.01	11.80	1.33	-1.51	4.94
ATTN_GEA	-.1234	.88	-3.32	2.93	6.26	1.18	.27	.05
ATTN_GA	-.1282	.90	-3.39	3.01	6.40	1.20	.28	.05
ATTN_GE	-.1233	.89	-3.41	3.01	6.42	1.20	.28	.04
ATTN_G	-.1280	.91	-3.52	2.83	6.35	1.16	.30	.04

Note. n = 8724. n = 5192 females. n = 3532 males. Memory composite score (MEMO). Executive function composite score (EXEC). Language composite score (LANG). Attention composite score (ATTN). Scores corrected for gender, education, and age (GEA). Scores corrected for gender, and age (GA). Scores corrected for gender, and education (GE). Scores corrected for gender only, also considered as the "raw" score (G).

Table 2

Descriptive Statistics Grouped by Dementia Severity Level

	Normal N = 6237			MCI N = 1505			Demented N = 982		
	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
CDR-SB	.10	.30	2.0	1.23	.90	5.12	4.90	1.70	6.50
MEMO_GEA	-.00	.78	5.98	-1.15	.94	5.12	-2.31	.81	4.69
MEMO_GA	-.00	.79	5.85	-1.15	.94	5.12	-2.34	.81	4.28
MEMO_GE	-.01	.81	5.71	-1.18	.94	5.32	-2.36	.80	4.23
MEMO_G	-.02	.82	5.59	-1.19	.95	5.00	-2.39	.80	5.80
EXEC_GEA	-.32	.80	6.32	-.71	1.06	6.72	-1.74	1.35	5.92
EXEC_GA	.01	.88	5.89	-.72	1.10	6.24	-1.80	1.36	6.40
EXEC_GE	-.01	.83	6.35	-.74	1.08	6.44	-1.80	1.36	6.34
EXEC_G	-.01	.87	6.71	-.76	1.15	5.99	-1.85	1.36	6.83
LANG_GEA	-.00	.80	7.93	-.84	1.07	10.80	-1.79	1.59	10.56
LANG _GA	.00	.84	8.35	-.85	1.08	11.01	-1.84	1.58	10.13
LANG _GE	.01	.82	8.37	-.87	1.09	10.63	-1.83	1.60	10.37
LANG _G	.02	.86	8.72	-.88	1.10	10.84	-1.88	1.59	11.80
ATTN_GEA	.01	.86	5.64	-.32	.82	5.83	-.65	.84	5.92
ATTN _GA	.01	.88	5.89	-.33	.83	5.83	-.69	.84	6.04
ATTN _GE	.02	.87	5.68	-.34	.82	5.90	-.68	.83	5.97
ATTN _G	.02	.89	5.66	-.35	.83	5.89	-.72	.83	6.35

Note. n = 8724. n = 5192 females. n = 3532 males.

Assumption Testing

A preliminary regression analyses was run to look for outliers identified using standardized and studentized residuals following model fitting. Single data points with residuals above or below 3.29 were considered for removal as they were unlikely to occur by chance (Field, 2013). The proportion of cases with standardized residuals greater or less than 1.96 were also examined because it was desired that 95% of data points fell within these values for a well-fitting model. Distances were calculated using Mahalanobis, Cook's, and Leverage values. Using a Chi-square table P = .001 and 4df,

the cutoff was 18.47. Cook's distance values over .00045877 could have been considered outliers and leverage values over .00114626 could also been considered for removal. No data violated all three markers and the data set was kept in its entirety without fear that outliers biased results because of the large size of the data set. Tests to see if the data met the assumption of collinearity indicated that it was not a concern. VIF and Tolerance values were examined because VIF > 10 and Tolerance < .02 could be problematic, none of the values violated those boundaries. The data also met the assumption of independent errors, Durbin Watson values that were all close to 2 and are included in Table 7.

Histograms and Q – Q plots were visually inspected (see Appendix). The histograms confirmed normality of the variables. All test score variables were roughly centered over 0 and the majority of the scores fell between – 2 and 2. Skew and kurtosis values are included in Table 1, values < 1 are generally considered acceptable but multiple linear regression does not require normally distributed variables (George & Mallery, 2016). The slight negative skew for the test variables was expected as individuals with impaired performance perform below normal. Q – Q plots were inspected for linearity and observed values adhered well to the line of best fit (expected values), they were not completely on the line, but close, so there were no obvious violations of this assumption. The departure from normality appeared more significant as the values of the CDR-SB increased (a higher dementia severity) which followed logic that the CDR-SB is more accurate at predicting normal cognition and less accurate at predicting demented people (see Appendix A). Homogeneity and homoscedasticity were examined using boxplots. The length of the boxes and their "whiskers" had

approximately the same spread so there were no obvious violations of this assumption. The data also met the assumption of non-zero variances as all values were greater than 0. Pearson correlations also showed no violations of collinearity between the outcome measure, the CDR-SB, as all values were less than .7. Table 3 contains the correlations for GEA model; Table 4 shows the GA model correlations; Table 5 shows the correlations for the GE model, and; Table 6 shows the correlations for the G (raw) model. The correlations have a negative relationship as expected. The lower the test scores, the higher the CDR-SB score, in other words people performed worse on the tests as their dementia severity increased.

Table 3

Pearson Correlations for Gender, Age, and Education Corrected Scores

	CDRSUM	MEMO_GEA	EXEC_GEA	ATTN_GEA	LANG_GEA
CDRSUM	1	-.645**	-.507**	-.240**	-.512**
MEMO_GEA	-.645**	1	.483**	.287**	.559**
EXEC_GEA	-.507**	.483**	1	.463**	.564**
ATTENT_GEA	-.240**	.287**	.463**	1	.305**
LANG_GEA	-.512**	.559**	.564**	.305**	1

Note. N = 8724. ** $p < .001$ (2-tailed).

Table 4

Pearson Correlations for Gender and Age Corrected Scores

	CDRSUM	MEMO_GA	EXEC_GA	ATTN_GA	LANG_GA
CDRSUM	1	-.645**	-.505**	-.248**	-.513**
MEMO_GA	-.646**	1	.501**	.309**	.574**
EXEC_GA	-.505**	.501**	1	.490**	.589**
ATTENT_GA	-.248**	.309**	.490**	1	.335**
LANG_GA	-.513**	.574**	.589**	.335**	1

Note. N = 8724. ** $p < .001$ (2-tailed).

Table 5

Pearson Correlations for Gender and Education Corrected Scores

	CDRSUM	MEMO_GE	EXEC_GE	ATTN_GE	LANG_GE
CDRSUM	1	-.651**	-.517**	-.255**	-.519**
MEMO_GE	-.651**	1	.511**	.308**	.582**
EXEC_GE	-.517**	.511**	1	.477**	.588**
ATTENT_GE	-.255**	.308**	.477**	1	.325**
LANG_GE	-.519**	.582**	.588**	.325**	1

Note. N = 8724. ** $p < .001$ (2-tailed).

Table 6

Pearson Correlations for Raw Scores

	CDRSUM	MEMO_G	EXEC_G	ATTN_G	LANG_G
CDRSUM	1	-.651**	-.515**	-.263**	-.520**
MEMO_G	-.651**	1	.528**	.329**	.595**
EXEC_G	-.515**	.528**	1	.503**	.611**
ATTENT_G	-.263**	.329**	.503**	1	.354**
LANG_G	-.520**	.595**	.611**	.354**	1

Note. N = 8724. ** $p < .001$ (2-tailed).

The bivariate correlations also supported the order of entry into the hierarchical regression equation for the testing the main hypotheses; memory composite scores showed the strongest relationship to the CDR-SB ($\alpha = -.65$, $p < .001$) so entering memory scores into the model first was supported; executive function and language composite scores both demonstrated moderate relationships with the CDR-SB ($\alpha = -.51$, $p < .001$) so they were entered second and third respectively, and; the attention composite scores

showed a weak yet still statistically significant relationship ($\alpha = .25$, $p < .001$) to the CDR – SB so entering these scores into the regression equation last was also supported.

Regressions to Predict Dementia Severity Rating by Norming Method

Hierarchical multiple regression was conducted to predict CDR-SB, a clinical measure of functional changes due to cognitive loss, from the memory, executive function, language, and attention test composite scores that were normed by 4 different methods. R^2 and AIC values were compared to analyze the extent to which the raw scores and demographically corrected scores influenced the ability of the tests to capture meaningful variation in the CDR-SB. See Table 7 for an overall summary of the significance of each of the models by norming method. The individual predictor variable results for each norming method are reported in Table 8.

The first hierarchical multiple regression was conducted to predict CDR-SB scores from memory scores, executive function scores, language scores, and attention scores adjusted for gender, age, and education. Model 1 showed memory test scores accounted for 41.6% of the CDR-SB variability, $R^2 = .416$, $F(1, 8722) = 6213.10$, $p < .001$. Model 2 showed executive function scores accounted for an additional 5% of the variability in the CDR-SB, $\Delta F(1, 8721) = 809.94$, $p < .001$, $\Delta R^2 = .05$. Model 3 showed that language had a smaller yet still significant effect by accounting for 1% of the variability in the CDR-SB, $\Delta F(1, 8720) = 169.06$, $p < .001$, $\Delta R^2 = .01$. Model 4 showed that attention scores were again significant even though they only captured .1% of the variance in the CDR-SB, $\Delta F(1, 8719) = 19.69$, $p < .001$, $\Delta R^2 = .001$. The complete model captured 47.7% of the variability in the CDR-SB, $R^2 = .477$, $p < .001$.

Further analysis demonstrated the significance of each predictor. Memory accounted for 42% of the CDR-SB score, β = -.65, $t(8722) = -78.82$, $p < .001$, $pr^2 = .42$. In other words, for every one unit decrease in memory scores there was an increase in dementia severity rating by .65 SD. In the second model executive function accounted for an additional 9% of the CDR-SB when controlling for memory, β = -.25, $t(8721) = -28.46$, $p < .001$, $pr^2 = .09$. In the third model language accounted for another 2% of the CDR-SB score over and above memory and executive function scores, β = -.13, $t(8720) = -13$, $p < .001$, $pr^2 = .02$. And in the fourth model attention was again a significant predictor, but only accounted for .22% of the CDR-SB score, β = .04, $t(8719) = 4.44$, $p < .001$, $pr^2 = .0022$.

A second hierarchical multiple regression was conducted to predict CDR-SB scores from memory test scores, executive function test scores, language test scores, and attention test scores that were adjusted for gender and age. Memory test scores accounted for 41.7% of the CDR-SB variability, $F(1, 8722) = 6235.38$, $p < .001$, $R^2 = .417$. Model 2 showed executive function scores accounted for an additional 4% of the variability in the CDR-SB, $\Delta F(1, 8721) = 809.94$, $p < .001$, $\Delta R^2 = .04$. Model 3 showed that language had a smaller yet still significant effect by accounting for .09% of the variability in the CDR-SB, $\Delta F(1, 8720) = 145.83$, $p < .001$, $\Delta R^2 = .009$. Model 4 showed that attention scores were again significant even though they only captured .2% of the variance in the CDR-SB, $\Delta F(1, 8719) = 27.08$, $p < .001$, $\Delta R^2 = .002$. The total model accounted for 47.1% of the variability in the CDR-SB score, $R^2 = .471$, $p < .001$.

The significance of the predictors for this model showed memory to have the same value as in the model corrected for gender, education, and age, β = -.65, $t(8722) = -78.96$, $p < .001$, $pr^2 = .42$. Executive function accounted for another 8% of the CDR-SB score when holding memory constant, β = -.24, $t(8721) = -26.62$, $p < .001$, $pr^2 = .08$. Language picked up another 2% of the CDR-SB score when controlling for memory and executive function scores, β = -.13, $t(8720) = -12.08$, $p < .001$, $pr^2 = .02$. Attention, although statistically significant, only accounted for an additional .31% of the CDR-SB score over and above the memory, executive function, and language scores, β = .05, $t(8719) = 5.20$, $p < .001$, $pr^2 = .0031$.

The third hierarchical multiple regression was conducted to predict CDR-SB scores from memory test scores, executive function test scores, language test scores, and attention test scores that were adjusted for gender and education. Memory test scores accounted for 42.3% of the CDR-SB variability, $F(1, 8722) = 6401.38$, $p < .001$, $R^2 = .423$. Executive function scores accounted for an additional 4.6% of the variability in the CDR-SB, $\Delta F(1, 8721) = 755.64$, $p < .001$, $\Delta R^2 = .046$. Language had a smaller yet still significant effect by accounting for .8% of the variability in the CDR-SB, $\Delta F(1, 8720) = 139.86$, $p < .001$, $\Delta R^2 = .008$. Lastly, attention scores were again significant even though they only captured .1% of the variance in the CDR-SB, $\Delta F(1, 8719) = 16.18$, $p < .001$, $\Delta R^2 = .001$. The total model accounted for 47.9% of the variability in the CDR-SB scores, $R^2 = .479$, $p < .001$.

The significance for each predictor adjusted for gender and education showed memory again accounting for 42% of the CDR-SB score, β = -.65, $t(8722) = -80.01$, $p <$

.001, pr^2 = .42. Executive function accounted for an additional 8% of the CDR-SB score over and above memory, β = -.25, $t(8721) = -27.49$, $p < .001$, pr^2 = .08. Language picked up an addition 2% of the CDR-SB score, β = -.12, $t(8720) = -11.83$, $p < .001$, pr^2 =.02. Attention was again significant while accounting for .18% of the CDR-SB score, β = .04, $t(8719) = 4.02$, $p < .001$, pr^2 = .0018.

The fourth hierarchical multiple regression was conducted to predict CDR-SB scores from raw scores, adjusted for gender only (raw scores). Memory test scores again accounted for 42% of the CDR-SB variability, $F(1, 8722) = 6420.78$, $p < .001$, R^2 = .424. Executive function scores accounted for an additional 4.1% of the variability in the CDR-SB scores, $\Delta F(1, 8721) = 662.44$, $p < .001$, ΔR^2 = .041. Language had a smaller yet still significant effect by accounting for .07% of the variability in the CDR-SB, $\Delta F(1, 8720) = 120.73$, $p < .001$, ΔR^2 = .007. Lastly, model 4 showed that attention scores were again significant even though they only captured .1% of the variance in the CDR-SB, $\Delta F(1, 8719) = 22.70$, $p < .001$, ΔR^2 = .001. The complete model accounted for 47.3% of the variability in the CDR-SB scores.R^2 = .473, $p < .001$.

The significance of the predictors for the raw scores showed memory again at 42%, β = -.65, $t(8722) = -80.13$, $p < .001$, pr^2 = .42. Executive function accounted for an additional 7% of CDR-SB scores, β = -.24, $t(8721) = -25.74$, $p < .001$, pr^2 = .07. Language accounted for 1% of the CDR-SB score, β = -.12, $t(8720) = -10.99$, $p < .001$, pr^2 =.01. And attention was again statistically significant while only accounting for .26% of the CDR-SB scores, β = .04, $t(8719) = 4.77$, $p < .001$, pr^2 = .0026.

The gender and education normed model captured the greatest amount of variance in the CDR-SB with 47.9% . The gender, education, and age normed model accounted for 47.7% of the variance in the CDR-SB. The raw scores accounted for 47.3% of the variance and the model normed for gender and age captured the least amount of variance 47.1% from the CDR-SB. The differences were small and did not fall into different effect size categorizations, but an argument can be made that any increase is of significant clinical value in dementia diagnostic evaluations. Each predictor was significant in every step of every regression for every norming method. Meaning no matter how the test scores were normed they all had important relationships with the CDR-SB. As expected, memory, executive functions, and language test scores significantly predicted the dementia severity rating in a negative direction, as test scores decreased, the dementia severity increased. However, an unexpected finding surfaced with the attention scores, while statistically significant, the contribution to the predicted outcome measure was much smaller than expected. The best explanation for the small effect size of the attention domain is that the ability to maintain attention is often well preserved until the later stages of AD. The majority of the sample group had normal cognition with inclusion criteria removing participants with a CDR-SB score above 9.5, meaning that individuals with moderate and severe dementia were excluded from this analysis. This decreased the lower end scores in the attention composite.

Age-adjusted scores weakened the ability of the model to capture the variance in the CDR-SB, and the theory that raw scores are superior to other norming methods could not be supported when using a NP battery approach (Hassenstab et al, 2016; Hessler et

al., 2014; Holtzer et al.,2008; O'Connell & Tuokko, 2010; Wyman-Chick et al., 2018). The findings clearly suggested that correction for education is best practice for processing NP test scores in an older population because the 2 models that included education correction were superior to the models that did not included education correction. The model that was only corrected for education, leaving age correction out completely, captured the most variability in the CDR-SB.

Table 7

Model Summary of Regressions for Prediction of CDR-SB Scores

		Total R^2	R^2 change	*df*	*F* Change	Total R^2	Durbin Watson
Model 1	1MEMO_GEA	.416	.416	1, 8722	6213.101	.416	
	2 EXEC_GEA	.466	.050	1, 8721	809.944	.466	
	3 LANG_GEA	.476	.010	1, 8720	169.055	.476	
	4 ATTN_GEA	.477	.001	1, 8719	19.692	.477	1.99
Model 2	1 MEMO_GA	.417	.417	1, 8722	6235.376	.417	
	2 EXEC_GA	.461	.044	1, 8721	708.728	.461	
	3 LANG_GA	.470	.009	1, 8720	145.833	.470	
	4 ATTN_GA	.471	.002	1, 8719	27.080	.471	1.98
Model 3	1 MEMO_GE	.423	.423	1, 8722	6401.380	.423	
	2 EXEC_GE	.469	.046	1, 8721	755.641	.469	
	3 LANG_GE	.478	.008	1, 8720	139.859	.478	
	4 ATTN_GE	.479	.001	1, 8719	16.182	.479	1.98
Model 4	1 MEMO_G	.424	.424	1, 8722	6420.780	.424	
	2 EXEC_G	.465	.041	1, 8721	662.440	.465	
	3 LANG_G	.472	.007	1, 8720	120.733	.472	
	4 ATTN_G	.473	.001	1, 8719	22.703	.473	1.98

Note. n = 8724. All p values were statistically significant $p < .001$. Memory composite score (MEMO). Executive function composite score (EXEC). Language composite score (LANG). Attention composite score (ATTN). Scores corrected for gender, education, and age (GEA). Scores corrected for gender, and age (GA). Scores corrected for gender, and education (GE). Raw scores (G) corrected for gender only.

Table 8

Significance of Predictors by Norming Method

	Model 1			Model 2			Model 3			Model 4		
Predictors	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE(B)*	*β*
MEMO_GEA	-.96	.01	-.65*	-.78	.01	-.52*	-.70	.01	-.47*	-.71	.01	-.46*
EXEC_GEA				-.40	.01	-.25*	-.32	.02	.20*	-.34	.02	-.22*
LANG_GEA							-.20	.02	-.13*	-.20	.02	-.13*
ATTN_GEA										.07	.02	.04*
MEMO_GA	-.95	.01	-.65*	-.77	.01	-.52*	-.70	.01	-.48*	-.70	.01	-.48*
EXEC_GA				-.36	.01	-.24*	-.29	.02	-.19*	-.32	.02	-.21*
LANG_GA							-.18	.02	-.13*	-.19	.02	-.13*
ATTN_GA										.09	.02	.05*
MEMO_GE	-.94	.01	-.65*	-.76	.01	-.52*	-.69	.01	-.48*	-.70	.01	.49*
EXEC_GE				-.38	.01	-.25*	-.30	.02	-.20*	-.33	.02	-.22*
LANG_GE							-.18	.02	-.12*	-.18	.02	-.13*
ATTN_GE										.07	.02	.04*
MEMO_G	-.93	.01	-.65*	-.75	.01	-.53*	-.69	.01	-.48*	-.69	.01	-.48*
EXEC_G				-.35	.01	-.24*	-.28	.02	-.19*	-.30	.02	-.21*
LANG_G							-.17	.02	-.12*	-.17	.02	-.12*
ATTN_G										.08	.02	.04*

Note: $n = 8724$. *$p < .001$. Memory composite score (MEMO). Executive function composite score (EXEC). Language composite score (LANG). Attention composite score (ATTN). Scores corrected for gender, education, and age (GEA). Scores corrected for gender, and age (GA). Scores corrected for gender, and education (GE). Raw scores (G) corrected for gender only.

Table 9

Comparison of Percentage Accounted for by Each Cognitive Domain (N = 8724)

Model	Memory Score +	Executive Score +	Language Score +	Attention Score =	Total
GEA	41.6%	5.0%	1%	.1%	47.7%
GA	41.7%	4.4%	.9%	.2%	47.1%
GE	42.3%	4.6%	.8%	.1%	47.9%
G	42.4%	4.1%	.7%	.1%	47.3%

Note: n = 8724. Scores corrected for gender, education, and age (GEA). Scores corrected for gender, and age (GA). Scores corrected for gender, and education (GE). Raw scores (G) corrected for gender only.

The models were also compared using AIC values by computing the delta AIC values as seen in Table 9. The lowest AIC value was the model corrected for gender and education, which when subtracted from the other AIC values provided the model delta value. The best model had the lowest AIC value unless another model had a lesser delta AIC (Burnham & Anderson, 2004). Burnham and Anderson (2004) determined that models with delta AIC ≤ 2 have substantial support and are comparable to the model with the lowest AIC value. Models with delta AIC ≤ 4 to 7 have considerably less support and those with delta AIC > 10 have essentially no support. This is the same story told by R^2 only clearer. Age correction of cognitive tests created an inferior model. The best model was gender and education corrected model with all the other models having essentially no support because the other delta AIC values were all > 10.

Table 10

AIC Values for All Regression Models

	AIC Value	Delta AIC Value
GEA	3373.47782	24.24
GA	3469.066734	122.83
GE	3346.239576	**
G	3433.620439	87.38

Note: n = 8724. *p < .001. Scores corrected for gender, education, and age (GEA). Scores corrected for gender, and age (GA). Scores corrected for gender, and education (GE). Raw scores (G) corrected for gender only.

Results of Main Hypotheses Testing

The first hypothesis was that age and education correction would increase the extent to which NP test scores would capture functional decline due to cognitive loss in dementia. This hypothesis was supported. A comparison of the gender, education, and age corrected model to the raw score model showed the gender, age, and education model captured more variance, R^2 = .477 than the raw scores, R^2 = .473 so the null hypothesis was rejected.

Subhypothesis 2 was that age correction would decrease the extent to which NP tests were able to capture functional decline due to cognitive loss in dementia. This hypothesis was also supported. A comparison of the gender and age corrected model to the raw score model showed the gender and age corrected scores, R^2 = .471 captured less variance than the raw scores, R^2 = .473 so the null hypothesis was rejected.

Subhypothesis 3 was that education correction would increase the extent to which NP tests was able to capture functional decline due to cognitive loss in dementia. The comparison between the gender and education scored model, R^2 = .479, showed that it

captured more variance in the CDR-SB than any other model so the null hypothesis was rejected. In short, the results of this hierarchical multiple regression showed that gender and education correction of test scores captured 47.9% of variance in CDR-SB, giving it a slight edge over the gender, education, and age corrected model which captured 47.7% of the variance in the CDR-SB. This was confirmed by the comparison of the AICs for the models (See Table 10). The best model was the one with the lowest AIC value, the model corrected for gender and education, and all the other models had essentially no support.

Summary

The overarching research question examined the effects of different types of demographic correction on the relationship between cognitive test scores and functional deterioration due to cognitive impairment as measured by the CDR-SB, a dementia severity measure. The first hypothesis, that age and education correction increased the extent to which NP test scores captured functional decline due to cognitive loss in dementia was supported. The first subhypotheses that age correction decreased the extent to which NP tests captured functional decline due to cognitive loss in dementia was also supported. The second subhypotheses, that education correction increased the extent to which NP tests captured functional decline due to cognitive loss in dementia was strongly supported. A comparison of the models showed that the education correction model captured the most variance, $R^2 = .479$, in the CDR-SB score. The results of this hierarchical multiple regression showed that education correction of test scores made for a superior model by accounting for 47.9% of variance in CDR-SB. The comparison of the

AIC values for each model confirmed these findings. The gender and education corrected scores provided the best model with the other models having essentially no support. Age correction alone of NP tests in an older population created an inferior model. The differences between the models corrected for education alone and the education and age correction model were small, and an argument can be made that any increase is of significant clinical value. However, the necessary amount of increase or increment that would make a significant clinical difference on a population level was beyond the scope of this research. Chapter 5 addresses further interpretation of these findings, as well as limitations of the study and recommendations for further research.

Chapter 5: Discussion, Conclusions, and Recommendations

The purpose of this study was to identify optimal normative methods for detecting pathologic cognitive changes in elderly individuals when using cognitive testing. Archival data from the NACC provided a large data set on which to perform a hierarchical multiple regression analyses to determine the strength of the relationships between cognitive test scores, normed 4 different ways, and the clinical dementia severity rating as measured by the CDR-SB. Total R^2 values and AIC values were compared to determine which norming method captured the greatest amount of variance, in other words which of the models garnered the most support. The results of the analysis revealed the differences between the models were small and did not fall into different effect size categories, but a clear hierarchy was established with education-corrected scores demonstrating superiority over the other demographic correction methods. Age correction weakened the models' ability to capture variance in the CDR-SB.

The main hypothesis was that age and education correction would increase the extent to which NP test scores captured functional decline due to cognitive loss in dementia. This hypothesis was supported. A comparison of the gender, education, and age scored model to the model composed of raw scores showed the demographically corrected model captured more variance, $R^2 = .477$ than the raw scores, $R^2 = .473$. The first subhypothesis was that age correction would decrease the extent to which NP tests captured functional decline due to cognitive loss in dementia. This hypothesis was also supported because the comparison between the gender and age corrected model and the model composed of raw scores showed gender and age corrected scores, $R^2 = .471$

captured less variance than the raw scores, R^2 = .473. The second subhypothesis was that education correction would increase the extent to which NP tests captured functional decline due to cognitive loss in dementia. Support was provided for this hypothesis as the gender and age corrected model, R^2 = .479, captured the most variance in the CDR-SB when compared to all other models. Age correction subtracted from the model's ability to capture variance. While the differences were small, they suggested the superiority of the gender and education corrected model over all others. The major limitation to these results is that further work is necessary to determine the extent of the difference in R^2 value that is required to make this clinically significant on a population level as no universally accepted criteria exists. Had the models fallen into different effect size categories the interpretation of the results would have been clearer. All the models had large effect sizes, R^2 values >.40 as per the conventions established by Cohen (1988). It can be argued that any increases in R^2 value is desirable because it may lead to greater diagnostic accuracy when evaluating for dementia. The current study confirmed that age-corrected scores are the least desirable method of norming in an older population. The findings of the current study also refute the idea that raw scores are superior to demographically corrected scores in NP testing when used in the service of dementia diagnostic evaluations that utilize a NP battery approach.

Interpretation of the Findings

The purpose of norming cognitive tests in the elderly is to provide a point of comparison for detecting deviant performance that would be indicative of a neurodegenerative disease such as AD. The findings of the current study provided a clear

hierarchy between the models; the education-corrected scores generated a superior model that captured 48% of the variance in the CDR-SB while the age-corrected model produced an inferior model that captured 47% of the variability in the CDR-SB. A comparison of the AIC values confirmed that gender and education corrected scores constructed the superior model. The next step would logically be to determine what useful clinical significance a 1% increase in R^2 translates into at a population level, but that is beyond the scope of the current study.

Sliwinisky et al., (1996) introduced the idea that conventional norms were not optimal for detecting deviant performance on cognitive tests because the norms contained individuals who were already in cognitive decline yet still performed within normal limits on testing. Their seminal work demonstrated that conventional norms; underestimated normal performance in the elderly; overestimated the variance in test scores; exaggerated cognitive decline due to normal aging, especially in the very old, and; produced norms that were less sensitive to detecting dementia making it more challenging to diagnose. Sliwinski et al. argued that correcting for age and education decreased the discriminative validity of their single memory test, and that the uncorrected raw scores were superior for detecting dementia. The present study challenged their findings by using different methods of norming to analyze the strength of test scores' relationships with the gold standard for clinical dementia severity ratings. The results of the current study partially supported their work, correcting for age and education did produce an inferior model if looking only at memory scores. Raw memory scores indeed captured more variation in the CDR-SB. However, when a battery approach was utilized and executive processes,

language, and attention scores were taken into consideration, the raw scores did not capture as much variation as models that included correction for education. It was clear that when using a typical NP battery approach, adding education correction produced a better model. Because a diagnosis between normal cognitive aging and dementia is never determined by a single cut off score in clinical practice, correcting for education is an essential component when processing scores.

In 2010 O'Connell and Tuokko expanded on Sliwinski et al.'s key idea that demographic corrections of cognitive test scores may not universally improve dementia classification accuracy. Their study concluded equivalent overall classification accuracy of demographically corrected scores and uncorrected test scores but the authors realized that their findings were of more importance when only one test was used, as in a dementia-screening evaluation, and of less clinical importance with a typical NP battery approach. O'Connell and Tuokko conceded numerous limitations in their data given its novel simulation methodology that necessitated replication to be of any clinical value. Although their conclusion was overall equivalence between demographically corrected scores and uncorrected test scores, they did mention that gender and education corrected scores showed slightly higher accuracy. More work needs to be done to determine the exact extent of the differences required to deem these small differences in findings as significant or non-significant. The future direction for research of this type should be to determine if this number translates into clinical significance at a population level.

Previous work by Sliwinski and colleagues (1996; 1997) focused on a single memory test score, while later work by Hessler et al. (2014) used a heterogeneous

composite score that represented multiple cognitive domains. The current study deemed it unlikely that cognitive domains would be affected by demographic variables in a uniform manner. Thus, theoretical groupings of particular tests were used to represent different cognitive domains (memory, executive function, language, and attention). This allowed for a first-of-its-kind examination of how each domain was affected by demographic corrections. Raw memory composite score captured the most variation in the CDR-SB when standing alone, but when the other domains were added and the model was examined as a whole, the raw scores underperformed the models that included correction for education (See Table 11). The executive and language composite scores appeared to benefit the most from gender, age, and education correction. While the attention domain received a slight increase from age correction but captured the same amount of variability with all other norming methods. The differences did not fall into different effect size categories leaving some work to be done on how meaningful the findings are on a practical level.

Table 10

Comparison of Percentage Accounted for by Each Cognitive Domain (N = 8724)

Model	Memory Score +	Executive Score +	Language Score +	Attention Score =	Total
GEA	41.6%	5.0%	1%	.1%	47.7%
GA	41.7%	4.4%	.9%	.2%	47.1%
GE	42.3%	4.6%	.8%	.1%	47.9%
G	42.4%	4.1%	.7%	.1%	47.3%

Note: n = 8724. Scores corrected for gender, education, and age (GEA). Scores corrected for gender, and age (GA). Scores corrected for gender, and education (GE). Raw scores (G) corrected for gender only.

Hessler et al. (2014) determined that both corrected and uncorrected scores were highly significant predictors of progression to dementia even when adjusted for age and education but observed that education-corrected scores had slightly higher predictive accuracy. Their overall conclusion was that the differences between the models were small and did not reach a level of clinical significance, but acquiesced further investigation was necessary because of the limitations of their study. They questioned the generalizability of their study because their sample size was composed of a small set of inpatients (n = 537) who were recruited from 3 general hospitals. Once a patient is hospitalized, there is a greater chance that their performance on cognitive testing is compromised which may decrease the variability in the sample and mask clear results in the analysis. The current study improved on this by using a large data set that included individuals from across the Nation with a robust normal group, individuals classified as having normal cognition, as well as those with mild to moderate cognitive impairment and the same patterns were uncovered. The current study supported their conclusions, education corrected scores enhanced the relationship between the predictors and the outcome variable and age-corrected scores were the worst predictors of the outcome measure.

In 2016 Quaranta et al. aimed to replicate the work of Hessler et al. (2014). Quaranta et al. (2016) failed to attain results that reached statistical significance and explicitly stated that their results did not *clearly* support age correction of test scores compromised their ability to predict progression to dementia, but acquiesced that, at least

theoretically, applying age norms in the diagnosis of MCI might partly decrease the prognostic value over the raw memory scores. Limitations cited by these authors included the recruitment of participants from general hospitals and not memory clinics, and also that the sample group was not homogeneous in etiology and included AD dementia, MD, and non-AD dementia. The current study addressed these limitations by using data from memory clinics and specified a primary diagnosis of AD dementia for inclusion criteria. A robust normal control group was also part of the current study design. The current study supported that age-corrected scores were inferior to raw scores and all other methods of demographic correction for capturing the amount of variability in the clinical dementia severity rating. A final thought on age correction, when revisiting the correlations between the cognitive composite test scores and one another, it appeared that even after scores were corrected for age and education, relationships still existed between the test scores and sample demographics. The strongest correlation was between the CDR-SB and age. It is common knowledge that the incidence of dementia increases with age, yet some may not consider the unintended consequence of age correcting test scores in an elderly population and removing the effects of a variable that is scientifically known to be correlated with dementia.

Strengths and Limitations

A strength of the present study is that it had a large amount of data utilizing the UDS-3NB, a standardized and clinically sound NP test battery endorsed by the National Alzheimer's Coordinating Center. The inclusion criteria specified only AD etiology which addressed one of limitations cited by a previous researcher who recognized that

having multiple etiologies of dementia was a limitation because different types of dementia present with different NP profiles. It also screened out participants in advanced stages of dementia because again, their testing profiles would be highly variable and less reliable. Yet, this large data set with its robust normal control group may have also functioned as a limitation because the number of individuals with normal cognition far outnumbered those who were demented and this may have obscured larger effect size differences between the models and diluted the final interpretation of these differences. Another limitation of this study was the lack of universally accepted criteria for determining the incremental difference required to make a clinically significant contribution to clinical practice. When analyzing the significance of each individual predictors' contribution to the overall models' ability to capture meaningful variations in the CDR-SB score, the differences between the models were small and did not fall into different effect size categories leaving the final interpretation quite ambiguous. In the context of dementia evaluations, it can always be argued that any increase in R^2 value leads to greater diagnostic accuracy and may be of practical significance at a population level. The recommendation for further research is determining how these differences translate at a population level, in other words what incremental increase confirms or disconfirms these findings as significant for clinical practice.

Implications

Improved diagnostic accuracy during the earliest stages of a neurodegenerative process is critically needed to move the field forward so disease-modifying interventions can succeed before too much irreversible damage is done (Ritchie et al., 2015; Sutphen et

al., 2015; Villemagne et al., 2012). Refinements in the science of cognitive testing that distinguish individuals who are experiencing the effects of normal aging from those in the process of developing neurodegenerative diseases gave rise to the question concerning the utility of raw scores versus demographically-corrected scores in norm-referenced cognitive testing of an older population. The current study explored the strength of the relationship between tests scored different ways and dementia severity with the ultimate purpose of finding which way of scoring was most diagnostically accurate. Using a typical NP battery approach with composite test scores representing major cognitive domains affected by AD, demographic corrections for gender and education, leaving age correction out completely, constructed the most accurate model for predicting the dementia severity rating and highlighted that standard normative corrections may be insufficient for removing the confounding effects of age, gender, and education.

Our ability to differentiate between normal cognitive aging and a neurodegenerative process earlier and with more accuracy has many implications for positive social change. If leaving out age correction increases the ability of testing to capture functional loss, it may allow for an earlier or more accurate diagnoses of a neurodegenerative process. This opens the window for the individual to earlier intervention, more time to provide evidence-based services that improve the individual's quality of life. Drug interventions may slow the progression of the disease while cognitive rehabilitation can maximize reserved cognitive resources by teaching compensatory strategies. Behavioral interventions such as diet and exercise increase

quality of life and prolong a person's independence. The individual also gains the benefit of participating in their own care and estate planning. Families benefit when they can develop and plan strategies to avoid disruption in employment, depletion of finances, and exacerbation of their own health issues from the emotional stress added by being a care giver. Society benefits because expenditures for an individual with dementia are 3 times the cost of care for people without dementia for the same age group causing a huge financial strain on our Medicare system. Delaying the onset of the disease for just one year saves resources on the costly and long course of this disease (AA, 2018; Dubois et al., 2016; Zissimopoulos et al., 2014).

Conclusion

Dementia currently affects approximately 50 million people worldwide; a number that is projected to grow to 82 million by 2030 and 152 million by 2050. It is the second largest cause of disability for individuals aged 70 years and older, and the seventh leading cause of death. Dementia imposes an estimated economic cost of approximately US $818 billion per year globally – or 1.1% of global gross domestic product. Left unaddressed, dementia could represent a significant barrier to social and economic development (WHO, 2018). Delaying the onset of the disease, even just for one year, has significant benefits (AA, 2018; Dubois et al., 2016; Hurd et al., 2013; Langa & Levine, 2014; Rockwood et al., 2014; Smith & Bondi, 2013; Sperling et al., 2011; Ward et al., 2013; Wei-Hong et al., 2017; Zissmopoulos et al., 2014). Any improvement in standardized testing is of value, even if it just detects one case of AD that might have gone undiagnosed. The results of the current study supported that best practice when

processing NP test scores in the service of a dementia diagnostic evaluation should include education correction. It can also be concluded that when we correct for age, we remove the effects of a variable that is systematically related to the outcome we are trying to predict simply because the incidence of AD increases with age. Clinicians need to consider this unintended consequence when utilizing test scores in the service of a dementia evaluation. Research focused on earlier and more accurate diagnoses is part of the formula leading to improvements in biomedical, psychological, and social interventions that have the potential to reduce the number of new cases by 10-20% and ease the physical, psychosocial, and financial hardships for individuals, their families, and developing nations (AA, 2018; WHO, 2018).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions* on Automatic Control, 19(6), 716-723. Retrieved from https://doi.org/10.1109/TAC.1974.1100705

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., … Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia, 7*(3), 270–279. Retrieved from https://doi.org/10.1016/j.jalz.2011.03.008

Allison, P. D. (1999). *Multiple regression: A primer*. Pine Forge Press.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Albert, M. S., Dekosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia, 7(3). 270-279. doi.org/10.1016/j.jalz.2011.03.008

Alzheimer's Association. (2015). Changing the trajectory of Alzheimer's disease: How treatment by 2025 saves lives and dollars. Retrieved from

https://www.alz.org/media/Documents/changing-the-trajectory-r.pdf

Alzheimer's Association. (2018). Alzheimer's disease facts and figures. Retrieved from https://www.alz.org/media/Documents/facts-and-figures-2018-r.pdf

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: Author.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Simon & Schuster.

Barnett, J. H., Lewis, L., Blackwell, A. D., & Taylor, M. (2014). Early intervention in Alzheimer's disease: A health economic study of the effects of diagnostic timing. *BMC Neurology*. Retrieved from https://doi.org/10.1186/1471-2377-14-101

Barona, A., Reynolds, C. R., & Chastain, R. (1984). A demographically based index of premorbid intelligence for the WAIS-R. *Journal of Consulting and Clinical Psychology, 52,* 885-887. doi.org/10.1037/0022-006X.52.5.885

Besser, L., Kukull, W., Knopman, D. S., Chui, H., Galasko, D., Weintruab, S., … Morris, J. C. (2018). Version 3 of the National Alzheimer's Coordinating Center's Uniform Data Set. *Alzheimer Disease and Associated Disorders, 32*(4), 351-358. doi: 10.1097/WAD.0000000000000279

Binet, A., & Simon, T. (1980). *The development of intelligence in children.* Vineland, NJ: Publications of the training school at Vineland.

Blom, K., Vaartjes, I., Peters, S. A., & Koek, H. L. (2014). The influence of vascular risk factors on cognitive decline in patients with Alzheimer's disease. *Maturitas,* 79(1), 96-99. doi.org/10.1016/j.maturitas.2014.06.017.

Bondi, M., Edmonds, E., Jak, A., Clark, L., Delano-Wood, L., McDonald, C. R.,…Salmon, D. (2014). Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates. *Journal of Alzheimer's Disease*, *42*(1), 275-289. doi:10.3233/JAD-140276

Borkowski, J., Benton, A., & Spreen, O. (1967). Word fluency and brain damage. *Neuropsychologia, 5*, 135-140.

Boyd, D., & Bee, H. (2019). *Lifespan development* (8th ed.). New York, NY: Pearson

Braak, H., & Braak, E. (1991). Neuropathological staging of Alzheimer-related changes. *Acta Neuropathologica*, 82, 239-259. Retrieved from http://info-centre.jenage.de/assets/pdfs/library/braak_braak_ACTA_NEUROPATHOL_1991.pdf

Bryant, B.R., & Brown, L. (1984). The why and how of special norms. *Remedial and Special Education, 5,* 52-61. doi: .org/10.1177/074193258400500415

Burke, W. J., Miller, J. P., Rubin, E. H., Morris, J. C., Coben, L. A., Duchek, J., ... & Berg, L. (1988). Reliability of the Washington University clinical dementia rating. *Archives of neurology*, *45*(1), 31-32.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*. Retrieved from https://doi.org/10.1177/0049124104268644

Cattell, J.M. (1890). Mental tests and measurements. *Mind, 15*, 317-381. Retrieved from http://www.jstor.org/stable/2247264

Cherry, B. J., Buckwalter, J. G., & Henderson, V. W. (2002). Better preservation of memory span relative to supraspan immediate recall in Alzheimer's disease. *Neuropsychologica, 40*(7), 846–852.

Chew, A., Kesler, E., & Sudduth, D. (1984). A practical example of how to establish local norms. *The Reading Teacher, 38*(2), 160-163. Retrieved from http://www.jstor.org/stable/20198720

Chou, K. L. (2019). Clinical manifestations of Parkinson disease. In A. F. Eichler (Ed.), *UpToDate.* Retrieved from https://uptodate.com/contents/clinical-manifestations-of-parkinson-disease

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed*.). Psychology Press.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. Retrieved from http://conservancy.umn.edu/handle/11299/37076

Cortina, J. M., Aguinis, H., & DeShon, R. P. (2017). Twilight of dawn or of evening? A century of research methods in the Journal of Applied Psychology. *Journal of Applied Psychology*, *102*(3), 274–290. https://doi.org/10.1037/apl0000163

Craft, S., Asthana, S., Cook, D. G., Baker, L. D., Cherrier, M., Purganan, K., … Krohn, A. J. (2003). Insulin dose-response effects on memory and plasma amyloid precursor protein in Alzheimer's disease: interactions with apolipoprotein E genotype. *Psychoneuroendocrinology*, *28*(6), 809–822.

Craft, S., Newcomer, J., Kanne, S., Dagogo-Jack, S., Cryer, P., Sheline, Y., … Alderson, A. (1996). Memory improvement following induced hyperinsulinemia in

Alzheimer's disease. Neurobiology of Aging, 17(1), 123–130. Retrieved from https://doi.org/10.1016/0197-4580(95)02002-0

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334. Retrieved from https://www.psychometricsociety.org/sites/default/files/cronbach_citation_classic_alpha.pdf

Donaghy, P. C., Taylor, J., O'Brien, J. T., Barnett, N., Olsen, K., Colloby, S. J., . . . Thomas, A. J. (2018). Neuropsychiatric symptoms and cognitive profile in mild cognitive impairment with Lewy bodies. *Psychological Medicine, 48*(14), 2384-2390. doi: 10.1017/S003329172000290

Donohue, M. C., Sperling, R. A., Salmon, D. P., Rentz, D. M., Raman, R., Thomas, R. G., … & Aisen, P. S. (2014). The preclinical Alzheimer cognitive composite. *JAMA Neurology*, *71*(8), 961–970. Retrieved from https://doi.org/10.1001/jamaneurol.2014.803

Dubois, B., Padovan, A., Scheltens, P., Rossi, A., & DellAgnello, G. (2016). Timely diagnosis for Alzheimer's disease: A literature review on benefits and challenges. *Journal of Alzheimer's Disease*, *49*(3), 617-631. doi: 10.3233/JAD-150692

Eckert. (2010). Age-related changes in processing speed: unique contributions of cerebellar and prefrontal cortex. *Frontiers in Human Neuroscience*. Retrieved from https://doi.org/10.3389/neuro.09.010.2010

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical

sciences. *Behavior Research Methods*, *39*(2), 175–191. Retrieved from https://doi.org/10.3758/BRM.41.4.1149

Field, A. (2013). *Discovering statistics using IBM SPSS statistics (4th ed.*). Sage Publications Ltd.

George, D., & Mallery, P. (2016). *IBM SPSS statistics 23 step by step: A simple guide and reference* (14 ed.). Boston: Routledge.

Grant, I., & Heaton, R. K. (2015). Ralph M. Reitan: A founding father of neuropsychology. *Archives of Clinical Neuropsychology: the official journal of the National Academy of Neuropsychologists*, *30*(8), 760-1. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4675827/

Gill, T. (2015). *Transcript from Rx The Quiet Revolution*. (D. Grubin, Interviewer). Retrieved from https://rxfilm.org/problems/silver-tsunami-united-states-healthcare-thomas-gill-yale-center-on-aging-interview/

Girvan, E. J., McIntosh, K., & Smolkowski, K. (2019). Tail, tusk, and trunk: What different metrics reveal about racial disproportionality in school discipline. *Educational Psychologist*, *54*(1), 40-59.

Gorno-Tempini, A. E., Hillis, S., Weintraub, A., Kertesz, M. Mendez, S. F. Cappa, J. M.,…Grossman, M. (2011). Classification of primary progressive aphasia and its variants. *Neurology*, 76(11), 1006-1014. doi:10.1212/WNL.0b013e31821103e6

Harada, C. N., Natelson Love, M. C., & Triebel, K. (2013). Normal Cognitive Aging. *Clinics in Geriatric Medicine*, *29*(4), 737–752. Retrieved from https://doi.org/10.1016/j.cger.2013.07.002

Harrell, F. H. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis* (Corrected ed.). New York: Springer.

Hassenstab, J., Chasse, R., Grabos, P., Benzinger, T. L. S., Fagan, A. M., Xiong, C. …& Morris, J. (2016). Certified normal: Alzheimer's disease biomarkers and normative estimates of cognitive functioning. Neurobiology of Aging. 43, 23-33. doi: 10.1016/j.neurobiolaging.2016.03.014

Heaton, R. K., Miller, S. W., Taylor, M. J., & Grant, I. (2004). *Revised comprehensive norms for an expanded Halstead-Reitan Battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults*: *professional manual.* Lutz, FL: PAR.

He, W., Goodkind, D., & Kowal, P. (2016). An aging world: 2015. Retrieved from https://www.census.gov/content/dam/Census/library/publications/2016/demo/p95-16-1.pdf

Hebert, L.E., Weuve, J., Scherr, P.A., & Evans, D.A. (2013). Alzheimer disease in the United States (2010 – 2050) estimated using the 2010 census. *Neurology*, *80*(19), 1778-1783. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3719424/

Hessler, J., Tucha, O., Förstl, H., Mösch, E., & Bickel, H. (2014). Age-correction of test scores reduces the validity of mild cognitive impairment in predicting progression to dementia. *PloS ONE,* 9(8), e106284. doi: 10.1371/journal.pone.0106284

Holtzer, R., Goldin, Y., Zimmerman, M., Katz, M., Buschke, H., & Lipton, R. B. (2008). Robust norms for selected neuropsychological tests in older adults. *Archives of*

Clinical Neuropsychology, *23*, 531-541. doi: 10.1016/j.acn.2008.05.004

Hurd, M.D., Martorell, P., Delavande, A., Mullen, K.J., & Langa, K.M. (2013). Monetary costs of dementia in the United States. *New England Journal of Medicine, 368,* 1326-1334. doi: 10.1056/NEJMsa1204629

Hughes, C., Berg, L., Danziger, W., Coben, L., & Martin, R. (1982). A New Clinical Scale for the Staging of Dementia. *British Journal of Psychiatry, 140*(6), 566-572. doi:10.1192/bjp.140.6.566

Ivanova, I., Salmon, D. P., & Gollan, T. H. (2013). The multilingual naming test in Alzheimer's disease: clues to the origin of naming impairments. *Journal of the International Neuropsychological Society: JINS*, *19*(3), 272–283. doi:10.1017/S1355617712001282

Jack, C. R., Barnes, J., Bernstein, M. A., Borowski, B. J., Brewer, J., Clegg, S.,… Weiner, M. (2015). Magnetic resonance imaging in Alzheimer's disease neuroimaging initiative 2. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *11*(7), 740-56. doi: 10.1016/j.jalz.2015.05.002

Jack, C. R., Jr, Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., … Trojanowski, J. Q. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet. Neurology*, *9*(1), 119–128. doi:10.1016/S1474-4422(09)70299-6

Jack, C. R., Knopman, D. S., Jagust, W. J., Petersen, R. J., Weiner, M. W., Aisen, P. S.,…Trojanowski, J. Q. (2013). Tracking pathophysiological processes in Alzheimer’s disease: An updated hypothetical model of dynamic biomarkers.

Lancet Neurology, *12*, 207-216. doi: 10.1016/S1474-4422(12)70291-0

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning with applications in R* (7th ed.). New York: Springer.

Kang, H. S., Kwon, J. H., Kim, S., Na, D. L., Kim, S. Y., Lee, J.,…Kim, D. K. (2016). Comparison of neuropsychological profiles in patients with Alzheimer's disease and mixed dementia. *Journal of the Neurological Sciences, 369,* 134-138. doi.org/10.1016

Kaplan, E., Goodglass, H., Weintraub, S., & Goodglass, H. (1983). *Boston naming test*. Philadelphia: Lea & Febiger.

Karantzoulis, S., & Galvin, J. E. (2011). Distinguishing Alzheimer's disease from other major forms of dementia. *Expert Review of Neurotherapeutics, 11*(11), 1579-91. Retrieved from https://doi.org/10.1586/ern.11.155

Karzmark, P., Heaton, R. K., Lehman, R. A. W., & Crouch, J. (1985). Utility of the Seashore Tonal Memory Test in neuropsychological assessment. *Journal of Clinical and Experimental Psychology*, *7*, 367-374. doi.org/10.1080/01688638508401270

Katsumata, Y., Mathews, M., Abner, E. L., Jicha, G. A., Caban-Holt, A., Smith, C. D., ... & Fardo, D. W. (2015). Assessing the discriminant ability, reliability, and comparability of multiple short forms of the Boston naming test in an Alzheimer's disease center cohort. *Dementia and geriatric cognitive disorders*, *39*(3-4), 215-227. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4374652/

Kral, V. A. (1962). Senescent forgetfulness: Benign and malignant. *Canadian Medical*

Association Journal, 86, 257-260. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1848846/pdf/canmedaj00930-0002.pdf

Langa, K. M., & Levine, D. A. (2014). The diagnosis and management of mild cognitive impairment: A clinical review. *JAMA, 312*(23), 2551–2561. doi:10.1001/jama.2014.13806

Leber, P. (1990). Guidelines for the clinical evaluation of antidementia drugs. First draft. Rockville, MD: US Food and Drug Administration. Retrieved from https://www.researchgate.net/publication/273131084_Guidelines_for_the_Clinical_Evaluation_of_antiDementia_Drugs_1990_draft/download

Lee, S. E. (2019). Frontotemporal dementia: Clinical features and diagnosis. In J. L. Wilerdink (Ed.), *UpToDate*. Retrieved from https://www.uptodate.com/contents/frontotemporal-dementia-clinical-features-and-diagnosis

Livingston, G., Sommerlad, A., Orgeta, V., Costafreda, S. G., Huntley, J., Ames, D., …Mukadam, N. (2017). Dementia prevention, intervention, and care. *The Lancet*, 390, 2673-2734. Retrieved from https://www.clinicalkey.com/#!/content/playContent/1-s2.0-S0140673617313636

Ma, M., Dorstyn, D., Ward, L., & Prentice, S. (2018). Alzheimer's disease and caregiving: a meta-analytic review comparing the mental health of primary carers to controls. *Aging & mental health*, *22*(11), 1395-1405.

Malek-Ahmadi, M., Powell, J. J., Belden, C. M., O'Connor, K., Evans, L., Coon, D. W.

& Nieri, W. (2015). Age- and education-adjusted normative data for the Montreal Cognitive Assessment (MoCA) in older adults age 70-99. *Aging, Neuropsychology, and Cognition*, *22*(6), 755-761. doi: 10.1080/13825585.2015.1041449

Mazaheri, A., Segaert, K., Olichney, J., Yang, J., Niu, Y., Shapiro, K., & Bowman, H. (2018). EEG oscillations during word processing predict MCI conversion to Alzheimer's disease. *Neuroimage Clinical. 17*. 188-197. Retrieved from https://doi.org/10.1016/j.nicl.2017.10.009

McKeith, I. G., Boeve, B. F., Dickson, D. W., Halliday, G., Taylor, J. P., Weintraub, D., … Kosaka, K. (2017). Diagnosis and management of dementia with Lewy bodies: Fourth consensus report of the DLB Consortium. *Neurology*, *89*(1), 88-100. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5496518/

McKhann, G. M., Knopman, D. S., Chertkow, H, Hyman, B. T., Jack, C. R., Kawas, C. H.,…Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia, 7*(3). 263-269. https://doi.org/10.1016/j.jalz.2011.03.005

Mesulam, M. M. (2013). Primary progressive aphasia and the language network: The 2013 H. Houston Merritt Lecture. *Neurology, 81*(5), 456-462. doi: 10.1212/WNL.0b013e31829d87df

Meyers, L. S., Gamst, G. C., & Guarino, A. J. (2016). *Applied multivariate research: Design and interpretation* (3rd ed.). Los Angeles: SAGE Publications, Inc.

Meyers, J. E., & Meyers, K. R. (1995). Rey complex figure test under four different administration procedures. *The Clinical Neuropsychologist, 9*(1), 63-67. doi: 10.1080/13854049508402059

Mitchell, A. J., Shiri-Feshki, M. (2009). Rate of progression of mild cognitive impairment to dementia: Meta-analysis of 41 robust inception cohort studies. *Acta Psychiatrica Scandinavica, 119*(4), 252-265. doi:10.1111/j.1600-0447.2008.01326.x

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment*. Oxford University Press.

Mitrushina, M., Drebing, C., & Uchiyama, C. (1994). The pattern of deficit in different memory components in normal aging and dementia of Alzheimer's type. *Journal of Clinical Psychology, 50*591-596. doi:10.1002/1097-4679(199407)50:4<591:AID-JCLP2270500415>3.0.CO;2-9

Möller, H., & Graeber, M. B. (1998). The case described by Alois Alzheimer in 1911. *European Archives of Psychiatry and Clinical Neuroscience, 248*(3), 111-122. doi.org/10.1007/s004060050027

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology, 43*(11), 2412-2414. Retrieved from http://dx.doi.org/10.1212/WNL.43.11.2412-a

Morris, J. C. (1997). Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. *International Psychogeriatrics, 9 Suppl 1*, 173–176; discussion 177-178.

Morris, J. C., Ernesto, C., Schafer, K., Coats, M., Leon, S., Sano, M., …& Woodbury, P. (1997). Clinical Dementia Rating training and reliability in multicenter studies: The Alzheimer's Disease cooperative study experience. *Neurology, 48*(6) 1508-1510. doi:10.1212/WNL.48.6.1508

Morris, John C., Weintraub, S., Chui, H. C., Cummings, J., Decarli, C., Ferris, S., … Kukull, W. A. (2006). The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer Disease Centers. *Alzheimer Disease and Associated Disorders*, *20*(4), 210–216. Retrieved from https://doi.org/10.1097/01.wad.0000213865.09806.92

Mortamais, M., Ash, J. A., Harrison, J., Kaye, J., Kramer, J., Randolph, C.,… Ritchie, K. (2017). Review article: Detecting cognitive changes in preclinical Alzheimer's disease: A review of its feasibility. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *13*, 468–492. doi: 10.1016/j.jalz.2016.06.2365.

Nathans, L. L., Oswald, F. L., & Nimon, K. (2012). *Interpreting multiple linear regression: A guidebook of variable importance*. *17*(9), 19.

National Academies of Sciences, Engineering, and Medicine. (2016). *Families caring for an aging America*. Washington, D.C.: The National Academies Press.

National Alzheimer's Coordinating Center. (2010). Research, data, consultation, collaboration. Retrieved from https://www.alz.washington.edu/WEB/researcher_home.html

O'Bryant, S. E., Lacritz, L. H., Hall, J., Waring, S. C., Chan, W., Khodr, Z. G., … Cullum, C. M. (2010). Validation of the new interpretive guidelines for the

Clinical Dementia Rating scale sum of boxes score in the National Alzheimer's Coordinating Center database. *Archives of Neurology, 67*(6). Retrieved from https://doi.org/10.1001/archneurol.2010.115

O'Bryant, S. E., Waring, S. C., Cullum, C. M., Hall, J., Lacritz, L., Massman, P. J., …& Doody, R. (2008). Staging dementia using Clinical Dementia Rating scale sum of boxes scores: A Texas Alzheimer's Research Consortium Study. *Archives of Neurology*, *65*(8), 1091–1095. Retrieved from https://doi.org/10.1001/archneur.65.8.1091

O'Connell, C., & Tuokko, H. (2010). Age corrections and dementia classification accuracy. *Archives of Clinical Neuropsychology, 25*, 126-138. doi: 10.1093/arclin/acp111

Olde-Rikkert, M. G., Tona, K. D., Janssen, L., Burns, A., Lobo, A., Robert, P., ... & Waldemar, G. (2011). Validity, reliability, and feasibility of clinical staging scales in dementia: a systematic review. *American Journal of Alzheimer's Disease & Other Dementias®*, *26*(5), 357-365. Retrieved from https://doi.org/10.1177/1533317511418954

Oosterhuis, H. E. M., Van der Ark, L. A., Sijtsma, K.(2016). Sample size requirements for traditional and regression-based norms. *Assessment*, *23*(2), 191–202. Retrieved from https://doi.org/10.1177%2F1073191115580638

Ossenkoppele, R., Pijnenburg, Y. A. L., Perry, D. C., Cohn-Sheehy, B. I., Scheltens, N. M. E., Vogel, J. W., … Rabinovici, G. D. (2015). The behavioural/dysexecutive variant of Alzheimer's disease: clinical, neuroimaging and pathological features.

Brain, *138*(9), 2732–2749. Retrieved from https://doi.org/10.1093/brain/awv191Peall, K. J., & Robertson, N. P. (2015). Biomarkers in Alzheimer's disease: Understanding disease trajectory and therapeutic targets. *Journal of Neurology*, *262*: 2195-2197. doi: 10.1007/s00415-015-7881-6

Petersen R., Smith, G., Ivnik, R., Tangalos, E., Schaid, D., Thibodeau, S.,…Kurland, L. (1995). Apolipoprotein E status as a predictor of the development of Alzheimer's disease in memory-impaired individuals. *JAMA. 273*(16).1274–1278.

Petersen, R., Smith, G., Waring, S., Ivnick, R., Tangalos, E., & Kokment, E. (1999). Mild cognitive impairment: Clinical characterization and outcome. *Archives of Neurology, The Journal of the American Medical Association*, *56* (3), 303-308.doi:10.1001/archneur.56.3.303

Petersen, R. C., Caracciolo, B., Brayne, C., Gauthier, S., Jelic, V., & Fratiglioni, L. (2014). Mild cognitive impairment: a concept in evolution. *Journal of Internal Medicine*, *275*(3). 214-228. doi: 10.1111/joim.12190

Pituch, K. A., & Stevens, J. P. (2015). *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS*. Routledge.

Possin, K. L., Laluz, V. R., Alcantar, O. Z., Miller, B. L., & Kramer, J. H. (2010). Distinct neuroanatomical substrates and cognitive mechanisms of figure copy performance in Alzheimer's disease and behavioral variant frontotemporal dementia. *Neuropsychologia*, *49*(1), 43–48. doi:10.1016/j.neuropsychologia.2010.10.026

Quaranta, D., Gainotti, G., Gabriella Vita, M., Lacidogna, G. Scaricamazza, E., Piccininni, C., Marra, C. (2016). Are raw scores on memory tests better than age- and education- adjusted scores for predicting progression from amnesic mild cognitive impairment to Alzheimer disease. *Current Alzheimer Research*, *13*(12), 1414-1420. doi: 10.2174/1567205013666160314145522

Ramirez-Gomez, L., Zheng, L., Reed, B., Kramer, J., Mungas, D., Zarow, C.,… Chui, H. (2017). Neuropsychological profiles differentiate Alzheimer disease from subcortical ischemic vascular dementia in an autopsy-defined cohort. *Dementia and Geriatric Cognitive Disorders*, *44*(1–2), 1–11. doi: 10.1159/000477344

Ramlall, S., Chipps, J., Bhigjee, A. I., & Pillay, B. J. (2014). Sensitivity and specificity of neuropsychological tests for dementia and mild cognitive impairment in a sample of residential elderly in South Africa. *South African Journal of Psychiatry*, *20*(4), 153-159. Retrieved from https://www.ajol.info/index.php/sajpsyc/article/viewFile/114810/104450

Rao, S. J. (2003). Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis. *Journal of the American Statistical Association*, *98*(461), 257–258. Retrieved from https://doi.org/10.1198/jasa.2003.s263

Ravscovsky, K., Hodges, J., Knopman, D., Mendex, M., Kramer, J., Neuhaus, J.,…Miller, B. (2011). Sensitivity of revised diagnostic criteria for the behavioral variant of frontotemporal dementia. *Brain*, *134*(Pt9), 2456-2477.

Riordan, H. (2017). Constructing composites to optimize cognitive outcomes. *Journal of*

Clinical Studies, *9*(2), 40–45.Retreived from https://www.worldwide.com/wp-content/uploads/2017/04/Constructing-Composites-to-Optimise-Cognitive-Outcomes.pdf

Ritchie, L. J., Frerichs, R. J., & Tuokko, H. (n.d.). Effective normative samples for the detection of cognitive impairment in older adults. *Clinical Neuropsychologist*, *21*(6), 863–874. doi: 10.1080/13854040701557239

Ritchie, K., Ritchie, C. W., Yaffe, K., Skoog, I., & Scarmeas, N. (2015). Review Article: Is late-onset Alzheimer's disease really a disease of midlife? *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, *1*, 122–130. doi: 10.1016/j.trci.2015.06.004

Rizzo, G., Arcuti, S., Copetti, M., Alessandria, M., Savica, R., Fontana, A.,... Logroscino, G. (2018). Accuracy of clinical diagnosis of dementia with Lewy bodies: A systematic review and meta-analysis. *Journal of Neurology, Neurosurgery and Psychiatry*, *89*(4), 358. doi: 10.1136/jnnp-2017-316844

Roalf, D. R., Moberg, P., Xie, S. X., Wolk, D. A., Moelter, S. T., & Arnold, S. E. (2013). Comparative accuracies of two common screening instruments for classification of Alzheimer's disease, mild cognitive impairment, and healthy aging. *Alzheimer's & Dementia, 9,* 529-537. doi: 10.1016/j.jalz.2012.10.001

Rockwood, K., Mitnitski, A., Richard, M., Kurth, M., Kesslak, P., & Abushakra, S. (2014). Neuropsychiatric symptom clusters targeted for treatment at earlier versus later stages of dementia. *International Journal of Geriatric Psychiatry*, *30*(4). 357-367. doi: org/10.1002/gps.4136

Rodnitzky, R. (2018). Cognitive impairment and dementia in Parkinson disease. In J. L. Wilterdink (Ed.)

Roth, D.L., Fredman, L., & Haley, W.E. (2015). Informal caregiving and its impact on health: A reappraisal from population based studies. *Gerontologist, 55*(2), 309-319.

Rossor, M. N., Fox, N. C., Mummery, C. J., Schott, J. M., & Warren, J. D. (2010). The diagnosis of young-onset dementia. *The Lancet Neurology*, *9*(8), 793–806. Retrieved from https://doi.org/10.1016/S1474-4422(10)70159-9

Salmon, D. P., & Bondi, M. W. (2009). Neuropsychological assessment of dementia. *Annual review of psychology*, *60*, 257-282. Retrieved online https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2864104/

Schoenberg, M.R., & Scott, J.G. (2011). *The little black book of neuropsychology*. New York, NY: Springer.

Sliwinski, M. (1996). The effects of preclinical dementia on estimates of normal cognitive functioning in aging. *The Journals of Gerontology Series B: Psychological and Social Sciences. 51B*(4). 217-225. doi: 10.1093/geronb/51B.4.P217

Sliwinski, M., Buschke, H., Stewart, W., Masur, D., & Lipton, R. (1997). The effect of dementia risk factors on comparative and diagnostic selective reminding norms. *Journal of the International Neuropsychological Society*, *3*(4), 317-326. doi:10.1017/S1355617797003172

Smith, G. E., & Bondi, M. (2013). *Mild cognitive impairment and dementia: Definitions,*

diagnosis, and treatment. New York, NY: Oxford University Press.

Smith, E. (2017). Clinical presentations and epidemiology of vascular dementia. *Clinical Science*, *131*(11), 1059-1068; doi.10.1042/CS20160607

Solway, E. (2017). Dementia caregivers: Juggling, delaying, and looking forward. Retrieved from https://www.healthyagingpoll.org/report/dementia-caregivers-juggling-delaying-and-looking-forward

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M.,... Phelps, C. H. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia, 7*(3). 280-292. doi.org/10.1016/j.jalz.2011.03.003

Stephan, B., Minett, T., Muniz-Terrera, G., Harrison, S. L., Matthews, F. E., Brayne, C. (2017). Neuropsychological profiles of vascular disease and risk of dementia: implications for defining vascular cognitive impairment no dementia (VCI-ND), *Age and Ageing*, *46*(5), 755–760. doi: 10.1093/ageing/afx016

Storandt, M., Grant, E. A., Miller, J. P., & Morris, J. C. (2006). Longitudinal course and neuropathologic outcomes in original vs revised MCI and in pre-MCI. *Neurology*, *67*(3), 467–473. Retrieved from https://doi.org/10.1212/01.wnl.0000228231.26111.6e

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests* (3rd ed.). New York, NY: Oxford University Press.

Sutphen C. L., Jasielec, M. S., Shah, A. R., Macy, E. M., Xiong, A. G., Vlassenko, A.

G.,…Fagan, A. M. (2015). Longitudinal cerebrospinal fluid biomarker changes in preclinical Alzheimer disease during middle age. *JAMA Neurol. 72*(9):1029–1042. doi:10.1001/jamaneurol.2015.1285

Svinicki, J. G. Tombari, M. L., & Needham, F. (1981). *Developing and interpreting local norms: Making test scores work for you*. Dallas, TX: DLM Teaching Resources.

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Tombaugh, T. N., Kozak, J., & Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Archives of clinical neuropsychology*, *14*(2), 167-177. Retrieved from https://doi.org/10.1093/arclin/14.2.167

U.S. Census Bureau. (2014). National population projections: Downloadable files. Retrieved from https://www.census.gov/data/datasets/2014/demo/popproj/2014-popproj.html

Villemagne, V. L., Burnham, S., Bourgeat, P., Brown, B., Ellis, K. A., & Salvado, O. (2012). Amyloid ß deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: A prospective cohort study. *The Lancet Neurology*. *12*(4), 357-367. doi.org/10.1016/S1474-4422(13)70044-9

Ward, A., Tardiff, S., Dye, C., & Arrighi, H. M. (2013). Rate of conversion from prodromal Alzheimer's disease to Alzheimer's dementia: A systematic review of the literature. *Dementia and Geriatric Cognitive Disorders*, *3*(1). doi.org/10.1159/000354370

Wei-Hong, C., Jin, W., Pei-Yuan, L., Liu, Y., Li, R., Hu, M., & Xiang-Jian, X. (2017). Carotid atherosclerosis and cognitive impairment in nonstroke patients. *Chinese Medical Journal, 130*(19), 2375-2379. doi: 10.4103/0366-6999.215331

Weintraub, S., Besser, L., Dodge, H. H., Teylan, M., Ferris, S., Goldstein, F. C., … Morris, J. C. (2018a). Version 3 of the Alzheimer Disease Centers' neuropsychological test battery in the Uniform Data Set (UDS). *Alzheimer disease and associated disorders, 32*(1), 10–17. doi:10.1097/WAD.0000000000000223

Weintraub, S., Carrillo, M. C., Farias, S. T., Goldberg, T. E., Hendrix, J. A., Jaeger, J., … & Randolph, C. (2018b). Measuring cognition and function in the preclinical stage of Alzheimer's disease. *Alzheimer's & Dementia : Translational Research & Clinical Interventions*, *4*, 64–75. Retrieved from https://doi.org/10.1016/j.trci.2018.01.003

Weintraub, S., Salmon, D., Mercaldo, N., Ferris, S., Graff-Radford, N. R., Chui, H., … Morris, J. C. (2009). The Alzheimer's Disease Centers' Uniform Data Set (UDS): the neuropsychologic test battery. *Alzheimer disease and associated disorders*, *23*(2), 91–101. doi:10.1097/WAD.0b013e318191c7dd

White, T. & Stern, R. (2003). *Neuropsychological Assessment Battery: Psychometric and technical manual*. Lutz, FL: PAR, Inc.

Wilkinson, G. S., & Robertson, G. J. (2006). WRAT4 Wide Range Achievement test: Professional manual. Lutz, FL: PAR, Inc.

Williams, M. M., Storandt, M., Roe, C. M., & Morris, J. C. (2013). Progression of

Alzheimer disease as measured by Clinical Dementia Rating sum of boxes scores. *Alzheimer's & Dementia : The Journal of the Alzheimer's Association*, *9*(1 0), S39–S44. Retrieved from https://doi.org/10.1016/j.jalz.2012.01.005

Wilson, R. S., Rosenbaum, G., Brown, G., Rourke, D., Whitman, D., & Gusell, J. (1978). An index of the premorbid intelligence. *Journal of Consulting and Clinical Psychology, 46*, 1554-1555. doi.org/10.1037/0022-006X.46.6.1554

Wimblad B, Palmer K, Kivipelto M, Jelic V, Fratiglioni L, Wahlund L,… Jack C. (2004). Mild cognitive impairment -- beyond controversies, towards a consensus: report of the international working group on mild cognitive impairment. *Journal of Internal Medicine*, *256*(3), 240–246.

Wittenberg, D., Possin, K., Rascovsky, K., Rankin, K., Miller, B., & Kramer, J. (2008) The early neuropsychological and behavioral characteristics of frontotemporal dementia. *Neuropsychology Review, 18*(1), 91-102. doi: 10.1007/s11065-008-9056-z

World Health Organization (2018). Dementia: A public health priority. Retrieved from https://www.who.int/mental_health/neurology/dementia/en/

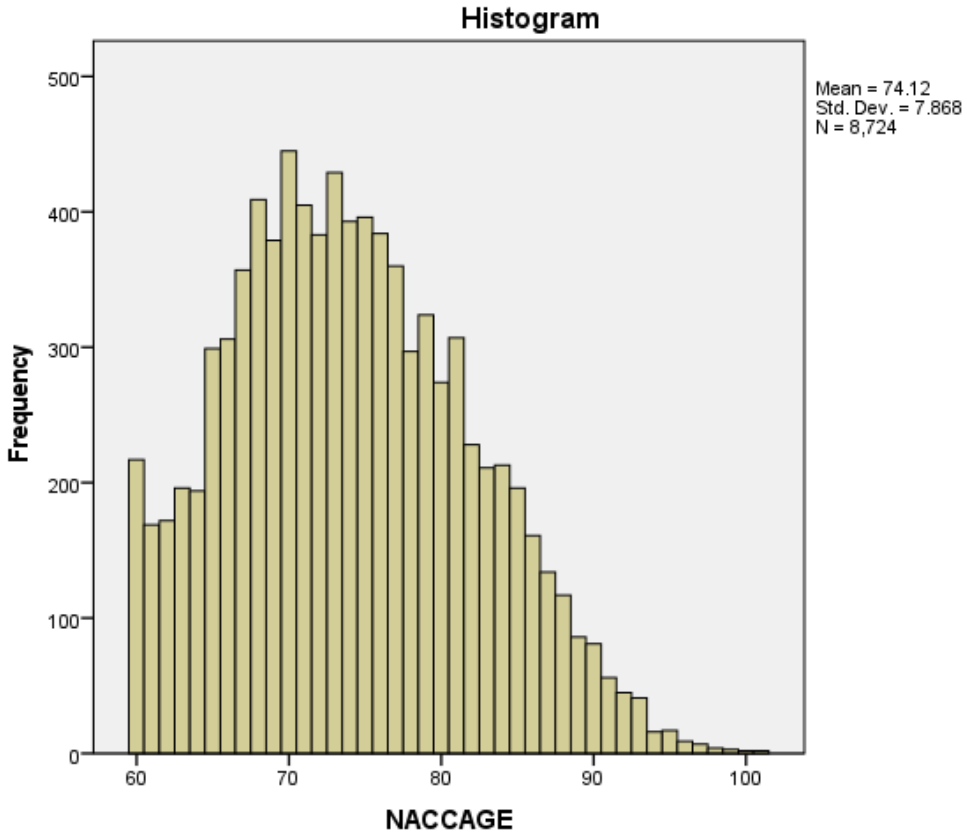
Wyman-Chick, K. A., Marin, P. K., Weintraub, D., Sperling, S. A., Erickson, L., O., Manning, C. A., & Barrett, M. J. (2018). Selection of normative group affects rates of mild cognitive impairment in Parkinson's disease. *Movement Disorders, 33*(5), 839-843. doi.org/10.1002/mds.27335

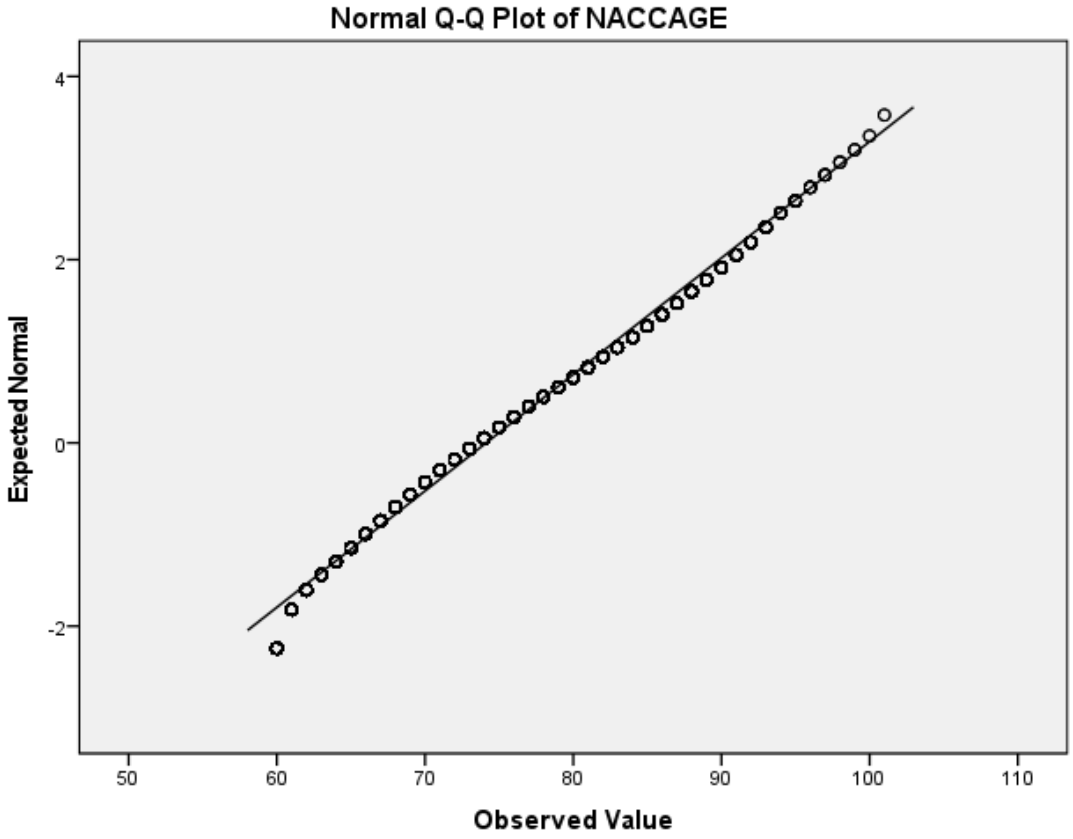
Zachary, R. A., Paulson, M. J., & Gorsuch, R. L. (1985). Estimating WAIS IQ from the Shipley Institute of Living Scale using continuously adjusted age norms. *Journal*

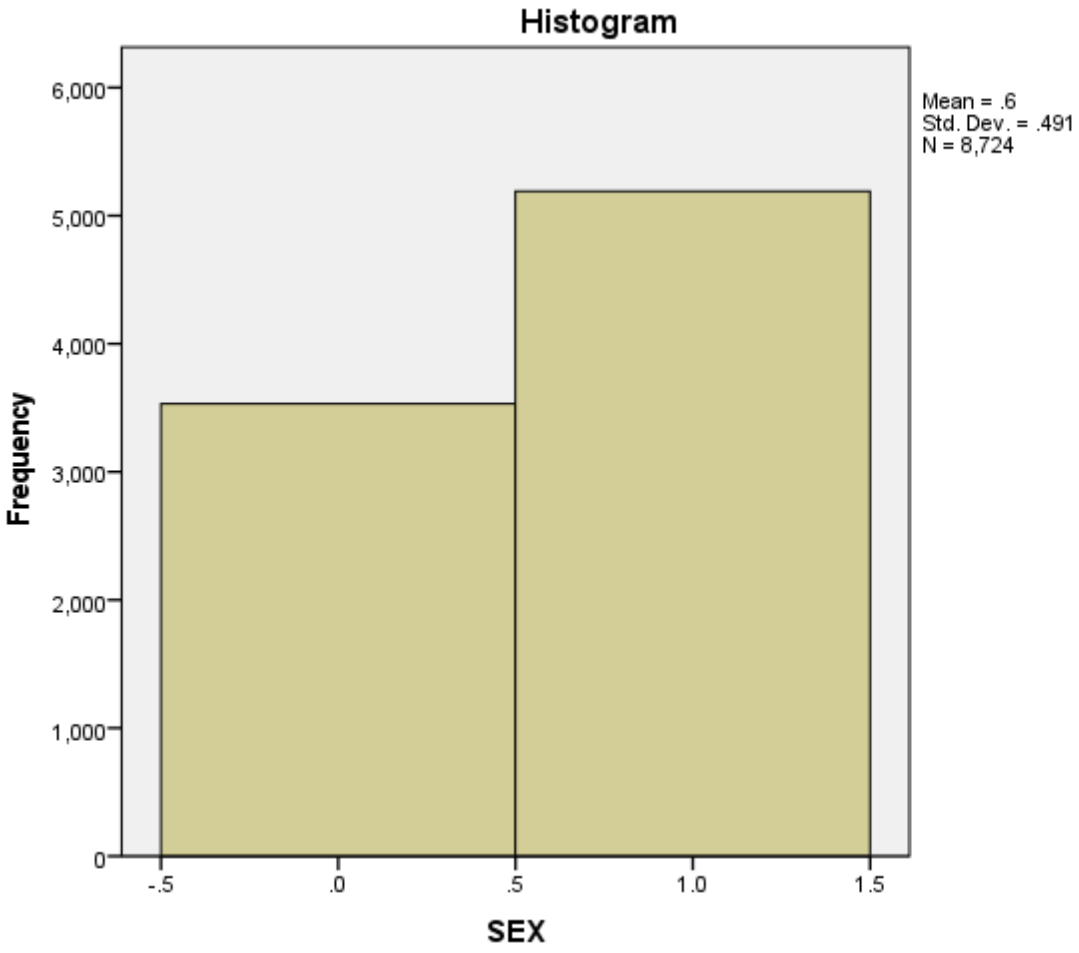
of Clinical Psychology, *41*(6), 820-831.

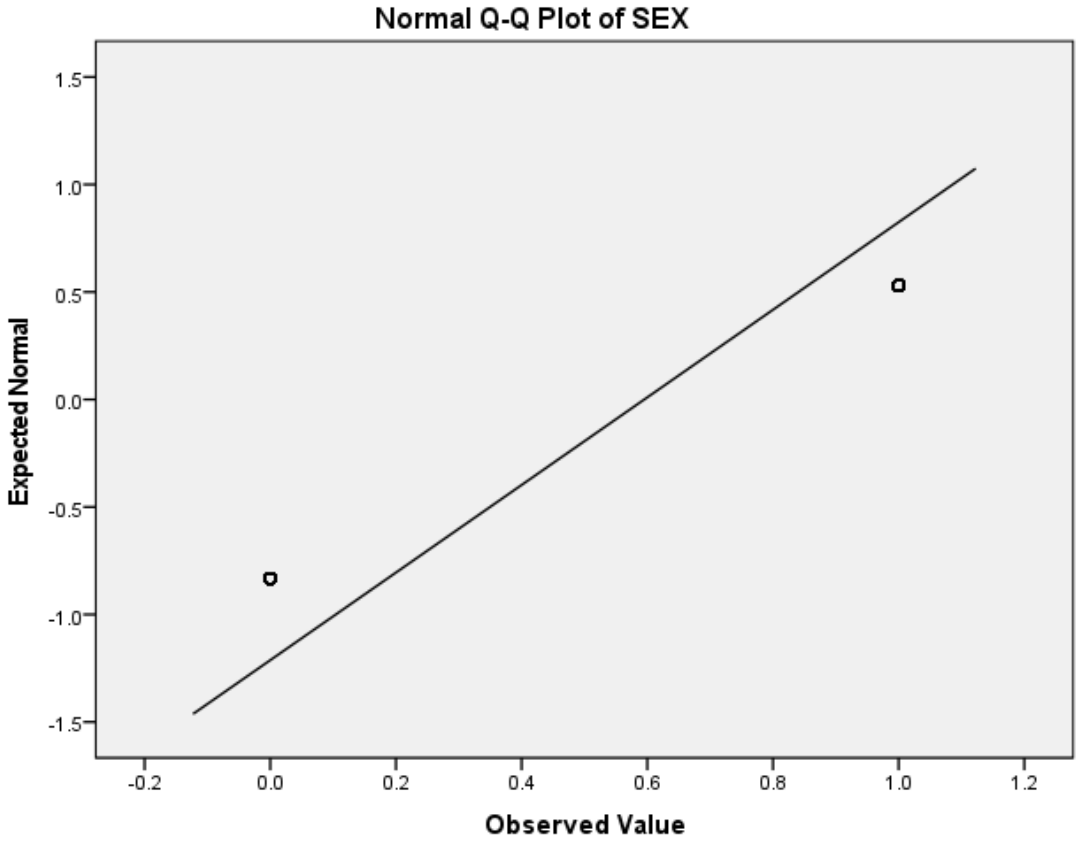
Zissimopoulos, J., Crimmins, E., & St. Clair, P. (2014). The value of delaying Alzheimer's disease onset. *Forum for Health Economics & Policy*, *18*(1), 25-39. doi: 10.1515/fhep-2014-0013

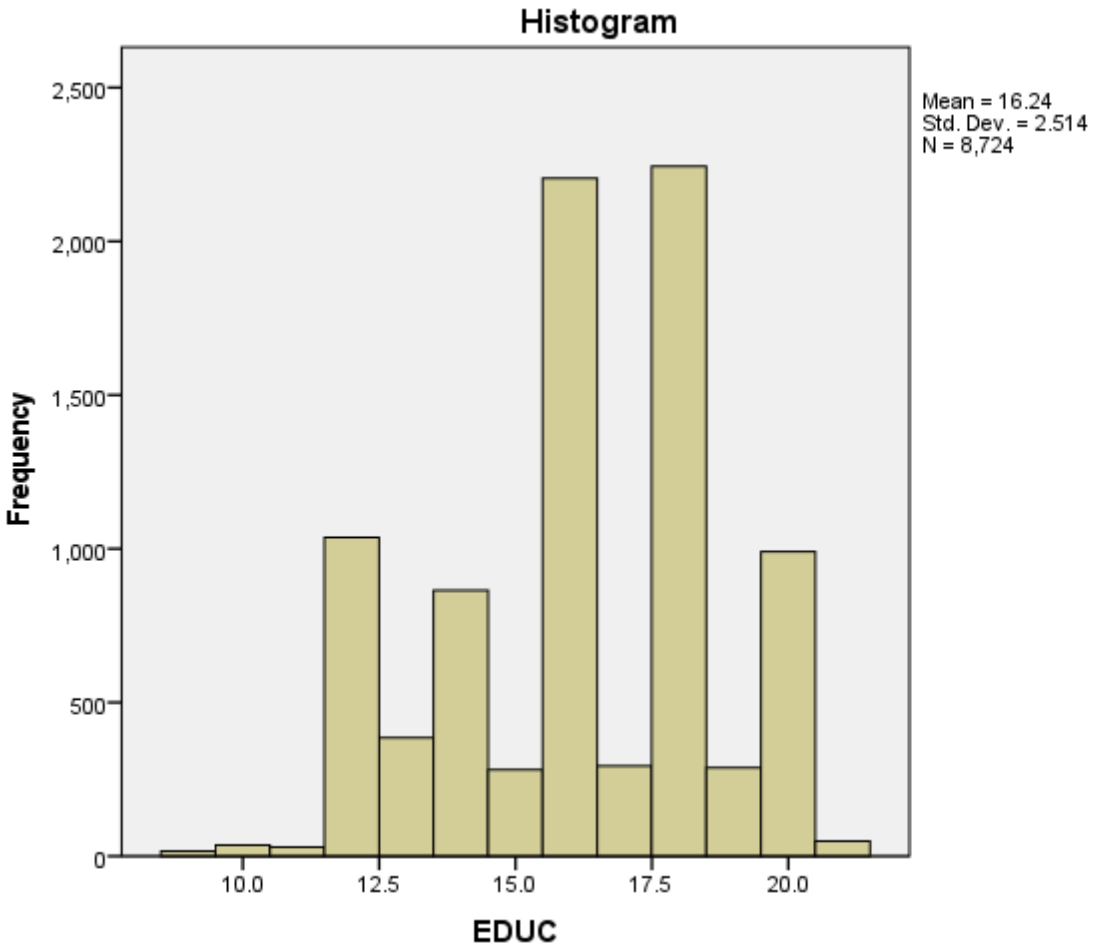
Appendix: Histograms and Q – Q Plots from SPSS

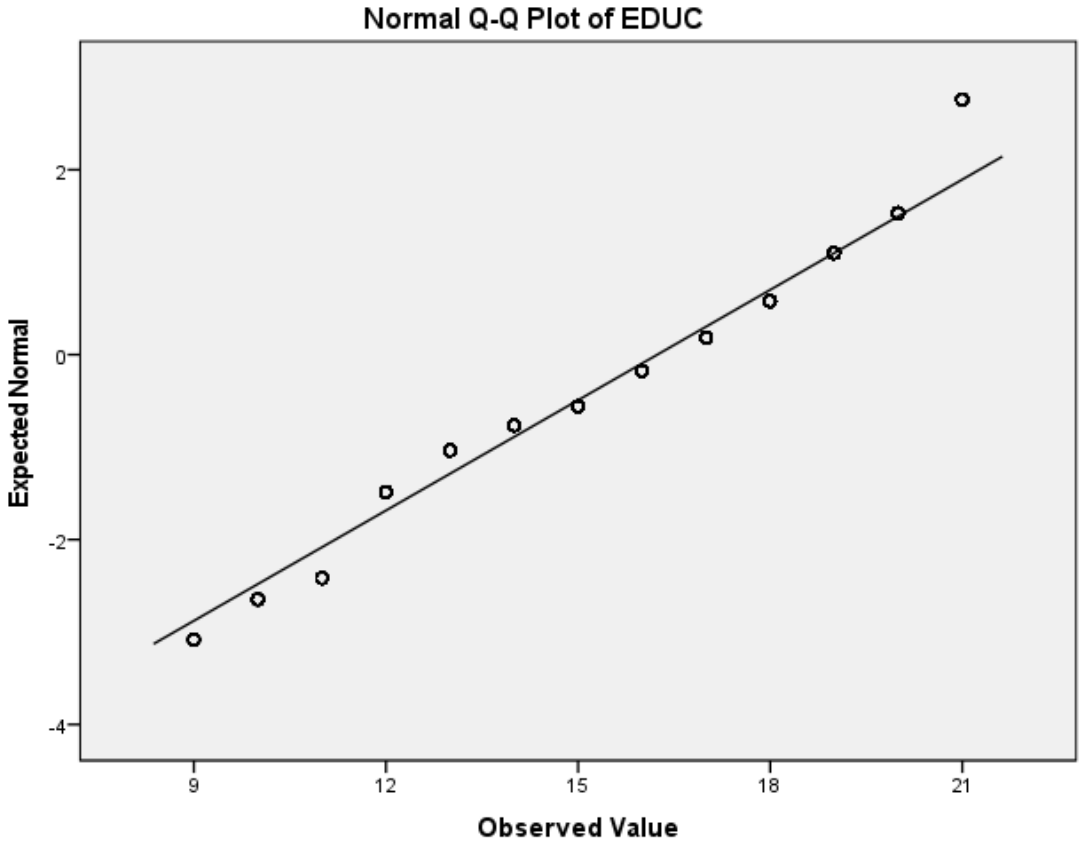


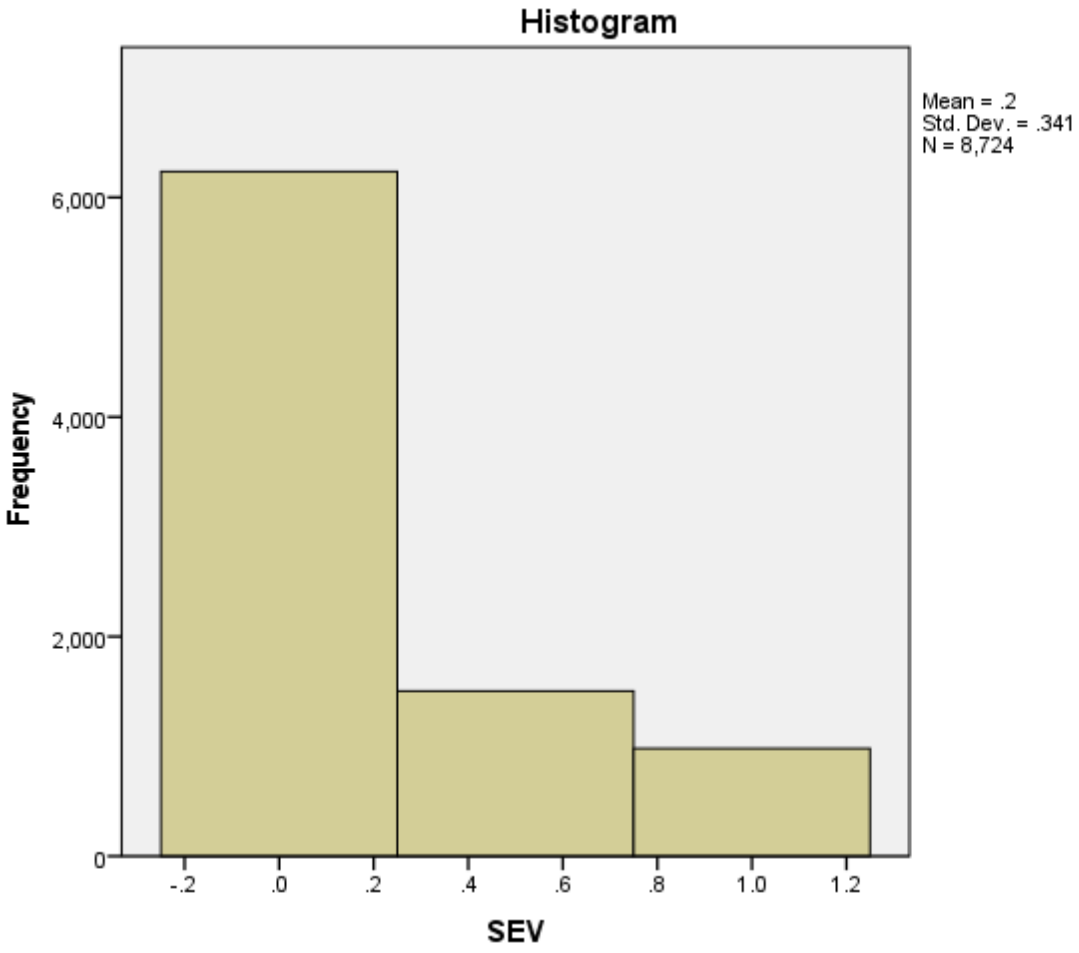

Normal Q-Q Plot of NACCAGE
Expected Normal
4
2
0
-2
50
60
70
80
90
100
110
Observed Value

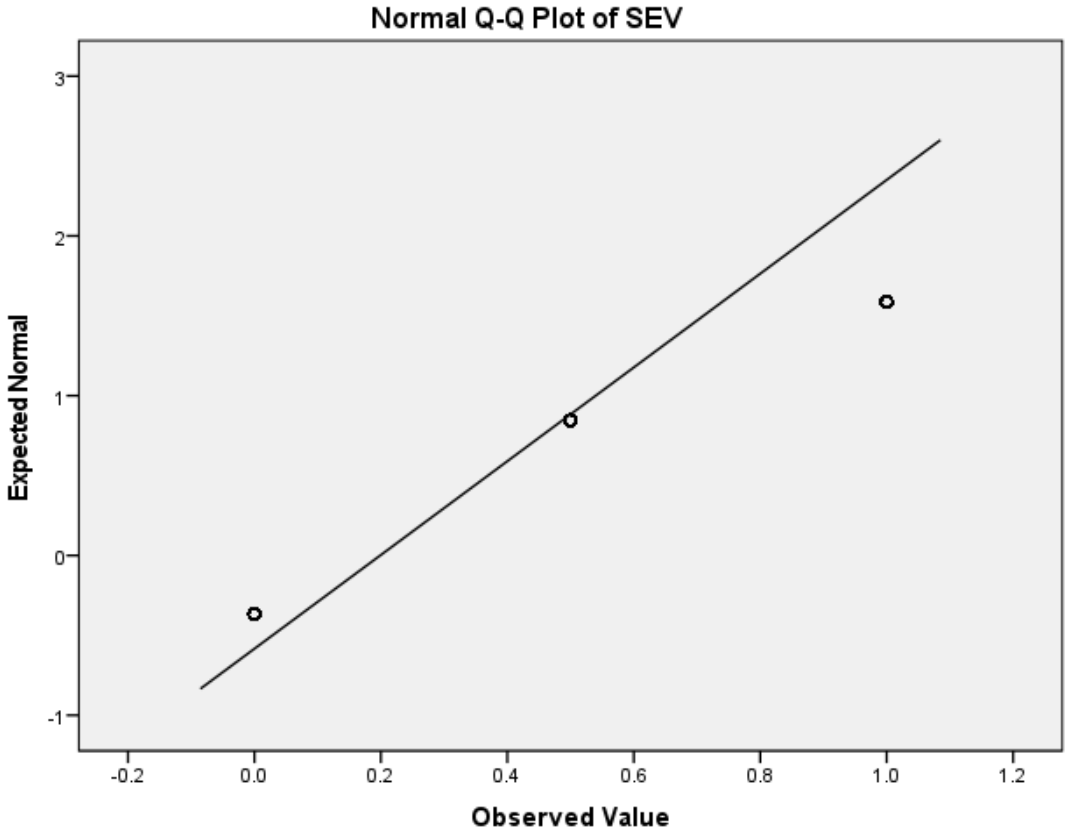

Histogram
Mean = .6
Std. Dev. = .491
N = 8,724
Frequency
6,000
5,000
4,000
3,000
2,000
1,000
0
-.5
.0
.5
1.0
1.5
SEX

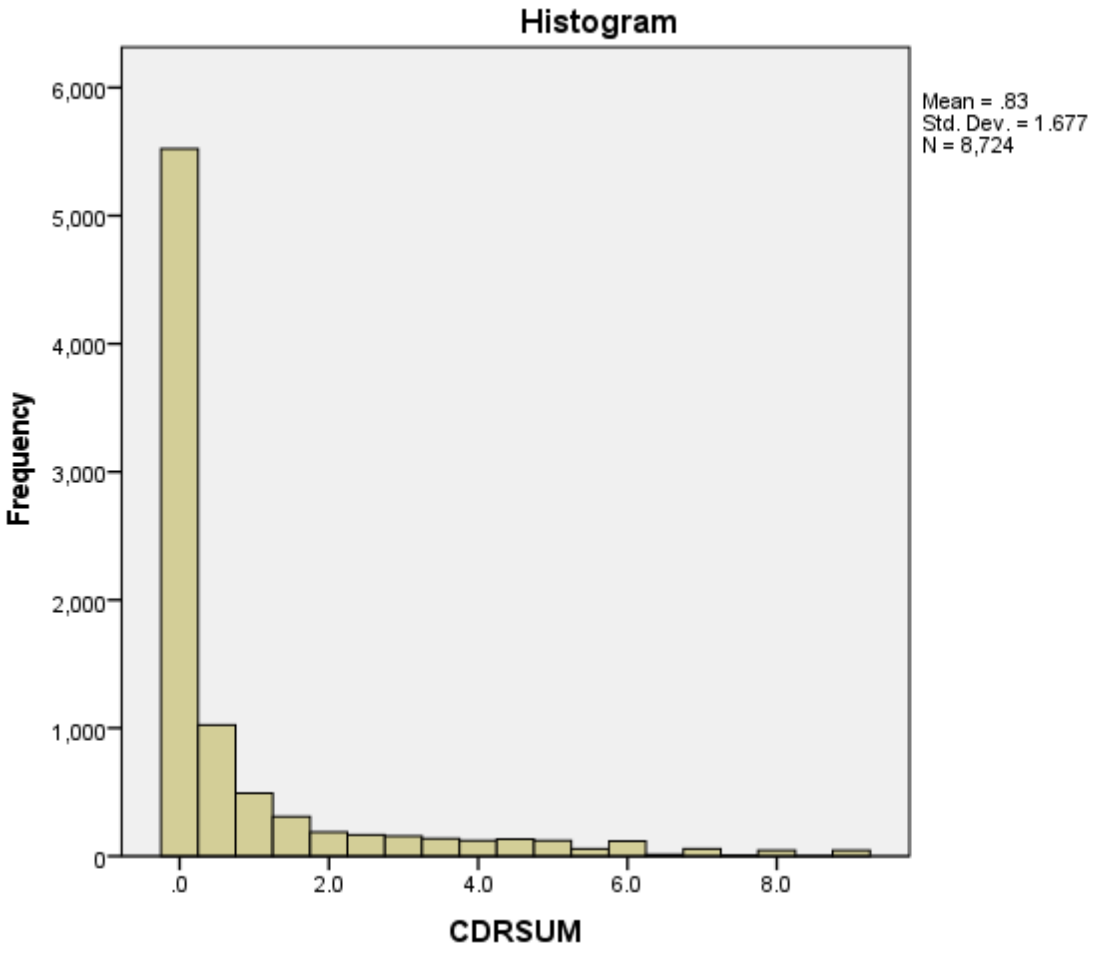

Normal Q-Q Plot of SEX
Expected Normal
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Observed Value

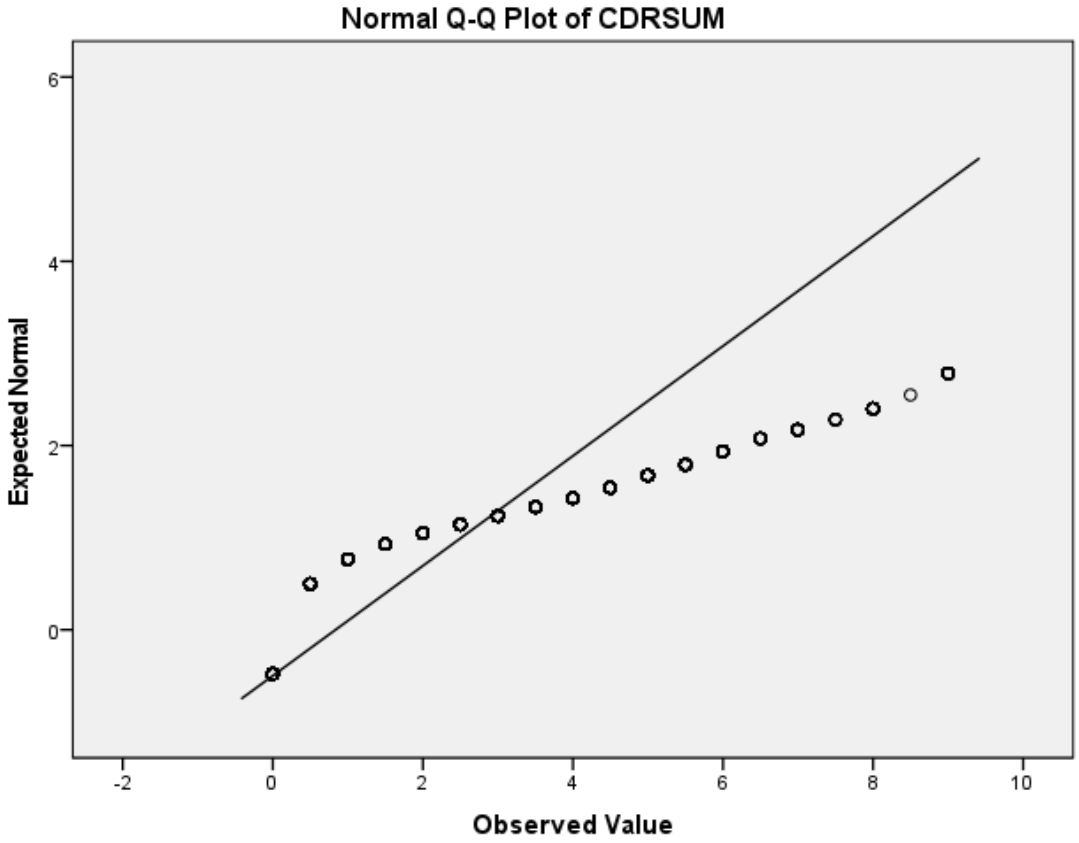

Histogram
Mean = 16.24
Std. Dev. = 2.514
N = 8,724
Frequency
2,500
2,000
1,500
1,000
500
0
10.0
12.5
15.0
17.5
20.0
EDUC

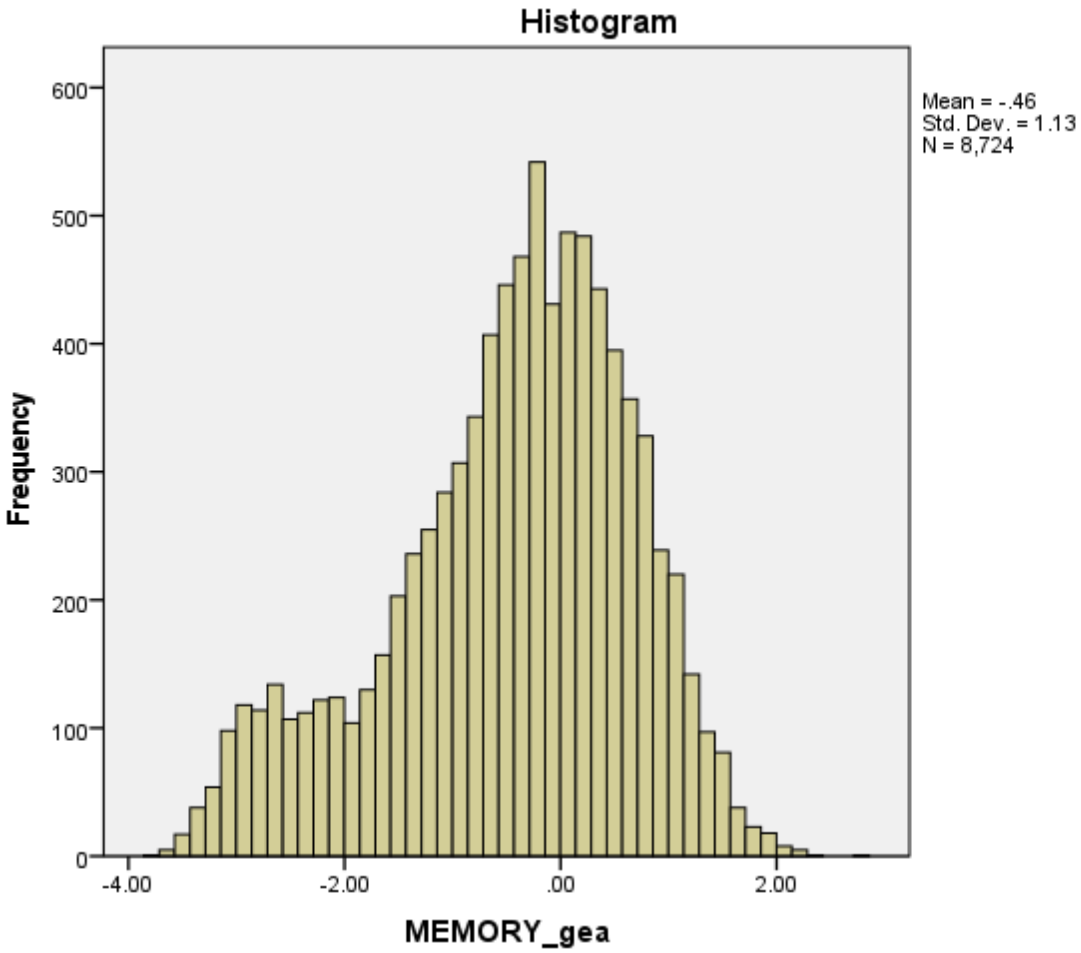

Normal Q-Q Plot of EDUC
Expected Normal
2
0
-2
-4
9
12
15
18
21
Observed Value

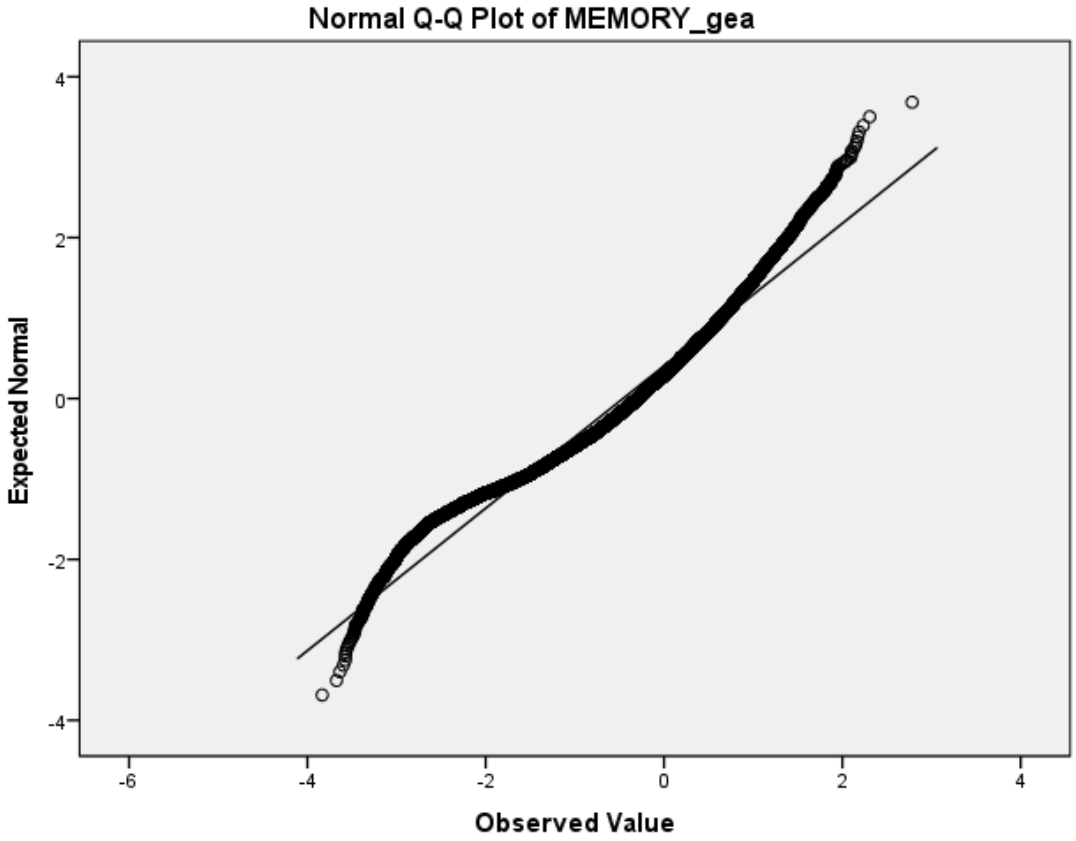

Histogram
Mean = .2
Std. Dev. = .341
N = 8,724
Frequency
6,000
4,000
2,000
0
-.2
.0
.2
.4
.6
.8
1.0
1.2
SEV

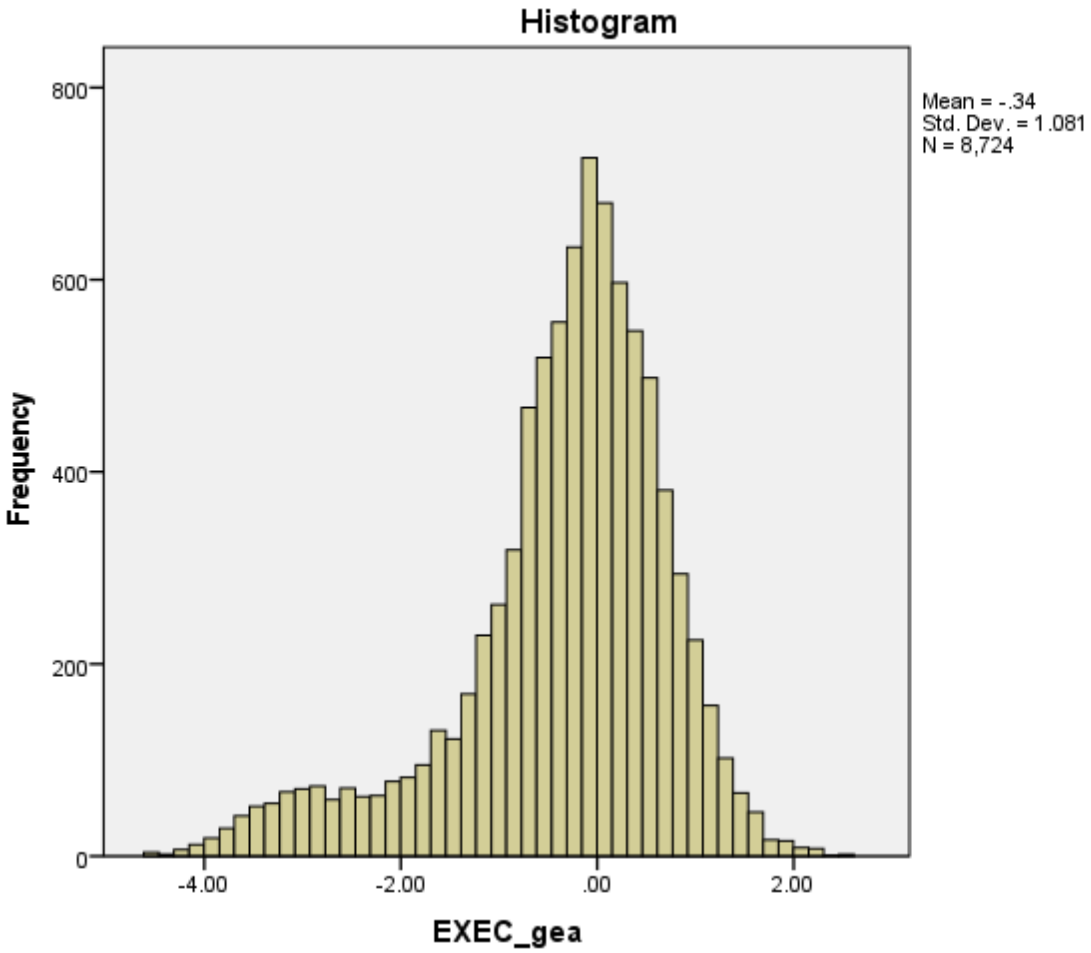

Normal Q-Q Plot of SEV
Expected Normal
Observed Value
3
2
1
0
-1
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

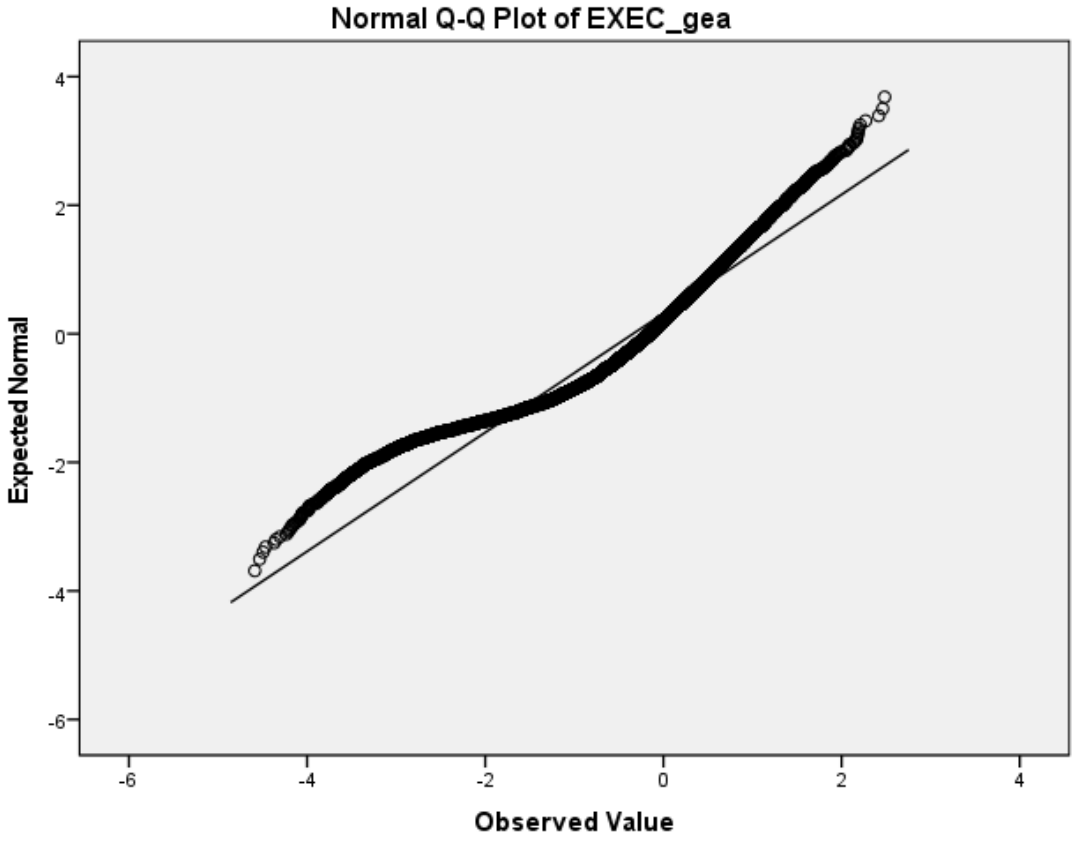

Histogram
Mean = .83
Std. Dev. = 1.677
N = 8,724
Frequency
6,000
5,000
4,000
3,000
2,000
1,000
0
.0
2.0
4.0
6.0
8.0
CDRSUM

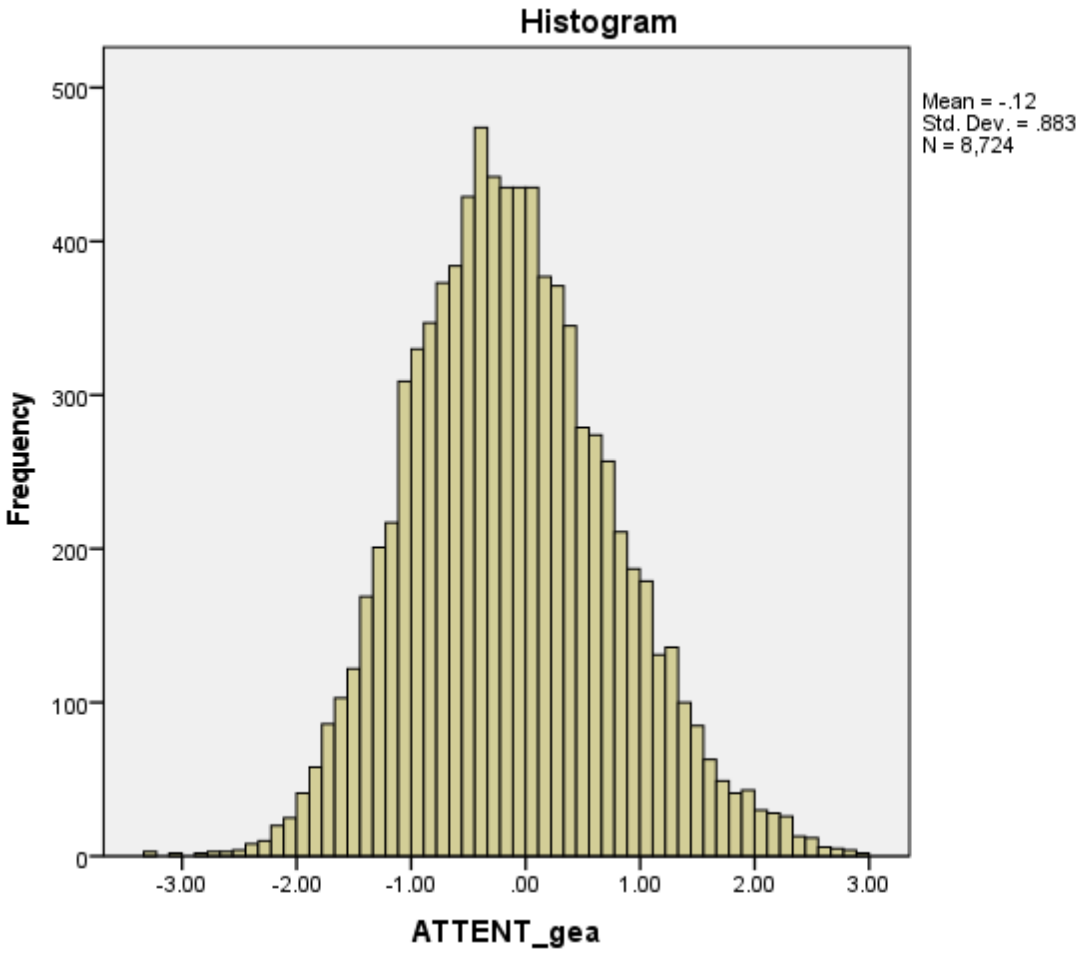

Normal Q-Q Plot of CDRSUM
Expected Normal
Observed Value
6
4
2
0
-2
0
2
4
6
8
10

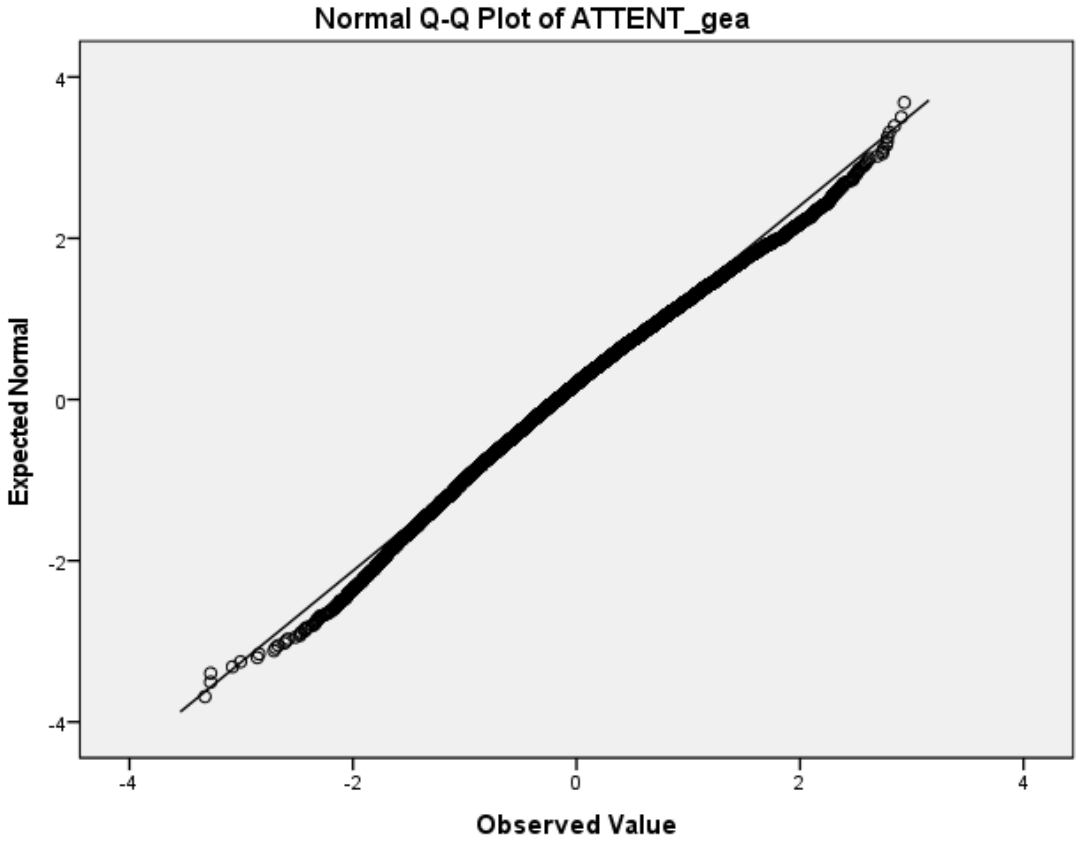

Histogram
Mean = -.46
Std. Dev. = 1.13
N = 8,724
Frequency
600
500
400
300
200
100
0
-4.00
-2.00
.00
2.00
MEMORY_gea

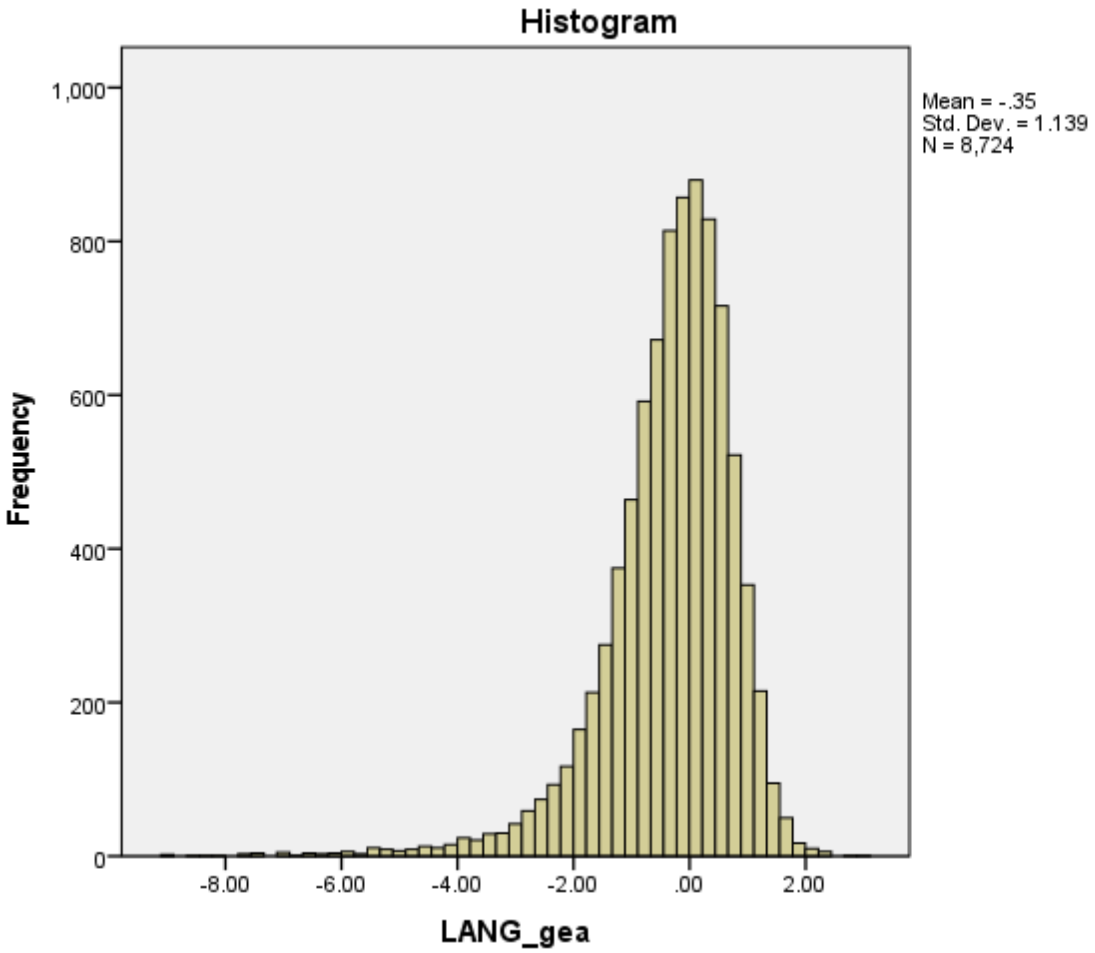

Normal Q-Q Plot of MEMORY_gea
Expected Normal
Observed Value
4
2
0
-2
-4
-6
-4
-2
0
2
4

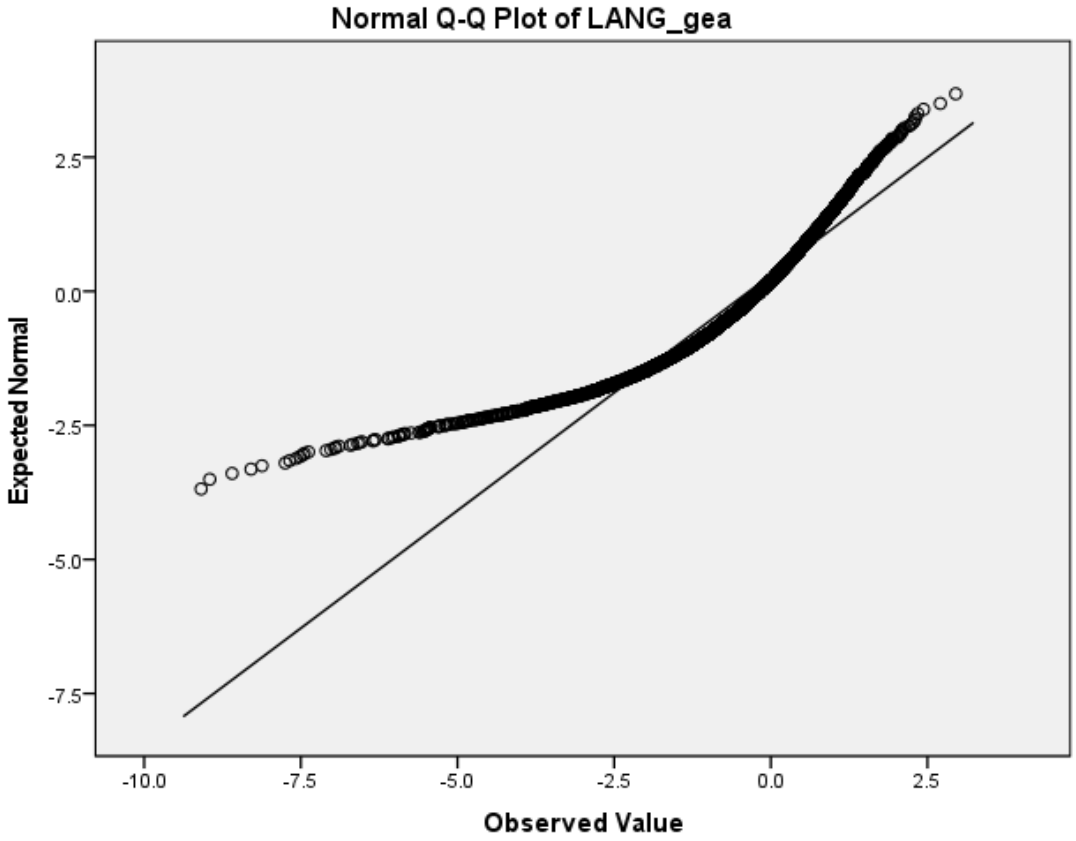

Histogram
Mean = -.34
Std. Dev. = 1.081
N = 8,724
Frequency
800
600
400
200
0
-4.00
-2.00
.00
2.00
EXEC_gea

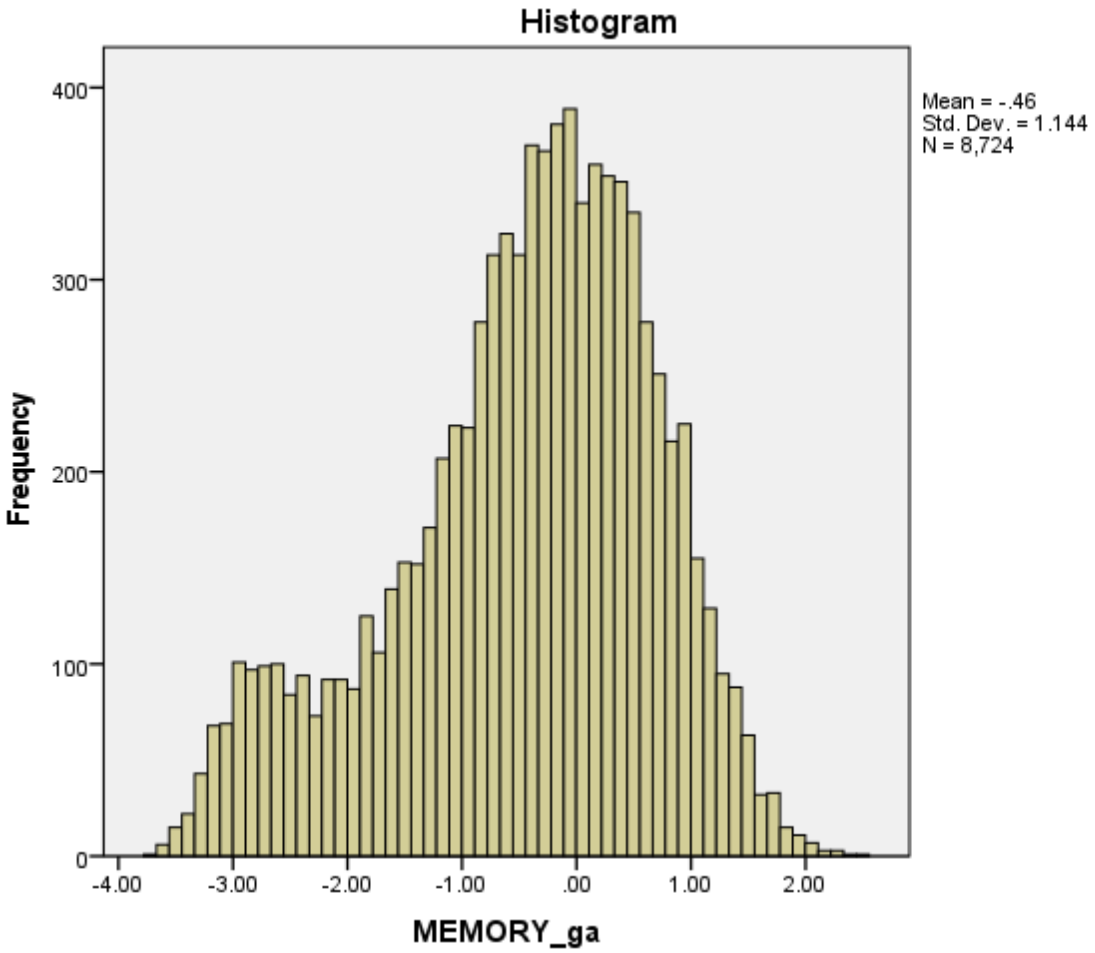

Normal Q-Q Plot of EXEC_gea
Expected Normal
Observed Value
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4

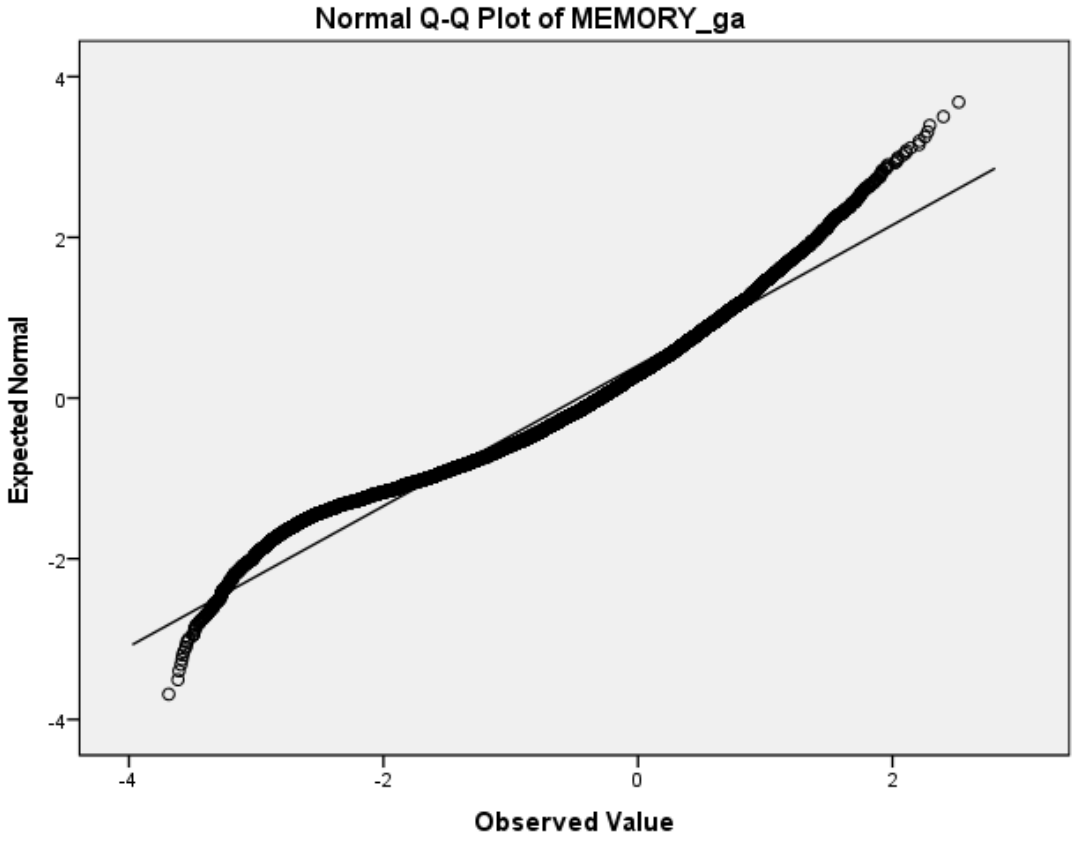

Histogram
Mean = -.12
Std. Dev. = .883
N = 8,724
Frequency
500
400
300
200
100
0
-3.00
-2.00
-1.00
.00
1.00
2.00
3.00
ATTENT_gea

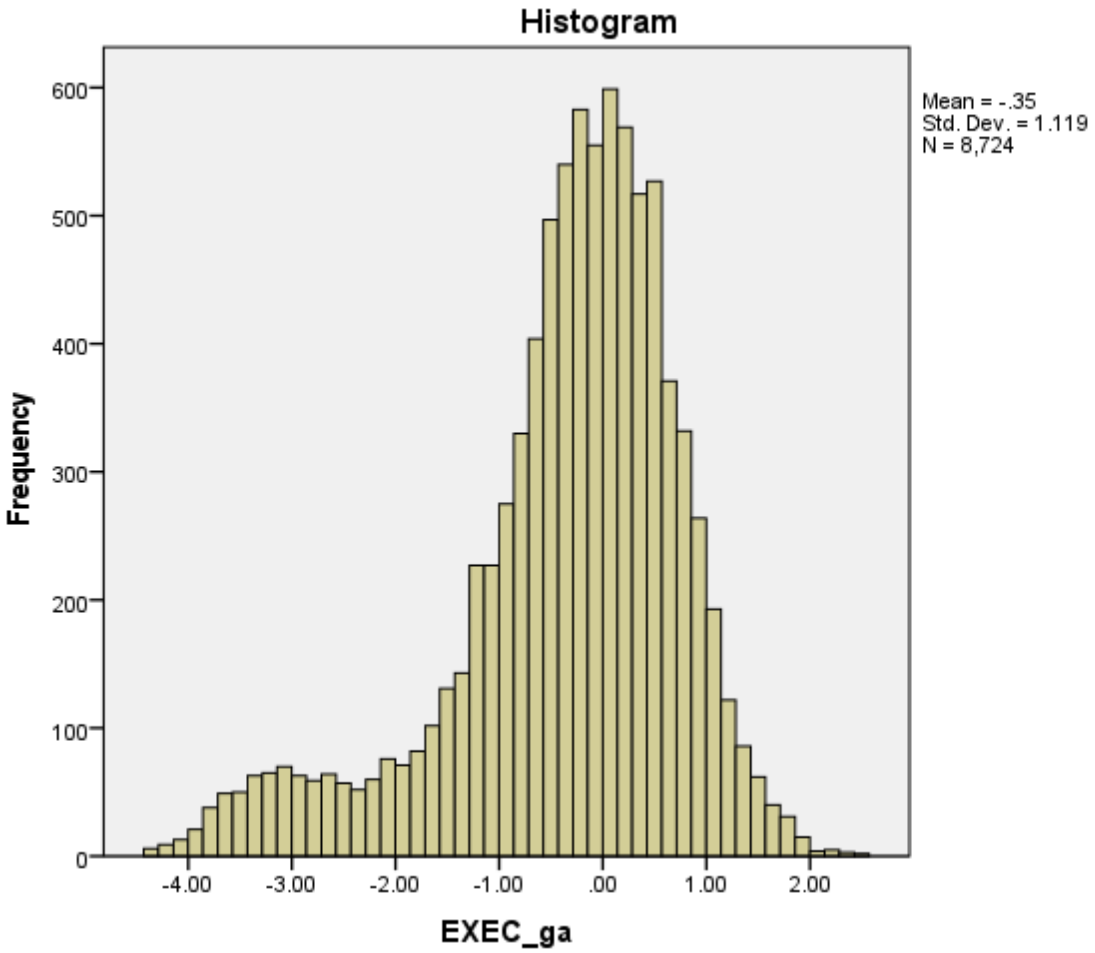

Normal Q-Q Plot of ATTENT_gea
Expected Normal
Observed Value
4
2
0
-2
-4
-4
-2
0
2
4

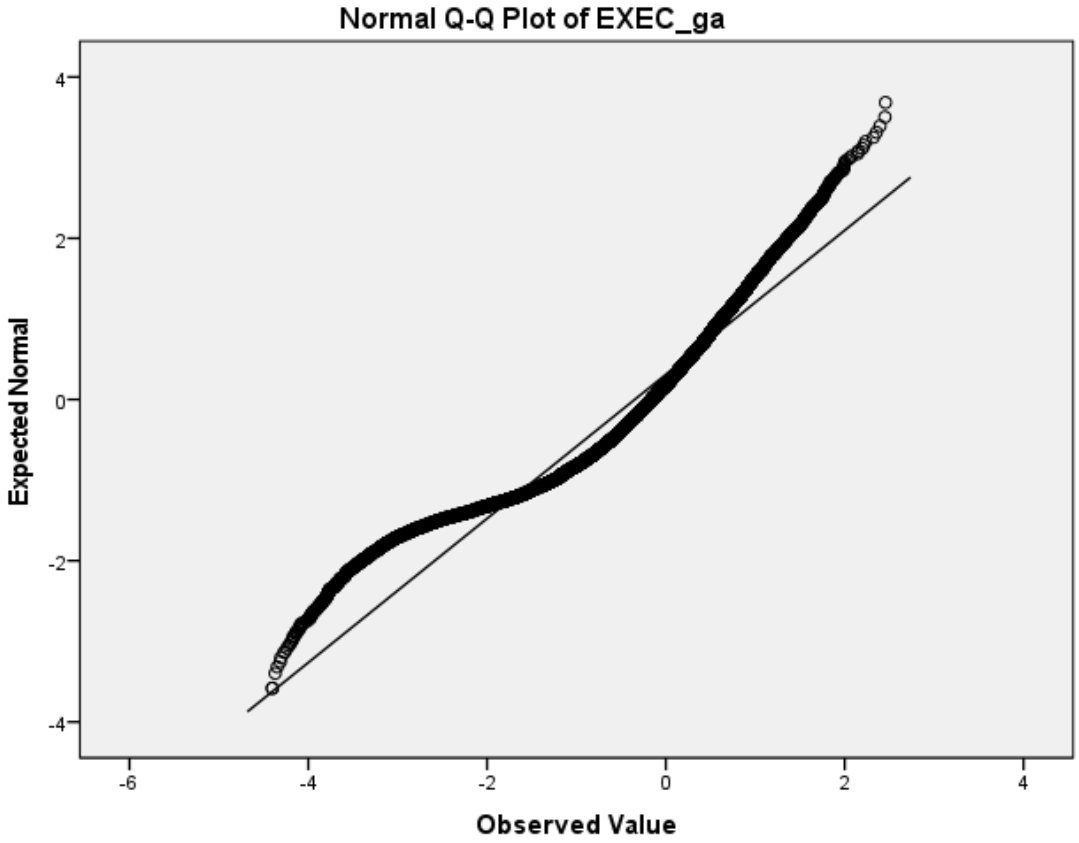

Histogram
Mean = -.35
Std. Dev. = 1.139
N = 8,724
Frequency
1,000
800
600
400
200
0
-8.00
-6.00
-4.00
-2.00
.00
2.00
LANG_gea

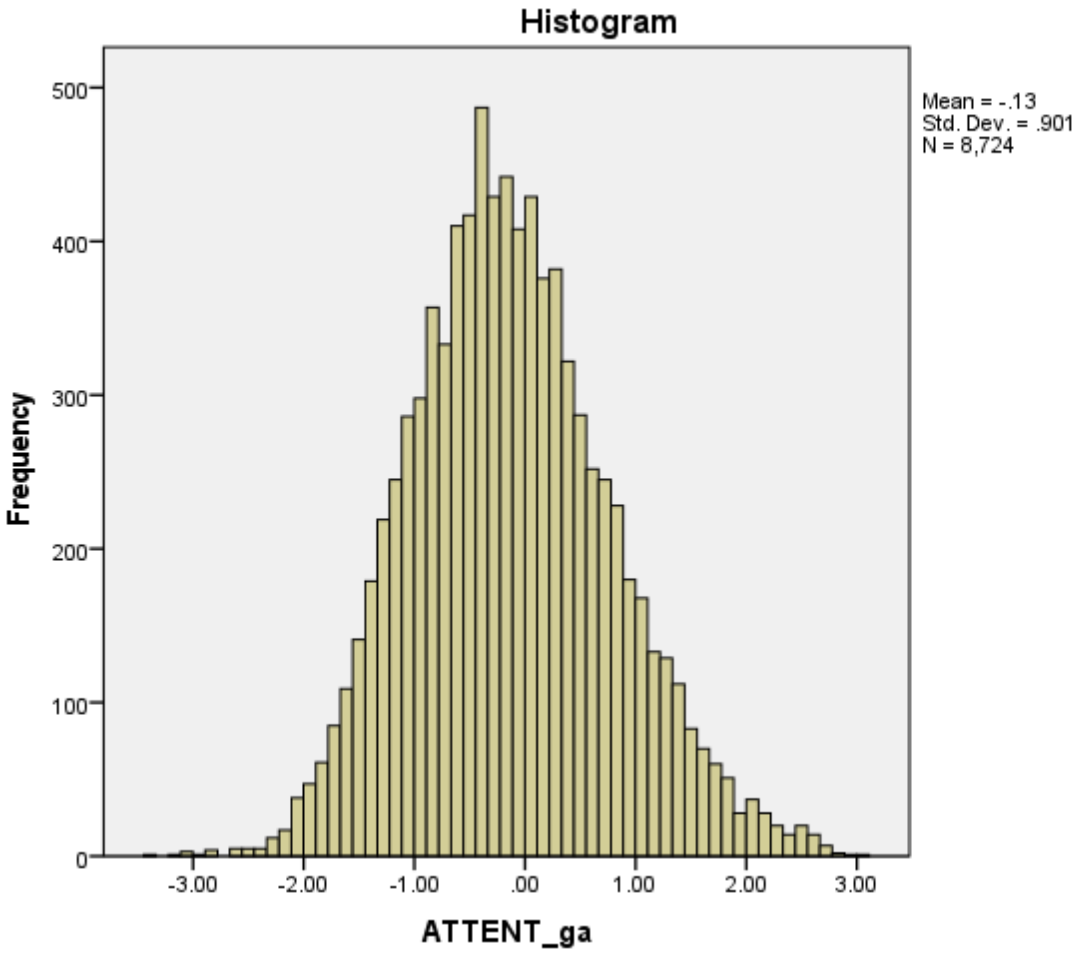

Normal Q-Q Plot of LANG_gea
Expected Normal
2.5
0.0
-2.5
-5.0
-7.5
-10.0
-7.5
-5.0
-2.5
0.0
2.5
Observed Value

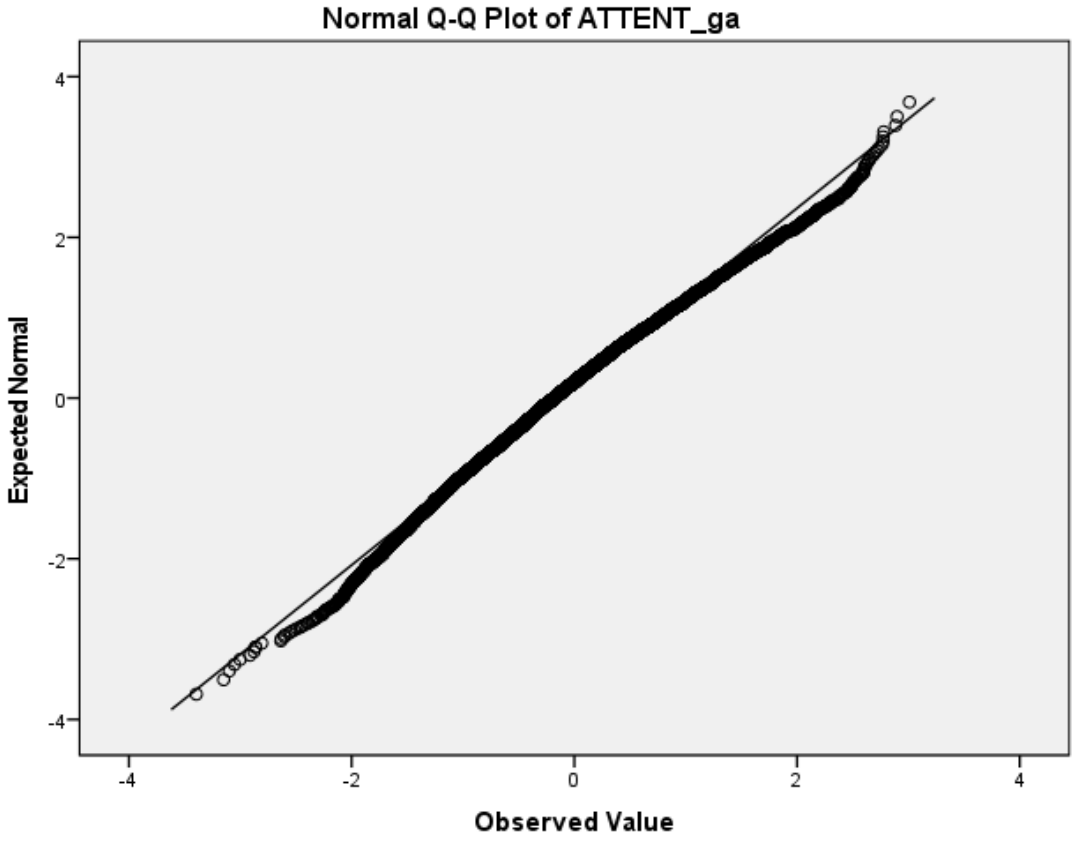

Histogram
Mean = -.46
Std. Dev. = 1.144
N = 8,724
Frequency
400
300
200
100
0
-4.00
-3.00
-2.00
-1.00
.00
1.00
2.00
MEMORY_ga

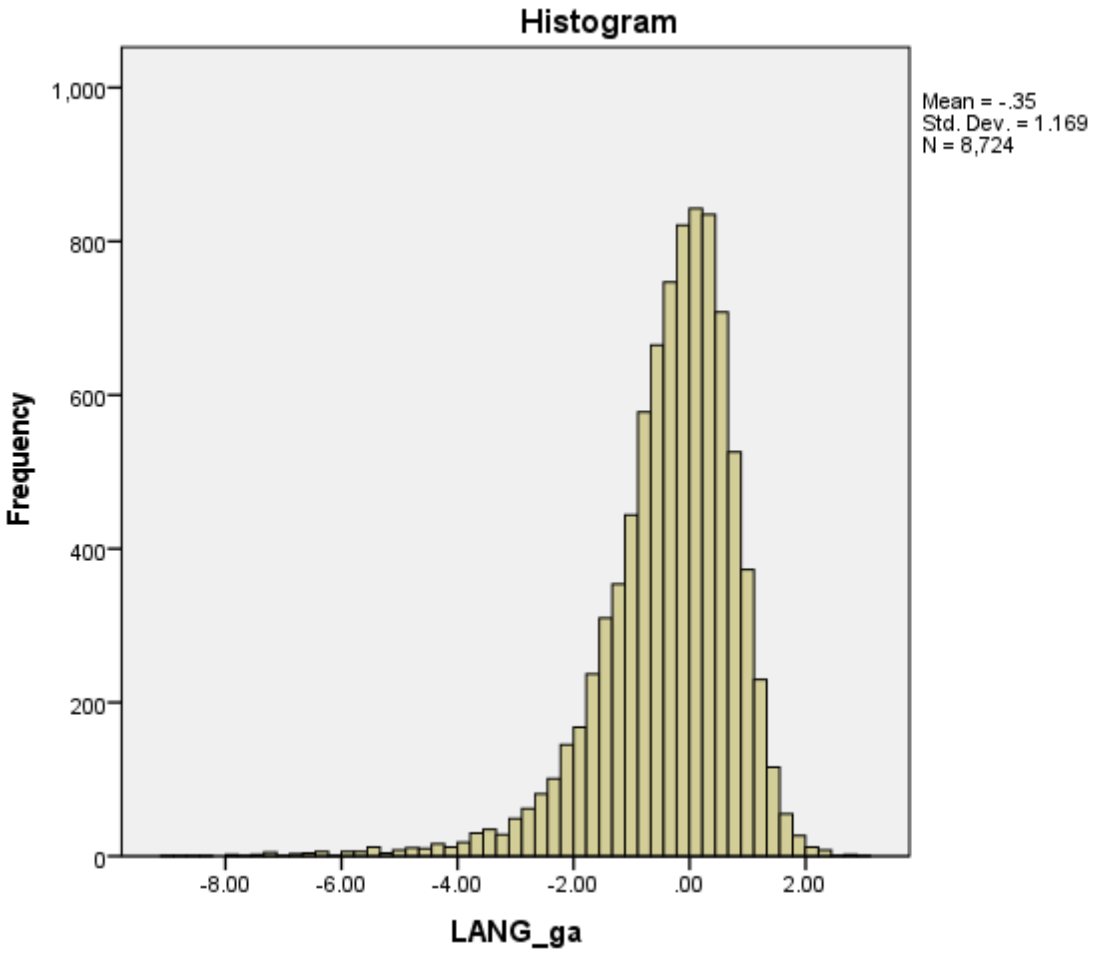

Normal Q-Q Plot of MEMORY_ga
Expected Normal
Observed Value
4
2
0
-2
-4
-4
-2
0
2

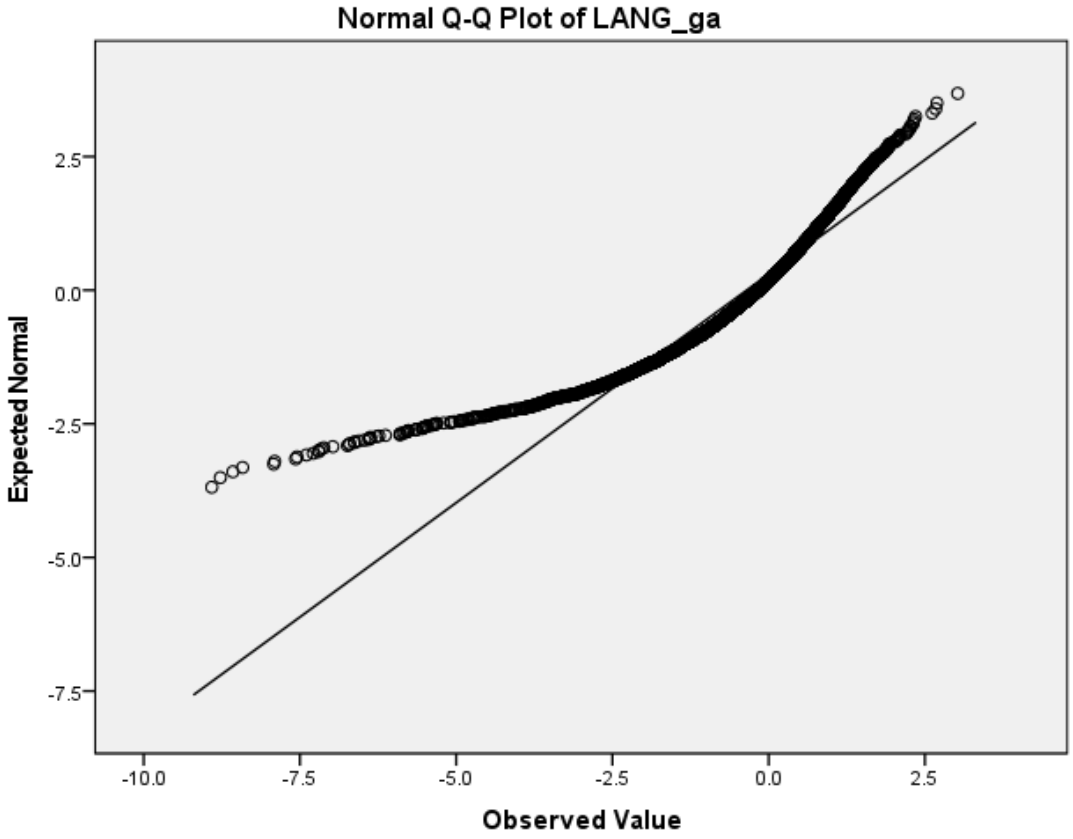

Histogram
Mean = -.35
Std. Dev. = 1.119
N = 8,724
Frequency
600
500
400
300
200
100
0
-4.00
-3.00
-2.00
-1.00
.00
1.00
2.00
EXEC_ga

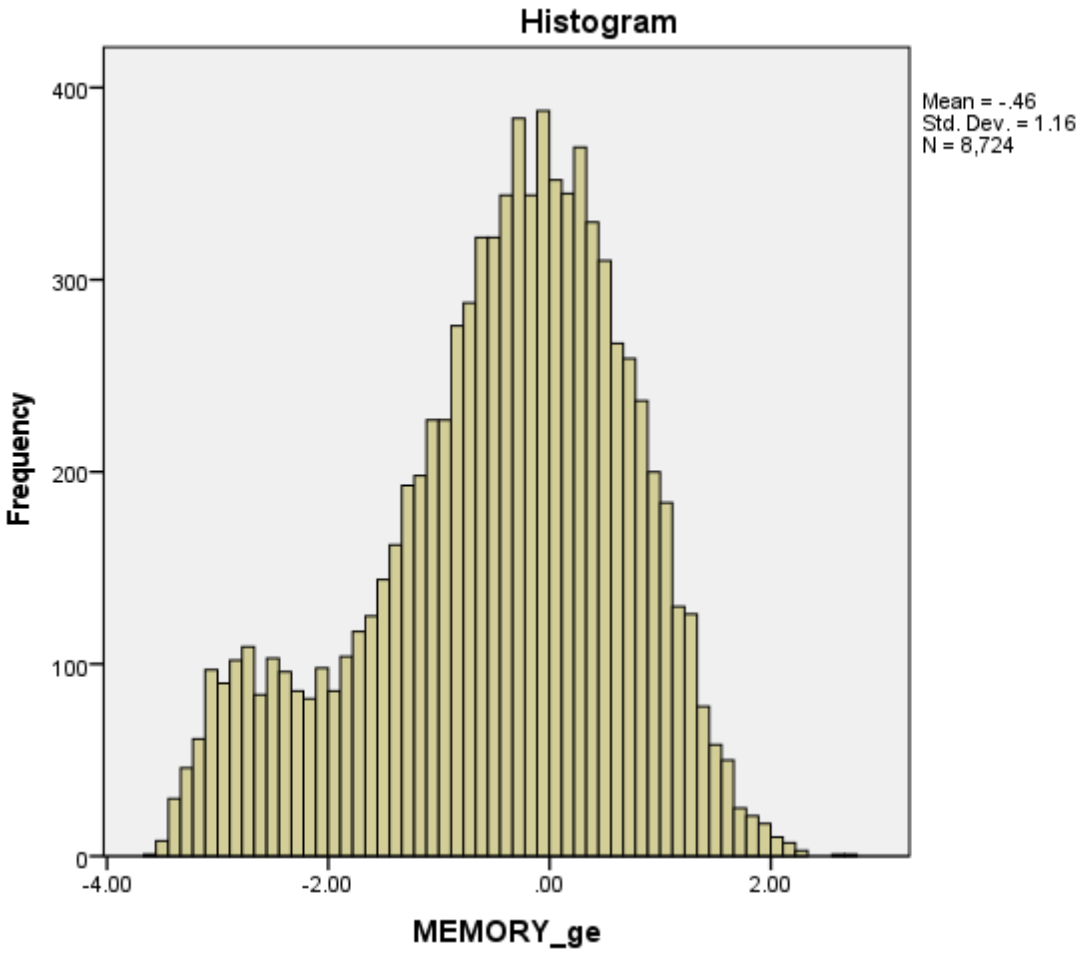

Normal Q-Q Plot of EXEC_ga
Expected Normal
Observed Value
4
2
0
-2
-4
-6
-4
-2
0
2
4

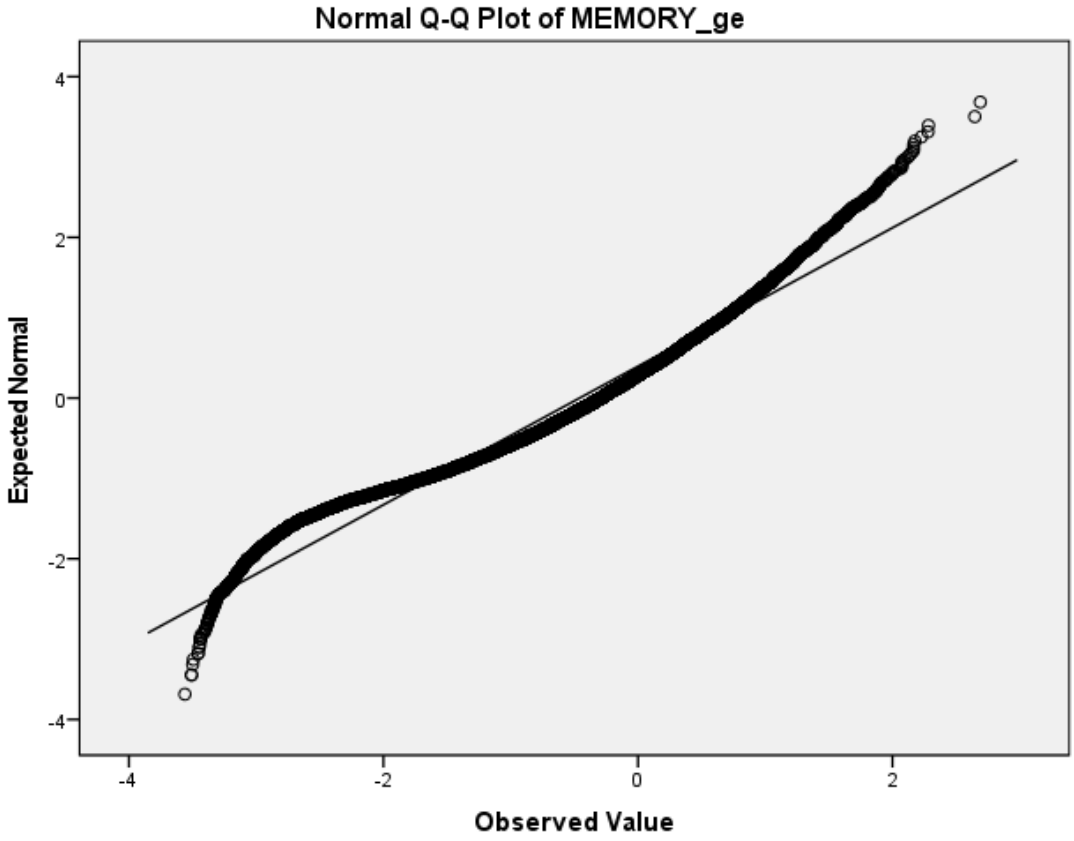

Histogram
Mean = -.13
Std. Dev. = .901
N = 8,724
Frequency
500
400
300
200
100
0
-3.00
-2.00
-1.00
.00
1.00
2.00
3.00
ATTENT_ga

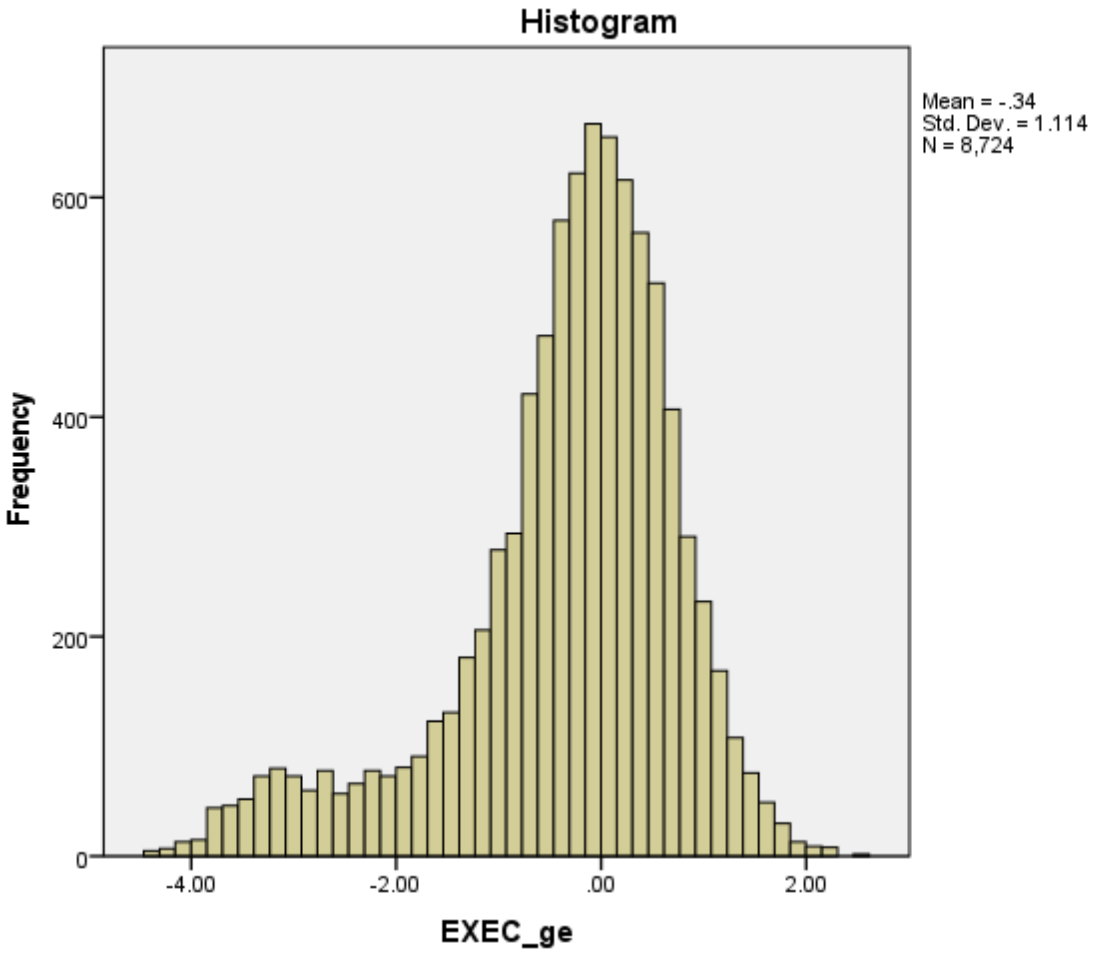

Normal Q-Q Plot of ATTENT_ga
Expected Normal
Observed Value
4
2
0
-2
-4
-4
-2
0
2
4

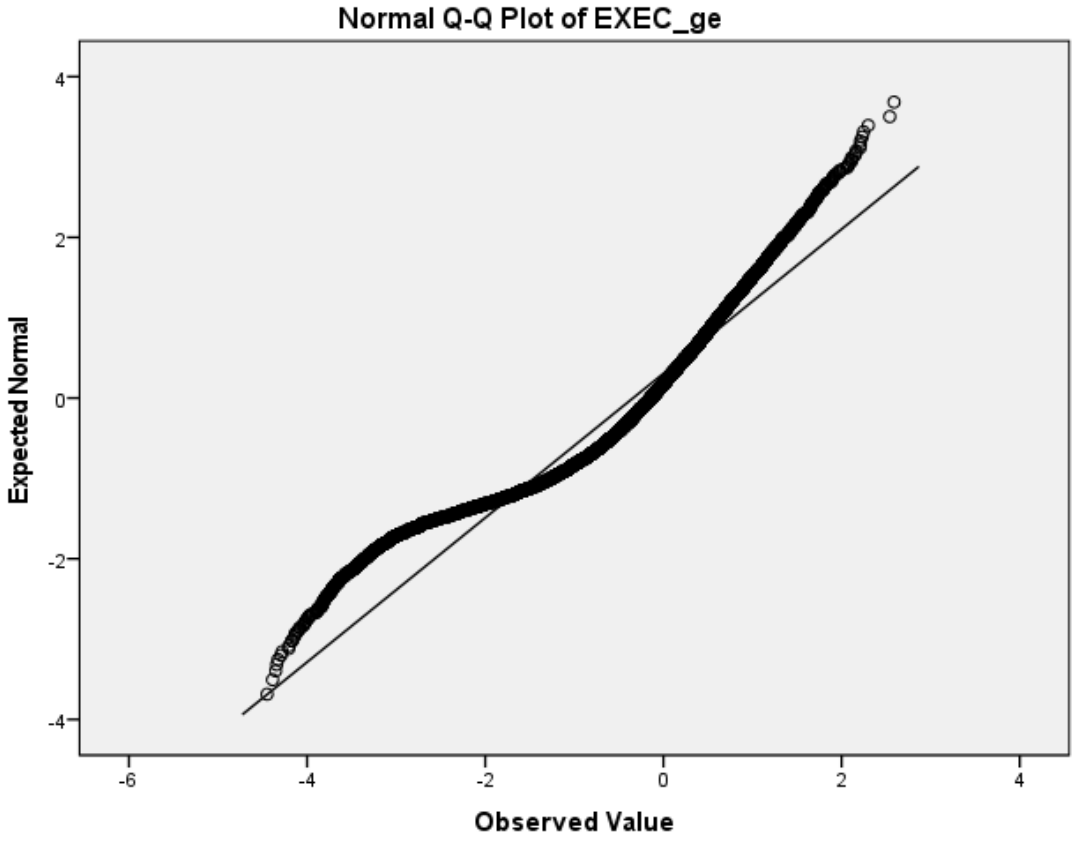

Histogram
Mean = -.35
Std. Dev. = 1.169
N = 8,724
Frequency
1,000
800
600
400
200
0
-8.00
-6.00
-4.00
-2.00
.00
2.00
LANG_ga

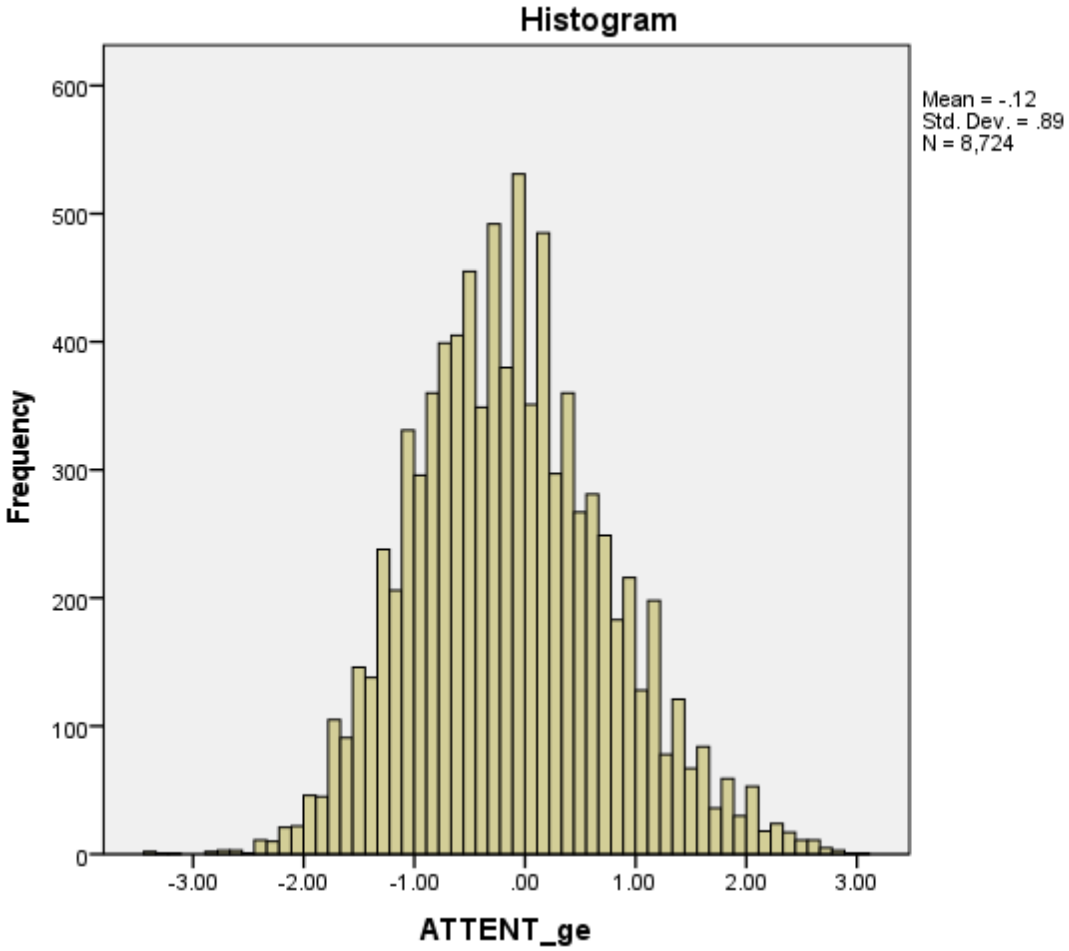

Normal Q-Q Plot of LANG_ga
Expected Normal
2.5
0.0
-2.5
-5.0
-7.5
-10.0
-7.5
-5.0
-2.5
0.0
2.5
Observed Value

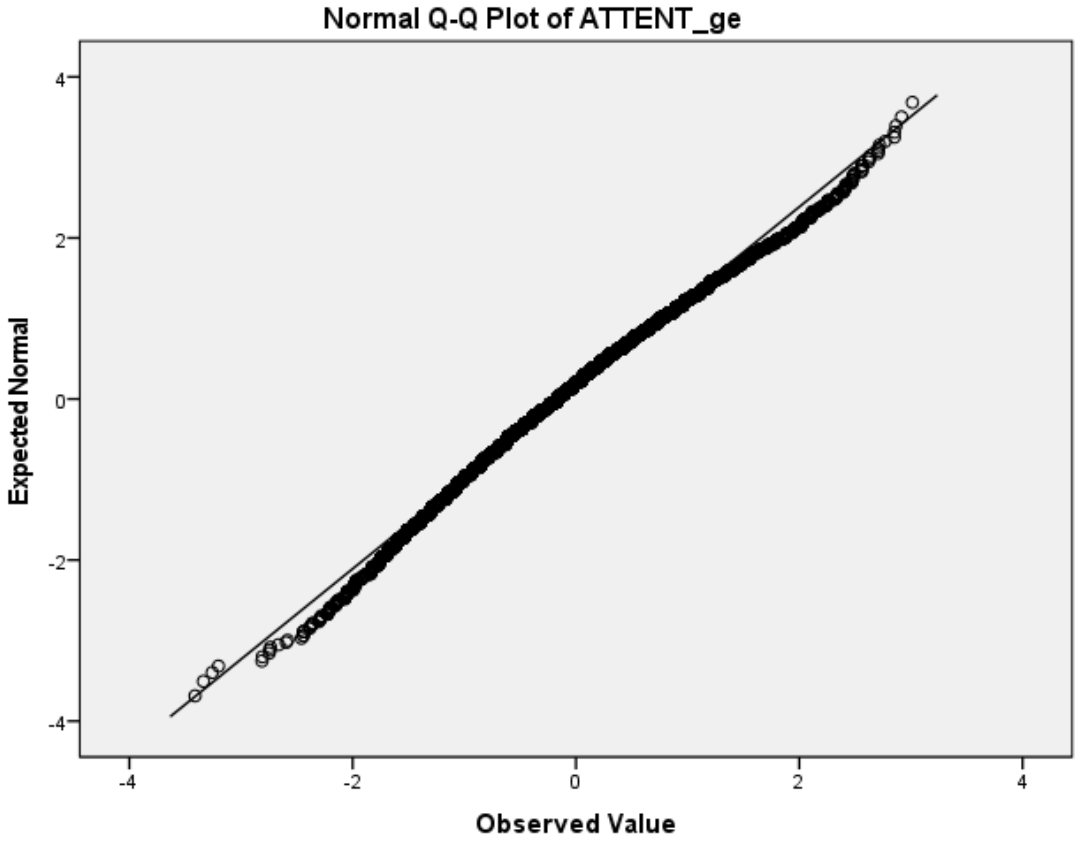

Histogram
Mean = -.46
Std. Dev. = 1.16
N = 8,724
Frequency
400
300
200
100
0
-4.00
-2.00
.00
2.00
MEMORY_ge

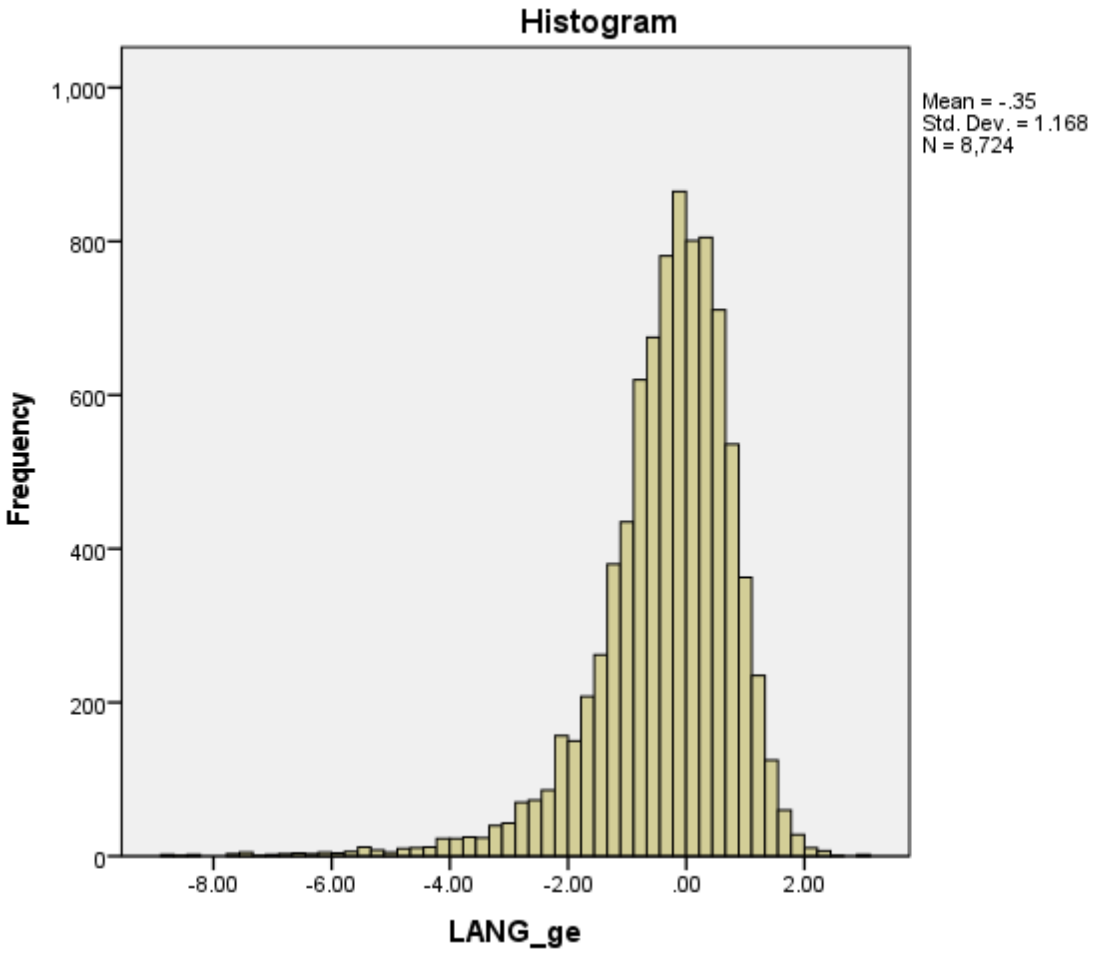

Normal Q-Q Plot of MEMORY_ge
Expected Normal
Observed Value
4
2
0
-2
-4
-4
-2
0
2

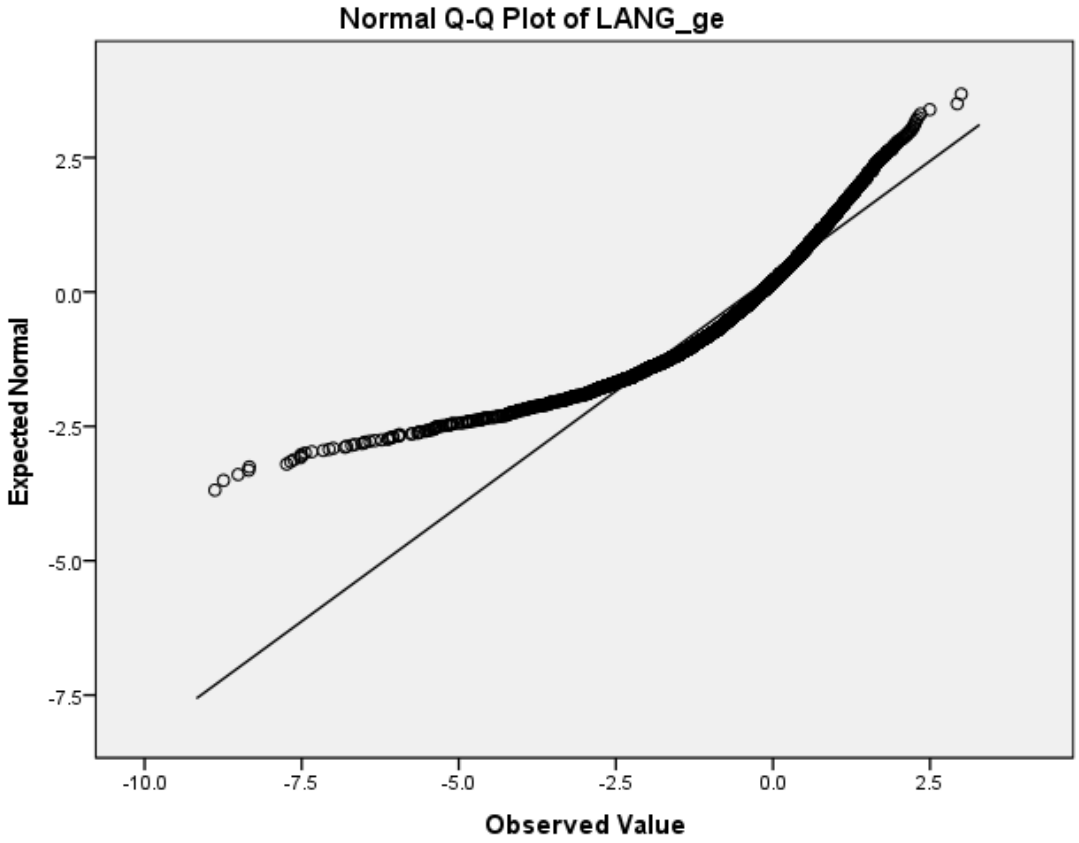

Histogram
Mean = -.34
Std. Dev. = 1.114
N = 8,724
Frequency
600
400
200
0
-4.00
-2.00
.00
2.00
EXEC_ge

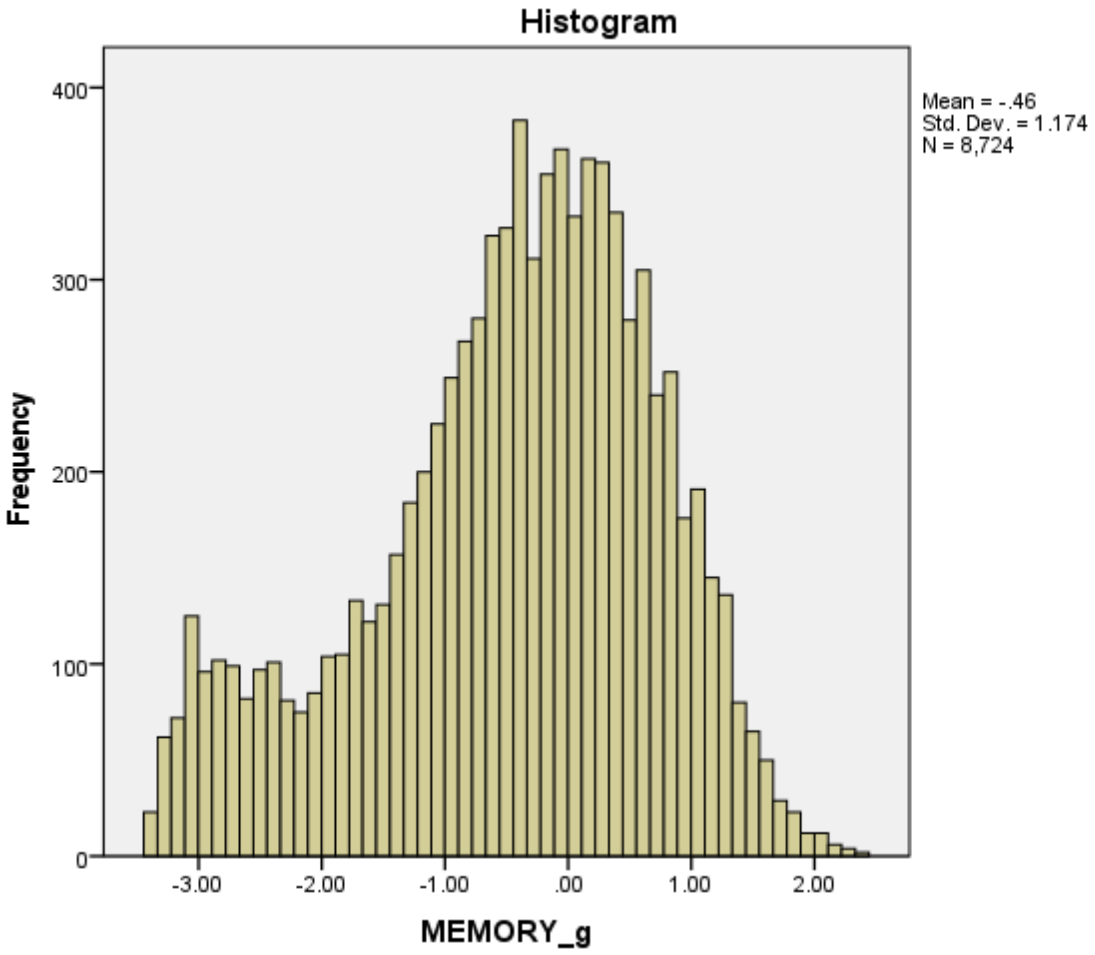

Normal Q-Q Plot of EXEC_ge
Expected Normal
Observed Value
4
2
0
-2
-4
-6
-4
-2
0
2
4

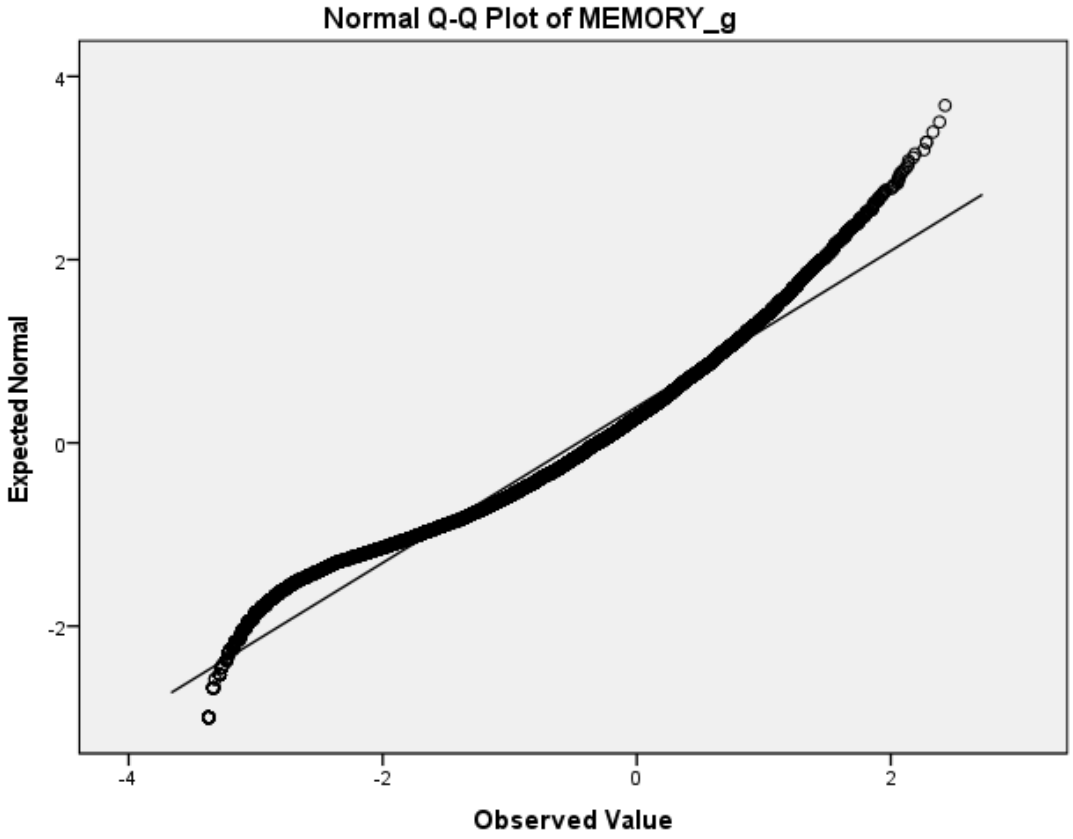

Histogram
Mean = -.12
Std. Dev. = .89
N = 8,724
Frequency
600
500
400
300
200
100
0
-3.00
-2.00
-1.00
.00
1.00
2.00
3.00
ATTENT_ge

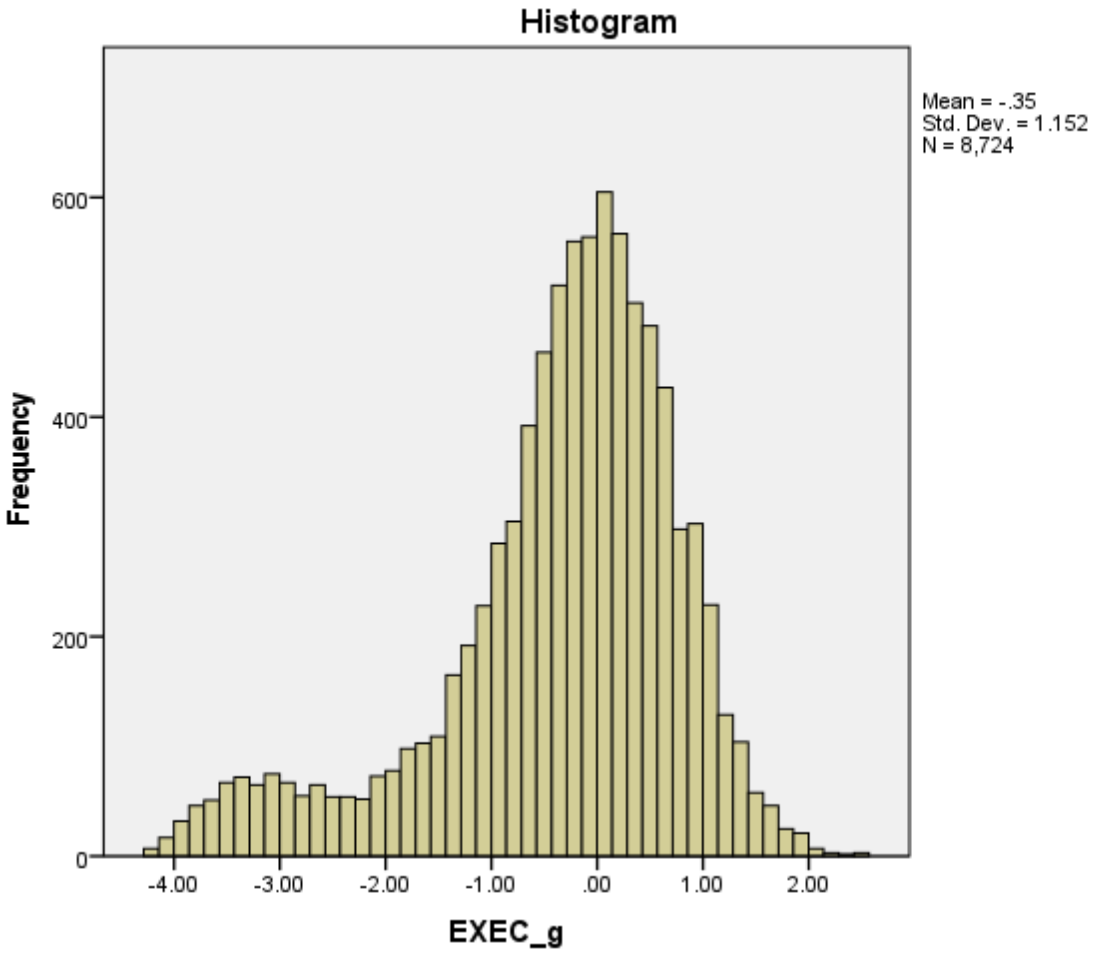

Normal Q-Q Plot of ATTENT_ge
Expected Normal
Observed Value
4
2
0
-2
-4
-4
-2
0
2
4

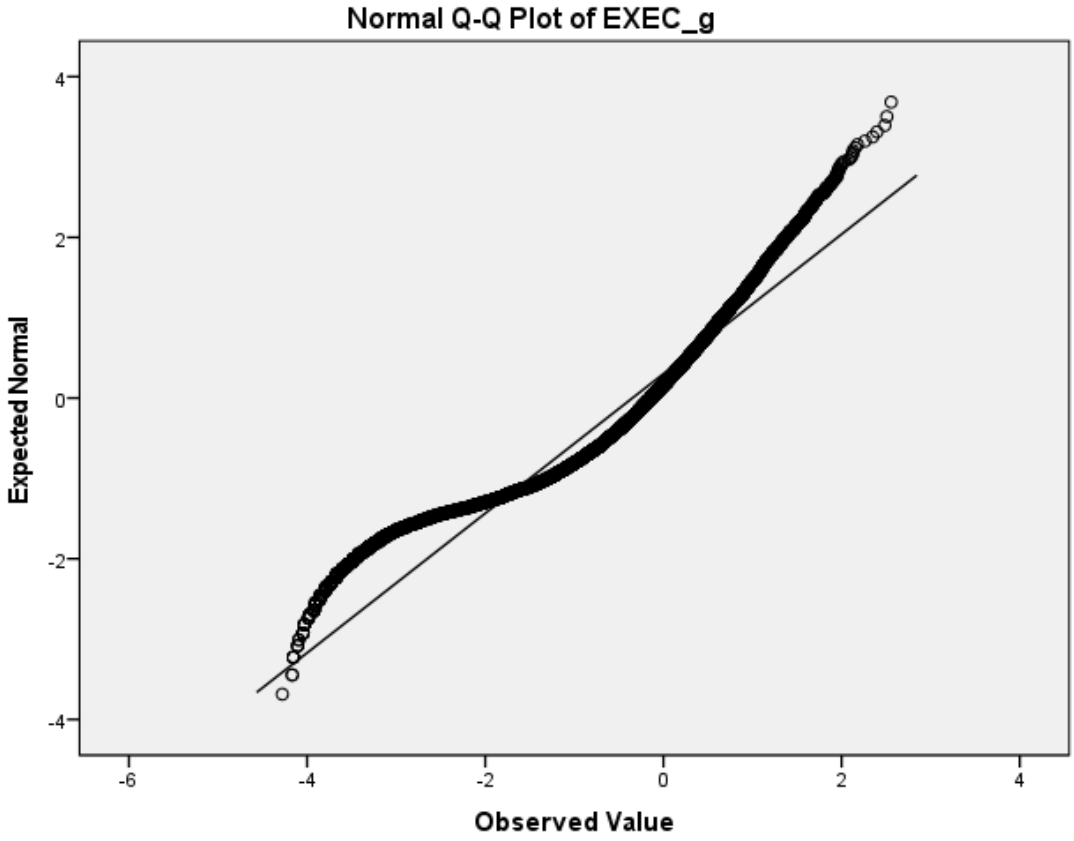

Histogram
Mean = -.35
Std. Dev. = 1.168
N = 8,724
Frequency
1,000
800
600
400
200
0
-8.00
-6.00
-4.00
-2.00
.00
2.00
LANG_ge

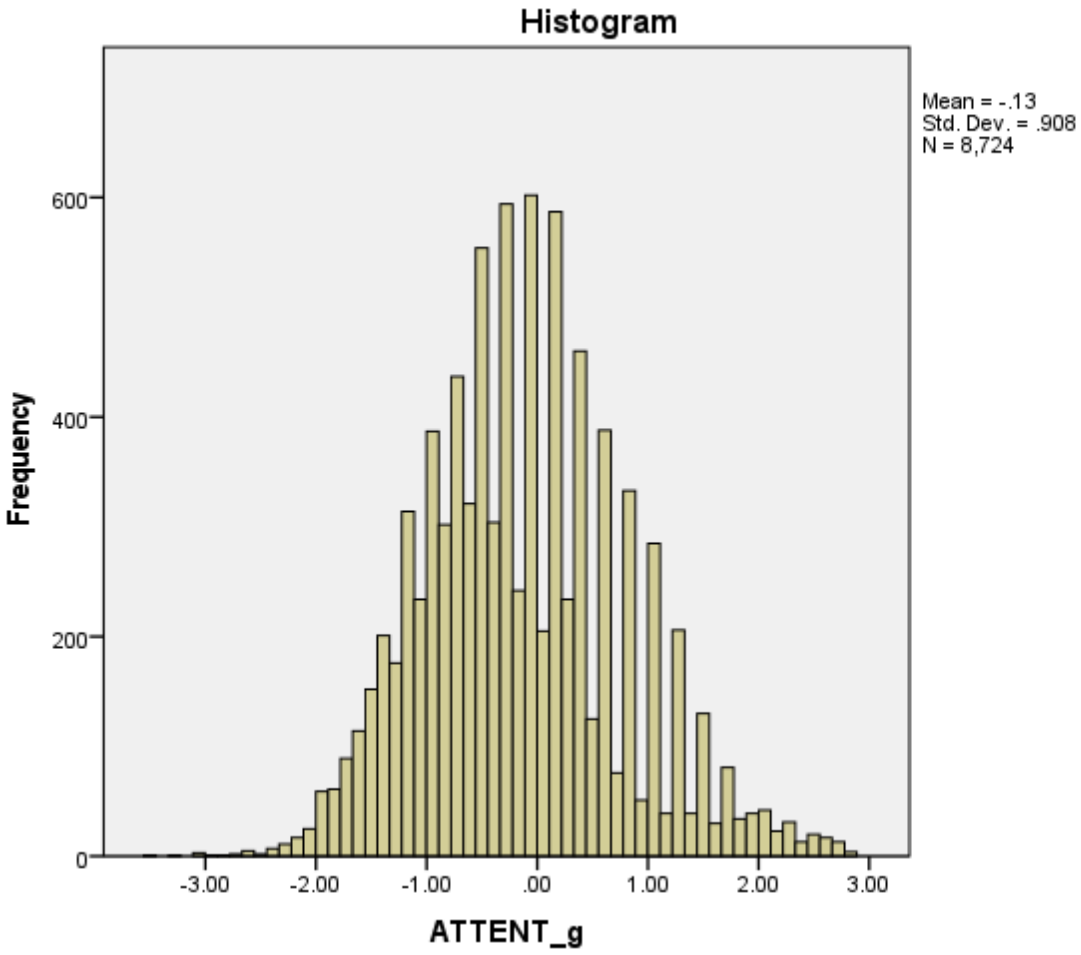

Normal Q-Q Plot of LANG_ge
Expected Normal
2.5
0.0
-2.5
-5.0
-7.5
-10.0
-7.5
-5.0
-2.5
0.0
2.5
Observed Value

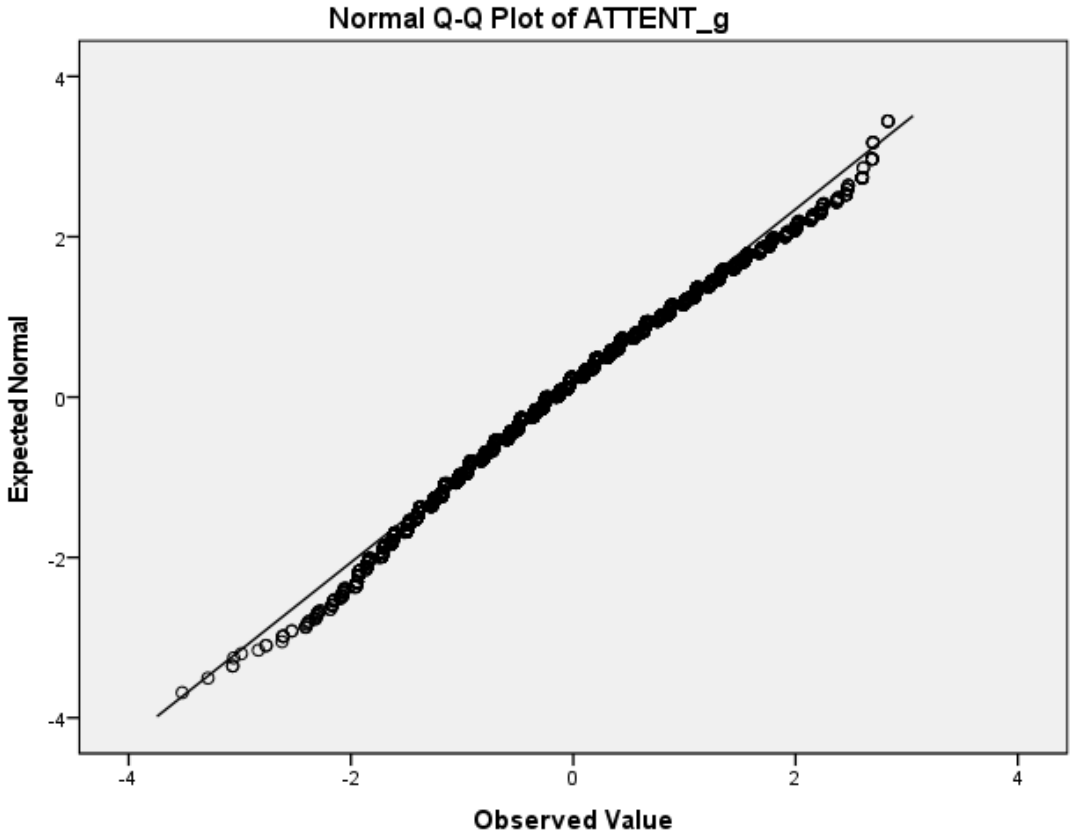

Histogram
Mean = -.46
Std. Dev. = 1.174
N = 8,724
Frequency
400
300
200
100
0
-3.00
-2.00
-1.00
.00
1.00
2.00
MEMORY_g

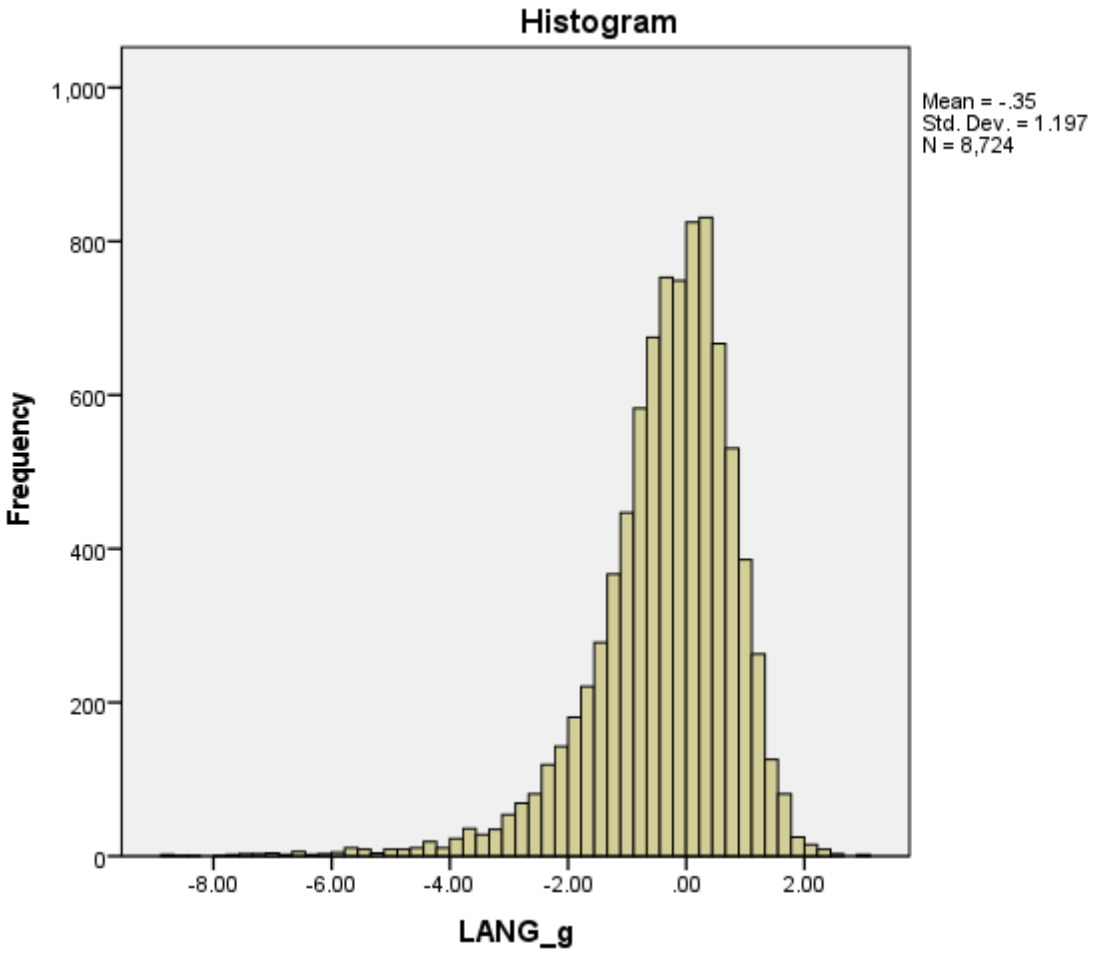

Normal Q-Q Plot of MEMORY_g
Expected Normal
Observed Value
4
2
0
-2
-4
-2
0
2

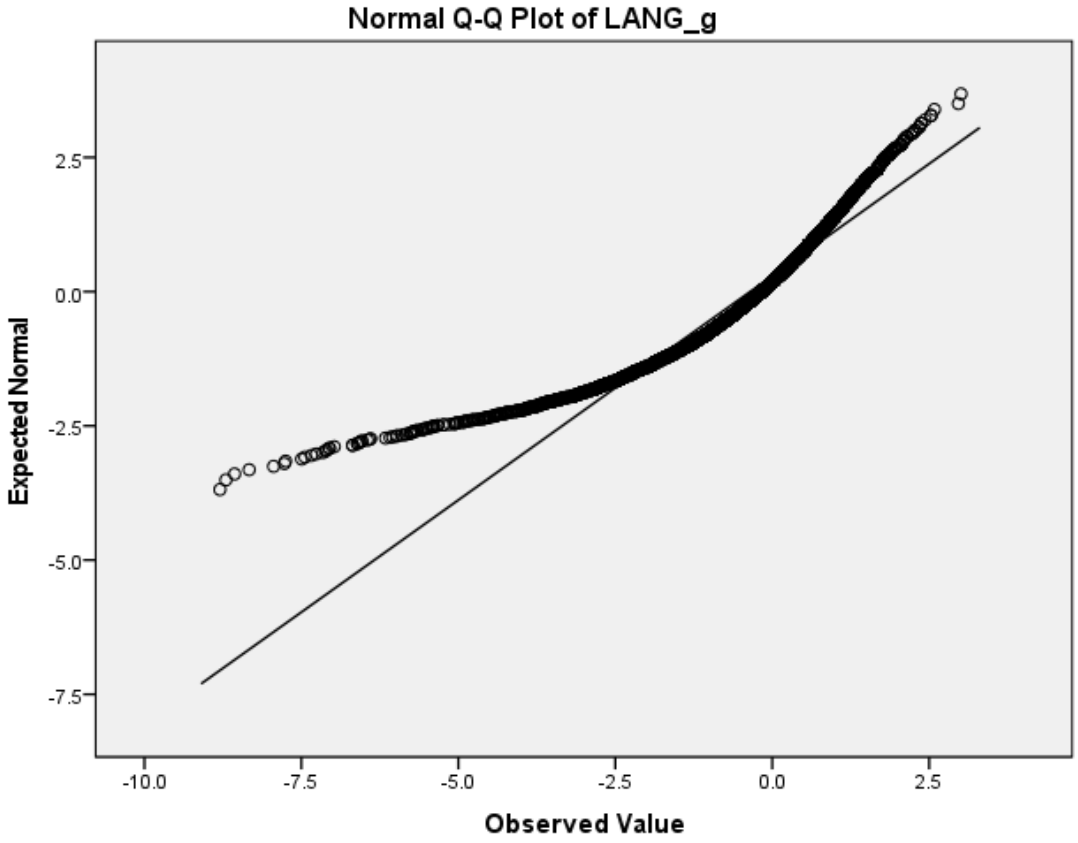

Histogram
Mean = -.35
Std. Dev. = 1.152
N = 8,724
Frequency
600
400
200
0
-4.00
-3.00
-2.00
-1.00
.00
1.00
2.00
EXEC_g


Normal Q-Q Plot of EXEC_g
Expected Normal
Observed Value
4
2
0
-2
-4
-6
-4
-2
0
2
4


Histogram
Mean = -.13
Std. Dev. = .908
N = 8,724
Frequency
600
400
200
0
-3.00
-2.00
-1.00
.00
1.00
2.00
3.00
ATTENT_g


Normal Q-Q Plot of ATTENT_g
Expected Normal
Observed Value
4
2
0
-2
-4
-4
-2
0
2
4


Histogram
Mean = -.35
Std. Dev. = 1.197
N = 8,724
Frequency
1,000
800
600
400
200
0
-8.00
-6.00
-4.00
-2.00
.00
2.00
LANG_g


Normal Q-Q Plot of LANG_g
Expected Normal
2.5
0.0
-2.5
-5.0
-7.5
-10.0
-7.5
-5.0
-2.5
0.0
2.5
Observed Value