8-2022

How Valuable are Clinical Neuropsychological Assessments? A Meta-analysis of Neuropsychological Tests with Comparison to Common Medical Tests and Treatments

Murphy Harrell
University of South Alabama, mnh1721@jagmail.southalabama.edu

How Valuable are Clinical Neuropsychological Assessments? A Meta-analysis of Neuropsychological Tests with Comparison to Common Medical Tests and Treatments

A Dissertation

Submitted to the Graduate Faculty
of the University of South Alabama in
partial fulfillment of the requirement
for the degree of

Doctorate of Philosophy

in

Clinical and Counseling Psychology

by
Murphy. N. Harrell
B.S., Appalachian State University, 2012
M.S., University of South Carolina Aiken, 2015
August 2022

ACKNOWLEDGMENTS

I send my sincerest gratitude to my mentor and Dissertation Chair Dr. Benjamin Hill, as well as my Dissertation Committee members Dr. Joseph M. Currier, Dr. Mark C. Yates, and Dr. Charlene Myers for their time, effort, and guidance they have afforded me during this project. I also acknowledge and thank Dr. Hill for not only introducing me to the original meta-analysis that sparked the idea for this updated project that serves to advocate for the field of neuropsychology, but also Dr. Stephen Aita for being available at short notice for consultation. This dissertation was also made possible by the tireless and high-quality efforts of Vasilios Ikonomou, Carolina Pereira, and Paola Ascencio-Ortiz who masterfully coded studies for this project. Finally, and certainly not last, the time and effort that was poured into this project could not have been possible without my incredibly supportive partner, mom, and friends to encourage me every step of the way and the best dogs in the world that kept me company through every late night and long weekend.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

AHM	Antihypertensive medication
AD	Alzheimer's Disease
APA	American Psychological Association
ADHD	Attention-deficit/hyperactivity disorder
CSF	Cerebrospinal Fluid
CTT	Classical Test Theory
ChEI	Cholinesterase inhibitor
CT	Computerized tomography scan
CI	Confidence Interval
CHD	Coronary Heart Disease
EF	Executive Functioning
GM	Gray Matter
HR	Hazard ratio
HC	Healthy Controls
IRT	Item-response theory
M	Arithmetic Mean
MRI	Magnetic resonance imaging
MCI	Mild Cognitive Impairment
MMSE	Mini-mental state examination
MoCA	Montreal Cognitive Assessment

MI	Myocardial infarction
NAB	Neuropsychological Assessment Battery
NPV	Negative predictive value
OR	Odds ratio
PPV	Positive predictive value
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PAWG	Psychological Assessment Work Group
RRR	Relative risk ratio
SES	Socioeconomic status
SD	Standard Deviation
TBI	Traumatic Brain Injury
WM	White Matter
WMH	White Matter Hypertensities

ABSTRACT

Harrell, Murphy, N., Ph.D., University of South Alabama, August 2022. How Valuable are Clinical Neuropsychological Assessments? A Meta-Analysis of Neuropsychological Tests with Comparison to Common Medical Tests and Treatments.
Chair of Committee: Benjamin D. Hill, Ph.D.

There has been a general decrease in neuropsychological assessments at a time when medical diagnostic technology and treatments have expanded, leading to a faulty assumption that medical tests and healthcare treatments provide more reliable or valid data than psychological assessments. A landmark report from the American Psychological Association's (APA) Psychological Assessment Work Group (PAWG) found that validity coefficients for many psychological tests were indistinguishable from those of medical tests (Meyer et al., 2001). An updated systematic review of the advancement in neuropsychological testing is essential to the continued advancement of the value of neuropsychological assessment in healthcare. This meta-analysis sought to (1) summarize effect sizes of neuroimaging to diagnose dementia, medications to treat chronic diseases, and neuropsychological tests to diagnose dementia and TBI, (2) determine the differences (if any) in effect sizes between medical domains, and (3) determine the differences (if any) in effect sizes between medical domains and neuropsychological tests. EBSCO networks were searched for original research examining the efficacy of neuroimaging for Alzheimer's Disease (AD),

neuropsychological tests for AD and traumatic brain injury (TBI), and medication to treat memory impairment and cardiovascular events between clinical and control samples. Studies were coded using a complex multi-comparison, outcome, and subgroup schema. Data were analyzed under random-effects modeling. Of 6,668 studies identified, 78 were retained for primary and ancillary meta-analyses (715 effect sizes extracted; 35,810 clinical and 42,964 control participants represented). Primary results indicated a significant difference between domains, such that neuroimaging (g = -1.603) and neuropsychological tests (g = -1.591) both yielded greater effect sizes than medication studies (g = -0.009]. Secondary results indicated the AD neuropsychological test effect size [g = -2.213) was significantly different than the TBI neuropsychological test efficacy [g = -0.649; $Q(1)$ = 42.821, p = 0.000]. Additionally, results indicated nonsignificant effect sizes for both memory impairment medications (g = -.052) and aspirin for cardiovascular events (g = .017). CONCLUSIONS: The diagnostic efficacy of neuroimaging and neuropsychological tests were both substantial and non-significantly different from one another. These findings provide clinicians and consumers with convincing evidence that neuropsychological tests are a reliable diagnostic tool for people with acquired and neurodegenerative brain disorders.

Keywords: neuropsychology, neuroimaging, medication, test efficacy

CHAPTER I

INTRODUCTION

The 21st century has brought drastic changes to healthcare, and clinical psychological assessment has not been immune to this medical evolution. Traditional comprehensive psychological assessments that could take many hours to days, and once defined psychological practice, fell victim to reimbursement constraints in the 1980s with the expansion of managed care plans, leading many psychologists to abandon long and high-cost evaluations (Piotrowski, 1999; Piotrowski, 2017). This decline in psychological assessments when medical diagnostic technology was beginning a sudden advance, particularly in imaging technology, led to a faulty assumption that medical tests provide more reliable or valid data than psychological assessments. However, a landmark report from the American Psychological Association's (APA) Psychological Assessment Work Group (PAWG) found that while psychological (including neuropsychological) and medical tests have varying degrees of validity, validity coefficients for many neuropsychological tests were indistinguishable from those of medical tests (Meyer et al., 2001). This psychological assessment white paper was highly influential but is now 20 years old. This dissertation aimed to replicate and extend the findings of the PAWG with an emphasis on findings published in the last five years.

In an examination of 144 meta-analyses and authoritative papers, Meyer et al. (2001) found evidence that disputed the common perception regarding popular psychological and medical assessments. Specifically, tests of cognitive abilities and personality traits all produced a range of validity coefficients that varied largely as a function of the criterion under consideration. Additionally, many projective and objective personality tests, as well as cognitive tests (e.g., Rorschach, MCMI, MMPI, and WAIS), produced medium to large effect sizes that were similar to popular medical tests such as Papanicolaou tests ("pap smears"), mammography, magnetic resonance imaging (MRI) and electrocardiograms. Finally, the PAWG concluded that many gold standard psychological tests work just as well as medical tests to detect the same outcome, such as neuropsychological testing detecting dementia on par with MRI (Meyer et al., 2001).

Despite the validation of psychological assessment Meyer et al. (2001) provided with the PAWG white paper, there were several noteworthy critiques of their methodology and interpretation of findings. These critiques ranged from the exclusion of significant medical tests that possibly negatively impacted inferences drawn about laboratory tests, creation of their own effect sizes (Garb, Klein, & Grove, 2002), incorrect use of base rates and selection ratios (Smith, 2002), and the lack of meta-analyses that supported the validity of assessment practices in either a forensic or a general clinical context (Hunsley, 2002). In response, Meyer et al. (2002) defended the application of a single effect size metric to all studies (they only used correlation values) because the same information could not be determined by odds ratio (OR) and relative risk reduction (RRR) magnitudes (e.g., Rosenthal, 1991). They also noted that research has found limited support for the use of RRRs and ORs within evidence-based medicine (e.g.,

Sackett, Deeks, & Altman, 1996) because they do not provide the most patient-relevant estimates of effect. Finally, Meyer et al. (2002) defended their use of validity coefficients as a "central foundation" for the merits of a test. However, they noted that clinicians should consider the validity of tests in terms of ongoing inference probabilities instead of relying on fixed base rates. This seminal article and subsequent rebuttal provided the framework for APA to establish assessment guidelines. However, an updated systematic review of the advancement in neuropsychological testing is essential to the continued advancement of the value of neuropsychological assessment in healthcare.

To provide an updated analysis of the efficacy of neuropsychological tests compared to medical tests, the current study aimed to systematically review recent neuropsychological testing as well as studies of medical tests and treatments. To improve on Meyer et al. (2001), analyses expanded beyond correlations as the singular proxy of test effect size by synthesizing recent studies on the topic of interest using meta-analysis. In addition to updating the seminal Meyer et al. (2001) paper on broad psychological test efficacy, specific neuropsychological tests were examined. In the following sections, a brief description of neuropsychological and medical tests are presented, followed by factors affecting common perceptions of medical and psychological science as well as statistics used in efficacy research. Next, effect sizes of contemporary medical and psychological tests across several diagnostic tests and categories are reviewed. Finally, the goals/hypotheses for an updated systematic review comparing the effect sizes of neuropsychological and medical tests are discussed.

CHAPTER II

LITERATURE REVIEW

2.1 History of Tests Within Psychology and Medicine

Broadly, science is defined as the systematic study of the physical and natural world through observation and experiment (Science & Technology, 2002). The branches of science are commonly divided into major groups, including formal sciences, natural sciences, and social sciences. The field of medicine is considered a natural science, whereas psychology falls under social sciences. Despite categorical differences, psychology and medicine are considered applied sciences, applying evidence-based knowledge to practical applications, such as inventions for advancing knowledge of disease and solving human behavior and experience problems (Roll-Hansen, 2009). Both psychological and medical tests are the application of scientific knowledge to screen, diagnose, and monitor health aspects.

2.1.1 Psychological Tests

A psychological test is a systematic method for observing a person's behavior or performance in the context of a numerical scale or category system (Anastasi & Urbina, 1997). Additionally, tests are used to measure differences between people or differences in the same person over time. Rudimentary psychological tests date back to 2200 B.C. as part of the Chinese imperial system in which officials were examined for fitness for duty

(Gregory, 2004). Psychological tests can be divided into several broad categories: intelligence and achievement tests, neuropsychological tests, personality tests, aptitude tests, and public employment safety tests (Gregory, 2004). For this study, neuropsychological tests will be reviewed in more detail.

Neuropsychological testing is a relatively straightforward process in which a test or scale is administered to obtain a specific score. A descriptive meaning can then be applied to the score using normative data (Gregory, 2004). Despite the overwhelming breadth of neuropsychological tests, core principles constitute sound tests, including standardization, objectivity, test norms, reliability, and validity (Schultz & Schultz, 2010). These core principles are key to examining neuropsychological tests' efficacy and will be discussed in more detail below. This brief history establishes the sound foundation that neuropsychological testing has in the field of psychology. However, a clinical question cannot be answered solely by a test score. Neuropsychological tests must be interpreted within a more comprehensive assessment to understand psychological health and disease more adequately

2.1.2 Psychological Assessment

A psychological assessment can include numerous tools, such as norm-referenced neuropsychological tests, self, and other report measures, interview information, school or medical records, and observational data that are integrated to answer clinical questions (American Psychological Association, 2013). The term *assessment* was used for the first time in 1943 by the U.S. Office of Strategic Services (Fernández-Ballesteros, 2002). This new assessment program for military officers represented an improvement in previous psychological testing. It described how the individual was able to act in various situations

(Du Bois, 1970). Subsequently, assessment has been used to link all psychology fields, becoming a generic term when decisions (e.g., classification, prediction, selection, counseling, intervention, or evaluation) about individuals must be made (Fernández-Ballesteros, 2002).

Neuropsychological assessment is a specialized subfield of psychological assessment that measures brain-behavior relationships. Some of the primary purposes of contemporary assessments are to (a) describe current functioning, including cognitive abilities, symptom severity, and capacity for independent living; (b) confirm or adjust the clinical impressions formed by less structured interactions; (c) to identify therapeutic needs, and create a therapeutic plan; (d) aid in the differential diagnosis of emotional, behavioral, and cognitive disorders; (e) monitor treatment over time to evaluate the success of interventions; (f) manage risk; and (g) provide skilled, assessment feedback as a therapeutic intervention in itself (Meyer et al., 2001).

Given the intricacy of neuropsychological assessments, the validity of individualized, contextually embedded inferences is incredibly complex. Each assessment method identifies useful data not available from other sources, and an assessment battery is likely to generate findings that, at least superficially, appear conflicting or contradictory (Meyer et al., 2001). Additionally, many different conditions can lead to an identical score on a particular test (Shea, 1985). Contextual factors (i.e., history, observed behavior, motivational context, etc.) play a large role in determining the final scores obtained on neuropsychological assessments. To integrate this information, psychologists must consider questions, symptoms, dynamics, and behaviors from multiple perspectives. Tests employed in other scientific disciplines (i.e., magnetic resonance imaging [MRI],

blood panel, and electroencephalogram [EEG]) are less affected by these factors (Meyer et al., 2001).

2.1.3 Medical Tests

A medical test is used to diagnose, screen for, or monitor the process and susceptibility of diseases and determine a course of treatment (Balogh, Miller, & Ball, 2015). The oldest known medical tests date before 400 BC (Berger, 1999). Medical tests are classified by specialty, including consulting room tests, cardiovascular, dermatology, otolaryngology (ENT), gastrointestinal, hematology, laboratory, neurological, and radiology. A diagnostic test is used to determine the presence of disease, usually due to patient-reported symptoms (Balogh et al., 2015). Screening tests are used to detect or predict the presence of disease in at-risk individuals due to population norms, family history, or workforce exposure (Balogh et al., 2015). Medical tests are used to monitor the progress of or response to medical treatment, including continuous monitoring of vital signs or repeated medical tests (Balogh et al., 2015). The most common medical tests include analysis of bodily fluids, including blood, urine, and cerebrospinal fluid; neuroimaging, such as X-rays, computerized tomography (CT), MRI, and positron emission tomography (PET); endoscopy; measurement of heart, brain, and lung function through an electrocardiogram (ECG), EEG, and pulmonary function tests; biopsies; and genetic testing (Merk Manuals, 2010).

Results of medical tests are often interpreted based on a reference range or a set of values, including upper and lower limits of a lab test based on a reference or healthy group (Jones & Barker, 2008). This range is usually defined as the set of values that 95 percent of the normal population falls within, determined by collecting data from vast

numbers of laboratory tests. The values often depend on factors such as age, sex, and specimen type (blood, urine, spinal fluid) and can also be influenced by situations such as fasting and exercise (Jones & Barker, 2008). Most frequently, patients undergo diagnostic medical tests due to chronic diseases, including skin disorders, osteoarthritis, joint disorders, back problems, metabolic disorders, and upper respiratory disease (St Sauver et al., 2013).

2.1.4 Medical Assessments

A comprehensive medical evaluation, or comprehensive physical exam, assesses all aspects of a person's health for medical screening purposes (Raffle & Gray, 2019). A comprehensive medical evaluation includes an updated medical history, including previous illnesses, surgeries, hospitalizations, medications, allergies, and family medical history; vital signs checks (i.e., blood pressure, heart rate, and respiratory rate); visual exam of the patient's physical appearance, speech, and walking ability; physical exam of eyes, ears, nose, throat, neurological functioning, abdominal palpation, and dermatological exam of skin and nails; laboratory tests; and neuroimaging (Raffle & Gray, 2019). The patient's medical history and sex, and age determine which tests are included. In recent decades, the efficacy of comprehensive medical examinations has come under debate. For example, a recent meta-analysis found that routine medical examinations did not measurably reduce the risk of illness or death and, conversely, could lead to excessive diagnosis and over-treatment (Krogsbøll et al., 2012). Conversely, Verghese et al. (2015) found that inadequate physical examinations account for oversights in care, such as increased missed or inaccurate diagnoses, unnecessary or delayed treatment, unnecessary cost and exposure to radiation, and complications caused

by treatment. These results indicated that while comprehensive medical screenings might not be beneficial in preventing disease, comprehensive diagnostic exams are crucial to accurate diagnosis and treatment.

Research has indicated that while both medical and neuropsychological tests are tools that provide scores, neuropsychological functioning requires a more comprehensive assessment of numerous contextual factors to adequately interpret test scores (Meyer et al., 2001). Conversely, these factors rarely have such significant effects on medical tests. As an example, a low full-scale I.Q. score on cognitive testing could indicate severe depression, a traumatic brain injury, or a developmental disability (Meyer et al., 2001). Additionally, as discussed above, comprehensive medical examinations are often used for screening purposes, whereas comprehensive psychological assessments are required for an accurate diagnosis.

The complexities of psychological assessments pose challenges for research on and communication of psychological test efficacy. Conversely, the interpretation of medical test results often indicates the presence or absence of disease with fewer alternative contributing factors. Because of this, trust in medical science is often stronger than in psychological science. Public trust in science is a critical component of how test efficacy research is regarded. The following section reviews the factors that affect the trust in medical and psychological science and tests.

2.2 Factors Affecting Public Attitudes Toward Science

It is recognized that there is a gap between scientific findings and people's acceptance of them. In an examination of the most salient factors in driving acceptance of

science (Funk, Rainie, & Page, 2015), higher education is correlated with greater knowledge about science and scientific processes (Kennedy & Hefferon, 2019). Furthermore, scientific knowledge has an independent effect in predicting varying attitudes on several science-related topics, even after controlling for demographic and political differences (Funk et al., 2015). Acceptance of scientific data is not based on educational attainment alone but also on people's political and religious beliefs and demographic variables, such as age, race, and gender (Funk et al., 2015). Demographic variables, statistical literacy, and beliefs about source credibility play important roles in understanding medical and psychological science acceptance.

2.2.1 Positive Attitudes Toward Medical Science

Laypeople believe numerous medical screening tests and preventative treatments to be strongly efficacious in preventing and treating disease. However, Meyer et al. (2001) examined 89 meta-analyses between 1988 and 1989 on medical tests and treatments, finding only low to medium effect sizes for many medical tests and treatments. These included: daily aspirin to reduce the risk of dying from a heart attack (r = .02); the impact of chemotherapy on breast cancer survival (r = .03); routine ultrasound on successful pregnancy outcomes (r = .01); the value of antihistamines for reducing sneezes and a runny nose (r = .11); mammogram screening and detection of breast cancer (r = .27); the effect of sleeping pills for short-term treatment of insomnia (r = .30); and the impact of Viagra on improved sexual functioning (r = .38) (Meyer et al., 2001). Nevertheless, despite these generally small effect sizes, general trust in medical tests and treatments remains high (Funk et al., 2019). Exploring public understanding of medical research and managed healthcare has indicated that factors

maintaining this excessive optimism in the efficacy of medical tests and treatments include blind trust in physicians, demographic variables, and inaccurate understanding of scientific methods and results.

Though there has been some decline in trust in healthcare across the United States (Huang et al., 2018), people generally continue to have a high degree of trust in medicine. A recent Pew Research Poll (Funk et al., 2019) found that 74% of Americans have a mostly positive view of medical providers. Additionally, while Americans trust medical researchers at high rates (i.e., 68% have positive views of medical researchers), only 32% of individuals believe medical researchers provide fair and accurate information (Funk et al., 2019). Interestingly, trust in healthcare appears to be mediated by demographic variables. Older Caucasian participants are more likely than their younger, minority counterparts of similar status to question medical advice (Meyer, Ward, & Jiwa, 2012; Smith, 2011). Overall, those with higher socioeconomic status (SES) are more likely to question medical advice than lower SES participants (Meyer et al., 2012). Finally, participants who perceived themselves at risk of a poor or uncertain outcome were unlikely to doubt medical advice, whereas primary care patients were more likely to doubt medical advice (Meyer et al., 2012).

Not only is trust in healthcare a strong maintaining factor in the assumed efficacy of medical tests and treatments, but research has also indicated that laypeople do not accurately understand statistical methods and results used in medical research (Gigerenzer et al., 2007). In a recent meta-analysis, participants rarely had accurate expectations of benefits and harms, and for many interventions, regardless of whether it was a treatment, test, or screen, they tended to overestimate its benefits and underestimate

potential harm (Hoffman & Marr, 2015). In addition to patient illiteracy, statistical illiteracy among physicians also contributes to physician overconfidence in medical research and tests. Studies have demonstrated that a substantial number of physicians do not correctly understand medical statistics (Gigerenzer et al., 2007; Wegwarth, Gaissmaier, & Gigerenzer, 2011), do not understand the scientific evidence for benefits and harms (Neuner-Jehle et al., 2011), and do not report them accurately to patients (Wegwarth & Gigerenzer, 2013). Counterintuitively, Wegwarth, Wagner, & Gigerenzer (2017) found that when physicians provided patients with inaccurate scientific information regarding the risks and benefits of cancer screening tests, patients were actually more likely to follow physician recommendations. This review illustrates the strong contribution of positional power and expectation to receptiveness to accept treatment advice, with valid information playing a lesser role in motivating treatment compliance.

In addition to increased scientific literacy (knowledge), individuals may need to shift their epistemic cognition (beliefs about the nature of knowledge) and epistemic trust (beliefs about source credibility) to accept scientific perspectives (Sinatra & Hofer, 2016). These cognitive processes influence learning, making medical decisions, and choosing which medical findings to trust (Sinatra & Hofer, 2016). Individuals move from an absolutist cognitive stance (knowledge viewed as objective, certain, and true) toward a multiplistic stance (knowledge viewed as subjective, based on interpretation and opinion) (Kuhn, Cheney, & Weinstock, 2000). Ideally, an evaluativist view is achieved where objectivity and subjectivity are integrated, knowledge is perceived as contingent and contextual, and research claims are critically evaluated (Kuhn, Cheney, & Weinstock,

2000). However, evaluativism may be relatively uncommon and correlated with increased education (Kuhn et al., 2000). People express absolutist processes more commonly, favoring certainty and misinterpreting the role of tentativeness in science (Sinatra & Hofer, 2016). This belief about the absoluteness of science creates considerable challenges in effectively communicating science to the public, especially for the media.

Although a great deal of opinion formation occurs in the direct exchange between health care professionals and patients, people also form their opinions and attitudes of scientific findings through media content (Kimmerle et al., 2015; Sapp et al., 2013). This becomes problematic when media sources do not accurately convey the tentativeness of most medical research findings (Jenson, 2008), as the general public favors research certainty (Bromme & Goldman, 2014; Chang, 2015). How the news communicates scientific findings—whether framed positively or negatively and presenting reliability as weak or strong—can influence perceptions of medical research findings (Kimmerle et al., 2015). However, this may be challenging, as research suggests that not all medical research includes limitations, especially when the studies were industry-supported (Ter Riet et al., 2013).

Overall, this information is critical to understanding the key factors that affect medical research credibility among the general public. Specifically, people tend to accept medical research findings because they have high trust in physicians and medical researchers, have poor health literacy to make evidence-based decisions that might differ from the physician's recommendation, and absorb media that promotes medical research as true without critical analysis of source credibility. While the public and physicians are

generally overconfident in the efficacy and validity of medical tests and treatments, there is a significant distrust by the public of psychological research. The following section considers how the factors contributing to overconfidence in medicine may also propel distrust in psychological research.

2.2.2 Negative Attitudes Toward Psychological Science

Psychological science elicits less respect from the general public than many "hard" sciences, like biology, chemistry, economics, medicine, and physics (Janda et al., 1998). The reasons for this are manifold, ranging from poor mental health literacy, external and internal distrust of psychological research and results, and the perpetuation of mental health stigma by the media and policymakers (Lilienfeld, 2012).

Mental health literacy is defined as knowledge about mental health, including recognizing, managing, and preventing mental health problems (Jorm, 2012). Research on layperson mental health literacy consistently reveals a widespread lack of knowledge of mental health disorders (Jorm, 2012), the role of psychologists versus psychiatrists (Patel, Caddy, & Tracy, 2018), and the use of psychological treatment for areas outside of mental health disorders (i.e., as reducing divorce rates, physical health problems, and improving organizational productivity) (Penn et al., 2008). In addition, mental health literacy varies across demographics such that older adults (Sørensen et al., 2015), men (Gibbons, Thorsteinsson, & Loi, 2015), and those with lower education (Fisher & Goldney, 2003) tend to have less knowledge about psychological disorders and are less likely to engage mental health treatment (Reavley, McCann, & Jorm, 2012).

Inadequate mental health literacy vastly affects the trust in psychological research and treatment. Laypeople generally distrust psychological science (Keil, Lockhart, &

Schlegel, 2010; Lilienfeld, 2012). For instance, only 30% of respondents believed psychology attempted to understand behavior through scientific research, and 41% believed that psychology is less rigorous than medical research (Penn et al., 2008). The public's negative attitudes about psychological science and mental health treatment have discouraged people from seeking mental health care and other recovery-oriented services (Corrigan et al., 2005).

Not only is distrust in psychology a strong maintaining factor in the assumed inefficacy of psychological tests and treatments, but research has also indicated that laypeople do not accurately understand statistical methods and results used in psychological research. For example, McPhetres & Pennycook (2020) showed participants graphs depicting a range of effect sizes in different formats. Even the largest effect examined (corresponding to a Cohen's d = .90) was considered small to moderate in size by laypeople, underscoring how the public does not have the requisite background to truly understand the significant effects of tests and interventions. Furthermore, psychological research often reports effect sizes ranging from .1 to .2 (Wilson & Golonka, 2012), further discrediting psychological research efficacy among laypeople. Not only are valid psychology statistics not well understood, but people tend to discount behavioral science when specific conclusions are undesirable (Munroe & Munroe, 2015). However, people are more likely to believe a neuroscience explanation of a psychological phenomenon than a purely psychological explanation (Greene & Cahill, 2012), even when the neuroscience explanation was irrelevant (Fernandez-Duque et al., 2015; Hopkins, Weisberg, & Taykor, 2016).

In fairness to the public, research on psychological methodology has found a significant problem with replicating psychological findings within social psychology (Stanley et al., 2018). Meta-analyses have indicated a skew in publication practices, such that studies with null results are often rejected for publication (Bakker, van Dijk, & Wicherts, 2012; Schimmack, 2012). This practice has been related to an increase in authors using questionable researcher practices known as "*p* hacking" to convert null findings to statistically significant findings (John et al., 2012; Mitchell, 2014), resulting in low generalizability of research findings (Ferguson, 2015). In addition, some of these practices are related to social psychologists having an advocacy narrative of findings they want to present, which further erodes public confidence in psychological research and hurts the view of psychology as an objective science. This further promotes distrust of the scientific rigor of psychological research (Schimmack, 2020).

These replication problems have received significant press coverage in recent years (e.g., Berezow, 2012; Ferguson, 2015; Gutting, 2012), perpetuating a public stigma against the validity of psychology as a science. Psychological researchers have begun tackling these issues with discussions about replication failures and self-criticisms of questionable research practices (Ferguson, 2015). The discussions have increasingly moved into public forums, such as blogs, Twitter, and Facebook discussion groups, making criticisms more accessible to journalists and the general public (Brumfiel, 2009). Some of this public discussion has arguably influenced policymakers' decisions regarding psychological science (Ferguson, 2015; Nyhan, 2014) and invited unfavorable comparisons to the "hard" sciences (Chambers, 2014).

Trust in medical research and treatment was high among laypeople, but trust in psychological treatment and research was poor. This differs by demographics, health literacy, research design interpretation, and treatment effects. Finally, the media plays an influential role in perpetuating trust in medical science and distrust in psychological science. Given the research on the poor statistical understanding of reporting methods for testing and treatment efficacy and the use of these analyses in the current systematic review, the following section briefly reviews the statistical methods for reporting psychological and medical research findings.

2.3 Evaluating Efficacy of Tests

In reviewing the medical and psychological literature, terms that deal with evaluating medical services, such as efficacy, effectiveness, and efficiency, are common. *Efficacy* evaluates how well a test, medication, program, or procedure works in an experimental or "ideal" setting, such as controlled clinical trials. *Effectiveness* assesses how well a test, medication, program, or procedure works under usual circumstances when applied to the general population. Finally, *efficiency* evaluates the costs and benefits of medical intervention (Burches & Burches, 2020). For this review, the scientific measurement of medical and neuropsychological test efficacy is explored.

2.3.1 Neuropsychological Efficacy Statistics

Psychometrics is the scientific study, including the development, interpretation, and evaluation of psychological tests and measures used to assess behavior variability and its relation to psychological phenomena. In evaluating the quality of neuropsychological measures, researchers are primarily concerned with test reliability

(i.e., consistency), validity (i.e., the accuracy of interpretations and use), and fairness (i.e., the equivalence of usage across groups) (Committee on Psychological Testing, 2015).

Reliability is the degree to which test scores are stable and consistent (Committee on Psychological Testing, 2015). Test scores are thought to be composed of true and error elements. Therefore, a standard error of measurement within a confidence interval (e.g., 95%) describes that a person's true score falls within a given range of test scores (Geisinger et al., 2013). There are generally four classes of reliability, *test-retest*: consistency of test scores over time; *inter-rater*: consistency of test scores across independent judges; *parallel or alternate forms*: consistency of scores across different forms of the test; and *internal consistency*: consistency of items intended to measure the same thing within the test (Committee on Psychological Testing, 2015). Within internal consistency, split-half reliability (two halves of a test are compared) is measured to determine an index of reliability. A common measure of internal consistency in neuropsychological research is Cronbach's α, which is the mean of all possible split-half correlations for an item set (Committee on Psychological Testing, 2015). A more reliable measure contributes relatively less error to statistical analyses (Suhr, 2003).

Many factors can affect the reliability of a test score. These include the time between two testing administrations, changes in subjects over time, or test-based factors (Committee on Psychological Testing, 2015). A test can yield reliable scores in one context and not in another, and inferences made from different reliability estimates are not interchangeable (Geisinger et al., 2013). A test's reliability is essentially an index of

the proportion of error variance in a test, as more error variance results in decreased reliability.

While the scores resulting from a test might have good reliability, this finding does not necessarily mean that the test scores are valid. *Validity* is defined as the degree to which a test measures what it is supposed to measure (Kelly, 1927). Validity does not refer to the measure or the test items being valid. Instead, validity refers to interpreting test scores grounded in psychological theory and empirical evidence and demonstrates a relationship between the test and what it claims to measure (Furr & Bacharach, 2013). The field of neuropsychology, and broadly psychology, is concerned with three primary types of validity: c*onstruct validity*: the degree to which an individual's test scores correlate with the theoretical concept the test is designed to measure; c*ontent validity*: the degree to which the test content represents the targeted subject matter; and c*riterion-related validity*: the degree to which the test's score correlates with other measurable, reliable, and relevant variables thought to measure the same construct (Sattler, 2014).

Within neuropsychological testing, other kinds of validity have also been examined: *diagnostic validity*: the degree to which psychological tests support an appropriate diagnosis; e*cological validity*: the degree to which test scores represent normal levels of functioning; and c*ultural validity*: the degree to which test content and procedures accurately reflect the socio-cultural context of the subjects being tested (Committee on Psychological Testing, 2015).

Psychological test efficacy research relies on test theories such as Classical Test Theory (CTT) and Item-Response Theory (IRT) to answer null hypothesis significance testing (NHST) through the interpretation of a *p*-value (Fisher, 1925). The *p*-value

determines if there is a significant difference between groups. The smaller the *p*-value, the stronger the evidence that the null hypothesis is rejected (McLeod, 2019). One of the many problems with NHST is that it encourages dichotomous thinking: either an effect is statistically significant or not (Kline, 2004). Using a *p*-value to merely determine if there is a significant difference between groups or test items does not provide useful information for clinical application. Effect sizes and confidence intervals (CI) provide more information about those clinical differences (Walker, 2007).

An effect size can be defined as a quantitative description of the magnitude of an effect (Kelley & Preacher, 2012). Effect sizes are particularly prominent in medical research, where the size of the treatment effect is important. In psychological research, effect sizes only more recently became required reporting standards (APA, 2010). There are many kinds of effect sizes, though most can be combined into groups: those measuring differences between groups and those that describe the strength of the association. Most prominently, Pearson's correlation coefficient *r*, Cohen's *d*, and odds ratios are used, though odds ratios are less common in psychological efficacy research. The correlation coefficient *r* indicates the amount of variance explained by the model and ranges from no relationship ($r = 0$) to a perfect relationship (1 or -1) (Walker, 2007). The classic rule of thumb for interpreting effect sizes suggests that an *r* of |.1| represents a 'small' effect size, |.3| represents a 'medium' effect size, and |.5| represents a 'large' effect size (Cohen, 1992). Another common measure of effect size is *d*, sometimes known as Cohen's *d*, which compares two means divided by the average of their standard deviations. Cohen suggested that d=0.2 be considered a 'small' effect size, 0.5 represents a 'medium' effect size, and 0.8 a 'large' effect size (Cohen, 1992).

While effect sizes contain more information than P values, it is also an estimate calculated from statistical inference (Lee, 2016). That is, an effect size estimated from a large sample is likely to be more accurate than one estimated from a small sample (Lee, 2016). Confidence intervals can provide a range of plausible values for both the mean and effect size estimate. For example, a 95% CI of the mean calculated from a sample implies that if the samples originate from the same population with the same extraction method, 95% of their CI ranges would include the population mean (Dekking et al., 2005). For all hypothetically sampled data from the same population and using the same sampling method, the population's effect would fall within 95% of the calculated 95% CIs for the effect size of these data (Lee, 2016). If this 95% CI contains "0," it indicates "statistical non-significance." Providing the effect size (point estimate) and CI (the precision of effects) is essential to understanding the magnitude of intended treatment effects (Lee, 2016).

In terms of diagnostic ability, tests are also examined for sensitivity and specificity and, subsequently, their predictive value. The sensitivity of a test is defined as the test's probability of correctly identifying all those who do indeed have that condition from among people who are known to have a condition (i.e., true positives) and not categorizing other people as not having the condition when in fact they *do* have it (i.e., avoiding false negatives) (Trevethan, 2017). A positive predictive value is the probability that people with a positive screening test result have the condition of interest. Specificity is the probability that a sign will be negative, given that the disorder is not present.

Conversely, specificity is defined as a screening test's probability of correctly identifying all those who do not have that condition *among people who are known not to*

have a condition (i.e., true negatives) and, at the same time, *not* categorizing some people as having the condition when in fact they do not have it (i.e., avoiding false positives) (Trevethan, 2017). A negative predictive value is the probability that people with a negative screening test result do not have the condition of interest (Trevethan, 2017). These metrics are often reported as percentages, although sometimes as decimal fractions, preferably accompanying 95% confidence intervals.

While sensitivity and specificity are not impacted by the base rate of the disorder in question, the base rate of the diagnosis in the population of interest should also be considered in clinical decision-making (Meehl & Rosen, 1955). Sensitivity and specificity are concerned with the predictive ability compared to a "gold standard" diagnostic test or best available reference standard. Within neuropsychological and medical assessment, there are few, if any, genuine "gold standard" tests (Brodsky & Lichtenstein, 2020). However, the use of randomized control trials as the reference standard of a test or medication efficacy (Meldrum, 2000) is much more prevalent in medical research than in neuropsychological research (Clay, 2010).

2.3.2 Medical Efficacy Statistics

To understand disease etiology and provide effective treatment for persons with a given disease, it is essential to accurately distinguish between patients who do not have the disease of interest through screening and diagnostic tests (Krousel-Wood, Chambers, & Muntner, 2006). However, it is important to assess the quality of these tests for accurate interpretation and clinical decision-making by using what have been termed medical and pharmaceutical statistics due to their heavy use in those areas (Dodge & Commenges, 2006).

Test validity (sensitivity and specificity as reviewed above) and predictive values (PPV and NPV as reviewed above) are critical components of quality medical tests. The tests that a provider chooses to diagnose or monitor a medical condition are based on their ability to distinguish whether the condition is present or not (i.e., sensitivity and specificity) (Krousel-Wood et al., 2006). A test's specificity is crucial when it is necessary to confirm a diagnosis that requires potentially dangerous therapy. Likewise, PPV and NPV are key in considering the probability of the disease given the patient's population (Krousel-Wood et al., 2006). In populations with high disease prevalence, there is greater confidence that a positive test result is a true positive and increased suspicion that a negative test result is a false negative. The reverse is true in populations where the prevalence of the disease is rare (Krousel-Wood et al., 2006). Therefore, adequate test validity and predictive power are crucial for effective healthcare decision-making.

Medical research also reports effect sizes, especially within meta-analysis reports. Much of the medical literature refers to effect size estimates such as relative risk (RR) and odds ratios (OR). Both provide a measure of the *strength* of the relationship between a factor and a disease or outcome. OR is the ratio of the odds of an event in one group compared to the odds of the event in the other group (Ranganathan, Pramesh, & Buse, 2015). The RR is the ratio of risk in an exposed group versus the risk of the event in the nonexposed group (Ranganathan et al., 2015). If the outcome is the same in both groups (RR = 1), it is indicative of no difference in risk (or odds) between the intervention and placebo or control group (Ranganathan et al., 2015). An RR (or OR) greater than 1.0 indicates an increase in risk (or odds) among the exposed compared to the unexposed

group, whereas an RR (or OR) less than one indicates a decrease in risk (or odds) in the exposed group (Ranganathan et al., 2015). Among 171 medical publications, reported ORs ranged from 1.17 to 290.00 and had a median value of 2.16 (Chen, Cohen, & Chen, 2010).

The relationship between OR and RR is complex and often used interchangeably, though incorrectly (Ranganathan et al., 2015). When there is an association between an exposure and an outcome, an OR exaggerates the estimate of the relationship (is farther from 1.0 than RR). Thus, when RR <1, OR is lower than RR. However, when RR is more than 1.0 OR is higher than RR when the outcome is rare (typically <10%), the value of OR is not too different from that of RR. The two can be used interchangeably, irrespective of whether the risk is lower or higher in the exposed group compared to the unexposed (Sedgwick, 2014). As event rates increase, the two ratios diverge and can no longer be used interchangeably (Ranganathan et al., 2015). Stated differently, when there is a low base rate of an event, OR and RR will be very similar, but as the base rate increases, OR will exaggerate the likelihood of an event compared to RR. Both OR and RR have confidence intervals (CI). If a 95% CI excludes the value one, the ratio is significant at $p<0.05$ (Krousel-Wood et al., 2006). Also reported, especially in randomized control trials, is the risk difference (RD), sometimes called absolute risk reduction. It is the difference in risk (probability) of an event between two groups. The RD provides a measure of the *public health impact* of the risk factor and focuses on the number of cases that could potentially be prevented by eliminating the risk factor.

In review, both psychological and medical research use significance testing to measure test efficacy. The use of effect sizes and confidence intervals are prevalent in

both types of research, though the use of odds ratios is more prevalent in medical research than *r* or *d* in psychological research. While there is some focus on sensitivity and specificity and predictive value in psychological tests, this is limited compared to the prevalence of these statistics in measuring medical test efficacy.

2.3.3 Approaches to Measuring Outcomes

In ideal cases, a test will be evaluated in a comprehensive clinical trial where every relevant outcome is assessed in a representative group of patients (Matchar, 2012). Unfortunately, such trials are not routinely performed in practice because they are often deemed unattainable (Trikalinos, Siebert, & Lau, 2009). More often, however, systematic reviews and meta-analyses are conducted. A systematic review is a protocol-driven comprehensive review and synthesis of data focusing on a topic or related key questions (Russel et al., 2009). A meta-analysis is a specialized statistical analysis used to combine the data derived from systematic reviews.

Meta-analyses generate estimates of a (weighted) average effect size, the dispersion of effect sizes, and the homogeneity (or heterogeneity) of the entire set of observed effect sizes and subgroups and support the exploration of the potential moderators (Suurmond, van Rhee, & Hak, 2017). Combined effect sizes from meta-analyses should only be used as an outcome of a group or subgroup homogeneity if the observed effect sizes are only for the domain defined by this specific group of populations (Suurmond et al., 2017). Because relevant heterogeneity is normally found in the social sciences, the main result of most meta-analyses is an insight into the dispersion of true effects (Suurmond et al., 2017). In those cases, meta-analysis functions as a tool for generating hypotheses about potential moderators of the effect.

2.3.4 Critique of Statistical Differences Across Disciplines

Though there are similarities in evaluating test efficacy between medical and neuropsychological research, including hypothesis testing, with effect sizes and confidence intervals, many differences contribute to the misinterpretation of research findings. Specifically, differences in the use and interpretation of effect sizes across medical and neuropsychological test research pose challenges for direct efficacy comparisons.

Research indicates a continued lag in teaching and the use of effect sizes in psychological courses (Funder & Ozer, 2019). Effect sizes such as *r* work ideally with continuous data found in psychological science. Much of the medical and epidemiological literature uses binomial data (Ferguson, 2009). Additionally, odds ratios are more often used with epidemiological data. This poses a problem when comparing effect sizes across the two disciplines. Ferguson (2009) found that effect sizes in medical research, such as relative risk, do not translate well into effect sizes such as r, such that translation often underestimates medical research effect sizes.

Additionally, conventional definitions of effect sizes are inadequate for interpreting most psychological phenomena (Funder & Ozer, 2019). For example, a meta-analysis of social and personality psychology literature found that the average effect-size *r* was .19 and that an *r* between .11 and .29 fell at the 25th and 75th percentiles, respectively (Gignac & Szodorai, 2016). The authors suggested recasting Cohen's guidelines in this light, such that correlations of .10, .20, and .30 could be considered *small, typical,* and *relatively large*, respectively (Gignac & Szodorai, 2016). Additionally, in the medical literature, odds ratios have often been confused with relative

risk. For example, a meta-analysis of articles in two medical journals found that 26% of the articles that used an odds ratio interpreted it as a risk ratio (Holcomb et al., 2001).

A major limitation in neuropsychological test research is the absence of "gold standard" tests (Tommasi, Ferrara, &, Saggino, 2018). In the absence of a perfect test, multiple imperfect tests are used to gain an estimate, increasing the risk of incorrect interpretation (Lesaffre, Speybroeck, & Berkvens, 2007). Before performing the test, it is necessary to add prior information, such as knowledge about the parameter of interest (e.g., the probability that a person is affected by a mental disorder). This previous knowledge constitutes the prior probability of the parameter, allowing for Bayesian analysis of the combination of this prior probability with the collected data to yield an estimate (Lesaffre et al., 2007). However, research has found that most neuropsychologists and other clinical diagnosticians either neglected or misused base rate information when presented in a similar format to that they would be expected to encounter the information (Labarge, McCaffrey, & Brown, 2003).

Additionally, despite neuroimaging being seen as an exact and highly sensitive, and specific technology, research has found high false-positive rates in neuroimaging literature due to low power caused by small sample sizes (Button et al., 2013; Yarkoni, 2009). Specifically, meta-analyses of fMRI research found that the most popular statistical analysis methods generate up to 70% false-positive results in null data (Eklund, Nichols, & Knutsson, 2016). Similar conclusions have been reached in analyses of diffusion-weighted MRI reconstructions of white matter pathways that have been reported to be dominated by false-positive outcomes (Maier-Hein et al., 2016).

2.4 Review of Effect Sizes in Medical Research

Given the high level of trust the public places in physicians discussed previously, people are apt to follow physician recommendations concerning diagnostic tests and treatments. As discussed above, Meyer et al. (2001) demonstrated small effect sizes for many medical tests that were considered highly effective 20 years ago. Research documenting low efficacy across medications and diagnostic tests is discussed.

2.4.1 Known Medical Risk Factors

In examining the effects of alcohol on cancer, Jayasekara et al. (2016) found a weak non-linear dose-response relationship for breast cancer and a positive linear dose-response relationship for upper aero-digestive tract (UADT) cancer and colorectal cancer. The pooled RRs were 1.28 (95% confidence interval, CI: 1.07-1.52) for breast, 2.83 (95% CI: 1.73-4.62) for UADT, 4.84 (95% CI: 2.51-9.32) for oral cavity and pharynx, 2.25 (95% CI: 1.49-3.42) for larynx, 6.71 (95% CI: 4.21-10.70) for esophageal and 1.49 (95% CI: 1.27-1.74) for colorectal cancer. It would appear that alcohol consumption has a huge effect on oral and pharyngeal cancers and esophageal cancer, with much smaller effects on other cancers. However, this information cannot be truly understood without considering the base rate of these cancers. The base rate of breast cancer is 13% in women, the base rate of UADT cancer is 8%, the base rate of colorectal cancer is 4%, the base rate for oral and pharyngeal cancer is 0.6%, the base rate for laryngeal cancers is 0.3%, and the base rate for esophageal cancer is 0.8% for men and 0.2% in women. Understanding this base rate information now lets the consumer understand that alcohol increases the rate of breast cancer from 13% to 16.6% in women, increases the rate of UADT cancer from 8% to 22.6%, increases the rate of colorectal cancer from 4% to 6%,

increases the rate of oral and pharyngeal cancer from 0.6% to 2.9%, increases the base rate for laryngeal cancers from 0.3% to 0.7%, and increases the base rate for esophageal cancer from 0.8% to 5.4% for men and 0.2% to 1.34% in women. With this information, it is clear that alcohol consumption poses a greater likelihood of causing breast or UADT cancer for an individual than for any of the other cancers, which are all low likelihood events, despite the RR associated with increased alcohol being much larger for the other cancers. This illustrates that the actual utility of medical advice for the patient is strongly connected to the base rate of an event and not just an increased OR or RR associated with a variable.

In regard to salt intake, a meta-analysis of five studies found adverse effects of salt intake on health outcomes, including mortality (HR per unit increase in sodium/potassium ratio was 1.13, 95% CI: 1.01-1.27, $p = .04$) (Cook, Appel, & Whelton, 2016); cardiovascular disease (HR: 1.36, 95% CI: 1.09-1.70; $p = 0.007$) (Mills et al., 2016); and kidney disease and initiation of dialysis (sub-hazard ratio (sHR) 1.17; 95% CI: 1.02-1.33) (Smyth et al., 2016). These are generally modest increases in outcomes associated with the level of salt consumption despite the widely held belief by individuals that salt is highly deleterious to their health, particularly when the effect is considered in actual terms of the number of new cases. Additionally, looking at cholesterol and the risk of stroke, a meta-analysis by Peters et al. (2016) indicated that raised total cholesterol is a risk factor for coronary heart disease (CHD) with pooled RR of 1.20 (95% CI: 1.16-1.24) in women and 1.24 (95% CI: 1.20-1.28) in men, resulting in a RRR of 0.96 (95% CI: 0.93-0.99) for CHD associated with a 1-mmol/L increase in total cholesterol. Again, these could be considered modest increases in risk for heart disease compared to how

individuals commonly view the association between cholesterol and cardiovascular health. Additionally, corresponding RRs for the risk of total stroke was not associated with cholesterol (1.01 [95% CI: 0.98-1.05]in women and 1.03 [95% CI: 1.00-1.05] in men), with a pooled RRR of 0.99 (95% CI: 0.93-1.04) indicating that cholesterol level had essentially no relationship with the likelihood of stroke.

Further, a meta-analysis (Sharp et al., 2011) of six longitudinal studies showed that hypertension was significantly associated with an increased risk of vascular dementia (OR: 1.59, CI: 1.29–1.95, $p<0.0001$) and roughly increased the odds of having vascular dementia by half. However, the yearly incidence of vascular dementia over 70 is approximately 0.9%, meaning that hypertension increased the rate of vascular dementia from 0.9% to 1.4% in those aged 70 and older. This would not result in a large increase in the absolute number of cases, particularly as considered by the layman. A similar association between hypertension and the risk of prevalent vascular dementia was found in the five cross-sectional studies (OR: 4.84, CI: 3.52–6.67, $p < 0.00001$) (Sharp et al., 2011). Additionally, Shah, Sutaria, & Vyas (2020) found that the pooled prevalence of stroke in patients with hypertension was 8.0% (95% CI: 5.1%–10.9%). The pooled unadjusted odds ratio of stroke in patients with hypertension compared to those without was 1.46 (95% CI: 1.07–1.99, I2 55.6, $n = 7$ studies) (Shah et al., 2020), meaning the effect of hypertension on stroke was a 2% increase in individuals having a stroke. Additionally, research on the association between risk factors and post-stroke dementia (PSD) indicated that six variables were PSD risk factors. The pooled relative ratio (RR) of atrial fibrillation was 1.68, previous stroke 1.59, myocardial infarction 1.40, hypertension 1.36, diabetes mellitus 1.25, and previous transient ischemic attack 1.25,

respectively (Surawan et al., 2017). These were statistically significant predictors of increased risk but were more modest effects than would likely be assumed by the general public.

Finally, compared to patients without obstructive sleep apnea (OSA), patients with OSA were significantly associated with cerebral small vessel disease MRI findings of white matter hyperintensity (WMH) and asymptomatic lacunar infarction (ALI) with a pooled OR of 2.31 (95% CI: 1.46-3.66) and 1.78 (95% CI: 1.06-3.01), respectively. However, there was no significant relationship between OSA and cerebral microbleeds (CMBs), with a pooled odds ratio (OR) of 2.15 (95% CI: 0.64-7.29) (Chokesuwattanaskul et al., 2020). Clearly, there is a connection between OSA and small vessel ischemic changes, but this may not be as large an absolute increase in white matter change and vascular events as the general public and healthcare providers believed.

Overall, these results indicate that most effect sizes for these commonly known associations between risk factors and health outcomes were less than the average reported OR of 2.16 found in the general medical literature (Chen et al., 2010). While the effect sizes for the associations such as the effects of alcohol consumption on the risk of cancer or the relationship between hypertension and developing vascular dementia were robust, the reviewed results suggest that effect sizes are generally small to medium-sized for many known medical risk factors and not out of line with effects seen in the clinical psychological literature.

2.4.2 Medication Efficacy

In a meta-analysis of community-based cohorts of people aged 55 years and older, Ding et al. (2020) examined the association between antihypertensive medications

(AHM) with incidents of dementia. There was evidence to show that those using any AHM had a reduced risk of developing dementia (HR: 0.88, 95% CI 0.79–0.98) and Alzheimer's disease (HR: 0.84, 0.73–0.97) (Ding et al., 2020). However, there was no evidence to support that a specific AHM drug class was more effective than others in lowering the risk of dementia, indicating that any AHM for hypertensive blood pressure might reduce the risk for dementia (Ding et al., 2020). Hughes et al. (2020) found similar results for the effect of AHM on reducing the risk of incidence of cognitive impairment or dementia (OR: 0.93, 95% CI: 0.88-0.98, absolute risk reduction: 0.39%, 95% CI: 0.09%-0.68%, $I2$ = 0.0%); lowering blood pressure was not significantly associated with a change in cognitive test scores.

There has been significant debate about the long-term use of aspirin to prevent stroke and cardiovascular events. For example, Bartolucci, Tendera, & Howard, (2011) conducted a meta-analysis of nine randomized control trials on aspirin's effects. The meta-analysis suggested aspirin can significantly decrease total cardiovascular events (OR: 0.865, 95% CI: 0.80–0.930; p<.001) and nonfatal myocardial infarction (MI) (OR: 0.813, CI: 0.667–0.992, p<0.042). Aspirin’s effects on decreasing stroke (OR: 0.919 95% CI: 0.828–1.021, p<0.116), cardiovascular mortality (OR: 0.956, 95% CI: 0.799–1.143, p<0.619), and all-cause mortality (OR: 0.945, 95% CI: 0.881–1.014, p<0.115) were nonsignificant (Bartolucci et al., 2011). Further, in a meta-analysis of 14 randomized control trials, Lei et al. (2016) found that aspirin use was associated with a modestly decreased risk of ischemic stroke than non-aspirin use (OR: 0.83, 95% CI: 0.74–0.93, p = 0.001). Specifically, the effect of aspirin on ischemic stroke in healthy adults was significant but not especially large (OR: 0.83, 95% CI: 0.74–0.94, I2 =

22%, $p = 0.002$). However, there was no difference in the risk of ischemic stroke between aspirin and non-aspirin groups for patients with cardiovascular diseases (OR: 0.75, 95% CI: 0.44–1.29, $p = 0.30$) (Lei et al., 2016). Additionally, it was determined that aspirin potentially increased the risk of serious bleeding events in both healthy individuals and patients with previous cardiovascular diseases, suggesting little protective benefit from aspirin (Lei et al., 2016).

Buckley & Salpeter (2015) systematically reviewed the risks and benefits of dementia medications. A review of 257 studies found that treatment with cholinesterase inhibitors (ChEIs) showed no significant difference in progression to dementia nor in global, cognitive, or neuropsychiatric outcomes for those with mild cognitive impairment (MCI) (RR: 0.67; 95% CI: 0.55-0.83). Additionally, there were significantly more adverse events in the ChEIs groups (RR: 1.09; 95% CI: 1.02-1.16), but no more serious adverse events or deaths (Russ & Morling, 2012). ChEIs produced small statistically significant improvements in cognitive, functional, and global outcomes for up to one year for those with mild to moderate Alzheimer's (OR: 1.56, 95% CI: 1.32-1.85) and Lewy body dementia (RR: 1.04; 95% CI: 0.43-1.65 to RR: 2.57; 95% CI: 0.90-4.23) (Birks, 2006; Wang et al., 2015). McShane et al. (2019) found differences in memantine efficacy in mild AD compared to moderate-to-severe AD. Overall, there is a small clinical benefit of memantine in moderate-to-severe AD, which occurs irrespective of whether they are also taking a ChEI, but no benefit in people with mild AD, which is the patient group where these medications are commonly prescribed despite problematic side effects.

Meta-analyses on effect sizes across various medication treatments indicate significant results, though most are consistently small. There were robust findings for the

efficacy of ChEIs in the treatment of MCI. While there are many well-established medications for treating medical disorders, this brief literature review highlights the importance of adequate effect size interpretation for adequate support of medication intervention.

2.4.3 Neuroimaging Efficacy

For the purposes of this review, the efficacy of neuroimaging in screening, diagnosis, and monitoring of disease trajectory will be reviewed. Identifying brain biomarkers of disease risk is limited by measurement reliability. Measuring brain activity using task functional MRI (fMRI) is a major focus of biomarker development; however, task fMRI's reliability has not been systematically evaluated. Elliot et al. (2020) found converging evidence demonstrating poor reliability of task-fMRI measures (mean intraclass correlation coefficient (ICC) = .397). The test-retest reliabilities of activity in a priori regions of interest across 11 common fMRI tasks were poor (ICCs = .067–.485) (Elliot et al., 2020).

Schmand et al. (2014) found that neuropsychological assessment is more accurate than MRI measures of brain atrophy for detecting disease progression in patients with MCI or AD. Measurement at baseline and two-year follow-up of hippocampal atrophy, cortical thickness, and performance on neuropsychological tests in memory clinic patients were collected. The composite neuropsychological score decreased by 0.6 SD in the impaired group and was unchanged in the normal group (Schmand et al., 2014). Estimates of required sample sizes to detect a 50% reduction in the rate of change were larger using the rate of hippocampal atrophy (n = 131) or cortical thickness (n = 488)

compared to change scores on neuropsychological assessment (n = 62) (Schmand et al., 2014).

In terms of the efficacy of imaging for detecting TBI severity, Amyot et al. (2015) found that CT, MRI, and transcranial doppler ultrasound (TCD) were determined to be the most useful modalities in the clinical setting. Specifically, in a review of publications using CT scans, evidence suggests the efficacy of CT scans for detecting moderate to severe TBIs, though less sensitive in detecting mild TBI. In mTBI, evidence suggests that CT scanning is of limited usefulness in the clinical evaluation of patients presenting to the ED, such that 24% of those who suffered a mild TBI displayed intracranial abnormalities on a CT scan (Jacobs et al., 2010).

The sensitivity of MRI compared to CT scans has been well documented. Orrison et al. (1994) documented that MRI's overall sensitivity for detecting abnormalities in acute head trauma was 96.4%, and CT was 63.4%. Additionally, research has found that an MRI improved a three-month outcome prediction better than a CT scan. Yuh et al. (2013) found that one or more contusions on MRI are associated with poor 3-month outcomes (OR: 4.5, p<.01) after adjusting for head CT findings and demographic, clinical, and socioeconomic factors. Conversely, CT evidence of subarachnoid hemorrhage was associated with a multivariate odds ratio of 3.5 ($p < 0.01$) for poorer 3-month outcomes after adjusting for demographic, clinical, and socioeconomic factors (Yuh et al., 2013).

TCD is an important tool for monitoring the course of TBI, evaluating the effect of medical treatment or intervention, forecasting, and identifying high-risk patients after TBI. Bouzat et al. (2016) found TCD thresholds had 80% sensitivity (95% CI: 56% -

94%) and 79% specificity (95% CI: 74%-83%) to predict worsening neurologic symptoms. The negative and positive predictive values of TCD were 98% (95% CI:96 to 100%) and 18% (95% CI: 11to 28%). In patients with mild TBI, the sensitivity and specificity of TCD were 91% (95% CI: 59% - 100%) and 80% (95% CI: 75% - 85%), respectively. The area under the ROC curve, including age and GCS, was significantly improved with the adjunction of TCD. On admission, patients with abnormal TCD showed a more altered disability rating on day 28 than those with normal TCD (Bouzat et al., 2016). Overall, research on neuroimaging efficacy is significant for the use of MRI, CT scans, and TCD to diagnose and monitor the course of AD and TBI.

2.5 Review of Effect Sizes of Neuropsychological Research

As discussed above, psychological research and treatment are often believed to be ineffective. However, Meyer et al. (2001) established that effect sizes for neuropsychological tests are at least as good as medical test effect sizes. Unfortunately, there is a paucity of research documenting the effect sizes of neuropsychological assessments as compared to more general psychological assessments. Therefore, the following is a streamlined analysis of the diagnostic efficacy of neuropsychological assessment.

2.5.1 Neuropsychological Test Efficacy

Neuropsychological assessment is the application of a variety of performance-based tests to assess functioning across multiple cognitive ability areas (Harvey, 2012). These include memory, attention, processing speed, reasoning, judgment, problem-solving, spatial, and language functions. Performance on tests is compared to reference

groups to determine whether performance is as or below expected (Harvey, 2012). Neuropsychological assessment is commonly indicated for developmental neurocognitive disorders (such as Attention-Deficit/Hyperactivity Disorder or ADHD), degenerative illnesses (dementia), or head injury (TBI).

Screening tests for cognitive decline or impairment are frequently used as a stand-alone cognitive test or part of a larger cognitive assessment battery. The Mini-Mental State Examination (MMSE) and Montreal Cognitive Assessment (MoCA) are two common screening measures for cognitive decline. Research has established both screening tests as good predictors of post-stroke cognitive impairment. For example, Salvadori et al. (2013) found that the MoCA baseline score is significantly associated with post-stroke cognitive impairment (OR:1.4, 95 % CI: 1.1–1.8). Research has found sensitivity and specificity for the MoCA to range from a sensitivity of 80.48% and specificity of 81.19% (Ciesielska et al., 2016); 86% sensitivity and 75% specificity (Shen et al., 2015); and 91.4 % sensitivity, 75.8 % specificity, 80 % positive predictive value, and 89.3 % negative predictive value (Salvadori et al., 2013). For the MMSE, research has found 66.34% sensitivity and 72.94% specificity (Ciesielska et al., 2016), 82% sensitivity, and 78% specificity (Shen et al., 2016).

In an examination of neuropsychological tests from the Neuropsychological Assessment Battery (NAB) to predict functional abilities (IADL) in dementia, Ashendorf et al. (2018) demonstrated that ROC curves showed the strongest prediction of IADL (AUC > 0.90) for memory measures (List Learning delayed recall, and Daily Living Memory delayed recall) and Daily Living Driving Scenes. Furthermore, at a predetermined level of specificity (95%), List Learning delayed recall (71%), and Daily

Living Memory delayed recall (88%) were the most sensitive (Ashendorf et al., 2018). In a longitudinal study of individuals over age 70 without dementia, Rabin et al. (2012) found that episodic memory tests, such as the Free and Cued Selective Reminding Test (FR-FCSRT) and immediate recall scores from the Logical Memory subtest of the Wechsler Memory Scale-Revised) contributed to the prediction of AD (FR-FCSRT HR: 3.25, 95% CI: 2.24–4.69, p <.001, Logical Memory HR: 1.54, 95% CI: 1.04–2.25, p<.03). Ewers et al. (2012) examined the accuracy of primary MRI and cerebrospinal fluid (CSF) biomarker candidates and neuropsychological tests for predicting the conversion from MCI to AD dementia. The best single predictors were right entorhinal cortex (prediction accuracy = 68.5% [95% CI: 59.5, 77.4]) and TMT-B test (prediction accuracy 64.6% [95% CI: 55.5, 73.4%]). In conclusion, short-term conversion to AD is predicted by single biomarkers with equal accuracy as a neuropsychological measure (Ewers et al., 2012).

Further, Summers & Sanders (2012) found that baseline neuropsychological test performance predicted the outcome of mild cognitive impairment with 86.3% accuracy at a 20-month follow-up. Specifically, impairments in visual episodic memory (PAL), verbal episodic memory (RAVLT), visual, immediate memory span (SSP), working memory (SWM), divided attention (CRT), and sustained attention and target detection (RVP A) differentiated between participants who developed dementia, recovered from MCI, or remained in stable MCI. Further, the measures accurately identified 100% of the cases that progressed to AD. Sixty-five percent of the recovered cases were correctly identified, with a further 10% classified as unimpaired controls. This resulted in the correct classification of recovery from MCI in 75% of cases. Of those with stable MCI,

88% were correctly classified using the same set of measures. The subset of measures resulted in a false positive rate of 25% for classifying recovered cases as persistent MCI and 4% of stable MCI as having progressed to AD. The measures had a false negative rate of misclassification of stable MCI of 8%. This study highlights the predictive reliability and validity of performance on multiple measures to diagnose AD instead of relying on only poor episodic memory test performance (Summers & Sanders, 2012).

In examination of the criterion validity of the NAB in patients with TBI (Donders & Levitt, 2012) found that patients with TBI were more than seven times more likely (n = 23, 43%) to have attention scores <80 than controls (n = 5, 9%), $\chi 2$ (n = 108) = 15.62, p < .0001, OR: 7.27 (90% CI: 2.97–17.80). Similarly, these patients were more than eight times more likely (n = 18, 33%) to have memory scores <80 than controls (n = 3, 6%), $\chi 2$ (n = 108) = 13.39, p < .0003, OR: 8.50 (90% CI: 2.87–25.19). Finally, patients with TBI were also more than four times more likely (n = 8, 15%) to have Executive Functions scores <80 than controls (n = 2, 4%), $\chi 2$ (n = 108) = 3.97, p < .05, OR: 4.52 (90% CI: 1.18–17.31) (Donders & Levitt, 2012).

Regarding effect sizes for neuropsychological tests to diagnose ADHD, Pauli-Pott & Becker (2011) found that corresponding neuropsychological tasks predicted concurrent and school-age ADHD symptoms. Mean effect sizes for response inhibition (r =.29), vigilance/arousal (r = .27), and delay aversion (r =.38) were of medium to large magnitude, the mean effect size for working memory was small (r = .18). Despite widespread recognition that ADHD is a chronic neurodevelopmental disorder in at least half of all cases, optimal diagnostic methods among adults remain elusive (Nikolas, Marshall, & Hoelzle, 2019). In the examination of the utility of an extensive

neuropsychological battery for diagnosing ADHD among adults, Nikolas et al. (2019) found that worse performance (indexed by higher scores) on CVLT short-delay free recall (OR: 1.4, 95% CI: 1.03-1.9), Salthouse Listening Span trials (OR: 1.4, 95% CI: 1.1-1.9), DKEFS CWIT inhibition/switching (OR: 1.5, 95% CI: 1.1-1.9), TOVA RT variability (OR: 1.4, 95% CI: 1.0-2.0), and TOVA omission errors ($\beta = .43$, $p = .01$, OR: 1.5, 95% [CI: 1.1-2.1]) all significantly and incrementally predicted membership in the ADHD group compared with the non-ADHD group, with an overall classification accuracy of 72.1% (Nikolas et al., 2019).

Generally, neuropsychological tests have been reported to have similar effect sizes and accuracy rates as medical tests described in the previous section. However, there is not a comprehensive synthesis of this literature. Therefore, for the current study, recent studies examining the efficacy of MRI to diagnose dementia, neuropsychological tests to diagnose dementia and TBI, and medication to treat memory impairment (cholinesterase inhibitors and memantine) and cardiovascular events (aspirin) were examined.

2.6 The Present Study

The overview of medical and psychological test effect size literature presented above demonstrated the need for continued test efficacy research. Investigators have examined the efficacy of medications, lifestyle choices, and neuroimaging for disease prevention and detection. There is strong evidence that the public generally overestimates the effectiveness of medical technologies and interventions. Additionally, there is considerable research on psychological test efficacy for diagnosing and severity of

neurocognitive disorders, daily functioning, psychopathology, and employee suitability. Meyer et al. (2001) reported on psychological test validity compared to medical test validity and found that psychological test validity was comparable to the validity of medical tests. However, this research examined mostly meta-analyses from before 1999. Additionally, the systematic review by Meyer et al. (2001) examined a broad range of medical and psychological diagnostic tests with limited ability to draw direct comparisons for similar diagnostic categories. Therefore, the current review drew from recent studies within specific diagnostic domains.

Aside from the need for an update to the Meyer et al. (2001) white paper, a greater issue motivated the conception of the present study. Research suggests a significant discrepancy in the credibility of medical tests versus psychological tests among consumers. Poling data suggests that consumers' trust in physicians and medical research is greater than that of psychological research. Additionally, research on test efficacy across the two disciplines utilizes different statistics and is reported differently by the media, leading to consumer confusion. In consideration of effect sizes across fields, the strength of association (*r*) or comparison of means (Cohen's *d*) is more common than odds ratios (OR) or risk ratios (RR), which are prominent in medical test efficacy (Ferguson, 2009). Further, effect sizes seen in the social sciences are often very small (Rosnow 2003), and there is no agreement on what magnitude of the effect is necessary to establish practical significance (Ferguson, 2009). This leads to difficulties in their interpretation. Similar to effect sizes seen more commonly in psychological research, interpretation of risk estimates in medical research is also context-dependent (Ferguson, 2009). However, the numerical size generated is often much larger (i.e., >1)

than the magnitude of psychological research effect sizes. This discrepancy furthers the doubt of efficacious psychological tests to medical tests among consumers. However, contemporary studies indicate that a test of psychological diagnosis and functioning produces similar effect sizes to medical tests. This quantitative synthesis of literature will serve as a powerful tool in utilizing neuropsychological tests for diagnosis and treatment.

Future information from this meta-analysis will likely benefit neuropsychologists and their patients alike, as research indicates the decline of neuropsychological tests despite their usefulness in diagnosis and treatment. This study holds promise in serving as a powerful tool that examines the consistency of effects across medical and neuropsychological test studies and assists both practitioners and scientists in determining just how important these tests are. To date, the Meyer et al. (2001) review continues to be the only synthesis of this literature.

2.7 Goals and Hypotheses

Goal One. The primary goal of this study was to compare the magnitude of the total effect size across medical tests, medications for chronic disease, and neuropsychological tests. That is, this goal sought to determine if there is a reliable difference in effect sizes between medical tests, medications, and neuropsychological tests. In accordance with this goal, it was hypothesized that:

1.1 Similar to findings reported from the Meyer et al. (2001) meta-analysis, effect sizes across medical tests, medications for chronic disease, and neuropsychological tests will not be significantly different.

Goal Two. The second goal was to index the average effect size within the two domains of medical interventions and technologies: neuroimaging and medications. In relation to this goal, it was hypothesized that:

2.1 Effects associated with neuroimaging would result in larger effect sizes than effects for medications for chronic disease.

Goal Three. The third goal of this study was to index the average effect size within the two clinical groups assessed with neuropsychological tests. To address this goal, it was hypothesized that:

3.1 Effects associated with neuropsychological tests for those diagnosed with AD would be comparable to effects for those diagnosed with a TBI.

CHAPTER III

METHODOLOGY

3.1 Study Identification

The formal search for prospective candidate articles was conducted using Academic Search Complete in addition to the following adjunctive databases: PsycINFO, PsycARTICLES, Embase, PubMed, CINAHL, ScienceDirect, and ProQuest Dissertations & Theses. To build the search term, *effect size,* or *efficacy*, was combined with the terms *test,* and *validity,* followed by the descriptors *neuropsych,* *cognition,* *dementia,* *Alzheimer's Disease,* head injury,*, and *diagnosis,* for neuropsychological test efficacy to diagnose to Alzheimer's Disease and TBI examined in the study. To provide a reasonable overview of the evidence on medical testing, *effect size,* or *efficacy*was combined with the terms *test,* and *validity,* followed by the descriptors *neuroimaging,* *MRI,* *medication*, *memantine,* *cholinesterase inhibitors,* *dementia,* *Alzheimer's Disease,* *mild cognitive impairment,* *diagnosis,* for each domain of medical test efficacy examined in the study. All semantically related variants of these keywords were included (e.g., head injury → acquired brain injury → traumatic brain injury → concussion → mild traumatic brain injury → TBI → mTBI etc.) In addition, the search was filtered to only display articles since 2016 from academic, peer-reviewed sources and the English language. Additional

studies were identified using a snowballing method of identifying candidate studies from reference sections of review papers and recent studies.

3.2 Study Selection Criteria

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were followed in carrying out this project (Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009). As relevant to this study, the satisfied PRISMA 2009 Checklist criteria/elements included (1) assigning an appropriate title to this project as being a meta-analysis; (2) provision of a structured summary (abstract included background, objectives, data sources, study eligibility criteria, total participants, synthesis methods, results, limitations, and conclusions/implications); (3) in the introduction, description of the rationale and objectives of the meta-analysis in the introduction; (4) in the methods section, detailed description search criteria, (5) information sources (i.e., databases), (6) search terms, (7) study selection/eligibility criteria, (8) data collection/extraction procedure, (9) specific data elements coded, (10) how bias was assessed (i.e., publication bias analyses), (11) summary of effect size estimate and discussion of how effect sizes were combined, and (12) secondary analysis plan (i.e., meta-regressions and subgroup analyses); (13) results section detailing final study sample, (14) study characteristics, (15) presentation of appropriate statistics (i.e., combined effect sizes, confidence intervals, measures of heterogeneity), (16) publication bias results, and (17) secondary analysis results; (18) a discussion section that summarizes the meta-analyses findings, (19) limitations, and (20) conclusions (Moher et al., 2009).

Studies identified from database searches and secondary obtainment were evaluated according to several criteria. According to the central goal of this study, investigations must feature the inclusion of a clinical group (i.e., individuals with dementia due to AD, MCI, TBI, and/or cardiovascular risk factors). Studies included should also include a control group. Next, studies must employ one of the following tests or treatments: a brain MRI in samples with a diagnosis of Alzheimer's Disease, medication for memory impairment (MCI or AD), aspirin for preventing or reducing the risk of cardiovascular events, or neuropsychological tests in samples with a diagnosis of AD or a TBI. Third, investigations must feature the inclusion of efficacy statistics, including effect sizes, confidence intervals, sensitivity, specificity, and/or error rates that reflect some analysis of cross-sectional and/or longitudinal analysis of differences between clinical and healthy individuals and/or correlation coefficients of the overall mean effect size relation. Finally, selected studies must provide sufficient descriptive statistical information (e.g., mean, *SD*) for each group.

3.3 Procedure

After finalizing the study selection for analysis, essential information from each was coded onto spreadsheets. To reduce bias and improve coding reliability, two trained graduate psychology students independently re-evaluated studies based on the inclusion criteria and coded them. The inter-coder agreement was examined and evaluated by a third trained graduate student to observe the level of reliability and consistency of coding between coders. All discovered discrepancies between coders were addressed on a case-

by-case basis by the third coder and reviewed by the whole research team to ensure 100% agreement across the final spreadsheet.

To appropriately evaluate each of the research aims, study information was conceptually organized hierarchically. The highest tier of coding included the type of test (i.e., neuroimaging, medication, or neuropsychological test). Next, subgroup diagnosis (i.e., AD, MCI, TBI, or cardiovascular condition) was coded. In the next level of coding, the outcome/measure utilized was coded for all studies to examine the breakdown of common measures and domains assessed. Coding for reported statistical information, specifically efficacy statistics reported, and descriptive information was obtained. Lastly, demographic (i.e., mean age of control and clinical samples, mean years of education of control and clinical samples, and % male) and study characteristics (i.e., cognitive domain, brain region, type of medication) were coded for each study to allow for subsequent moderation analyses (i.e., estimation of 'third variable' effects). This coding paradigm resulted in two levels of analysis according to specificity, which ranged from broad domain (i.e., neuroimaging, medication, or neuropsychological test) to the specific diagnosis within each domain (e.g., AD or TBI for the neuropsychological test) with the ability to examine moderating effects of several study/methodologic factors. The analysis plan for primary and secondary (i.e., moderation) analyses is discussed below.

3.4 Data Analysis

Quantitative meta-synthesis of selected studies was performed using Comprehensive Meta-Analysis (v3) software. To account for presumed heterogeneity between and within studies, all analyses were modeled by way of random effects

(Borenstein et al., 2009; Field & Gillett, 2010; Hunter & Schmidt, 2004). In general, analysis of studies occurred across two levels of granularity such that effect sizes were computed for (1) aggregated test or treatment grouplets (i.e., neuroimaging, medication, or neuropsychological test) and (2) when feasible (i.e., $k \geq 2$), at the level of individual conditions within each domain (e.g., AD or TBI for the neuropsychological test). In addition, in line with the random-effects model used for this study's analyses, within-group estimates of $\tau 2$ (i.e., the total variance between studies; Huedo-Medina et al., 2006) were pooled (Higgins & Green, 2011). Stated differently, a common among-study variance component across subgroups was assumed.

3.4.1 Effect Size Computation and Combination

The main index of effect size used in this study was Hedge's *g*. Its computation was based on available demographic, descriptive, and/or inferential statistics (Borenstein et al., 2009). Ideally, Hedge's *g* is optimally estimated using control and clinical sample size (*n*), *M*, and *SD*. For the instances when studies did not report sufficient descriptive statistical data (i.e., partial information), effect sizes were estimated using inferential statistical information, including test statistics and *p* values (when non-specific *p*-values were given, they were set to conservative defaults of .05 for significant results and .999 for non-significant results; Borenstein et al., 2009). Hedges' *g* is analogous to Cohen's *d* in that their unit of measurement is in *SDs*. Unique to this meta-analytic metric is its correction for study sample size, which functions to correct for possible overestimation bias in study effect sizes (Zakzanis, 2001). Interpretation of Hedges' *g* followed the convention proposed by Cohen (1988), such that 0.2 is small, 0.5 is moderate, and 0.8 is large.

For studies with complex data structures (i.e., multiple outcomes) that reported multiple effect sizes, they were combined using the mean of the selected outcomes, and the study was used as the unit of analysis. In other words, for studies in which subgroups produced multiple effect sizes across multiple outcomes, effect sizes were not treated independently (as the assumption of independence of outcomes would not be met) but rather pooled/averaged (Raudenbush, 2009; Shadish & Haddock, 2009). Following conservative recommendations from Borenstein et al. (2009), random-effects modeling was used for all primary and secondary (i.e., moderation) analyses. In addition, stratification of effect sizes by clinical aggregate enabled formal evaluation of all hypotheses (i.e., indexing the effect size differences across and within each domain).

3.4.2 Assessing Heterogeneity

For all primary and sub-meta-analyses, heterogeneity of effect sizes was assessed in several ways. Specifically, Cochran's *Q* test was used to determine whether the degree of observed between-study heterogeneity (i.e., true variation) was above or below expectation based on within-study error (i.e., sampling error; Borenstein et al., 2009; Higgins et al., 2003). Due to underlying statistical assumptions (i.e., the *Q* statistic's inherent comparison of observed heterogeneity to what is expected due to pure sampling error in the theoretical distribution containing *all possible studies*), fixed-effects modeling was used to estimate heterogeneity (Borenstein et al., 2009). Heterogeneity was also examined through the *I2* statistic, which broadly reflects the ratio of true heterogeneity (i.e., variability due to true differences between studies) to total variation. In other words, the *I2* statistic provides an approximation of the proportion of observed variance that reflects real differences in effect size (due to between-study differences as

opposed to sampling error; Higgins et al., 2003). As *I2* reflects a proportion or ratio, it is multiplied by 100 in the default equation to present as a percentage (%)—the percentage of between-study heterogeneity that reflects true heterogeneity (Huedo-Medina et al., 2006). According to guidelines from Higgins and Thompson (2002), degree/proportion of true heterogeneity derived from *I2* is qualified as follows: 25% = small heterogeneity, 50% = medium heterogeneity, and 75% = large heterogeneity. Others have postulated that any *I2* value exceeding 75 is thought to reflect considerable heterogeneity (Higgins & Green, 2011). Thus, a non-significant *Q* test and an *I2* value of 0 would indicate that effect sizes are broadly homogenous and that all observed between-study variability in effect sizes is broadly due to sampling error (i.e., within-study error; Borenstein et al., 2009; Huedo-Medina et al., 2006).

3.4.3 Assessing Publication Bias

At the broadest sense, the funnel plot of effect sizes (*x*-axis) distributed across standard error (*y*-axis) was visually inspected for preliminary evidence of publication bias (i.e., gross asymmetry toward the positive end of the *x*-axis; Sutton, 2009). In addition, for each main and sub-meta-analysis with sufficient studies (i.e., $k \geq 3$), Egger's regression intercept was examined to determine whether there was a linear relation between effect size and standard error (Egger et al., 1997). A one-sided significance test was used for these analyses as a positive relation between standard error and effect size is particularly concerning for publication bias in the present study. Such a relation reflects a systematic relation between study sample size (i.e., standard error) and effect size, with smaller studies producing disproportionate large effect estimates (and the inverse true for large sample studies; Sutton, 2009). Lastly, Duval and Tweedie's (2000a, 2000b) trim-

and-fill procedure was used to test and subsequently correct for publication bias. In simple terms, this methodology remedies gross asymmetry in the funnel plot by imputing missing studies (hypothetically due to publication bias) in the right field of the funnel plot as well as trimming off studies (extreme outliers) on the left field of the funnel plot in order to restore symmetry to the center of the funnel plot (Sutton, 2009). The Trim and Fill procedure augments studies (i.e., imputation and trimming), resulting in an adjusted combined effect size estimate and corrected funnel plot. The random-effects model was used in searching for missing studies (Borenstein et al., 2009).

3.4.4 Assessing Continuous and Categorical Moderators

According to several authorities on meta-analysis, a minimum of 10 studies should accompany each moderator variable characteristic being modeled (Borenstein et al., 2009; Higgins & Green, 2011). This ideal was upheld as best as possible for the current study, and exceptions were made on a minimal and outstanding basis (e.g., one to two studies shy of the standard). Continuous and categorical moderators/covariates were formally assessed with statistical methods at the domain-broad and domain-specific for studies to maximize statistical power for each analysis. In addition, heterogeneity was only further explored using these moderation analyses when prompted by analyses showing significant between-study heterogeneity as assessed by *Q* and *I2* (Higgins & Green, 2011).

The moderating effect of continuous covariates were assessed using the meta-regression procedure, where effect sizes are regressed upon covariates to determine whether they predict significant heterogeneity/variance in effect sizes (Hedges & Pigott, 2004; Higgins & Green, 2011). This analysis allows for either single (i.e., univariate) or

simultaneous (i.e., multivariate) examination of potential continuous moderators (Hedges & Pigott, 2004). All regression analyses were modeled under random effects and estimated using the method of moments approach (Raudenbush, 2009). The following continuous moderators were coded for: clinical age (*M*), control age (*M*), education clinical (*M* years), education control (*M* years), % male clinical, and % male control. Continuous moderator pairs (e.g., clinical age and control age) were linked in order to combine them into a superordinate covariate (e.g., age), but this procedure was only done if each separate covariate had sufficient studies (i.e., $k \geq 10$; Borenstein et al., 2009; Higgins & Green, 2011). Thus, when continuous moderators had sufficient studies, they were linked together and run simultaneously in a multivariate meta-regression model.

Finally, the following categorical moderator variables were coded: cognitive domain (i.e., verbal learning and memory, visual learning and memory, language, attention, working memory, processing speed, executive functioning, visuospatial skills, motor skills, global cognition), brain region (i.e., cingulate cortex, cortical grey matter, frontal lobe, hippocampus, occipital lobe, parietal lobe, temporal lobe, white matter lesions, whole-brain volume), and type of medication (i.e., aspirin and medication for dementia (i.e., donepezil or memantine). When there were enough studies (i.e., $k \geq 10$ within each level of the categorical moderator), subgroup analysis was conducted to formally analyze categorical variables using a group-by and compare function (Borenstein et al., 2009; Higgins & Green, 2011). A *Q* statistic was used for pairwise comparisons of effect sizes across levels of continuous moderator variables, and a significant finding indicated a significant difference in combined effect sizes.

3.5 Power Analysis

The literature addressing fundamental statistical power (i.e., $1 - \beta$) questions about meta-analyses is sparse. Because of the nature of meta-analysis research, unique parameters determine the number of studies necessary for achieving satisfactory statistical power. Examples of such parameters are heterogeneity between studies, expected average effect size(s), and average per-group sample sizes of each study (Valentine, Pigott, & Rothstein, 2010). Additionally, Meyer et al. (2001) did not apply a power analysis to determine how many studies to include. Consistent with this research and the primary purpose of this paper to compare well-known tests and technologies in the medical and psychological literature, a power analysis was not included.

CHAPTER IV

RESULTS

4.1 Study Characteristics

The initial search yielded 6,668 results and was reduced to 6,522 after removing duplicate hits. Following the first screening phase, 6,240 records were excluded for irrelevancy based on the abstract. Two hundred eighty-two studies were selected for an in-depth review to determine eligibility. From these, 210 additional records were excluded for the following reasons: no clinical or control group (k = 53); wrong assessment tools used (k = 48); wrong outcomes (k = 28); wrong patient population (k = 32); insufficient statistical information/data provided (k = 26); and inappropriate study type (k = 23). Four additional eligible studies were identified through reference sections of other studies reviewed (i.e., snowballing). Following this detailed review of results, 72 studies satisfied inclusion criteria and were selected for meta-analysis, and coding yielded 681 effect sizes for extraction. Included studies by broad domain aggregate were as follows: neuropsychology = 32, neuroimaging = 21, and medication = 19 (*note*: 5 studies provided effect sizes for both neuropsychological and neuroimaging data for AD clinical groups). Three TBI neuropsychological studies reported multiple groups based on injury severity (Baum et al., 2017, Steward et al., 2017, Tulsky et al., 2017). Each group was analyzed as an individual study with independent outcomes by TBI severity, increasing

the total number of studies to 75. See *Appendix A* for a detailed flowchart of the study selection process (and reasons for study exclusion).

Across the total sample of selected studies (k = 75, 681 effect sizes), Hedge's g was estimated with full statistical information (i.e., clinical/control M, SD, and n) for 63 studies (84% of all included studies) and a total of 586 extracted effect sizes (86% of all effect sizes). When only partial statistical information was available, effect sizes (Hedge's g) were estimated under the following alternative circumstances: p-values and clinical/control n (k = 4, 31 effect sizes) and clinical/control events and clinical/control n (k = 10, 64 effect sizes). In the broadest sense, only 4% (31) of all the included effect sizes were extracted under the crudest conditions—when papers provided minimum statistical information for effect size estimation: p-values and clinical/control n. Alternatively stated, 96% (650) of the total number of effect sizes extracted from included studies provided complete (i.e., M, SD, and n for clinical and control groups) or near-complete (e.g., events and sample sizes for both groups), statistical information for effect size (Hedge's g) estimation. The more statistical information provided, the fewer assumptions are made in back-interpolating effect sizes (Borenstein et al., 2009).

4.2 Broad-Domain Findings

Seventy-five studies (including independent subgroups within studies reported above), yielding 681 unique effect sizes, were represented in the primary meta-analysis (i.e., comparing effect sizes across domains: neuropsychological tests, neuroimaging, and medications). There were sufficient studies to determine whether effect sizes differed

across domains. Analysis revealed that combined effect sizes significantly differed as a function of broad-domain findings [$Q(2) = 78.081$, $p < .001$]. See *Figure 1* for visual plot of this analysis.

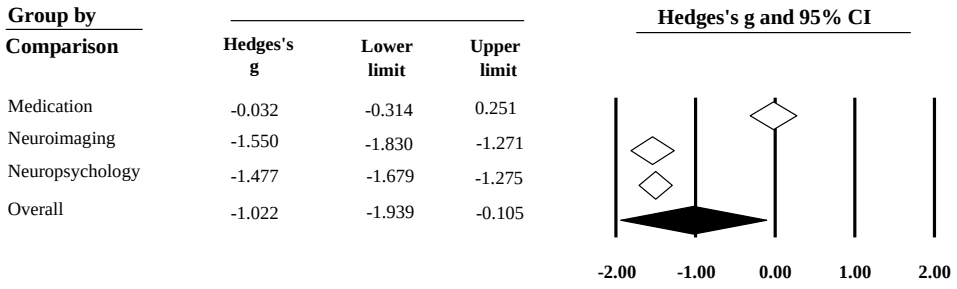

Figure 1. Subgroup analysis of studies across broad-domain aggregates.

According to pairwise comparisons, the neuroimaging aggregate combined effect size [Hedges g = -1.670] was not significantly different than the neuropsychological test efficacy [Hedges g = -1.548; Q(1) = 0.213, p = 0.644] but was significantly greater than the medication test efficacy [Hedges g = -0.030; Q(1) = 104.006, p <0.001]. Further, neuropsychological tests aggregate combined effect size [Hedges g = -1.468] was also significantly greater than medication test efficacy [Hedges g = -0.032; Q(1) = 73.358, p <0.001]. More detailed information on these comparisons is displayed below. See Table 1 for full subgroup analysis results. Below, domain-specific findings are displayed.

Table 1.

Categorical Moderators for Broad Domain Studies

Categorical Moderator (*k*)	***Q***	***df***	***p***	Pairwise
Broad Domain (78)[g]	78.08	2	<.001	---
Neuroimaging vs. Neuropsychological	0.21	1	0.644	MRI = Neuropsych
Neuroimaging vs. Medication	104.01	1	<.001	MRI > Meds
Neuropsychological vs. Medication	73.36	1	<.001	Neuro > Meds

4.3 Neuroimaging Aggregate Findings

Across the 21 studies examining brain MRIs between clinical (n = 1695) and control (n = 1507) samples, 171 effect sizes were extracted for analysis. Meta-analysis resulted in a significant large, combined effect size [Hedge's g = -1.623, 95% CI = -1.973 to -1.273, $p < .001$]. Specifically, clinical samples diagnosed with Alzheimer's Disease had greater atrophy across brain regions than controls (see *Figure 2*). There was a significant degree of between-study heterogeneity [Q(21) = 298.655, p < .001, I2 = 93.303]. Visual inspection of the funnel plot indicated that two studies (Machado et al., 2020 and Park et al., 2017) with smaller sample sizes yielded larger (in the negative direction), more unstable effect sizes, and perhaps should be interpreted with caution. Similarly, Egger's regression intercept [t(19) = 3.52, $p_{(one\text{-}tailed)}$ = .001] indicated significant publication bias (see *Figure 3*). However, the Trim and Fill procedure did not identify any studies likely missing due to publication bias.

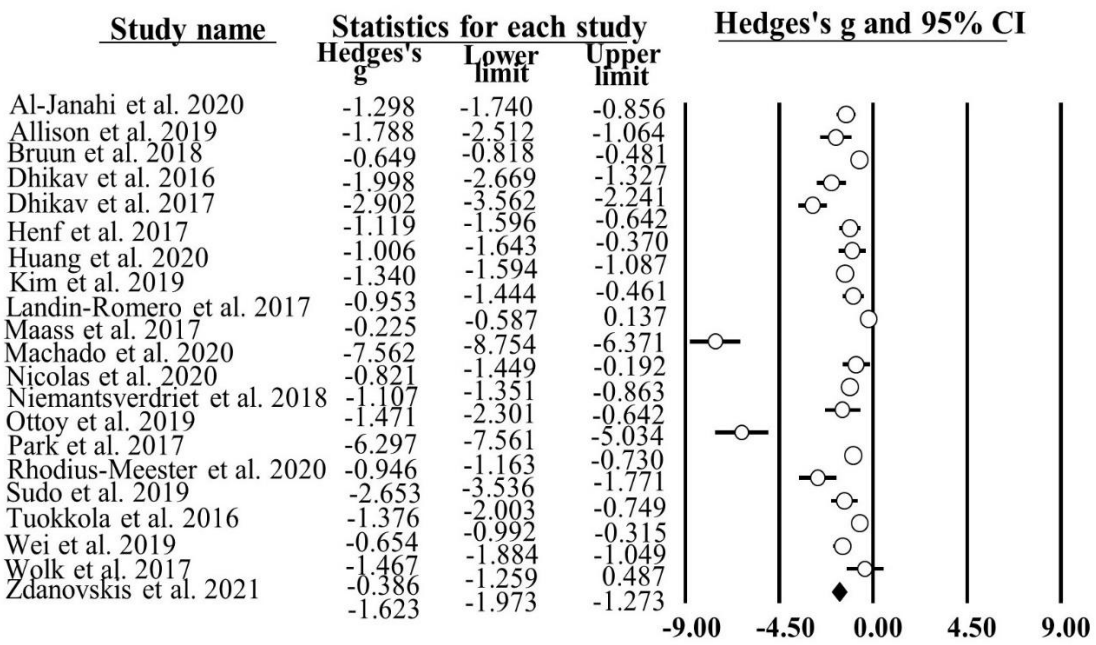

Figure 2. Forest plot for neuroimaging studies.

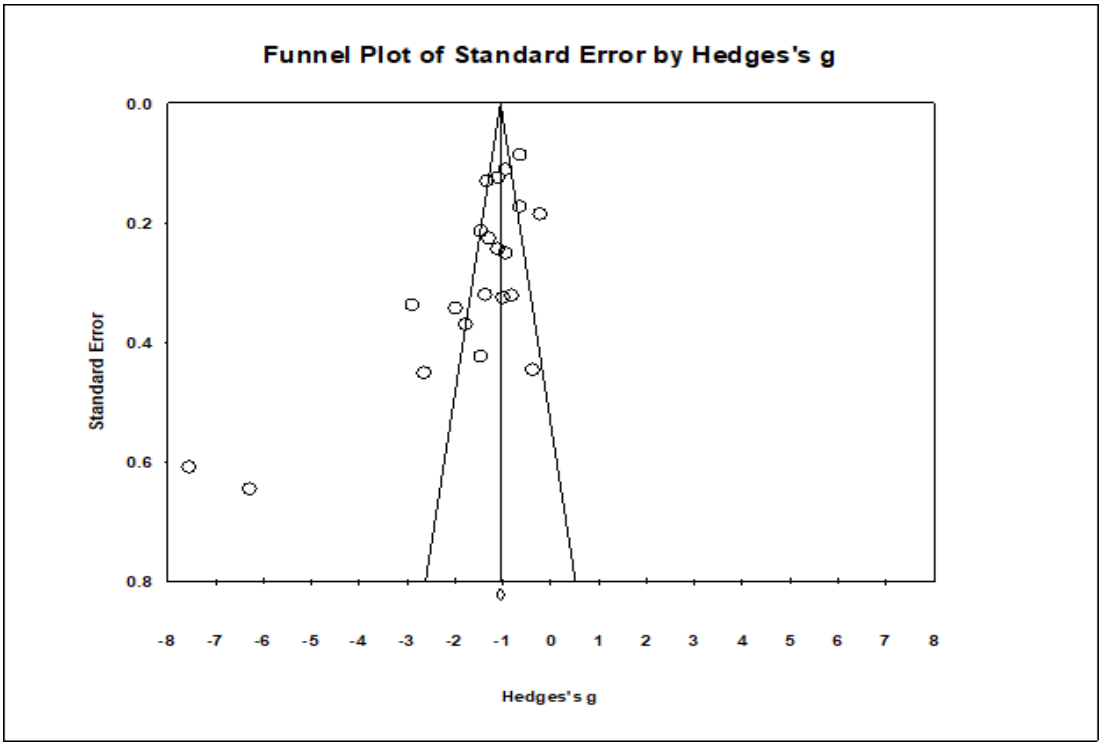

Figure 3. Funnel plot of standard error by Hedge's *g* for neuroimaging studies.

For broad neuroimaging studies, multivariate meta-regressions were performed on the following variables with sufficient studies for linking: age (k = 21), education (k = 13), and sex (k = 19). Age [$Q(2) = 5.18$, $R^2 = .00$, $p = .0750$], education [$Q(2) = 4.59$, $R^2 = .00$, $p = 0.1007$], nor sex [$Q(2) = 1.65$, $R^2 = .00$, $p = .4374$] were statically significant. Education trended toward statistical significance but did not account for much variance. Specifically, analyses indicated that clinical education [$\beta = 0.403$, $SE = 0.189$, $z = 2.13$, $p = 0.033$] was a significant positive predictor of effect sizes. Namely, studies with more educated clinical groups yielded greater effect sizes. See Table 2 for full meta-regression analysis results. Meta-regression plot for the significant education analysis is displayed on *Figure 4*.

Table 2.

Meta-Regressions for Continuous Variables in Neuroimaging Studies

Variable (*k*)	***R²***	**β**	***SE***	***z***	***p***
Age[a] (21)					
Clinical Age (21)	< .01	-.038	.038	-0.99	.32
Control Age (21)		-.066	.038	-1.75	.08
Education[b] (13)					
Clinical Education (13)	<.01	.403	.189	2.13	.03
Control Education (13)		-.380	.235	-1.62	.11
Sex (19)[c]					
% Male Clinical (19)	< .01	-.027	.022	-1.27	.20
% Male Control (19)		.016	.024	0.64	.52

Note. [a]$Q(2) = 5.18$, $p = .0750$. [b] $Q(2) = 4.59$, $p = 0.1007$. [c] $Q(2) = 1.65$, $p = .4374$.

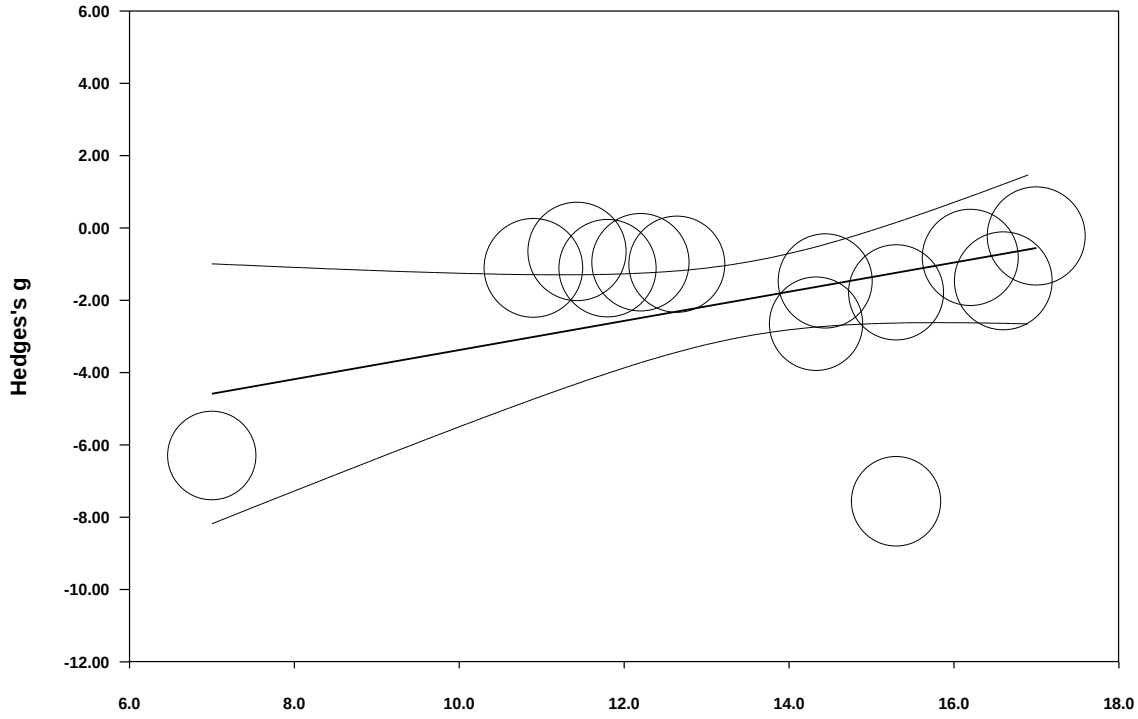

Figure 4. Regression of Hedge's *g* on clinical education for neuroimaging studies.

There were insufficient studies to formally examine the moderating effect of brain region measurement on MRI effect sizes. While comparisons could not be made, there were enough studies to report on effect sizes for the following brain region measurement between clinical and control groups. Meta-analysis on 17 studies measuring hippocampus resulted in a significant large, combined effect size [Hedge's g = -2.458, 95% CI = -3.040 to -1.875, $p < .001$]. There was a significant degree of within subgroup (hippocampus measurement) heterogeneity [$Q(16) = 410.586$, $p < .001$, $I2 = 96.103$]. Nine studies measuring temporal lobe resulted in a significant large, combined effect size [Hedge's g = -1.741, 95% CI = -2.333 to -1.249, $p < .001$]. There was a considerable

degree of within subgroup heterogeneity [$Q(8) = 113.563$, $p < .001$, $I2 = 92.955$]. See Table 3 for full subgroup analysis results.

Table 3.
Subgroup, Combined Effect Size, and Heterogeneity Statistics for Brain Regions Measured

Aggregate (*n*)	**# es (*k*)**	**Hedge's *g* (*p*)**	**95% CI**	***Q* (*p*)**	***I*²**
Hippocampus (1015)	48 (17)	-2.46 (< .001)	-3.04- (-)1.88	410.59 (< .001)	96.10
Temporal Lobe (990)	34 (9)	-1.74 (< .001)	-2.33- (-)1.25	113.56 (< .001)	92.96
Total (1372)[a]	**82 (19)**	**-2.14 (< .001)**	**-2.61- (-)1.68**	**389.506(< .001)**	**95.38**

Note. # es = Number of Effect Sizes.
[a]Controls = 1405.

4.4 Neuropsychological Tests Aggregate Findings

Across the 40 studies examining neuropsychological tests between clinical (n = 2161) and control (n = 2605) samples, 331 effect sizes were extracted for analysis. Meta-analysis resulted in a significant large, combined effect size [Hedge's $g = -1.563$, 95% CI = -1.895 to -1.230, $p < .001$]. Specifically, clinical samples (combined TBI and AD groups) had worse performance on neuropsychological tests compared to controls (see *Figure 5*). According to pairwise comparisons, the AD neuropsychological test effect size [Hedges $g = -2.213$] was significantly different than the TBI neuropsychological test efficacy [Hedges $g = -0.649$; $Q(1) = 42.821$, $p = 0.000$]. There was a significant degree of between-study heterogeneity [$Q(39) = 860.715$, $p < .001$, $I^2 = 95.469$]. Visual inspection of the funnel plot indicated that one study (Reas et al., 2017) with a smaller sample size yielded a significantly larger (in the negative direction), more unstable effect size and perhaps should be interpreted with caution. Similarly, Egger's regression intercept [$t(38) = 2.60$, $p_{(one\text{-}tailed)} = .007$] indicated significant publication bias (see *Figure*

6). However, the Trim and Fill procedure did not identify any studies likely missing due to publication bias.

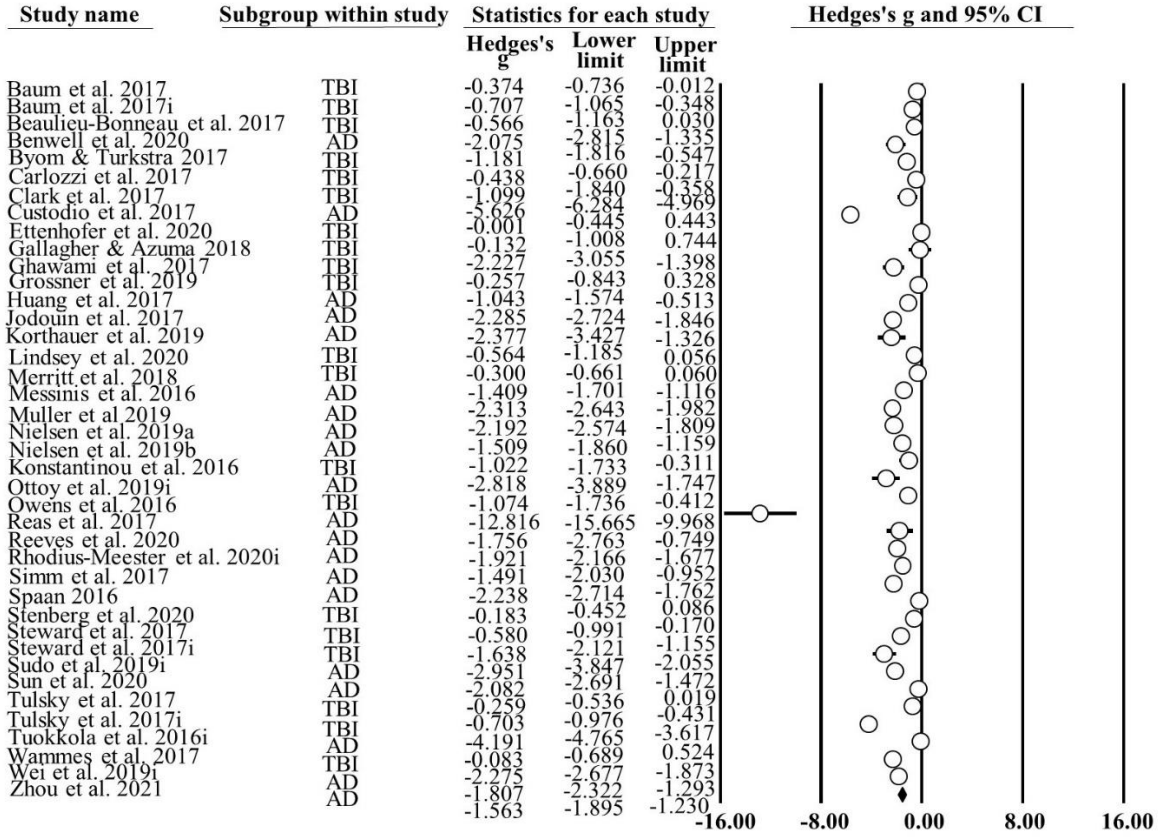

Figure 5. Forest plot for neuropsychological studies.
Note. "i" after an article denotes multiple comparisons used within study.

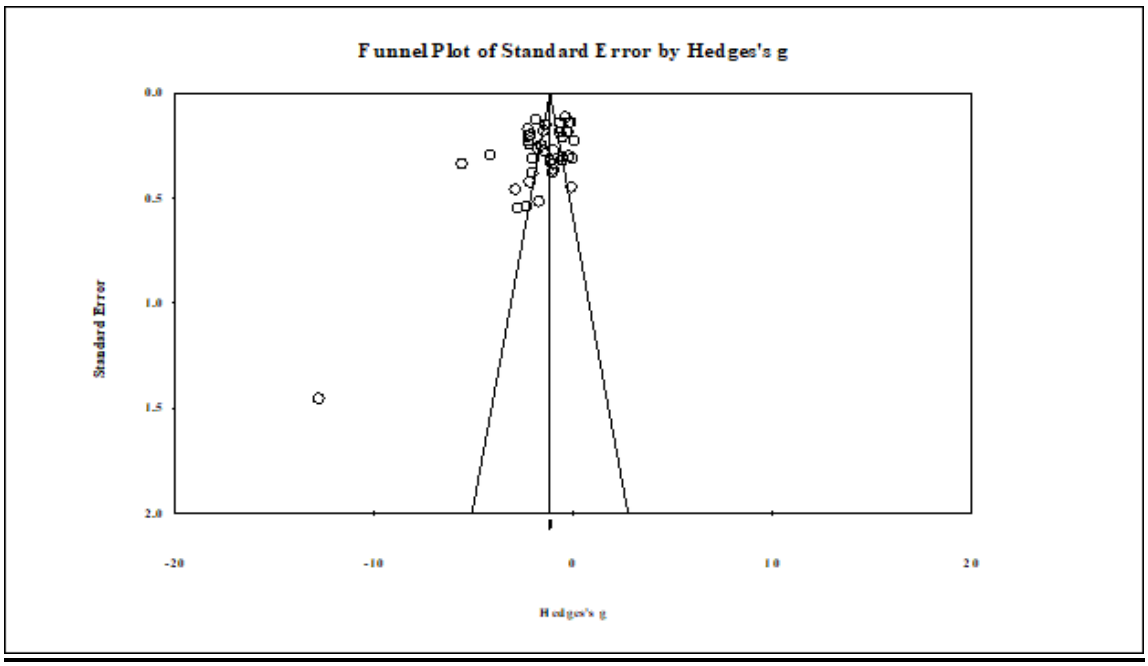

Figure 6. Funnel plot of standard error by Hedge's *g* for broad neuropsychological studies.

For broad neuropsychological test studies, multivariate meta-regressions were performed on the following variables with sufficient studies for linking: age (k = 37), education (k = 31), and sex (k = 30). Age accounted for a moderate portion of true heterogeneity between neuropsychological test studies age [$Q(2) = 34.78$, $R^2 = .47$, $p < .001$]. Within this model, analyses indicated that neither clinical age [β = -0.044, SE = 0.0316, z = -1.41, p = 0.1590] nor control age [β = 0.0072, SE = 0.0341, z = 0.21, p = 0.8338] were individually significant predictors of effect size. Sex accounted for a small portion of true heterogeneity between neuropsychological test studies age [$Q(2) = 8.48$, $R^2 = .22$, p = 0.0143]. Within this model, analyses indicated that neither clinical sex [β = 0.0167, SE = 0.0127, z = 1.31, p = 0.1898] nor control sex [β = 0.0129, SE = 0.0111, z = 1.16, p = 0.2473] were individually significant predictors of effect size. Education was

not statistically significant [$Q(2) = 3.11$, $R^2 = .23$, $p = 0.2111$]. See Table 4 for full meta-regression analysis results.

Table 4.

Meta-Regressions for Continuous Variables in Neuroimaging Studies

Variable (*k*)	R^2	β	*SE*	*z*	*p*
Age[a] (37)					
Clinical Age (37)	.47	-.044	.032	-1.41	.16
Control Age (37)		-.007	.034	.21	.83
Education[b] (28)					
Clinical Education (31)	.23	.286	.165	.78	.43
Control Education (29)		-.281	.164	-.95	.34
Sex (36)[c]					
% Male Clinical (36)	.22	.017	.013	1.31	.19
% Male Control (38)		.013	.011	1.16	.25

Note. [a] $Q(2) = 34.78$, $p < .001$. [b] $Q(2) = 0.98$, $p = 0.6138$. [c] $Q(2) = 8.48$, $p = 0.0143$.

4.4.1 Neuropsychological Tests for Alzheimer's Disease

Across the 20 studies examining neuropsychological tests between Alzheimer's Disease dementia (n = 1144) and control (n = 1524) samples, 161 effect sizes were extracted for analysis. Meta-analysis resulted in a significant large effect size [Hedge's $g = -2.483$, 95% CI = -2.906 to -2.059, $p < .001$]. Specifically, clinical samples diagnosed with AD had worse performance on neuropsychological tests than controls (see *Figure 7*). There was a significant degree of between-study heterogeneity [$Q(19) = 280.975$, $p < .001$, $I^2 = 93.238$]. Visual inspection of the funnel plot indicated that one study (Reas et al., 2017) with a smaller sample size yielded a significantly larger (in the negative direction), more unstable effect size, and perhaps should be interpreted with caution.

Similarly, Egger's regression intercept [$t(18) = 2.40$, $p_{(one\text{-}tailed)} = .013$] indicated significant publication bias (see *Figure 8*). However, the Trim and Fill procedure did not identify any studies likely missing due to publication bias.

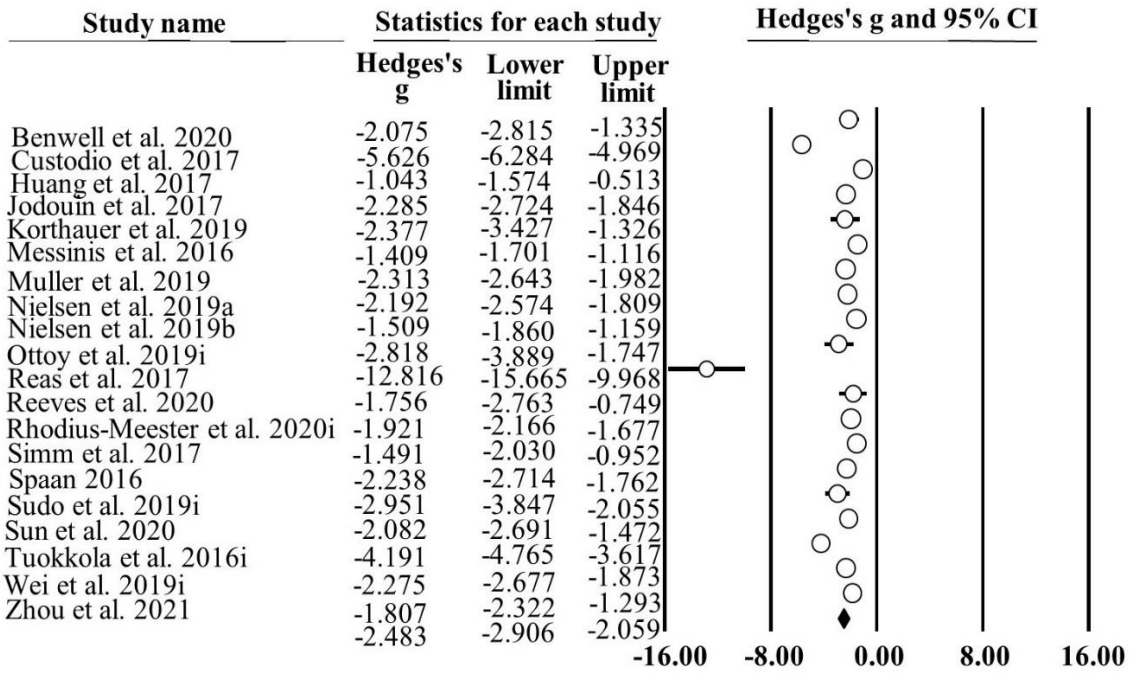

Figure 7. Forest plot for AD neuropsychological studies.
Note. "i" after an article denotes multiple comparisons used within study.

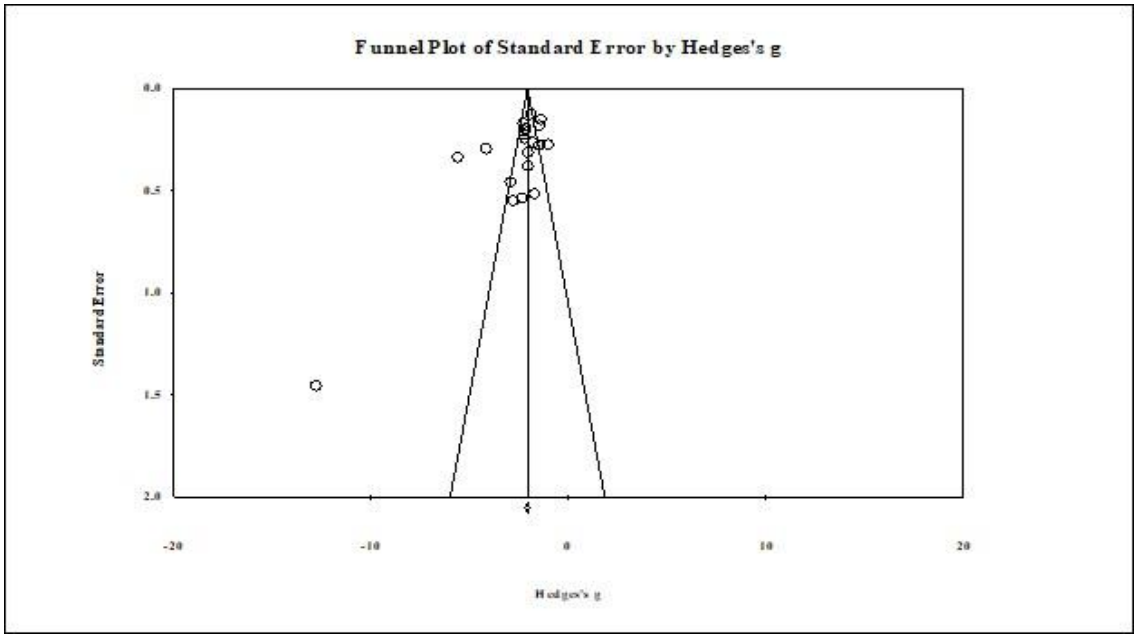

Figure 8. Funnel plot of standard error by Hedge's *g* for AD neuropsychological studies.

For AD neuropsychological studies, multivariate meta-regressions were performed on the following variables with sufficient studies for linking: age (k =20), education (k = 14), and sex (k = 19). Education accounted for a small portion of true heterogeneity between neuropsychological test studies [$Q(2) = 8.66$, $R^2 = .15$, $p = 0.0132$], though neither control education [$\beta = -0.1098$, $SE = 0.1574$, $z = -0.70$, $p = 0.4854$] or clinical education [$\beta = -0.0883$, $SE = 0.1689$, $z = -0.52$, $p = 0.6012$] alone were statistically significant. Neither age [$Q(2) = 0.21$, $R^2 = .00$, $p = 0.9016$] nor sex [$Q(2) = 3.13$, $R^2 = .00$, $p = 0.2090$] was statistically significant. See Table 5 for full meta-regression analysis results.

Table 5.

Meta-Regressions for Continuous Variables in AD Neuropsychology Studies

Variable (*k*)	R^2	β	*SE*	*z*	*p*
Age[a] (20)					
Clinical Age (20)	< .01	-.029	.082	-.35	.73
Control Age (20)		.023	.052	.45	.65
Education[b] (14)					
Clinical Education (15)	.15	-.088	.169	-.52	.60
Control Education (14)		-.110	.157	-.70	.49
Sex (19)[c]					
% Male Clinical (19)	< .01	-.044	.026	-1.71	.09
% Male Control (20)		.003	.021	.12	.90

Note. [a] $Q(2) = 0.21$, $p = 0.9016$. [b] $Q(2) = 8.66$, $p = 0.0132$. [c] $Q(2) = 3.13$, $p = 0.2090$.

There were insufficient studies to formally examine the moderating effect of the cognitive domain on neuropsychological test efficacy within an AD population. While comparisons could not be made, there were enough studies to report on effect sizes for the following cognitive domains between clinical and control groups. Meta-analysis on 14 studies measuring executive functioning resulted in a significant large effect size [Hedge's $g = -2.011$, 95% CI = -2.458 to -1.563, $p < .001$]. There was a significant degree of heterogeneity [$Q(13) = 175.280$, $p < .001$, $I^2 = 92.583$]. Ten studies measuring language resulted in a significant large effect size [Hedge's $g = -2.596$, 95% CI = -3.998 to -1.194, $p < .001$]. There was a large degree of within subgroup heterogeneity [$Q(9) = 591.676$, $p < .001$, $I^2 = 98.479$]. Fifteen studies measuring verbal learning and memory resulted in a significant large effect size [Hedge's $g = -3.642$, 95% CI = -4.368 to -2.915, $p < .001$]. There was a large degree of within subgroup heterogeneity [$Q(14) = 328.202$, $p < .001$, $I^2 = 95.734$]. See Table 6 for full subgroup analysis results.

Table 6.
Subgroup, Combined Effect Size, and Heterogeneity Statistics for Cognitive Domains in AD Neuropsychology Studies

Aggregate (*n*)	**# es (*k*)**	**Hedge's *g* (*p*)**	**95% CI**	***Q* (*p*)**	***I*²**
Executive Functions (687)	51 (14)	-2.01 (< .001)	-2.46- (-)1.56	175.28 (< .001)	92.58
Language (419)	16 (10)	-2.60 (< .001)	-4.00- (-)1.19	591.68 (< .001)	98.48
Verbal Learning and Memory (847)	53 (15)	-3.64 (< .001)	-4.37- (-)2.92	328.20 (< .001)	95.73
Total (1038)[a]	**120 (20)**	**-2.62 (< .001)**	**-3.70- (-)2.08**	**442.32 (< .001)**	**95.70**

Note. # es = Number of Effect Sizes.
[a]Controls = 1387.

4.4.2 Neuropsychological Tests for TBI

Across the 20 studies examining neuropsychological tests between those with a TBI (n = 1016 and control (n = 1081) samples, 170 effect sizes were extracted for analysis. Meta-analysis resulted in a significant medium effect size [Hedge's g = -0.613, 95% CI = -0.804 to -0.442, $p < .001$]. Specifically, clinical samples diagnosed with a TBI had worse performance on neuropsychological tests compared to controls (see *Figure 9*). There was a moderate degree of between-study heterogeneity [$Q(19) = 71.671$, $p < .001$, $I^2 = 73.490$]. Visual inspection of the funnel plot indicated overall symmetry, with one study (Ghawami et al., 2017) with a smaller sample size yielded a slightly larger (in the negative direction), more unstable effect size, and perhaps should be interpreted with caution. Similarly, Egger's regression intercept [$t(18) = 2.00$, $p_{(one\text{-}tailed)} = .03$] indicated significant publication bias (see *Figure 10*). However, the Trim and Fill procedure failed to identify any studies likely missing due to publication bias.

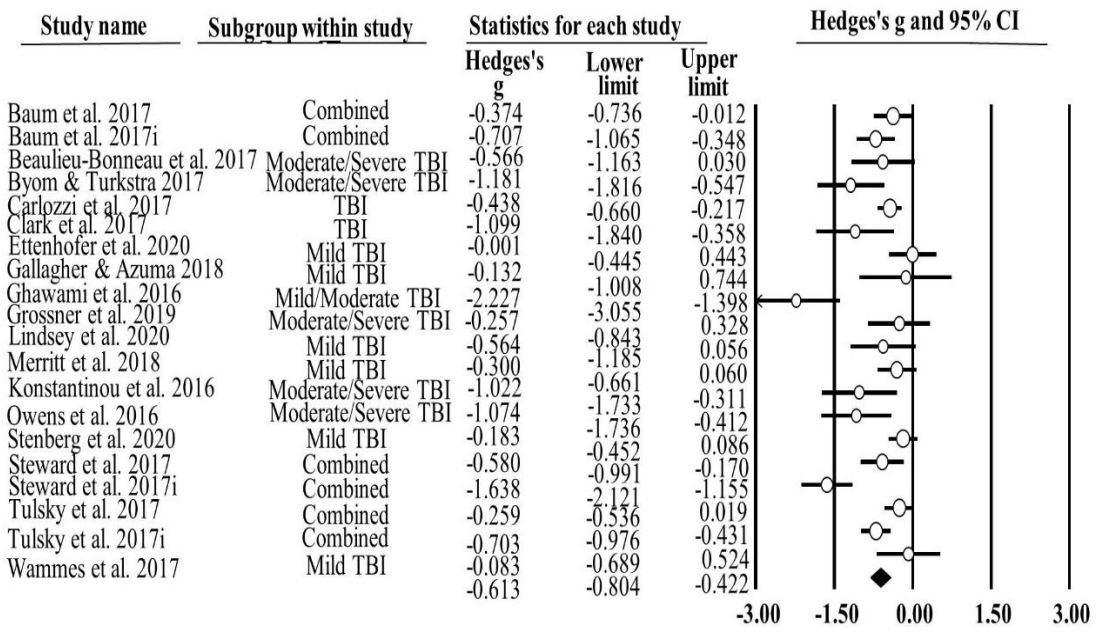

Figure 9. Forest plot for TBI neuropsychological studies.
Note. "i" after an article denotes multiple comparisons used within study.

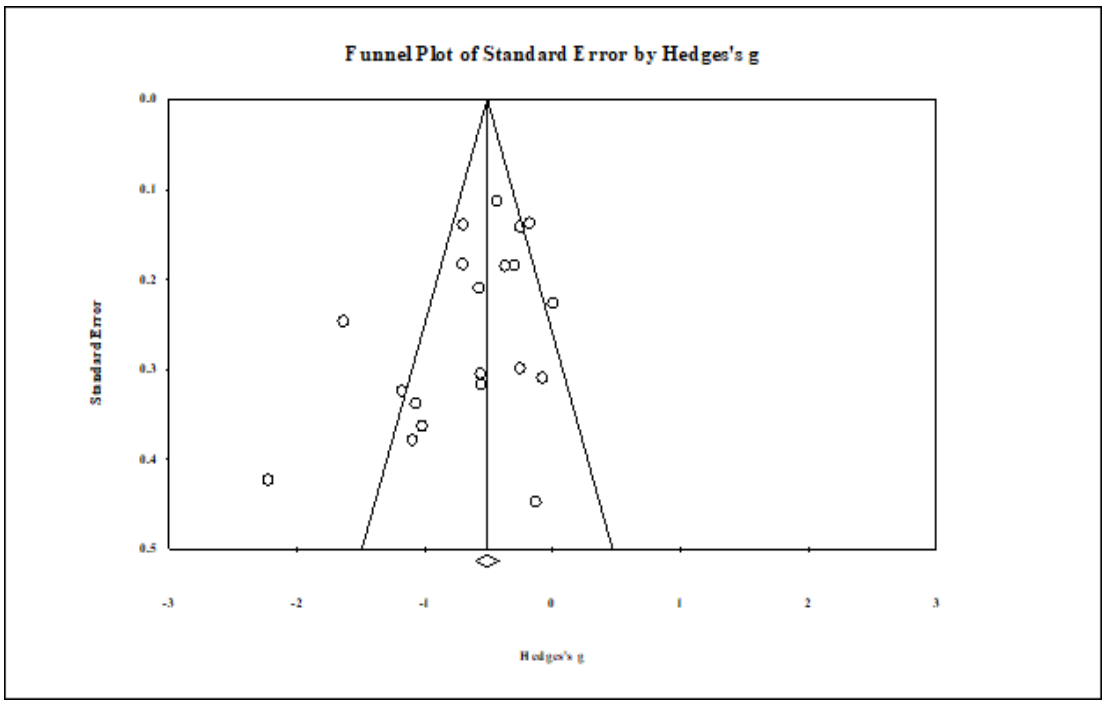

Figure 10. Funnel plot of standard error by Hedge's *g* for TBI neuropsychological studies

For TBI neuropsychological studies, multivariate meta-regressions were performed on the following variables with sufficient studies for linking: age (k = 17), education (k = 10), and sex (k = 11). Education was statistically significant in explaining a moderate degree of heterogeneity [$Q(2) = 12.11$, $R^2 = .50$, $p = 0.0023$], but neither clinical [$\beta = 0.3364$, $SE = 0.1745$, $z = 1.93$, $p = 0.0538$] nor control [$\beta = -0.0898$, $SE = 0.1779$, $z = -.51$, $p = 0.6135$] groups alone were statistically significant. Neither age [$Q(2) = .94$, $R^2 = .00$, $p = 0.6251$] nor sex [$Q(2) = 5.52$, $R^2 = .04$, $p < .0632$] was significant in explaining observed heterogeneity between studies. See Table 7 for full meta-regression analysis results.

Table 7.

Meta-Regressions for Continuous Variables in TBI Neuropsychology Studies

Variable (*k*)	***R²***	**β**	***SE***	***z***	***p***
Age[a] (17)					
Clinical Age (17)	< .01	.033	.034	.97	.33
Control Age (18)		.032	.036	-.89	.38
Education[b] (14)					
Clinical Education (16)	.50	.336	.175	1.93	.05
Control Education (15)		-.089	.178	-.51	.61
Sex (17)[c]					
% Male Clinical (17)	.04	-.023	.012	-1.95	.05
% Male Control (18)		-.002	.007	.26	.80

Note. [a] $Q(2) = .94$, $p = 0.6251$. [b] $Q(2) = 12.11$, $p = 0.0023$. [c] $Q(2) = 5.52$, $p < .0632$.

Subgroup analysis to examine combined effect size differences across TBI severity was not conducted, as the number of studies fell below the recommended k of 10, (Mild TBI k = 6, Mild/Moderate k = 1, Moderate/Severe k = 5, Combined Severity k

= 6). There were insufficient studies to formally examine the moderating effect of the cognitive domain on neuropsychological test efficacy within a TBI population. While comparisons could not be made, there were enough studies to report on effect sizes for the following cognitive domains between clinical and control groups. Meta-analysis on 18 studies measuring executive functioning resulted in a significant moderate effect size [Hedge's g = -0.685, 95% CI = -0.880 to -0.491, $p < .001$]. There was a moderate degree of heterogeneity [$Q(17) = 65.728$, $p < .001$, $I^2 = 74.136$]. Thirteen studies measuring processing speed resulted in a significant moderate, combined effect size [Hedge's g = -.645, 95% CI = -0.879 to -0.411, $p < .001$]. There was a moderate degree heterogeneity [$Q(12) = 54.789$, $p < .001$, $I^2 = 78.098$]. See Table 8 for full subgroup analysis results.

Table 8.
Subgroup, Combined Effect Size, and Heterogeneity Statistics for Cognitive Domains in TBI Neuropsychology Studies

Aggregate (*n*)	**# es (*k*)**	**Hedge's *g* (*p*)**	**95% CI**	***Q* (*p*)**	***I*²**
Executive Functions (1001)	68 (18)	-.69 (< .001)	-.88- (-).49	65.728 (< .001)	74.14
Processing Speed (774)	28 (13)	-.65 (< .001)	-.88- (-).41	54.79 (< .001)	78.10
Total (1001)[a]	**96 (18)**	**-.69 (< .001)**	**-89- (-).49**	**69.73 (< .001)**	**75.62**

Note. # es = Number of Effect Sizes.
[a]Controls = 1048.

4.5 Medication Aggregate Findings

Across the 19 studies examining medications between clinical (n = 32333) and control (n = 39122) samples, 178 effect sizes were extracted for analysis. Meta-analysis resulted in a small, nonsignificant effect size [Hedge's g = -.024, 95% CI = -0.105 to 0.058, p = 0.570]. There was a moderate degree of between-study heterogeneity [Q(17)

= 41.862, $p < .001$, $I^2 = 57.002$]. There was no indication of publication bias across these studies based on Egger's regression intercept [$t(17) = 0.42$, $p_{(one\text{-}tailed)} = .34$], visual inspection of the funnel plot (see *Figure 11*), and the Trim and Fill procedure.

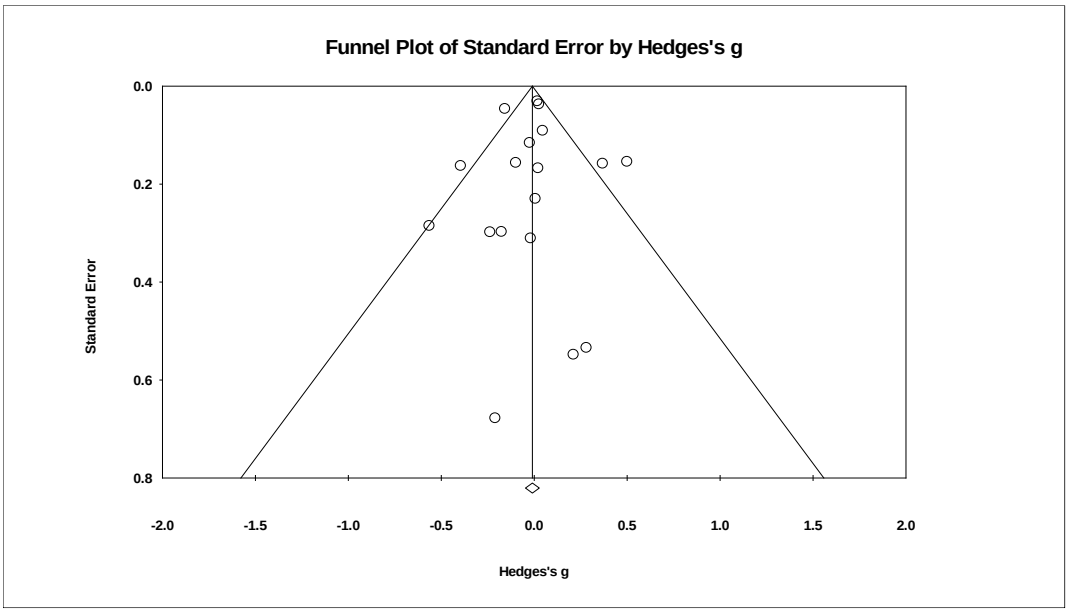

Figure 11. Funnel plot of standard error by Hedge's *g* for medications

For broad medication studies, multivariate meta-regressions were performed on the following variables with sufficient studies for linking: age ($k = 16$) and sex ($k = 16$), as there was not sufficient studies for education. Neither age [$Q(2) = 0.41$, $R^2 = .00$, $p = 0.8153$] nor sex were statistically significant [$Q(2) = 0.08$, $R^2 = .00$, $p = 0.9624$] contributors to heterogeneity found across medication studies. See Table 9 for full meta-regression analysis results.

Table 9.

Meta-Regressions for Continuous Variables in Broad Medication Studies

Variable (*k*)	R^2	β	*SE*	*z*	*p*
Age[a] (16)					
Clinical Age (16)	< .01	.001	.038	.04	.97
Control Age (17)		-.008	.034	-.24	.81
Sex[b] (16)					
% Male Clinical (16)	< .01	-.007	.024	-.21	.79
% Male Control (17)		.007	.024	.28	.78

Note. [a] $Q(2) = 0.41$, $p = 0.8153$. [b] $Q(2) = 0.08$, $p = 0.9624$.

4.5.1 Medication for Memory Impairment Findings

Across the 10 studies examining medication for memory impairment only between clinical (n = 1777) and control (n = 1568) samples, 126 effect sizes were extracted for analysis. Meta-analysis resulted in a small, nonsignificant effect size [Hedge's g = -.052, 95% CI = -0.237 to 0.133, p = 0.580] and degree of heterogeneity was moderate [$Q(8) = 34.139$, $p = 0.000$, $I^2 = 73.637$]. Specifically, there was no difference in performance on cognitive tests for those diagnosed with MCI or AD who were prescribed medication compared to those who were not (see *Figure 12*). There was no indication of publication bias across these studies based on Egger's regression intercept [$t(8) = 0.51$, $p_{(one\text{-}tailed)} = .31$], visual inspection of the funnel plot (see *Figure 13*), and the Trim and Fill procedure.

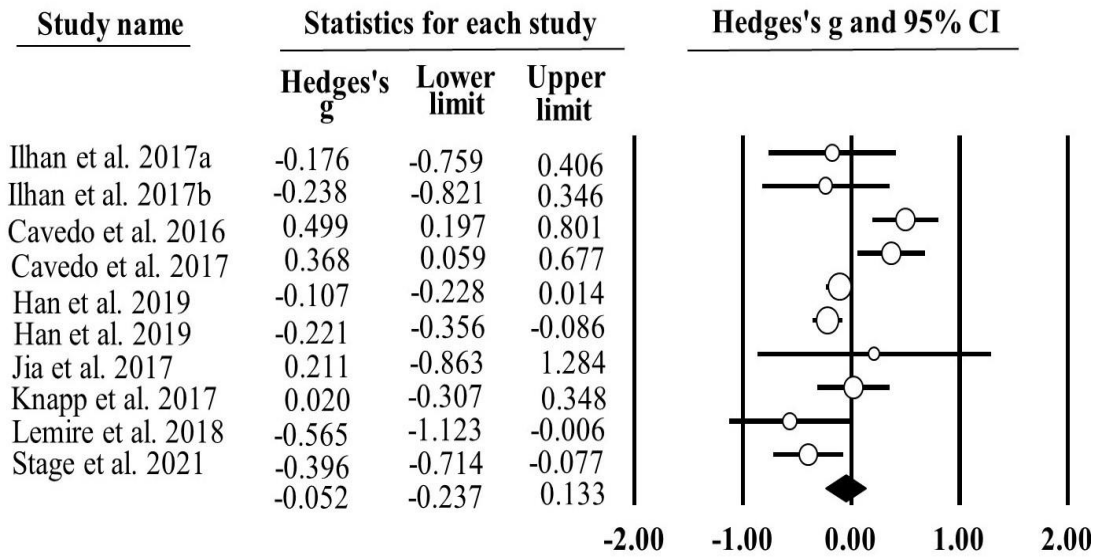

Figure 12. Forest plot for memory impairment medication studies.

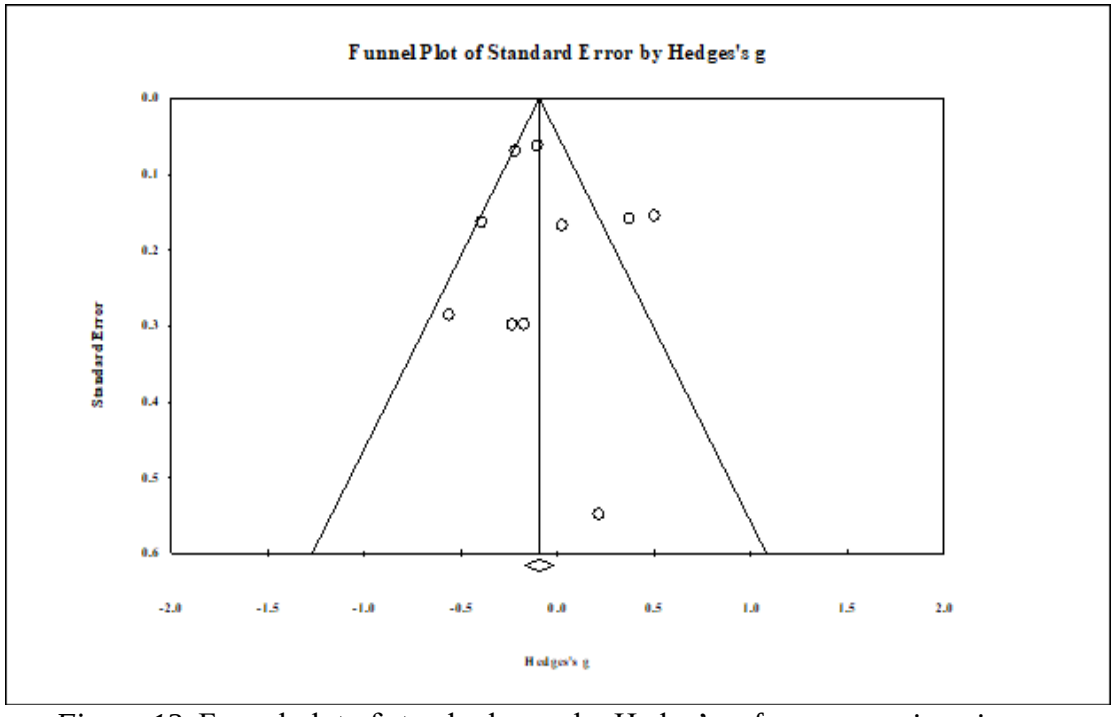

Figure 13. Funnel plot of standard error by Hedge's *g* for memory impairment medication studies

There were insufficient studies to formally examine the moderating effect of types of medication prescribed for memory impairment (cholinesterase inhibitors and memantine) nor to formally examine the impact of medication on measures of cognitive functioning due to the number of studies falling below the recommended *k* of 10 (cholinesterase inhibitors $k = 4$; memantine $k = 3$). Further, due to an insufficient number of studies (i.e., no continuous moderators had valid data for ≥10 studies), meta-regressions were not performed to further evaluate sources of heterogeneity.

4.5.2 Aspirin Findings

Across the nine studies examining aspirin for cardiovascular events only between clinical (n = 30556) and control (n = 37554) samples, 52 effect sizes were extracted for analysis. Meta-analysis resulted in a small, nonsignificant effect size [Hedge's $g = .017$, 95% CI = -0.025 to 0.060, $p = 0.426$] and degree of heterogeneity was small [$Q(8) = 1.202$, $p = 0.997$, $I2 = 0.000$]. Specifically, there was no difference in the number of cardiovascular events between those prescribed aspirin and those who were not (see *Figure 14*). There was no indication of publication bias across these studies based on Egger's regression intercept [$t(7) = 0.75$, $p_{(one\text{-}tailed)} = .24$], visual inspection of the funnel plot (see *Figure 15*), and the Trim and Fill procedure. Further, due to an insufficient number of studies (i.e., no continuous moderators had valid data for ≥10 studies), meta-regressions were not performed to further evaluate sources of heterogeneity.

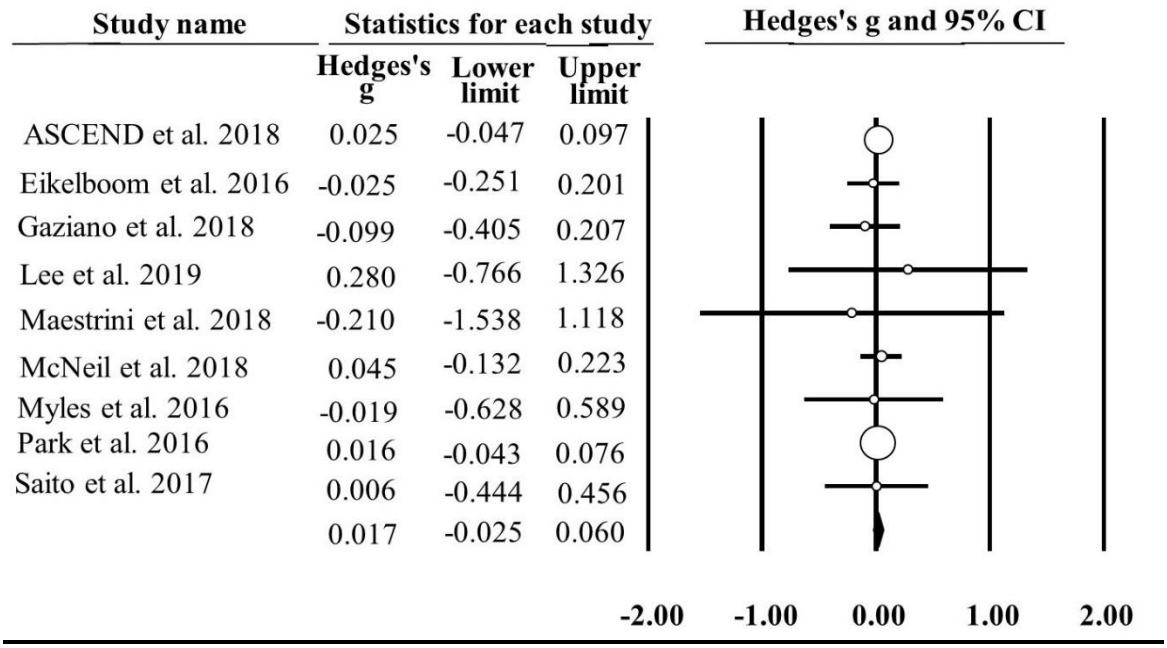

Figure 14. Forest plot for Aspirin studies.

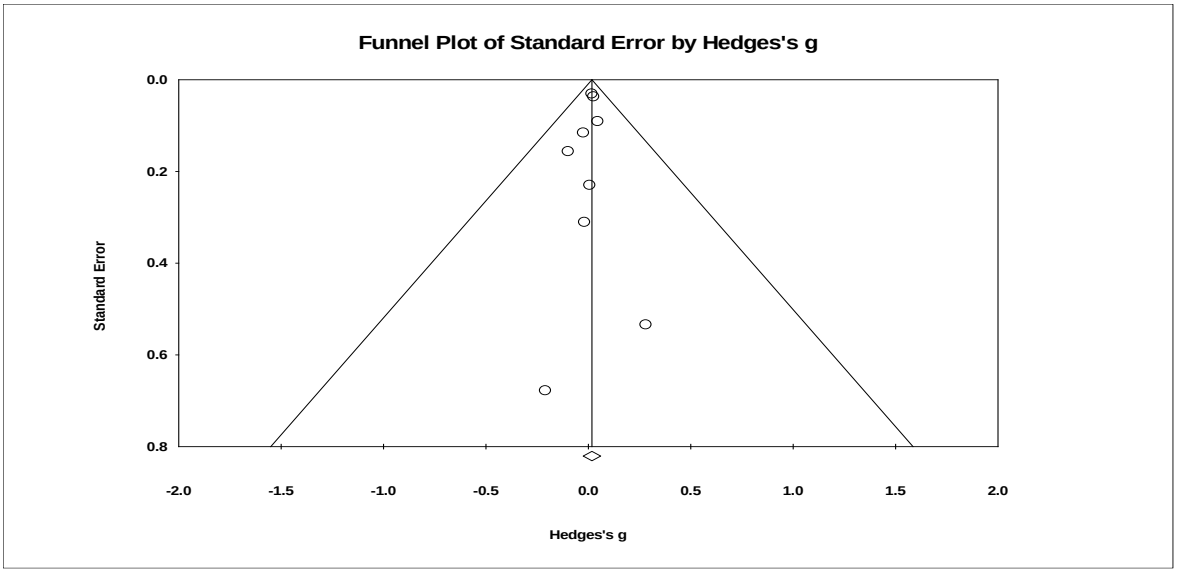

Figure 15. Funnel plot of standard error by Hedge's *g* for Aspirin for cardiovascular

CHAPTER V

DISCUSSION

The past 50 years have brought drastic changes to healthcare and an emphasis on cost management and efficiency. This negatively affected the use of lengthy traditional comprehensive psychological assessments (Piotrowski, 1999; Piotrowski, 2017). Additionally, there has been a general increase in medical literacy due to the internet and expanded access to healthcare (Kimmerle et al., 2015; Sapp et al., 2013). Unfortunately, these advances have been preferentially associated with "medicine" and have culturally resulted in the faulty assumption that medical tests provide more reliable or valid data than psychological assessments (Chambers, 2014). However, in an examination of 144 meta-analyses and authoritative papers, the APA PAWG found that while psychological and medical tests have varying degrees of validity, validity coefficients for many psychological tests were indistinguishable from those of medical tests (Meyer et al., 2001). Specifically, the workgroup concluded that many gold standard psychological tests work just as well as medical tests to detect the same outcome, such as neuropsychological testing detecting dementia on par with MRI (Meyer et al., 2001). This dissertation aimed to replicate the findings of the PAWG with an emphasis on the efficacy of common neuropsychological tests, neuroimaging, and medication studies published in the last five years.

The scope of this meta-analysis was constrained to the three following main study questions: **Q1.** Is there a significant difference in effect sizes between medical tests, medications, and neuropsychological tests? **Q2.** Is there a significant difference in effect sizes within medical interventions and technologies? **Q3.** Is there a significant difference in effect sizes of neuropsychological tests between different clinical groups? This comprehensive meta-analysis of studies sought to address these central study questions with three overarching goals: (1) index the broad effect size of the difference in medical tests, medications, and neuropsychological tests; (2) determine the impact of measured brain regions and type of medication has on effect sizes for neuroimaging, and medication studies, respectively; and (3) determine the impact of clinical group and cognitive domains measured within neuropsychological studies. With the exception of the first goal (i.e., the broad domain comparison), domain-specific studies were meta-analyzed separately to accomplish the other study goals, including finer subgroup analyses.

5.1 Goal One: Are Effect Sizes Different Between Neuropsychological Tests, Medical Tests, and Medications?

The first overarching goal of this study was to index the magnitude of effect sizes across broad-domain findings for neuropsychological tests, neuroimaging, and medication. An a priori directional prediction was made that no significant difference would be observed across the three broad domains. Meta-analysis of all included studies partly confirmed this hypothesis. There was a statistically non-significant difference between the large effect sizes obtained for neuroimaging (Hedges g = -1.670) and

neuropsychological test efficacy [Hedges g = -1.548; $Q(1)$ = 0.213, p = 0.644]. However, there was a significant difference between neuroimaging and medication test efficacy [Hedges g = -0.030; $Q(1)$ = 104.006, p <0.001], as well as neuropsychological tests and medication test efficacy [$Q(1)$ = 73.358, p <0.001]. Stated differently, the diagnostic efficacy of neuroimaging and neuropsychological tests were both substantial and non-significantly different from one another. The treatment efficacy of included medications was non-significant and substantially smaller than the other two broad domains.

5.2 Goal Two: Do Moderating Factors Impact Effect Sizes of Medical Domains?

The second broad goal of this study (**G2**) was to study the possible factors impacting test efficacy across selected medical diagnostics and treatments. For **H2.1**, it was broadly hypothesized that neuroimaging studies would yield larger effect sizes than medication studies. As mentioned in the broad-domain findings, the effect size for broad medications was nonsignificant. Subgroup analysis revealed that both medication for memory impairment (Hedges g = -.052) and aspirin (Hedges g = .017) yielded small, nonsignificant effect sizes. Therefore, hypothesis two was confirmed, such that effect sizes for neuroimaging studies (Hedges g = -1.623) were larger than for medication studies (Hedges g = -.024).

Next, a significant large degree of between-study heterogeneity was observed in neuroimaging studies ($I2$ = 93.30), respectively, reflecting considerable “true” heterogeneity (i.e., true between-study differences), that is beyond what would be expected by sampling error alone. This finding warranted further investigation of heterogeneity using coded continuous and categorical moderators. Meta-regression

analyses indicated that clinical group education level (p = 0.033) trended toward statistical significance, but education overall did not account for much variance (*R2* = .00). Age and sex did not account for any variance in neuroimaging effect size findings. This pattern of meta-regression findings may demonstrate the cognitive reserve effect of education on brain atrophy, at least in the early stages of decline (Nyberg et al., 2021).

Regarding categorical moderators, there were insufficient studies within each domain to conduct formal analyses of measurements of each brain region. However, there were sufficient studies to analyze effect sizes for some specific brain regions measured by neuroimaging. A meta-analysis of 17 studies measuring the hippocampus resulted in a significant large, combined effect size [Hedge's g = -2.458]. Nine studies measuring the temporal lobe resulted in a significant large, combined effect size [Hedge's g = -1.741]. These findings are consistent with hippocampal atrophy as one of the most established and validated core biomarkers of AD and disease progression (Pini et al., 2016), as well as temporal lobe atrophy being one of the earliest changes seen in the brains of those with AD (Visser et al., 1999).

5.3 Goal Three: Do Moderating Factors Impact Effect Sizes of Neuropsychological Tests?

The final goal was to index the magnitude of effect sizes between the two clinical groups within the neuropsychological test studies. An a priori directional prediction was made that no significant difference would be observed between neuropsychological tastes for those with AD versus those with a TBI. Meta-analysis of included studies did not support this hypothesis, such that AD neuropsychological test effect size [Hedges g = -

2.213] was significantly different than the TBI neuropsychological test efficacy [Hedges g = -0.649; $Q(1)$ = 42.821, p = 0.000]. There was a significant large degree of between-study heterogeneity observed in neuropsychological studies ($I2$ = 95.47), respectively, reflecting considerable "true" heterogeneity (i.e., true between-study differences) that is beyond what would be expected by sampling error alone. This finding warranted further investigation of heterogeneity using coded continuous and categorical moderators.

Meta-regression analyses indicated that age accounted for the greatest amount of true heterogeneity ($R2$ = .47), followed by sex ($R2$ = .22); however, neither clinical nor control age or sex were independently significant predictors of the observed variance. These meta-regression models were run separately due to power limitations. This pattern of meta-regression findings may demonstrate an increased risk of AD with age (2022 Alzheimer's disease facts and figures, 2022) that is being captured by neuropsychological tests. While education was not significant at the broad neuropsychological domain level, it was a significant predictor of observed variance for both the AD studies ($R2$ = .15) and TBI studies ($R2$ = .50). However, neither clinical nor control education levels were independently significant predictors of observed variance for neither AD nor TBI studies. Additionally, the cognitive reserve effect of education has been consistently associated with healthier cognitive aging, reduced risk/rate of cognitive decline/dementia, and a proxy for cognitive endowment/brain reserve (Stern, 2009).

There were insufficient studies within each domain regarding categorical moderators to conduct formal analyses of assessed cognitive domains. However, there were sufficient studies to analyze effect sizes for some cognitive domains assessed by

neuropsychological tests. Within the AD clinical group, a meta-analysis of 14 studies measuring executive functioning resulted in a significant large, combined effect size [Hedge's *g* = -2.011], and 10 studies measuring language skills resulted in a significant large, combined effect size [Hedge's *g* = -2.596], and 15 studies measuring verbal learning and memory skills resulted in a significant large, combined effect size [Hedge's *g* = -3.642]. These results are consistent with known neuropsychological profile of patients with AD, with the earliest impairments seen in verbal learning and memory, with impairment in other areas as the disease progresses, including executive functioning and language skills (Weintraub, Wicklund, & Salmon, 2012). A meta-analysis of 18 studies measuring executive functioning within the TBI clinical group resulted in a significant moderate combined effect size [Hedge's *g* = -.685], and 13 studies measuring processing speed resulted in a significant moderate combined effect size [Hedge's *g* = -.645]. These results are consistent with the known neuropsychological profile of patients with TBI, with impairments in executive functioning (i.e., cognitive flexibility, planning, and inhibition), processing speed, and attention (Azouvi et al., 2017; Donders & Levitt, 2012).

5.4 Limitations

While this project had a number of strengths inherent in meta-analyses, it was not without limitations. First, contrary to the PRISMA Checklist (Moher et al., 2009), this study was not formally registered, and, aside from statistical analyses examining publication bias, study quality was not indexed. Incorporating a study quality index could have been utilized in the selection criteria (e.g., excluding studies falling below a certain

threshold on a study quality measure) to strengthen inclusion/exclusion criteria. Additionally, having a study quality index as a moderator could have been used to determine whether effect sizes varied as a function of study quality. However, the current project's methodological strengths of using multiple independent coders, stringent inclusion criteria, and thorough analysis of publication bias mitigated this limitation. Next, while valid reasons were considered for not including unpublished manuscripts/data, doing so may have inflated the risk of publication bias. Nevertheless, while both neuroimaging and neuropsychological test domains yielded significant evidence of publication bias, this was most likely due to a select number of studies with small sample sizes, as analyses failed to identify any studies likely missing due to publication bias.

The relatively small number of medication studies (especially within the medication subgroups) compared to the other two domains and the high heterogeneity seen between studies adversely impacted statistical power. Because of this, formal analysis of moderators for medication studies was only conducted at the broad level (i.e., across all medications) and limited to the continuous variables—there were insufficient medication studies within the two subgroups for categorical variables to be formally analyzed. Beyond medication studies, there were too few studies within each broad domain to run categorical meta-regression on brain region-specific neuroimaging findings or cognitive domain-specific neuropsychological test findings. It is worthy to note that the limited number of studies is not due to sparse literature but instead was limited by the inclusion criteria of research within the last five years. There was an abundance of efficacy studies of aspirin and medications for memory impairment prior to

2016 that were excluded from this meta-analysis to provide the most current update on the existing literature across the three broad domains. Separate from power-limiting moderator analyses, the relatively smaller number of studies in broad domains and clinical subgroups reduces the generalizability/external validity of findings.

Prominent heterogeneity observed across analyses also compounded statistical power. However, as stated above, the problem of insufficient studies (for certain sub-analyses) greatly restricted the ability to explore such heterogeneity further. Using the Cochrane Collaboration guidelines (Higgins et al., 2011) of a minimum of 10 studies for each moderation characteristic modeled, heterogeneity could not be further explored (i.e., with moderation analyses) in several sub-groups/aggregates. Aside from the neuropsychological tests for the TBI subgroup and medication domain, which still yielded moderate heterogeneity, high heterogeneity was generally observed across all analyses. Heterogeneity was especially high in broad domain neuroimaging and neuropsychological studies, likely reflecting considerable methodological differences in these studies. Meta-regressions indicated that age and sex, and to a lesser degree, education accounted for a significant amount of this heterogeneity. Nevertheless, the lack of studies precluded further analysis of sources of heterogeneity. Most broad and sub-analyses resulted in significant combined effect sizes despite these limitations.

5.5 Implications and Future Directions

Notwithstanding the limitations listed above, the results of this dissertation made a significant contribution by demonstrating the similar efficacy of neuropsychological testing to equivalent medical diagnostics and treatments. In the broadest sense, this

comprehensive meta-analysis offered an updated *quantitative consensus* for several questions that the APA PAWG set out to explore in 2001. In addition, this dissertation provided important insights into the efficacy of two of the most frequent patient requests: diagnosis and treatment.

At a broad-domain level and consistent with the most important goal of this study (**G1)**, no significant difference was found between neuropsychological test studies and neuroimaging studies in diagnostic efficacy. Additionally, there was a significant difference between neuropsychological test studies and medication studies, such that neuropsychological studies evidenced a greater, significant effect size. However, it is important to note the relatively smaller number of studies within the medication domain, though this was somewhat moderated by the large number of participants resulting from medication clinical trials. Additionally, there was support for **G2,** as there was also a significant difference between neuroimaging studies and medication studies, such that neuroimaging studies evidenced a greater, significant effect size. This is consistent with Meyer et al. (2001) findings that the correlational effect size of taking aspirin and a subsequent cardiac event was much lower than the effect size for the use of MRI and neuropsychological tests to differentiate between dementia and controls.

These findings not only replicate aspects of the Meyer et al. (2001) landmark meta-analysis indicating no substantial difference in MRI and neuropsychological assessment to diagnose AD but also provide evidence for the efficacy of neuropsychological tests to diagnose TBI on par with neuroimaging. Interestingly, while neuropsychological tests were significant for both clinical groups, there was no support for the third goal **(G3).** There was a significant difference between the effect sizes

between the two clinical groups. Neuropsychological tests to diagnose AD were significantly larger than tests to diagnose TBI. This may, in part, be due to a difference in the level of severity of AD versus TBI at the time of evaluation and/or battery of tests administered.

Though the secondary meta-regression and subgroup analyses were not a central focus of the study, they provided useful information. None of the continuous moderators coded accounted for any significant proportion of variance for neuroimaging studies. This suggests there are likely other moderating factors, such as the reliability of raters when using visual assessment atrophy scales versus measurement software, which may warrant further study. Research has found a significant difference in results based on the type of software (Gronenschild et al., 2012) and independent or group of coders (Harpter et al., 2015; Victoroff et al., 1994) when interpreting brain MRI data.

For broad neuropsychological studies, age and sex accounted for a significant portion of the variance between studies, highlighting the overall lower age of and predominantly male participants of the TBI studies. Further, education accounted for a small to large amount of variability within both AD and TBI clinical groups. These findings are consistent with cognitive reserve literature showing that higher educational attainment is associated with general cognitive health, protection from cognitive dysfunction, and slowed the progression of neurodegenerative cognitive decline (Stern, 2009). Further, when considering categorical moderators (i.e., specific cognitive domains), results were consistent with significant research showing verbal learning and memory difficulties among those with AD (Bäckman et al., 2005; Rabin et al., 2009). However, findings also showed significant difficulties with executive functioning and

language skills, which could represent mixed etiology of dementia, and/or later progression of the disease. Future meta-analyses examining longitudinal neuropsychological studies of AD and/or coding for the length of time since diagnosis should be considered.

Finally, the medication studies yielded a nonsignificant result, both at the broad-domain level and across the two subgroups. While the relatively few studies certainly impact this finding, this is moderated by a large number of participants in the clinical trials. These results suggest that neither aspirin nor memory impairment medications significantly reduce cardiac events or improve cognitive functioning. Daily aspirin has been prescribed to prevent cardiac events or strokes since the 1970s; however, the US Preventative Services Task Force (USPSTF) recently recommended that people ages 60 and older should not start taking a daily aspirin due to the risk of internal bleeding cancels out the benefits of preventing heart problems (USPSTF, 2021). The current findings are in line with the new guidelines and should be considered when determining treatments for cardiac events. Conversely, Cochrane reviews for memory impairment medication indicated that people with AD might experience small benefits in cognitive functioning from donepezil (Birks & Harvey, 2018), demonstrate at least an initial improvement in global cognition to galantamine (Loy & Schneider, 2007), had a slower cognitive decline and ADLs with rivastigmine (Birks et al., 2015), and small clinical benefit in global cognition in moderate to severe AD with memantine (McShane et al., 2019). Of note, all studies used in this dissertation were published after the Cochrane reviews and constitute a new contribution to this literature.

The current findings have both experimental and clinical applicability. While the vast majority of studies in this analysis provided means and standard deviations, resulting in the same effect size (Hedges *g*) across domains, this is often not the case. Research on test efficacy across the two disciplines often utilizes different statistics and is reported differently by the media, leading to consumer confusion. In consideration of effect sizes across fields, effect sizes seen in the social sciences are often very small (Rosnow, 2003), and there is no agreement on what magnitude of effect is necessary to establish practical significance (Ferguson, 2009). Conversely, the numerical size of medical test efficacy effect sizes is often much larger (i.e., >1) than the magnitude of psychological research effect sizes. This leads to difficulties in their interpretation and furthers the doubt of efficacious psychological tests to medical tests among consumers. Future medical and psychological research should consider providing full statistical information (i.e., clinical/control *M*, *SD*, and *n*) such that a common effect size can be generated, and a more direct comparison can be made across domains.

In the case of clinical applicability, this quantitative synthesis of the literature provides clinicians and consumers with convincing evidence that neuropsychological tests are a reliable diagnostic tool for people with both acquired and neurodegenerative brain disorders. There is strong evidence that the public generally overestimates the effectiveness of medical technologies and interventions (Funk et al., 2019). However, Meyer et al. (2001) found that psychological test validity was comparable to the validity of medical tests. This updated meta-analysis drew from recent studies within specific diagnostic domains and successfully replicated findings that the diagnostic efficacy of neuropsychological tests was comparable to the diagnostic efficacy of neuroimaging. In

contrast, effect sizes for well-known and commonly prescribed medications such as aspirin and memory impairments were nonsignificant. The evidence supports the diagnostic efficacy of both neuropsychological tests and neuroimaging, with less evidence to support the treatment efficacy of the included medications.

This meta-analysis lends to the efficacy of neuropsychological tests for consumers and provides a further argument for insurance re-valuation of neuropsychological evaluations. Unfortunately, insurance coverage of comprehensive neuropsychological assessments continues to decline, while most standard medical diagnostic tests, such as MRIs for suspicion of dementia, are covered by insurance. However, research has found that the cost of comprehensive neuropsychological assessments was less than the standard costs of an MRI by approximately $1,500 (Kurniadi & Davis, 2021). Further, research has found that neuropsychological evaluations significantly decrease healthcare spending and utilization (Lichtenstein, Linnea, & Maerlender, 2018; Van Kirk et al., 2013), potentially due to early identification and increased community support and treatment (Barnett et al., 2014; Leibson et al., 2015). This research demonstrates that the lower cost of neuropsychological tests, decreased long-term healthcare cost through early neuropsychological evaluation and diagnosis, and comparable diagnostic efficacy between MRI and neuropsychological tests found in the current study warrants continued advocacy for a re-valuation of insurance coverage and reimbursement of neuropsychological tests. This would reduce the financial burden of consumers undergoing a neuropsychological evaluation and boost the incremental value of neuropsychological tests in an evolving healthcare arena (Glen et al., 2020).

The scope of the current study was broad. Important relations were not examined to retain some semblance of focus and consistency with the original Meyer et al. (2001) article. Particularly the relation between other commonly used medical imaging (e.g., CT scans, fMRI, PET), other common presentations examined by neuropsychology (e.g., ADHD, stroke, movement disorders), and other commonly prescribed medications (e.g., hypertensive medications, statins, thyroid, diabetes, anti-depressants) would benefit from further study. Given limited support for the coded moderators in contributing to the observed heterogeneity between studies, future studies should consider other moderators, such as time since diagnosis, comorbid diagnoses, and race and ethnicity. Finally, future meta-analyses examining the comparable efficacy of neuropsychology and medical diagnostic tools and treatments should consider analyzing longitudinal studies, as the cross-sectional design of this meta-analysis limits the generalizability of the long-term efficacy of these healthcare tools.

REFERENCES

Asterisks (*) denote articles included in meta-analyses.

*Allison, S. L., Koscik, R. L., Cary, R. P., Jonaitis, E. M., Rowley, H. A., Chin, N. A., Zetterberg, H., Blennow, K., Carlsson, C. M., Asthana, S., Bendlin, B. B., & Johnson, S. C. (2019). Comparison of different MRI-based morphometric estimates for defining neurodegeneration across the Alzheimer's disease continuum. *NeuroImage. Clinical*, *23*, 101895. https://doi.org/10.1016/j.nicl.2019.101895

*Al-Janahi, E., Ponirakis, G., Al Hamad, H., Vattoth, S., Elsotouhy, A., Petropoulos, I. N., Khan, A., Gad, H., Chandran, M., Sankaranarayanan, A., Ramadan, M., Elorrabi, M., Gadelseed, M., Tosino, R., Gawhale, P. V., Arasn, A., Alobaidi, M., Khan, S., Manikoth, P., Hamdi, Y., … Malik, R. A. (2020). Corneal Nerve and Brain Imaging in Mild Cognitive Impairment and Dementia. *Journal of Alzheimer's disease: JAD*, *77*(4), 1533–1543. https://doi.org/10.3233/JAD-200678

American Psychological Association [APA] (2010). *Publication Manual of the American Psychological Association*, 6th Edn. Washington, DC: American Psychological Association.

American Psychological Association. (2013). *Understanding psychological testing and assessment.* http://www.apa.org/topics/psychological-testing-assessment

Amyot, F., Arciniegas, D. B., Brazaitis, M. P., Curley, K. C., Diaz-Arrastia, R., Gandjbakhche, A., Herscovitch, P., Hinds, S. R., 2nd, Manley, G. T., Pacifico, A., Razumovsky, A., Riley, J., Salzer, W., Shih, R., Smirniotopoulos, J. G., & Stocker, D. (2015). A Review of the Effectiveness of Neuroimaging Modalities for the Detection of Traumatic Brain Injury. *Journal of neurotrauma*, *32*(22), 1693–1721. https://doi.org/10.1089/neu.2013.3306

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall/Pearson Education.

*ASCEND Study Collaborative Group, Bowman, L., Mafham, M., Wallendszus, K., Stevens, W., Buck, G., Barton, J., Murphy, K., Aung, T., Haynes, R., Cox, J., Murawska, A., Young, A., Lay, M., Chen, F., Sammons, E., Waters, E., Adler, A., Bodansky, J., Farmer, A., … Armitage, J. (2018). Effects of Aspirin for Primary Prevention in Persons with Diabetes Mellitus. *The New England journal of medicine*, *379*(16), 1529–1539. https://doi.org/10.1056/NEJMoa1804988

Ashendorf, L., Alosco, M. L., Bing-Canar, H., Chapman, K. R., Martin, B., Chaisson, C. E., Dixon, D., Steinberg, E. G., Tripodis, Y., Kowall, N. W., & Stern, R. A. (2018). Clinical Utility of Select Neuropsychological Assessment Battery Tests in Predicting Functional Abilities in Dementia. *Archives of clinical neuropsychology: the official journal of the National Academy of Neuropsychologists*, *33*(5), 530–540. https://doi.org/10.1093/arclin/acx100

Azouvi, P., Arnould, A., Dromer, E., & Vallat-Azouvi, C. (2017). Neuropsychology of traumatic brain injury: An expert overview. *Revue neurologique*, *173*(7-8), 461–472. https://doi.org/10.1016/j.neurol.2017.07.006

Bäckman, L., Jones, S., Berger, A. K., Laukka, E. J., & Small, B. J. (2005). Cognitive impairment in preclinical Alzheimer's disease: a meta-analysis. *Neuropsychology*, *19*(4), 520–531. https://doi.org/10.1037/0894-4105.19.4.520

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on psychological science: a journal of the Association for Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Balogh, E. P., Miller, B. T., Ball, J. R., Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, & The National Academies of Sciences, Engineering, and Medicine (Eds.). (2015). *Improving Diagnosis in Health Care*. National Academies Press (US).

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*, *14*, 101. https://doi.org/10.1186/1471-2377-14-101

Bartolucci, A. A., Tendera, M., & Howard, G. (2011). Meta-analysis of multiple primary prevention trials of cardiovascular events using aspirin. *The American journal of cardiology*, *107*(12), 1796–1801. https://doi.org/10.1016/j.amjcard.2011.02.325

*Baum, C. M., Wolf, T. J., Wong, A., Chen, C. H., Walker, K., Young, A. C., Carlozzi, N. E., Tulsky, D. S., Heaton, R. K., & Heinemann, A. W. (2017). Validation and clinical utility of the executive function performance test in persons with traumatic brain injury. *Neuropsychological rehabilitation*, *27*(5), 603–617. https://doi.org/10.1080/09602011.2016.1176934

*Beaulieu-Bonneau, S., Fortier-Brochu, É., Ivers, H., & Morin, C. M. (2017). Attention following traumatic brain injury: Neuropsychological and driving simulator data, and association with sleep, sleepiness, and fatigue. *Neuropsychological Rehabilitation*, *27*(2), 216–238. https://doi-org.libproxy.usouthal.edu/10.1080/09602011.2015.1077145

*Benwell, C., Davila-Pérez, P., Fried, P. J., Jones, R. N., Travison, T. G., Santarnecchi, E., Pascual-Leone, A., & Shafi, M. M. (2020). EEG spectral power abnormalities and their relationship with cognitive dysfunction in patients with Alzheimer's disease and type 2 diabetes. *Neurobiology of aging*, *85*, 83–95. https://doi.org/10.1016/j.neurobiolaging.2019.10.004

Berezow, A. (2012). Why psychology isn't science. The L.A. Times. Retrieved from http://articles.latimes.com/2012/jul/13/news/la-olblowback-pscyhology-science-20120713

Berger D. (1999). A brief history of medical diagnosis and the birth of the clinical laboratory. Part 1--Ancient times through the 19th century. *MLO: medical laboratory observer*, *31*(7), 28–40.

Birks J. (2006). Cholinesterase inhibitors for Alzheimer's disease. *The Cochrane database of systematic reviews*, (1), CD005593. https://doi.org/10.1002/14651858.CD005593

Birks JS, Chong LY, Grimley Evans J. Rivastigmine for Alzheimer's disease [published online ahead of print, 2015 Sep 22]. *Cochrane Database Syst Rev*. 2015;9(9):CD001191. doi:10.1002/14651858.CD001191.pub4

Birks, J. S., & Harvey, R. J. (2018). Donepezil for dementia due to Alzheimer's disease. *The Cochrane database of systematic reviews*, *6*(6), CD001190. https://doi.org/10.1002/14651858.CD001190.pub3

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex, UK: John Wiley & Sons.

Bouzat, P., Almeras, L., Manhes, P., Sanders, L., Levrat, A., David, J. S., Cinotti, R., Chabanne, R., Gloaguen, A., Bobbia, X., Thoret, S., Oujamaa, L., Bosson, J. L., Payen, J. F., Asehnoune, K., Pes, P., Lefrant, J. Y., Mirek, S., Albasini, F., Scrimgeour, C., … TBI-TCD Study Investigators (2016). Transcranial Doppler to Predict Neurologic Outcome after Mild to Moderate Traumatic Brain Injury. *Anesthesiology*, *125*(2), 346–354. https://doi.org/10.1097/ALN.0000000000001165

Brodsky, S. L., & Lichtenstein, B. (2020). The Gold Standard and the Pyrite Principle: Toward a Supplemental Frame of Reference. *Frontiers in psychology*, *11*, 562. https://doi.org/10.3389/fpsyg.2020.00562

Bromme, R., & Goldman, S. R. (2014). The public's bounded understanding of science. *Educational Psychologist*, *49*(2), 59-69. DOI: 10.1080/00461520.2014.921572

Brumfiel G. (2009). Science journalism: Supplanting the old media?. *Nature*, *458*(7236), 274–277. https://doi.org/10.1038/458274a

*Bruun, M., Rhodius-Meester, H., Koikkalainen, J., Baroni, M., Gjerum, L., Lemstra, A. W., Barkhof, F., Remes, A. M., Urhemaa, T., Tolonen, A., Rueckert, D., van Gils, M., Frederiksen, K. S., Waldemar, G., Scheltens, P., Mecocci, P., Soininen,

H., Lötjönen, J., Hasselbalch, S. G., & van der Flier, W. M. (2018). Evaluating combinations of diagnostic tests to discriminate different dementia types. *Alzheimer's & dementia (Amsterdam, Netherlands), 10*, 509–518. https://doi.org/10.1016/j.dadm.2018.07.003

Buckley, J. S., & Salpeter, S. R. (2015). A Risk-Benefit Assessment of Dementia Medications: Systematic Review of the Evidence. *Drugs & aging*, *32*(6), 453–467. https://doi.org/10.1007/s40266-015-0266-9

Burches E, Burches M (2020) Efficacy, Effectiveness and Efficiency in the Health Care: The Need for an Agreement to Clarify its Meaning. Int Arch Public Health Community Med 4:035. doi.org/10.23937/2643-4512/1710035

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

*Byom, L., & Turkstra, L. S. (2017). Cognitive task demands and discourse performance after traumatic brain injury. *International journal of language & communication disorders*, *52*(4), 501–513. https://doi.org/10.1111/1460-6984.12289

*Carlozzi, N. E., Goodnight, S., Casaletto, K. B., Goldsmith, A., Heaton, R. K., Wong, A., Baum, C. M., Gershon, R., Heinemann, A. W., & Tulsky, D. S. (2017). Validation of the NIH Toolbox in Individuals with Neurologic Disorders. *Archives of clinical neuropsychology: the official journal of the National Academy of Neuropsychologists*, *32*(5), 555–573. https://doi.org/10.1093/arclin/acx020

*Cavedo, E., Dubois, B., Colliot, O., Lista, S., Croisile, B., Tisserand, G. L., Touchon, J., Bonafe, A., Ousset, P. J., Rouaud, O., Ricolfi, F., Vighetto, A., Pasquier, F., Galluzzi, S., Delmaire, C., Ceccaldi, M., Girard, N., Lehericy, S., Duveau, F., Chupin, M., … Hippocampus Study Group (2016). Reduced Regional Cortical Thickness Rate of Change in Donepezil-Treated Subjects With Suspected Prodromal Alzheimer's Disease. *The Journal of clinical psychiatry*, *77*(12), e1631–e1638. https://doi.org/10.4088/JCP.15m10413

*Cavedo, E., Grothe, M. J., Colliot, O., Lista, S., Chupin, M., Dormont, D., Houot, M., Lehéricy, S., Teipel, S., Dubois, B., & Hampel, H. (2017). Reduced basal forebrain atrophy progression in a randomized Donepezil trial in prodromal Alzheimer's disease. *Scientific reports*, *7*(1), 1-10. https://doi.org/10.1038/s41598-017-09780-3

Chambers, C. (2014). Physics envy: Do "hard" sciences hold the solution to the replication crisis in psychology? The Guardian. Retrieved from http://www.theguardian.com/science/head-quarters/2014/jun/10/physicsenvy-do-hard-sciences-hold-the-solution-to-the-replication-crisis-inpsychology

Chang, C. (2015). Motivated Processing: How People Perceive News Covering Novel or Contradictory Health Research Findings. Science Communication, 37(5), 602-634. doi:10.1177/1075547015597914

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics—Simulation and Computation*, *39*(4), 860-864. DOI: 10.1080/03610911003650383

Chokesuwattanaskul, A., Lertjitbanjong, P., Thongprayoon, C., Bathini, T., Sharma, K., Mao, M. A., Cheungpasitporn, W., & Chokesuwattanaskul, R. (2020). Impact of obstructive sleep apnea on silent cerebral small vessel disease: a systematic review and meta-analysis. *Sleep medicine*, *68*, 80–88. https://doi.org/10.1016/j.sleep.2019.11.1262

Ciesielska, N., Sokołowski, R., Mazur, E., Podhorecka, M., Polak-Szabela, A., & Kędziora-Kornatowska, K. (2016). Is the Montreal Cognitive Assessment (MoCA) test better suited than the Mini-Mental State Examination (MMSE) in mild cognitive impairment (MCI) detection among people aged over 60? Meta-analysis. Czy test Montreal Cognitive Assessment (MoCA) może być skuteczniejszy od powszechnie stosowanego Mini-Mental State Examination (MMSE) w wykrywaniu łagodnych zaburzeń funkcji poznawczych u osób po 60. roku życia? Metaanaliza. *Psychiatria polska*, *50*(5), 1039–1052. https://doi.org/10.12740/PP/45368

*Clark, A. J., Anderson, N. D., Nalder, E., Arshad, S., & Dawson, D. R. (2017). Reliability and construct validity of a revised Baycrest Multiple Errands Test. *Neuropsychological rehabilitation*, *27*(5), 667–684. https://doi.org/10.1080/09602011.2015.1117981

Clay, R. A. (2010, September). The pitfalls of randomized controlled trials. *Monitor on Psychology*, *41*(8). http://www.apa.org/monitor/2010/09/trials

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Cohen J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155–159. https://doi.org/10.1037//0033-2909.112.1.155

Committee on Psychological Testing, Including Validity Testing, for Social Security Administration Disability Determinations, Board on the Health of Select Populations, & Institute of Medicine. (2015). *Psychological Testing in the Service of Disability Determination*. National Academies Press (US).

Cook, N. R., Appel, L. J., & Whelton, P. K. (2016). Sodium Intake and All-Cause Mortality Over 20 Years in the Trials of Hypertension Prevention. *Journal of the American College of Cardiology*, *68*(15), 1609–1617. https://doi.org/10.1016/j.jacc.2016.07.745

Corrigan, P., Kerr, A., & Knudsen, L. (2005). The stigma of mental illness: explanatory models and methods for change. Applied and Preventive Psychology, 11, 179-190. doi:10.1016/j.appsy.2005.07.001

*Custodio, N., Lira, D., Herrera-Perez, E., Montesinos, R., Castro-Suarez, S., Cuenca-Alfaro, J., & Valeriano-Lorenzo, L. (2017). Memory Alteration Test to Detect Amnestic Mild Cognitive Impairment and Early Alzheimer's Dementia in Population with Low Educational Level. *Frontiers in aging neuroscience*, *9*, 278. https://doi.org/10.3389/fnagi.2017.00278

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., & Meester, L. E. (2005). *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media.

*Dhikav, V., Duraisamy, S., Anand, K. S., & Garga, U. C. (2016). Hippocampal volumes among older Indian adults: Comparison with Alzheimer's disease and mild cognitive impairment. *Annals of Indian Academy of Neurology*, *19*(2), 195–200. https://doi.org/10.4103/0972-2327.176863

*Dhikav, V., Duraiswamy, S., & Anand, K. S. (2017). Correlation between hippocampal volumes and medial temporal lobe atrophy in patients with Alzheimer's disease. *Annals of Indian Academy of Neurology*, *20*(1), 29–35. https://doi.org/10.4103/0972-2327.199903

Ding, J., Davis-Plourde, K. L., Sedaghat, S., Tully, P. J., Wang, W., Phillips, C., Pase, M. P., Himali, J. J., Gwen Windham, B., Griswold, M., Gottesman, R., Mosley, T. H., White, L., Guðnason, V., Debette, S., Beiser, A. S., Seshadri, S., Ikram, M. A., Meirelles, O., Tzourio, C., … Launer, L. J. (2020). Antihypertensive medications and risk for incident dementia and Alzheimer's disease: a meta-analysis of individual participant data from prospective cohort studies. *The Lancet. Neurology*, *19*(1), 61–70. https://doi.org/10.1016/S1474-4422(19)30393-X

Dodge, Y., & Commenges, D. (Eds.). (2006). *The Oxford dictionary of statistical terms*. Oxford University Press on Demand.

Donders, J., & Levitt, T. (2012). Criterion validity of the neuropsychological assessment battery after traumatic brain injury. *Archives of clinical neuropsychology: the official journal of the National Academy of Neuropsychologists*, *27*(4), 440–445. https://doi.org/10.1093/arclin/acs043

DuBois, P. H. (1970). *A history of psychological testing.* Boston: Allyn & Bacon.

Duval S., & Tweedie R. (2000a). A nonparametric ‘‘trim and fill’’ method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95,* 89-98.

Duval S., & Tweedie R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics 56*, 455-463.

Egger, M., Davey, S. G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*, 629-634. doi:10.1136/bmj.315.7109.629

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(28), 7900–7905. https://doi.org/10.1073/pnas.1602413113

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological science*, *31*(7), 792–806. https://doi.org/10.1177/0956797620916786

*Eikelboom, J. W., Kearon, C., Guyatt, G., Sessler, D. I., Yusuf, S., Cook, D., Douketis, J., Patel, A., Kurz, A., Allard, R., Jones, P. M., Dennis, R. J., Painter, T. W., Bergese, S. D., Leslie, K., Wijeysundera, D. N., Balasubramanian, K., Duceppe, E., Miller, S., Diedericks, J., … Devereaux, P. J. (2016). Perioperative Aspirin for Prevention of Venous Thromboembolism: The PeriOperative ISchemia Evaluation-2 Trial and a Pooled Analysis of the Randomized Trials. *Anesthesiology*, *125*(6), 1121–1129. https://doi.org/10.1097/ALN.0000000000001352

*Ettenhofer, M. L., Gimbel, S. I., & Cordero, E. (2020). Clinical validation of an optimized multimodal neurocognitive assessment of chronic mild TBI. *Annals of clinical and translational neurology, 7*(4), 507–516. https://doi.org/10.1002/acn3.51020

Ewers, M., Walsh, C., Trojanowski, J. Q., Shaw, L. M., Petersen, R. C., Jack, C. R., Jr, Feldman, H. H., Bokde, A. L., Alexander, G. E., Scheltens, P., Vellas, B., Dubois, B., Weiner, M., Hampel, H., & North American Alzheimer's Disease Neuroimaging Initiative (ADNI) (2012). Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. *Neurobiology of aging*, *33*(7), 1203–1214. https://doi.org/10.1016/j.neurobiolaging.2010.10.019

Ferguson, C. J. (2009). Is psychological research really as good as medical research? Effect size comparisons between psychology and medicine. *Review of General Psychology*, *13*(2), 130-136. doi:10.1037/a0015103

Ferguson C. J. (2015). "Everybody knows psychology is not a real science": Public perceptions of psychology and how we can improve our relationship with policymakers, the scientific community, and the general public. *The American psychologist*, *70*(6), 527–542. https://doi.org/10.1037/a0039405

Fernández-Ballesteros, R. (2002). How should psychological assessment be considered? *American Psychologist, 57*, 138–139.

Fernandez-Duque, D., Evans, J., Christian, C., & Hodges, S. D. (2015). Superfluous neuroscience information makes explanations of psychological phenomena more

appealing. *Journal of cognitive neuroscience*, *27*(5), 926–944. https://doi.org/10.1162/jocn_a_00750

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*, 665-694. doi:10.1348/000711010X502733

Fisher, R. (1925). *Statistical Methods for Research Workers, First Edition.* Edinburgh: Oliver and Boyd.

Fisher, L. J., & Goldney, R. D. (2003). Differences in community mental health literacy in older and younger Australians. *International journal of geriatric psychiatry*, *18*(1), 33–40. https://doi.org/10.1002/gps.769

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Funk, C., Hefferon, M., Kennedy, B., & Johnson, C. (2019). Trust and mistrust in Americans' views of scientific experts. *Pew Research Center*.

Funk, C., Rainie, L., & Page, D. (2015). Public and scientists' views on science and society. *Pew Research Center*, *29*.

Furr, R. M., & Bacharach, V. R. Psychometrics: An Introduction. 2013.

*Gallagher, K., & Azuma, T. (2018). Analysis of Story Recall in Military Veterans With and Without Mild Traumatic Brain Injury: Preliminary Results. *American journal of speech-language pathology*, *27*(1S), 485–494. https://doi.org/10.1044/2017_AJSLP-16-0208

Garb, H. N., Klein, D. F., & Grove, W. M. (2002). Comparison of medical and psychological tests. *The American psychologist*, *57*(2), 137–138. https://doi.org/10.1037//0003-066x.57.2.137

*Gaziano, J. M., Brotons, C., Coppolecchia, R., Cricelli, C., Darius, H., Gorelick, P. B., Howard, G., Pearson, T. A., Rothwell, P. M., Ruilope, L. M., Tendera, M., Tognoni, G., & ARRIVE Executive Committee (2018). Use of aspirin to reduce risk of initial vascular events in patients at moderate risk of cardiovascular disease (ARRIVE): a randomised, double-blind, placebo-controlled trial. *Lancet (London, England)*, *392*(10152), 1036–1046. https://doi.org/10.1016/S0140-6736(18)31924-X

Geisinger, K. F., Bracken, B. A., Carlson, J. F., Hansen, J. I. C., Kuncel, N. R., Reise, S. P., & Rodriguez, M. C. (2013). *APA handbook of testing and assessment in psychology, Vol. 3: Testing and assessment in school psychology and education* (pp. vii-694). American Psychological Association.

*Ghawami, H., Sadeghi, S., Raghibi, M., & Rahimi-Movaghar, V. (2017). Executive functioning of complicated-mild to moderate traumatic brain injury patients with frontal contusions. *Applied neuropsychology. Adult*, *24*(4), 299–307. https://doi.org/10.1080/23279095.2016.1157078

Gibbons, R. J., Thorsteinsson, E. B., & Loi, N. M. (2015). Beliefs and attitudes towards mental illness: an examination of the sex differences in mental health literacy in a community sample. *PeerJ*, *3*, e1004. https://doi.org/10.7717/peerj.1004

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping Doctors and Patients Make Sense of Health

Statistics. *Psychological science in the public interest: a journal of the American Psychological Society*, *8*(2), 53–96. https://doi.org/10.1111/j.1539-6053.2008.00033.x

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102,* 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Glen, T., Hostetter, G., Roebuck-Spencer, T. M., Garmoe, W. S., Scott, J. G., Hilsabeck, R. C., … Espe-Pfeifer, P. (2020). Return on Investment and Value Research in Neuropsychology: A Call to Arms. *Archives of Clinical Neuropsychology*, *35*(5), 459–468. https://doi.org/10.1093/arclin/acaa010

Greene, E., & Cahill, B. S. (2012). Effects of neuroimaging evidence on mock juror decision making. *Behavioral sciences & the law*, *30*(3), 280–296. https://doi.org/10.1002/bsl.1993

Gregory, R. J. (2004). *Psychological testing: History, principles, and applications.* Allyn & Bacon.

Gronenschild, E. H., Habets, P., Jacobs, H. I., Mengelers, R., Rozendaal, N., van Os, J., & Marcelis, M. (2012). The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. *PloS one*, *7*(6), e38234. https://doi.org/10.1371/journal.pone.0038234

*Grossner, E. C., Bernier, R. A., Brenner, E. K., Chiou, K. S., Hong, J., & Hillary, F. G. (2019). Enhanced default mode connectivity predicts metacognitive accuracy in

traumatic brain injury. *Neuropsychology*, *33*(7), 922–933. https://doi-org.libproxy.usouthal.edu/10.1037/neu0000559

Gutting, G. (2012). How reliable are the social science? The New York Times. Retrieved from http://opinionator.blogs.nytimes.com/2012/05/ 17/how-reliable-are-the-social-sciences/

*Han, J. Y., Besser, L. M., Xiong, C., Kukull, W. A., & Morris, J. C. (2019). Cholinesterase Inhibitors May Not Benefit Mild Cognitive Impairment and Mild Alzheimer Disease Dementia. *Alzheimer disease and associated disorders*, *33*(2), 87–94. https://doi.org/10.1097/WAD.0000000000000291

Harper, L., Barkhof, F., Fox, N. C., & Schott, J. M. (2015). Using visual rating to diagnose dementia: a critical evaluation of MRI atrophy scales. *Journal of neurology, neurosurgery, and psychiatry*, *86*(11), 1225–1233. https://doi.org/10.1136/jnnp-2014-310090

Harvey P. D. (2012). Clinical applications of neuropsychological assessment. *Dialogues in clinical neuroscience*, *14*(1), 91–99. https://doi.org/10.31887/DCNS.2012.14.1/pharvey

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9*, 426-445. doi:10.1037/1082-989X.9.4.426

*Henf, J., Grothe, M. J., Brueggen, K., Teipel, S., & Dyrba, M. (2017). Mean diffusivity in cortical gray matter in Alzheimer's disease: The importance of partial volume correction. *NeuroImage. Clinical*, *17*, 579–586. https://doi.org/10.1016/j.nicl.2017.10.005

Higgins, J. P., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savovic, J., Schulz, K. F., Weeks, L., Sterne, J. A., Cochrane Bias Methods Group, & Cochrane Statistical Methods Group (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ (Clinical research ed.)*, *343*, d5928. https://doi.org/10.1136/bmj.d5928

Higgins, J. P. T., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions version 5.1.0*. Retrieved from https://www.handbook.cochrane.org.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539-1558. doi:10.1002/sim.1186

Higgins, J., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analysis. *BMJ, 327*, 557-560.

Hoffmann, T. C., & Del Mar, C. (2015). Patients' expectations of the benefits and harms of treatments, screening, and tests: a systematic review. *JAMA internal medicine*, *175*(2), 274–286. https://doi.org/10.1001/jamainternmed.2014.6016

Holcomb, W. L., Jr, Chaiworapongsa, T., Luke, D. A., & Burgdorf, K. D. (2001). An odd measure of risk: use and misuse of the odds ratio. *Obstetrics and gynecology*, *98*(4), 685–688. https://doi.org/10.1016/s0029-7844(01)01488-0

Hopkins, E. J., Weisberg, D. S., & Taylor, J. (2016). The seductive allure is a reductive allure: People prefer scientific explanations that contain logically irrelevant reductive information. *Cognition*, *155*, 67–76. https://doi.org/10.1016/j.cognition.2016.06.011

Huang, E. C., Pu, C., Chou, Y. J., & Huang, N. (2018). Public Trust in Physicians-Health Care Commodification as a Possible Deteriorating Factor: Cross-sectional

Analysis of 23 Countries. *Inquiry: a journal of medical care organization, provision and financing*, *55*, 46958018759174. https://doi.org/10.1177/0046958018759174

*Huang, S. F., Liu, C. K., Chang, C. C., & Su, C. Y. (2017). Sensitivity and specificity of executive function tests for Alzheimer's disease. *Applied neuropsychology. Adult*, *24*(6), 493–504. https://doi.org/10.1080/23279095.2016.1204301

*Huang, L., Chen, K., Hu, X., & Guo, Q. (2020). Differential Atrophy in the Hippocampal Subfield Volumes in Four Types of Mild Dementia. *Frontiers in neuroscience*, *14*, 699. https://doi.org/10.3389/fnins.2020.00699

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: *Q* statistic or I^2 index? *Psychological Methods, 11*, 193-206. doi:10.1037/1082-989X.11.2.193

Hughes, D., Judge, C., Murphy, R., Loughlin, E., Costello, M., Whiteley, W., Bosch, J., O'Donnell, M. J., & Canavan, M. (2020). Association of Blood Pressure Lowering With Incident Dementia or Cognitive Impairment: A Systematic Review and Meta-analysis. *Jama*, *323*(19), 1934-1944. https://doi.org/10.1001/jama.2020.4249

Hunsley, J. (2002). Psychological testing and psychological assessment: A closer examination. *American Psychologist, 57*, 139– 140.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

* Ilhan Algin, D., Dagli Atalay, S., Ozkan, S., Ozbabalik Adapinar, D., & Ak Sivrioz, I. (2017). Memantine improves semantic memory in patients with amnestic mild

cognitive impairment: A single-photon emission computed tomography study. *The Journal of international medical research*, *45*(6), 2053–2064. https://doi.org/10.1177/0300060517715166

*Ilhan Algin, D., Dagli Atalay, S., ÖZKAN, S., & Ozbabalik Adapinar, D. (2017). The Effect of Memantine on Cognitive Performance with Amnestic Mild Cognitive Impairment. Journal of Neurological Sciences-Turkish, *34*(3).

Jacobs, B., Beems, T., Stulemeijer, M., van Vugt, A. B., van der Vliet, T. M., Borm, G. F., & Vos, P. E. (2010). Outcome prediction in mild traumatic brain injury: age and clinical variables are stronger predictors than CT abnormalities. *Journal of neurotrauma*, *27*(4), 655-668. https://doi.org/10.1089/neu.2009.1059

Janda, L. H., England, K., Lovejoy, D., & Drury, K. (1998). Attitudes toward psychology relative to other disciplines. *Professional Psychology: Research and Practice, 29*(2), 140–143. https://doi.org/10.1037/0735-7028.29.2.140

Jayasekara, H., MacInnis, R. J., Room, R., & English, D. R. (2016). Long-term alcohol consumption and breast, upper aero-digestive tract and colorectal cancer risk: a systematic review and meta-analysis. *Alcohol and alcoholism*, *51*(3), 315-330.

Jensen, J. D. (2008). Scientific uncertainty in news coverage of cancer research: Effects of hedging on scientists' and journalists' credibility. *Human communication research*, *34*(3), 347-369. https://doi.org/10.1111/j.1468-2958.2008.00324.x

*Jia, J., Wei, C., Jia, L., Tang, Y., Liang, J., Zhou, A., Li, F., Shi, L., & Doody, R. S. (2017). Efficacy and Safety of Donepezil in Chinese Patients with Severe Alzheimer's Disease: A Randomized Controlled Trial. *Journal of Alzheimer's disease: JAD*, *56*(4), 1495–1504. https://doi.org/10.3233/JAD-161117

*Jodouin, K. A., O'Connell, M. E., & Morgan, D. G. (2017). RBANS memory percentage retention: No evidence of incremental validity beyond RBANS scores for diagnostic classification of mild cognitive impairment and dementia and for prediction of daily function. *Applied Neuropsychology. Adult, 24*(5), 420–428. https://doi-org.libproxy.usouthal.edu/10.1080/23279095.2016.1189425

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, *23*(5), 524-532. https://doi.org/10.1177/0956797611430953

Jones, G., & Barker, A. (2008). Reference intervals. *The Clinical biochemist. Reviews, 29 Suppl 1*(Suppl 1), S93–S97.

Jorm, A. F. (2012). Mental health literacy: Empowering the community to take action for better mental health. *American psychologist*, *67*(3), 231. https://doi.org/10.1037/a0025957

Keil, F. C., Lockhart, K. L., & Schlegel, E. (2010). A bump on a bump? Emerging intuitions concerning the relative difficulty of the sciences. *Journal of experimental psychology. General*, *139*(1), 1–15. https://doi.org/10.1037/a0018319

Kelley, T. L. (1927). *Interpretation of educational measurements. New York*: Macmillan.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17,* 137–152. https://doi.org/10.1037/a0028086.

Kennedy, B., & Hefferon, M. (2019). What Americans Know about Science: Science Knowledge Levels Remain Strongly Tied to Education; Republicans and Democrats Are about Equally Knowledgeable. *Pew Research Center*.

*Kim, M., Snowden, S., Suvitaival, T., Ali, A., Merkler, D. J., Ahmad, T., Westwood, S., Baird, A., Proitsi, P., Nevado-Holgado, A., Hye, A., Bos, I., Vos, S., Vandenberghe, R., Teunissen, C., Ten Kate, M., Scheltens, P., Gabel, S., Meersmans, K., … Legido-Quigley, C. (2019). Primary fatty amides in plasma associated with brain amyloid burden, hippocampal volume, and memory in the European Medical Information Framework for Alzheimer's Disease biomarker discovery cohort. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *15*(6), 817–827. https://doi-org.libproxy.usouthal.edu/10.1016/j.jalz.2019.03.004

Kimmerle, J., Flemming, D., Feinkohl, I., & Cress, U. (2015). How laypeople understand the tentativeness of medical research news in the media: An experimental study on the perception of information about deep brain stimulation. *Science Communication*, *37*(2), 173-189. https://doi.org/10.1177/1075547014556541

Kline, R. B. (2004). Beyond significance testing: reforming data analysis methods in behavioral research. American Psychological Association, Washington, DC.

* Knapp, M., King, D., Romeo, R., Adams, J., Baldwin, A., Ballard, C., Banerjee, S., Barber, R., Bentham, P., Brown, R. G., Burns, A., Dening, T., Findlay, D., Holmes, C., Johnson, T., Jones, R., Katona, C., Lindesay, J., Macharouthu, A., McKeith, I., … Howard, R. (2017). Cost-effectiveness of donepezil and memantine in moderate to severe Alzheimer's disease (the DOMINO-AD trial). *International journal of geriatric psychiatry*, *32*(12), 1205–1216. https://doi.org/10.1002/gps.4583

*Konstantinou, N., Pettemeridou, E., Seimenis, I., Eracleous, E., Papacostas, S. S., Papanicolaou, A. C., & Constantinidou, F. (2016). Assessing the Relationship between Neurocognitive Performance and Brain Volume in Chronic Moderate-Severe Traumatic Brain Injury. *Frontiers in neurology*, *7*, 29. https://doi.org/10.3389/fneur.2016.00029

*Korthauer, L. E., Salmon, D. P., Festa, E. K., Galasko, D., & Heindel, W. C. (2019). Alzheimer's disease and the processing of uncertainty during choice task performance: Executive dysfunction within the Hick-Hyman law. *Journal of Clinical & Experimental Neuropsychology*, *41*(4), 380–389.

Krogsbøll, L. T., Jørgensen, K. J., Grønhøj Larsen, C., & Gøtzsche, P. C. (2012). General health checks in adults for reducing morbidity and mortality from disease: Cochrane systematic review and meta-analysis. *BMJ (Clinical research ed.)*, *345*, e7191. https://doi.org/10.1136/bmj.e7191

Krousel-Wood, M. A., Chambers, R. B., & Muntner, P. (2006). Clinicians' Guide to Statistics for Medical Practice and Research: Part I. *The Ochsner Journal*, *6*(2), 68.

Kuhn, D., Cheney, R., & Weinstock, M. (2000). The development of epistemological understanding. *Cognitive development*, *15*(3), 309-328.

Kurniadi, N. E., & Davis, J. J. (2021). Comparison shopping: Is neuropsychological evaluation more expensive than neuroimaging?. *The Clinical neuropsychologist*, 1–12. Advance online publication. https://doi.org/10.1080/13854046.2021.1976839

Labarge, A. S., McCaffrey, R. J., & Brown, T. A. (2003). Neuropsychologists' abilities to determine the predictive value of diagnostic tests. *Archives of Clinical Neuropsychology, 18*(2), 165-175.

*Landin-Romero, R., Kumfor, F., Leyton, C. E., Irish, M., Hodges, J. R., & Piguet, O. (2017). Disease-specific patterns of cortical and subcortical degeneration in a longitudinal study of Alzheimer's disease and behavioural-variant frontotemporal dementia. *NeuroImage, 151*, 72–80. https://doi.org/10.1016/j.neuroimage.2016.03.032

Lee D. K. (2016). Alternatives to P value: confidence interval and effect size. *Korean journal of anesthesiology, 69*(6), 555–562. https://doi.org/10.4097/kjae.2016.69.6.555

*Lee, K.Y., Shin, D.I., Her, S.H., Han, S.H., Ahn, Y., Kim, D., Choi, D.J., Kwon, H.M., Gwon, H., Rha, S., Jo, S., Lim, S.C., Myong, J., & Baek, S.H. (2019). Aspirin Has a Neutral Effect in Preventing Future Cardiovascular Events in Vasospastic Angina. *Cardiovascular Prevention and Pharmacotherapy*.

Lei, H., Gao, Q., Liu, S. R., & Xu, J. (2016). The benefit and safety of aspirin for primary prevention of ischemic stroke: a meta-analysis of randomized trials. *Frontiers in Pharmacology, 7*, 440. https://doi.org/10.3389/fphar.2016.00440

Leibson, C. L., Long, K. H., Ransom, J. E., Roberts, R. O., Hass, S. L., Duhig, A. M., Smith, C. Y., Emerson, J. A., Pankratz, V. S., & Petersen, R. C. (2015). Direct medical costs and source of cost differences across the spectrum of cognitive decline: a population-based study. *Alzheimer's & dementia : the journal of the*

Alzheimer's Association, 11(8), 917–932. https://doi.org/10.1016/j.jalz.2015.01.007

*Lemire, P., Brangier, A., Beaudenon, M., Duval, G. T., & Annweiler, C. (2018). Cognitive changes under memantine according to vitamin D status in Alzheimer patients: An exposed/unexposed cohort pilot study. *The Journal of steroid biochemistry and molecular biology, 175*, 151-156. https://doi.org/10.1016/j.jsbmb.2016.12.019

Lesaffre, E., Speybroeck, N., & Berkvens, D. (2007). Bayes and diagnostic testing. *Veterinary parasitology*, *148*(1), 58-61.

Lichtenstein, J. D., Linnea, K. S., & Maerlender, A. C. (2018). Patterns of referral in high school concussion management programs: A pilot study of consultants from different disciplines. *Applied neuropsychology. Child, 7*(4), 334–341. https://doi.org/10.1080/21622965.2017.1340158

Lilienfeld, S. O. (2012). Public skepticism of psychology: Why many people perceive the study of human behavior as unscientific. *American Psychologist, 67*(2), 111–129. https://doi.org/10.1037/a0023963

*Lindsey, H. M., Lalani, S. J., Mietchen, J., Gale, S. D., Wilde, E. A., Faber, J., MacLeod, M. C., Hunter, J. V., Chu, Z. D., Aitken, M. E., Ewing-Cobbs, L., & Levin, H. S. (2020). Acute pediatric traumatic brain injury severity predicts long-term verbal memory performance through suppression by white matter integrity on diffusion tensor imaging. *Brain imaging and behavior*, *14*(5), 1626–1637. https://doi.org/10.1007/s11682-019-00093-9

Loy, C., & Schneider, L. (2006). Galantamine for Alzheimer's disease and mild cognitive impairment. *The Cochrane database of systematic reviews*, (1), CD001747. https://doi.org/10.1002/14651858.CD001747.pub3

Maier-Hein, K. H., Neher, P., Houde, J. C., Côté, M. A., Garyfallidis, E., Zhong, J., ... & Reddick, W. E. (2016). Tractography-based connectomes are dominated by false-positive connections. *biorxiv*, 084137. DOI:10.1101/084137

*Machado, A., Ferreira, D., Grothe, M. J., Eyjolfsdottir, H., Almqvist, P. M., Cavallin, L., Lind, G., Linderoth, B., Seiger, Å., Teipel, S., Wahlberg, L. U., Wahlund, L.-O., Westman, E., & Eriksdotter, M. (2020). The cholinergic system in subtypes of Alzheimer's disease: an in vivo longitudinal MRI study. *Alzheimer's Research & Therapy*, *12*(1), 1–11. https://doi-org.libproxy.usouthal.edu/10.1186/s13195-020-00620-7

Mackintosh, N., & Mackintosh, N. J. (2011). *IQ and human intelligence*. Oxford University Press.

* Maestrini, I., Altieri, M., Di Clemente, L., Vicenzini, E., Pantano, P., Raz, E., Silvestrini, M., Provinciali, L., Paolino, I., Marini, C., Di Giuseppe, M., Russo, T., Federico, F., Coppola, C., Prontera, M. P., Mezzapesa, D. M., Lucivero, V., Parnetti, L., Sarchielli, P., Peducci, M., … Di Piero, V. (2018). Longitudinal Study on Low-Dose Aspirin versus Placebo Administration in Silent Brain Infarcts: The Silence Study. *Stroke research and treatment*, *2018*, 7532403. https://doi.org/10.1155/2018/7532403

*Maass, A., Landau, S., Baker, S. L., Horng, A., Lockhart, S. N., La Joie, R., Rabinovici, G. D., Jagust, W. J., & Alzheimer's Disease Neuroimaging Initiative (2017).

Comparison of multiple tau-PET measures as biomarkers in aging and Alzheimer's disease. *NeuroImage, 157*, 448–463. https://doi.org/10.1016/j.neuroimage.2017.05.058

Matchar, D. B. (2012). Introduction to the Methods Guide for Medical Test Reviews. In S. M. Chang (Eds.) et. al., *Methods Guide for Medical Test Reviews*. Agency for Healthcare Research and Quality (US).

McLeod, S. A. (2019, May 20). *What a p-value tells you about statistical significance*. Simply Psychology. www.simplypsychology.org/p-value.html

* McNeil, J. J., Nelson, M. R., Woods, R. L., Lockery, J. E., Wolfe, R., Reid, C. M., Kirpach, B., Shah, R. C., Ives, D. G., Storey, E., Ryan, J., Tonkin, A. M., Newman, A. B., Williamson, J. D., Margolis, K. L., Ernst, M. E., Abhayaratna, W. P., Stocks, N., Fitzgerald, S. M., Orchard, S. G., … ASPREE Investigator Group (2018). Effect of Aspirin on All-Cause Mortality in the Healthy Elderly. *The New England journal of medicine*, *379*(16), 1519–1528. https://doi.org/10.1056/NEJMoa1803955

McPhetres, J., & Pennycook, G. (2020, April 2). Lay people are unimpressed by the effect sizes typically reported in psychological science. https://doi.org/10.31234/osf.io/qu9hn

McShane, R., Westby, M. J., Roberts, E., Minakaran, N., Schneider, L., Farrimond, L. E., Maayan, N., Ware, J., & Debarros, J. (2019). Memantine for dementia. *The Cochrane database of systematic reviews*, *3*(3), CD003154. https://doi.org/10.1002/14651858.CD003154.pub6

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological bulletin*, *52*(3), 194–216. https://doi.org/10.1037/h0048070

Merck Manuals. (2010). *Common Medical Tests*. Merck Manuals Consumer Version. https://www.merckmanuals.com/home/resources/common-medical-tests/common-medical-tests

Meldrum M. L. (2000). A brief history of the randomized controlled trial. From oranges and lemons to the gold standard. *Hematology/oncology clinics of North America*, *14*(4), 745–vii. https://doi.org/10.1016/s0889-8588(05)70309-9

*Merritt, V. C., Clark, A. L., Crocker, L. D., Sorg, S. F., Werhane, M. L., Bondi, M. W., Schiehser, D. M., & Delano-Wood, L. (2018). Repetitive mild traumatic brain injury in military veterans is associated with increased neuropsychological intra-individual variability. *Neuropsychologia*, *119*, 340–348. https://doi-org.libproxy.usouthal.edu/10.1016/j.neuropsychologia.2018.08.026

*Messinis, L., Nasios, G., Mougias, A., Politis, A., Zampakis, P., Tsiamaki, E., Malefaki, S., Gourzis, P., & Papathanasopoulos, P. (2016). Age and education adjusted normative data and discriminative validity for Rey's Auditory Verbal Learning Test in the elderly Greek population. *Journal of clinical and experimental neuropsychology*, *38*(1), 23–39. https://doi.org/10.1080/13803395.2015.1085496

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment. A review of evidence and issues. *The American psychologist*, *56*(2), 128–165.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2002). Amplifying issues related to psychological testing and assessment. *The American psychologist*, *57*(2), 140–141.

Meyer, S. B., Ward, P. R., & Jiwa, M. (2012). Does prognosis and socioeconomic status impact on trust in physicians? Interviews with patients with coronary disease in South Australia. *BMJ open*, *2*(5), e001389. https://doi.org/10.1136/bmjopen-2012-001389

Mills, K. T., Chen, J., Yang, W., Appel, L. J., Kusek, J. W., Alper, A., Delafontaine, P., Keane, M. G., Mohler, E., Ojo, A., Rahman, M., Ricardo, A. C., Soliman, E. Z., Steigerwalt, S., Townsend, R., He, J., & Chronic Renal Insufficiency Cohort (CRIC) Study Investigators (2016). Sodium Excretion and the Risk of Cardiovascular Disease in Patients With Chronic Kidney Disease. *JAMA*, *315*(20), 2200–2210. https://doi.org/10.1001/jama.2016.4447

Mitchell, J. (2014). *On the evidentiary emptiness of failed replications*. Working paper. http://jasonmitchell. fas. harvard. edu/Papers/Mitchell_failed_science_2014. pdf Letöltve: 2018.01. 22.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine*, *6*, e1000097. doi:10.1371/journal.pmed1000097

*Müller, S., Herde, L., Preische, O., Zeller, A., Heymann, P., Robens, S., Elbing, U., & Laske, C. (2019). Diagnostic value of digital clock drawing test in comparison with CERAD neuropsychological battery total score for discrimination of patients

in the early course of Alzheimer's disease from healthy individuals. *Scientific reports*, *9*(1), 3543. https://doi.org/10.1038/s41598-019-40010-0

Munro, G. D., & Munro, C. A. (2015). Soft vs. Hard Psychological Science in the Courtroom. *Jury Expert*, *27*, 26.

* Myles, P. S., Smith, J. A., Forbes, A., Silbert, B., Jayarajah, M., Painter, T., Cooper, D. J., Marasco, S., McNeil, J., Bussières, J. S., Wallace, S., & ATACAS Investigators of the ANZCA Clinical Trials Network (2016). Stopping vs. Continuing Aspirin before Coronary Artery Surgery. *The New England journal of medicine*, *374*(8), 728–737. https://doi.org/10.1056/NEJMoa1507688

Neuner-Jehle, S., Senn, O., Wegwarth, O., Rosemann, T., & Steurer, J. (2011). How do family physicians communicate about cardiovascular risk? Frequencies and determinants of different communication formats. *BMC family practice*, *12*, 15. https://doi.org/10.1186/1471-2296-12-15

*Nicolas, B., Alessandra, D., Daniela, P., Osman, R., Sara, T., Giovanni B, F., Valentina, G., & Alzheimer's Disease Neuroimaging Initiative (2020). Basal forebrain metabolism in Alzheimer's disease continuum: relationship with education. *Neurobiology of aging*, *87*, 70–77. https://doi.org/10.1016/j.neurobiolaging.2019.11.013

*Nielsen, T. R., Segers, K., Vanderaspoilden, V., Beinhoff, U., Minthon, L., Pissiota, A., Bekkhus-Wetterberg, P., Bjørkløf, G. H., Tsolaki, M., Gkioka, M., & Waldemar, G. (2019a). Validation of a brief Multicultural Cognitive Examination (MCE) for evaluation of dementia. *International journal of geriatric psychiatry*, *34*(7), 982–989. https://doi.org/10.1002/gps.5099

*Nielsen, T. R., Segers, K., Vanderaspoilden, V., Beinhoff, U., Minthon, L., Pissiota, A., Bekkhus-Wetterberg, P., Bjørkløf, G. H., Tsolaki, M., Gkioka, M., & Waldemar, G. (2019b). Validation of a European Cross-Cultural Neuropsychological Test Battery (CNTB) for evaluation of dementia. *International journal of geriatric psychiatry*, *34*(1), 144–152. https://doi.org/10.1002/gps.5002

*Niemantsverdriet, E., Ribbens, A., Bastin, C., Benoit, F., Bergmans, B., Bier, J. C., Bladt, R., Claes, L., De Deyn, P. P., Deryck, O., Hanseeuw, B., Ivanoiu, A., Lemper, J. C., Mormont, E., Picard, G., Salmon, E., Segers, K., Sieben, A., Smeets, D., Struyfs, H., … Engelborghs, S. (2018). A Retrospective Belgian Multi-Center MRI Biomarker Study in Alzheimer's Disease (REMEMBER). *Journal of Alzheimer's disease: JAD*, *63*(4), 1509–1522. https://doi.org/10.3233/JAD-171140

Nikolas, M. A., Marshall, P., & Hoelzle, J. B. (2019). The role of neurocognitive tests in the assessment of adult attention-deficit/hyperactivity disorder. *Psychological assessment*, *31*(5), 685–698. https://doi.org/10.1037/pas0000688

Nyberg, L., Magnussen, F., Lundquist, A., Baaré, W., Bartrés-Faz, D., Bertram, L., Boraxbekk, C. J., Brandmaier, A. M., Drevon, C. A., Ebmeier, K., Ghisletta, P., Henson, R. N., Junqué, C., Kievit, R., Kleemeyer, M., Knights, E., Kühn, S., Lindenberger, U., Penninx, B., Pudas, S., … Fjell, A. M. (2021). Educational attainment does not influence brain aging. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(18), e2101644118. https://doi.org/10.1073/pnas.2101644118

Nyhan, B. (2014). To get more out of science, show the rejected research. New York Times. Retrieved from http://www.nytimes.com/2014/09/19/ upshot/to-get-more-out-of-science-show-the-rejected-research.html?abt 0002&abg0

Orrison, W. W., Gentry, L. R., Stimac, G. K., Tarrel, R. M., Espinosa, M. C., & Cobb, L. C. (1994). Blinded comparison of cranial CT and MR in closed head injury evaluation. *American journal of neuroradiology*, *15*(2), 351-356.

*Ottoy, J., Niemantsverdriet, E., Verhaeghe, J., De Roeck, E., Struyfs, H., Somers, C., Wyffels, L., Ceyssens, S., Van Mossevelde, S., Van den Bossche, T., Van Broeckhoven, C., Ribbens, A., Bjerke, M., Stroobants, S., Engelborghs, S., & Staelens, S. (2019). Association of short-term cognitive decline and MCI-to-AD dementia conversion with CSF, MRI, amyloid- and ^{18}F-FDG-PET imaging. *NeuroImage. Clinical*, *22*, 101771. https://doi.org/10.1016/j.nicl.2019.101771

*Owens, J. A., Spitz, G., Ponsford, J. L., Dymowski, A. R., Ferris, N., & Willmott, C. (2016). White matter integrity of the medial forebrain bundle and attention and working memory deficits following traumatic brain injury. *Brain and Behavior*, *7*(2), e00608. https://doi-org.libproxy.usouthal.edu/10.1002/brb3.608

Owora, A. H. (2018). Diagnostic validity and clinical utility of HbA1c tests for type 2 diabetes mellitus. *Current diabetes reviews*, *14*(2), 196.

* Park, J. M., Kang, K., Cho, Y. J., Hong, K. S., Lee, K. B., Park, T. H., Lee, S. J., Ko, Y., Han, M. K., Lee, J., Cha, J. K., Kim, D. H., Kim, D. E., Kim, J. T., Choi, J. C., Yu, K. H., Lee, B. C., Lee, J. S., Lee, J., Gorelick, P. B., … CRCS-5 Investigators (2016). Comparative Effectiveness of Prestroke Aspirin on Stroke

Severity and Outcome. *Annals of neurology*, *79*(4), 560–568. https://doi.org/10.1002/ana.24602

*Park, J. E., Park, B., Kim, S. J., Kim, H. S., Choi, C. G., Jung, S. C., Oh, J. Y., Lee, J. H., Roh, J. H., Shim, W. H., & Alzheimer's Disease Neuroimaging Initiative (ADNI) (2017). Improved Diagnostic Accuracy of Alzheimer's Disease by Combining Regional Cortical Thickness and Default Mode Network Functional Connectivity: Validated in the Alzheimer's Disease Neuroimaging Initiative Set. *Korean journal of radiology*, *18*(6), 983–991. https://doi.org/10.3348/kjr.2017.18.6.983

Patel, K., Caddy, C., & Tracy, D. K. (2018). Who do they think we are? Public perceptions of psychiatrists and psychologists. *Advances in Mental Health*, *16*(1), 65-76. DOI: 10.1080/18387357.2017.1404433

Pauli-Pott, U., & Becker, K. (2011). Neuropsychological basic deficits in preschoolers at risk for ADHD: a meta-analysis. *Clinical psychology review*, *31*(4), 626–637. https://doi.org/10.1016/j.cpr.2011.02.005

Penn, Schoen and Berland Associates. (2008). *American Psychological Association benchmark study*. New York: Author.

Peters, S. A., Singhateh, Y., Mackay, D., Huxley, R. R., & Woodward, M. (2016). Total cholesterol as a risk factor for coronary heart disease and stroke in women compared with men: A systematic review and meta-analysis. *Atherosclerosis, 248,* 123–131. https://doi.org/10.1016/j.atherosclerosis.2016.03.016

Pini, L., Pievani, M., Bocchetta, M., Altomare, D., Bosco, P., Cavedo, E., Galluzzi, S., Marizzoni, M., & Frisoni, G. B. (2016). Brain atrophy in Alzheimer's Disease and aging. *Ageing research reviews*, *30*, 25–48. https://doi.org/10.1016/j.arr.2016.01.002

Piotrowski C. (1999). Assessment practices in the era of managed care: current status and future directions. *Journal of clinical psychology*, *55*(7), 787–796. https://doi.org/10.1002/(sici)1097-4679(199907)55:7<787::aid-jclp2>3.0.co;2-u

Piotrowski, C. (2017). Neuropsychological testing in professional psychology specialties: Summary findings of 36 studies (1990-2016) in applied settings. *Journal of the Indian Academy of Applied Psychology*, *43*(1), 134.

Rabin, L. A., Paré, N., Saykin, A. J., Brown, M. J., Wishart, H. A., Flashman, L. A., & Santulli, R. B. (2009). Differential memory test sensitivity for diagnosing amnestic mild cognitive impairment and predicting conversion to Alzheimer's disease. *Neuropsychology, development, and cognition. Section B, Aging, neuropsychology and cognition*, *16*(3), 357–376. https://doi.org/10.1080/13825580902825220

Rabin, L. A., Wang, C., Katz, M. J., Derby, C. A., Buschke, H., & Lipton, R. B. (2012). Predicting Alzheimer's disease: neuropsychological tests, self-reports, and informant reports of cognitive difficulties. *Journal of the American Geriatrics Society*, *60*(6), 1128–1134. https://doi.org/10.1111/j.1532-5415.2012.03956.x

Raffle, A. E., & Gray, J. M. (2019). *Screening: evidence and practice*. Oxford University Press, USA.

Raina, P., Santaguida, P., Ismaila, A., Patterson, C., Cowan, D., Levine, M., Booker, L., & Oremus, M. (2008). Effectiveness of cholinesterase inhibitors and memantine for treating dementia: evidence review for a clinical practice guideline. *Annals of internal medicine, 148*(5), 379–397. https://doi.org/10.7326/0003-4819-148-5-200803040-00009

Ranganathan, P., Pramesh, C. S., & Buyse, M. (2015). Common pitfalls in statistical analysis: Clinical versus statistical significance. *Perspectives in clinical research, 6*(3), 169–170. https://doi.org/10.4103/2229-3485.159943

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 295-315).

*Reas, E. T., Hagler, D. J., Jr, White, N. S., Kuperman, J. M., Bartsch, H., Cross, K., Loi, R. Q., Balachandra, A. R., Meloy, M. J., Wierenga, C. E., Galasko, D., Brewer, J. B., Dale, A. M., & McEvoy, L. K. (2017). Sensitivity of restriction spectrum imaging to memory and neuropathology in Alzheimer's disease. *Alzheimer's research & therapy*, *9*(1), 55. https://doi.org/10.1186/s13195-017-0281-7

Reavley, N. J., McCann, T. V., & Jorm, A. F. (2012). Mental health literacy in higher education students. *Early intervention in psychiatry*, *6*(1), 45–52. https://doi.org/10.1111/j.1751-7893.2011.00314.x

*Reeves, S., Williams, V., Costela, F. M., Palumbo, R., Umoren, O., Christopher, M. M., Blacker, D., & Woods, R. L. (2020). Narrative video scene description task discriminates between levels of cognitive impairment in Alzheimer's disease. *Neuropsychology*, *34*(4), 437–446. https://doi.org/10.1037/neu0000621

*Rhodius-Meester, H., van Maurik, I. S., Koikkalainen, J., Tolonen, A., Frederiksen, K. S., Hasselbalch, S. G., Soininen, H., Herukka, S. K., Remes, A. M., Teunissen, C. E., Barkhof, F., Pijnenburg, Y., Scheltens, P., Lötjönen, J., & van der Flier, W. M. (2020). Selection of memory clinic patients for CSF biomarker assessment can be restricted to a quarter of cases by using computerized decision support, without compromising diagnostic accuracy. *PloS one*, *15*(1), e0226784. https://doi.org/10.1371/journal.pone.0226784

Roll-Hansen, N. (2009). *Why the distinction between basic (theoretical) and applied (practical) research is important in the politics of science*. London School of Economics and Political Science, Contingency and Dissent in Science Project.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Newbury Park, CA: Sage.

Rosnow R. L. (2003). Effect sizes for experimenting psychologists. *Canadian journal of experimental psychology = Revue canadienne de psychologie experimentale*, *57*(3), 221–237. https://doi.org/10.1037/h0087427

Russ, T. C., & Morling, J. R. (2012). Cholinesterase inhibitors for mild cognitive impairment. *The Cochrane database of systematic reviews*, *2012*(9), CD009132. https://doi.org/10.1002/14651858.CD009132.pub2

Russell, R., Chung, M., Balk, E. M., Atkinson, S., Giovannucci, E. L., Ip, S., Taylor Mayne, S., Raman, G., Ross, A. C., Trikalinos, T., West, K. P., Jr, & Lau, J. (2009). *Issues and Challenges in Conducting Systematic Reviews to Support Development of Nutrient Reference Values: Workshop Summary: Nutrition Research Series, Vol. 2*. Agency for Healthcare Research and Quality (US).

* Saito, Y., Okada, S., Ogawa, H., Soejima, H., Sakuma, M., Nakayama, M., Doi, N., Jinnouchi, H., Waki, M., Masuda, I., Morimoto, T., & JPAD Trial Investigators (2017). Low-Dose Aspirin for Primary Prevention of Cardiovascular Events in Patients With Type 2 Diabetes Mellitus: 10-Year Follow-Up of a Randomized Controlled Trial. *Circulation*, *135*(7), 659–670. https://doi.org/10.1161/CIRCULATIONAHA.116.025760

Sackett, D. L., Deeks, J. J., & Altman, D. G. (1996). Down with odds ratios!. *BMJ Evidence-Based Medicine*, *1*(6), 164.Salgado, J. F., & Moscoso, S. (2019). Meta-Analysis of the Validity of General Mental Ability for Five Performance Criteria: Hunter and Hunter (1984) Revisited. *Frontiers in psychology*, *10*, 2227. https://doi.org/10.3389/fpsyg.2019.02227

Salvadori, E., Pasi, M., Poggesi, A., Chiti, G., Inzitari, D., & Pantoni, L. (2013). Predictive value of MoCA in the acute phase of stroke on the diagnosis of mid-term cognitive impairment. *Journal of neurology*, *260*(9), 2220–2227. https://doi.org/10.1007/s00415-013-6962-7

Sapp, S. G., Korsching, P. F., Arnot, C., & Wilson, J. J. (2013). Science communication and the rationality of public opinion formation. *Science Communication*, *35*(6), 734-757. doi:10.1177/1075547013480491

Sattler, J. M. (2014). *Foundations of behavioral, social and clinical assessment of children*. Jerome M. Sattler, Publisher, Incorporated.

St Sauver, J. L., Warner, D. O., Yawn, B. P., Jacobson, D. J., McGree, M. E., Pankratz, J. J., Melton, L. J., 3rd, Roger, V. L., Ebbert, J. O., & Rocca, W. A. (2013). Why patients visit their doctors: assessing the most prevalent conditions in a defined

American population. *Mayo Clinic proceedings*, *88*(1), 56–67. https://doi.org/10.1016/j.mayocp.2012.08.020

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*(4), 551-566. https://doi.org/10.1037/a0029487

Schmand, B., Rienstra, A., Tamminga, H., Richard, E., van Gool, W. A., Caan, M. W., & Majoie, C. B. (2014). Responsiveness of magnetic resonance imaging and neuropsychological assessment in memory clinic patients. *Journal of Alzheimer's disease: JAD*, *40*(2), 409–418. https://doi.org/10.3233/JAD-131484

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological bulletin*, *124*(2), 262. https://doi.org/10.1037/0033-2909.124.2.262

Schultz, D., & Shultz, S.E. (2010). Psychology and work today. Upper Saddle River, NJ: Pearson Education.

Science and Technology. (2015). Oxford Reference. Retrieved July 15, 2020, from https://www.oxfordreference.com/page/134

Sedgwick, P. (2014). Relative risks versus odds ratios. *BMJ*, *348*.

Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect size. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 257-277). New York: Russel Sage Foundation.

Shah, T. G., Sutaria, J. M., & Vyas, M. V. (2020). The association between pulmonary hypertension and stroke: a systematic review and meta-analysis. *International journal of cardiology*, *303*, 78. https://doi.org/10.1016/j.ijcard.2019.12.020

Sharp, S. I., Aarsland, D., Day, S., Sønnesyn, H., Alzheimer's Society Vascular Dementia Systematic Review Group, & Ballard, C. (2011). Hypertension is a potential risk factor for vascular dementia: systematic review. *International journal of geriatric psychiatry*, *26*(7), 661–669. https://doi.org/10.1002/gps.2572

Shea, V. (1985). Overview of the assessment process. In C. S. Newmark (Ed.), *Major psychological assessment instruments* (pp. 1-10). Boston: Allyn & Bacon.

Shen, Y. J., Wang, W. A., Huang, F. D., Chen, J., Liu, H. Y., Xia, Y. L., Han, M., & Zhang, L. (2016). The use of MMSE and MoCA in patients with acute ischemic stroke in clinical. *The International journal of neuroscience*, *126*(5), 442–447. https://doi.org/10.3109/00207454.2015.1031749

*Simm, L. A., Jamieson, R. D., Ong, B., Garner, M., & Kinsella, G. J. (2017). Making sense of self in Alzheimer's disease: reflective function and memory. *Aging & mental health*, *21*(5), 501–508. https://doi.org/10.1080/13607863.2015.1120706

Sinatra, G. M., & Hofer, B. K. (2016). Public understanding of science: Policy and educational implications. *Policy Insights from the Behavioral and Brain Sciences*, *3*(2), 245-253. doi:10.1177/2372732216656870

Smith D. A. (2002). Validity and values: monetary and otherwise. *The American psychologist*, *57*(2), 136–137. https://doi.org/10.1037//0003-066x.57.2.136

Smith, D. (2011). Health care consumer's use and trust of health information sources. *Journal of Communication in Healthcare*, *4*(3), 200-210.

Smyth, A., Griffin, M., Yusuf, S., Mann, J. F., Reddan, D., Canavan, M., Newell, J., & O'Donnell, M. (2016). Diet and Major Renal Outcomes: A Prospective Cohort Study. The NIH-AARP Diet and Health Study. *Journal of renal nutrition : the official journal of the Council on Renal Nutrition of the National Kidney Foundation*, *26*(5), 288–298. https://doi.org/10.1053/j.jrn.2016.01.016

Sørensen, K., Pelikan, J. M., Röthlin, F., Ganahl, K., Slonska, Z., Doyle, G., Fullam, J., Kondilis, B., Agrafiotis, D., Uiters, E., Falcon, M., Mensing, M., Tchamov, K., van den Broucke, S., Brand, H., & HLS-EU Consortium (2015). Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). *European journal of public health*, *25*(6), 1053–1058. https://doi.org/10.1093/eurpub/ckv043

*Spaan, P. E. J. (2016). Episodic and semantic memory impairments in (very) early Alzheimer's disease: The diagnostic accuracy of paired-associate learning formats. *Cogent Psychology*, *3*(1), Article 1125076. https://doi.org/10.1080/23311908.2015.1125076

* Stage, E., Svaldi, D., Sokolow, S., Risacher, S. L., Marosi, K., Rotter, J. I., Saykin, A. J., Apostolova, L. G., & Alzheimer's Disease Neuroimaging Initiative (2021). Prescribing cholinesterase inhibitors in mild cognitive impairment-Observations from the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & dementia (New York, N. Y.)*, *7*(1), e12168. https://doi.org/10.1002/trc2.12168

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological bulletin*, *144*(12), 1325–1346. https://doi.org/10.1037/bul0000169

Stern, Y. (2009). Cognitive reserve. *Neuropsychologia, 47*, 2015-2028. doi:10.1016/j.neuropsychologia.2009.03.004

*Stenberg, J., Håberg, A. K., Follestad, T., Olsen, A., Iverson, G. L., Terry, D. P., Karlsen, R. H., Saksvik, S. B., Karaliute, M., Ek, J. A. N., Skandsen, T., & Vik, A. (2020). Cognitive Reserve Moderates Cognitive Outcome After Mild Traumatic Brain Injury. *Archives of Physical Medicine and Rehabilitation, 101*(1), 72–80. https://doi-org.libproxy.usouthal.edu/10.1016/j.apmr.2019.08.477

*Steward, K. A., Novack, T. A., Kennedy, R., Crowe, M., Marson, D. C., & Triebel, K. L. (2017). The Wechsler Test of Adult Reading as a Measure of Premorbid Intelligence Following Traumatic Brain Injury. *Archives of clinical neuropsychology: the official journal of the National Academy of Neuropsychologists*, *32*(1), 98–103. https://doi.org/10.1093/arclin/acw081

*Sudo, F. K., de Souza, A. S., Drummond, C., Assuncao, N., Teldeschi, A., Oliveira, N., Rodrigues, F., Santiago-Bravo, G., Calil, V., Lima, G., Erthal, P., Bernardes, G., Monteiro, M., Tovar-Moll, F., & Mattos, P. (2019). Inter-method and anatomical correlates of episodic memory tests in the Alzheimer's Disease spectrum. *PloS one*, *14*(10), e0223731. https://doi.org/10.1371/journal.pone.0223731

Suhr, D. (2003). Reliability, exploratory & confirmatory factor analysis for the scale of athletic priorities. *Diambil pada tanggal*, *2*, 274-28.

Summers, M. J., & Saunders, N. L. (2012). Neuropsychological measures predict decline to Alzheimer's dementia from mild cognitive impairment. *Neuropsychology*, *26*(4), 498–508. https://doi.org/10.1037/a0028576

*Sun, T., Xie, T., Wang, J., Zhang, L., Tian, Y., Wang, K., Yu, X., & Wang, H. (2020). Decision-Making Under Ambiguity or Risk in Individuals With Alzheimer's Disease and Mild Cognitive Impairment. *Frontiers in psychiatry, 11*, 218. https://doi.org/10.3389/fpsyt.2020.00218

Surawan, J., Areemit, S., Tiamkao, S., Sirithanawuthichai, T., & Saensak, S. (2017). Risk factors associated with post-stroke dementia: a systematic review and meta-analysis. *Neurology international*, *9*(3), 7216. https://doi.org/10.4081/nir.2017.7216

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 435-452).

Suurmond, R., van Rhee, H., & Hak, T. (2017). Introduction, comparison, and validation of Meta-Essentials: A free and simple tool for meta-analysis. *Research synthesis methods*, *8*(4), 537–553. https://doi.org/10.1002/jrsm.1260

Ter Riet, G., Chesley, P., Gross, A. G., Siebeling, L., Muggensturm, P., Heller, N., Umbehr, M., Vollenweider, D., Yu, T., Akl, E. A., Brewster, L., Dekkers, O. M., Mühlhauser, I., Richter, B., Singh, S., Goodman, S., & Puhan, M. A. (2013). All that glitters isn't gold: a survey on acknowledgment of limitations in biomedical studies. *PloS one*, *8*(11), e73623. https://doi.org/10.1371/journal.pone.0073623

Tommasi, M., Ferrara, G., & Saggino, A. (2018). Application of Bayes' Theorem in Valuating Depression Tests Performance. *Frontiers in psychology*, *9*, 1240. https://doi.org/10.3389/fpsyg.2018.01240

Trevethan R. (2017). Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Frontiers in public health, 5,* 307. https://doi.org/10.3389/fpubh.2017.00307

Trikalinos, T. A., Siebert, U., & Lau, J. (2009). Decision-Analytic Modeling to Evaluate Benefits and Harms of Medical Tests—Uses and Limitations. In *Medical Tests-White Paper Series*. Agency for Healthcare Research and Quality (US).

*Tulsky, D. S., Carlozzi, N. E., Holdnack, J., Heaton, R. K., Wong, A., Goldsmith, A., & Heinemann, A. W. (2017). Using the NIH Toolbox Cognition Battery (NIHTB-CB) in individuals with traumatic brain injury. *Rehabilitation Psychology*, *62*(4), 413–424. https://doi-org.libproxy.usouthal.edu/10.1037/rep0000174

*Tuokkola, T., Koikkalainen, J., Parkkola, R., Karrasch, M., Lötjönen, J., & Rinne, J. O. (2016). Visual rating method and tensor-based morphometry in the diagnosis of mild cognitive impairment and Alzheimer's disease: a comparative magnetic resonance imaging study. *Acta radiologica (Stockholm, Sweden: 1987)*, *57*(3), 348–355. https://doi.org/10.1177/0284185115584656

US Preventative Services Task Force (October, 2021). Aspirin Use to Prevent Cardiovascular Disease: Preventive Medication. Retrieved from: https://www.uspreventiveservicestaskforce.org/uspstf/draft-recommendation/aspirin-use-to-prevent-cardiovascular-disease-preventive-medication#:~:text=In%202016%2C%20the%20USPSTF%20recommended,dose%20aspirin%20daily%20for%20at

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Sciences, 35,* 215-247. doi:10.3102/1076998609346961

Vankirk, K. K., Horner, M. D., Turner, T. H., Dismuke, C. E., & Muzzy, W. (2013). CE hospital service utilization is reduced following neuropsychological evaluation in a sample of U.S. veterans. *The Clinical neuropsychologist*, *27*(5), 750–761. https://doi.org/10.1080/13854046.2013.783122

Verghese, A., Charlton, B., Kassirer, J. P., Ramsey, M., & Ioannidis, J. P. (2015). Inadequacies of Physical Examination as a Cause of Medical Errors and Adverse Events: A Collection of Vignettes. *The American journal of medicine*, *128*(12), 1322–4.e3. https://doi.org/10.1016/j.amjmed.2015.06.004

Victoroff, J., Mack, W. J., Grafton, S. T., Schreiber, S. S., & Chui, H. C. (1994). A method to improve interrater reliability of visual inspection of brain MRI scans in dementia. *Neurology*, *44*(12), 2267–2276. https://doi.org/10.1212/wnl.44.12.2267

Visser, P. J., Scheltens, P., Verhey, F. R., Schmand, B., Launer, L. J., Jolles, J., & Jonker, C. (1999). Medial temporal lobe atrophy and memory dysfunction as predictors for dementia in subjects with mild cognitive impairment. *Journal of neurology*, *246*(6), 477–485. https://doi.org/10.1007/s004150050387

Walker, I. (2007). *Effect sizes*. Statistics for Psychology. Retrieved from: https://people.bath.ac.uk/pssiw/stats2/page2/page14/page14.html

*Wammes, J. D., Good, T. J., & Fernandes, M. A. (2017). Autobiographical and episodic memory deficits in mild traumatic brain injury. *Brain and cognition*, *111*, 112–126. https://doi.org/10.1016/j.bandc.2016.11.004

Wang, H. F., Yu, J. T., Tang, S. W., Jiang, T., Tan, C. C., Meng, X. F., Wang, C., Tan, M. S., & Tan, L. (2015). Efficacy and safety of cholinesterase inhibitors and memantine in cognitive impairment in Parkinson's disease, Parkinson's disease dementia, and dementia with Lewy bodies: systematic review with meta-analysis and trial sequential analysis. *Journal of neurology, neurosurgery, and psychiatry*, *86*(2), 135–143. https://doi.org/10.1136/jnnp-2014-307659

Wegwarth, O., & Gigerenzer, G. (2013). Less is more: Overdiagnosis and overtreatment: evaluation of what physicians tell their patients about screening harms. *JAMA internal medicine*, *173*(22), 2086–2087. https://doi.org/10.1001/jamainternmed.2013.10363

Wegwarth, O., Gaissmaier, W., & Gigerenzer, G. (2011). Deceiving numbers: survival rates and their impact on doctors' risk communication. *Medical decision making: an international journal of the Society for Medical Decision Making*, *31*(3), 386–394. https://doi.org/10.1177/0272989X10391469

Wegwarth, O., Wagner, G. G., & Gigerenzer, G. (2017). Can facts trump unconditional trust? Evidence-based information halves the influence of physicians' non-evidence-based cancer screening recommendations. *PloS one*, *12*(8), e0183024. https://doi.org/10.1371/journal.pone.0183024

*Wei, M., Shi, J., Ni, J., Zhang, X., Li, T., Chen, Z., Zhou, M., Zhang, L., Tan, Z., Wang, Y., & Tian, J. (2019). A new age-related cutoff of medial temporal atrophy scale on MRI improving the diagnostic accuracy of neurodegeneration due to Alzheimer's disease in a Chinese population. *BMC geriatrics*, *19*(1), 59. https://doi.org/10.1186/s12877-019-1072-8

Weintraub, S., Wicklund, A. H., & Salmon, D. P. (2012). The neuropsychological profile of Alzheimer disease. *Cold Spring Harbor perspectives in medicine*, *2*(4), a006171. https://doi.org/10.1101/cshperspect.a006171

Wilson, A., & Golonka, S. (2012). The small effect size effect: Why do we put up with small effects? Retrieved from http://psychsciencenotes .blogspot.com/2012/08/the-small-effect-size-effect-why-do-we.html

*Wolk, D. A., Das, S. R., Mueller, S. G., Weiner, M. W., Yushkevich, P. A., & Alzheimer's Disease Neuroimaging Initiative (2017). Medial temporal lobe subregional morphometry using high resolution MRI in Alzheimer's disease. *Neurobiology of aging*, *49*, 204–213. https://doi.org/10.1016/j.neurobiolaging.2016.09.011

Yarkoni T. (2009). Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power-Commentary on Vul et al. (2009). *Perspectives on psychological science: a journal of the Association for Psychological Science*, *4*(3), 294–298. https://doi.org/10.1111/j.1745-6924.2009.01127.x

Yuh, E. L., Mukherjee, P., Lingsma, H. F., Yue, J. K., Ferguson, A. R., Gordon, W. A., Valadka, A. B., Schnyer, D. M., Okonkwo, D. O., Maas, A. I., Manley, G. T., & TRACK-TBI Investigators (2013). Magnetic resonance imaging improves 3-month outcome prediction in mild traumatic brain injury. *Annals of neurology*, *73*(2), 224–235. https://doi.org/10.1002/ana.23783

Zakzanis K. K. (2001). Statistics to tell the truth, the whole truth, and nothing but the truth: formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. *Archives of clinical*

neuropsychology: the official journal of the National Academy of Neuropsychologists, 16(7), 653–667.

*Zdanovskis, N., Platkājis, A., Kostiks, A., Grigorjeva, O., & Karelis, G. (2021). Cerebellar Cortex and Cerebellar White Matter Volume in Normal Cognition, Mild Cognitive Impairment, and Dementia. *Brain sciences, 11*(9), 1134. https://doi.org/10.3390/brainsci11091134

*Zhou, Z., Zhu, R., Shao, W., Zhang, S. J., Wang, L., Ding, X. J., & Peng, D. T. (2021). Changes in Resting-State Functional Connectivity of Cerebellum in Amnestic Mild Cognitive Impairment and Alzheimer's Disease: A Case-Control Study. *Frontiers in systems neuroscience, 15*, 596221. https://doi.org/10.3389/fnsys.2021.596221

2022 Alzheimer's disease facts and figures. (2022). *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 10.1002/alz.12638. Advance online publication. https://doi.org/10.1002/alz.12638

Appendix A: Flow Chart of Study Screening and Selection

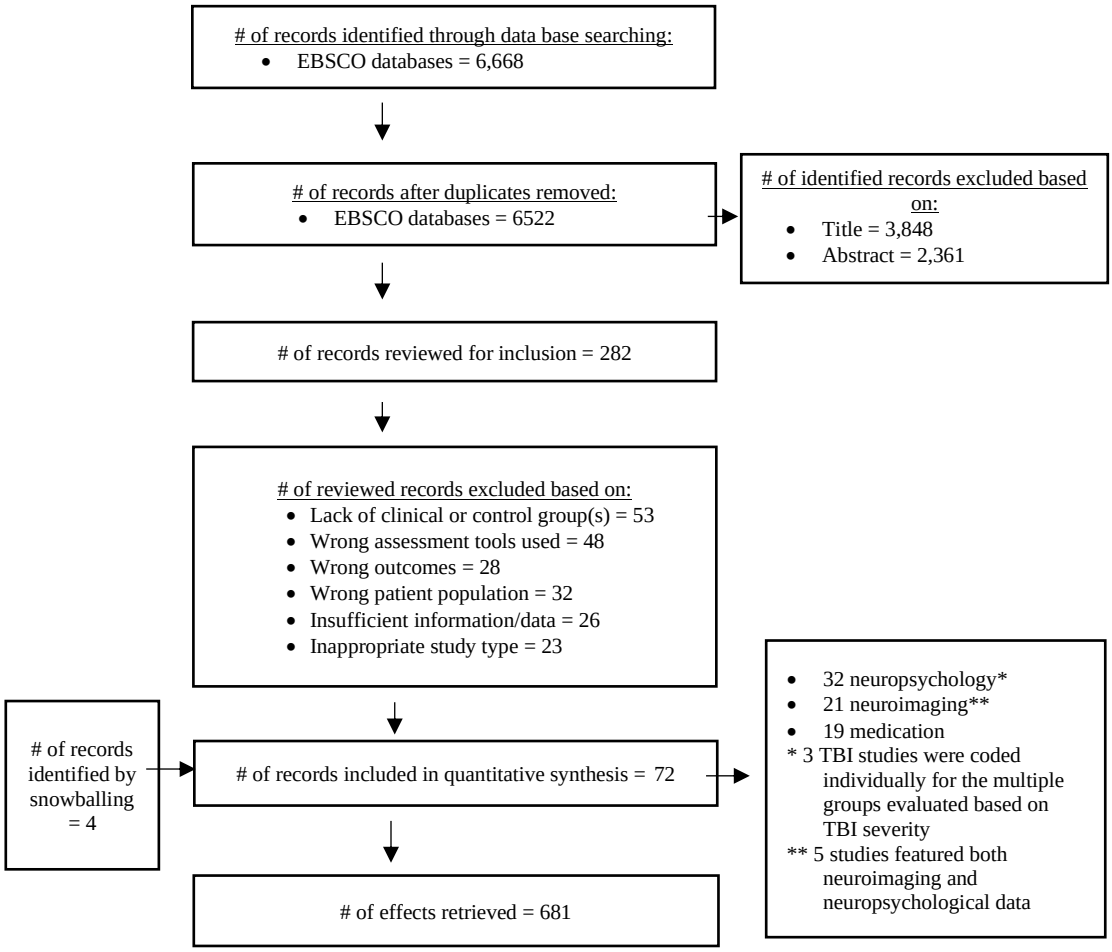

Figure 16. Flow Chart of Study Screening and Selection

BIOGRAPHICAL SKETCH

Murphy N. Harrell, the daughter of Randy Harrell and Clara Murphy, was raised in Greensboro, North Carolina. Currently, she is a fifth-year doctoral student in the Combined-Integrated Clinical and Counseling Psychology PhD program at the University of South Alabama. In the Fall of 2012, she attended Appalachian State University in Boone, North Carolina and received her Bachelor of Science degree in Psychology in May 2012. She then entered graduate school at University of South Carolina Aiken in Aiken, South Carolina in August 2012. She obtained her Master of Science in Applied Clinical Psychology in May 2015. She anticipates graduating with her PhD in Clinical and Counseling Psychology in August of 2022 with a focus in clinical neuropsychology. Upon completion of her doctorate, she expects to have a career as a clinical neuropsychologist where she will have the ability to advance the field of neuropsychology among psychiatric populations.