8-15-2013

Battery Construction, Age Bias, and the Detection of Neurological Impairment: A Survey of North American Clinical Neuropsychologists

Kristina Lynne Talbert
Indiana University of Pennsylvania

BATTERY CONSTRUCTION, AGE BIAS, AND THE DETECTION OF NEUROLOGICAL IMPAIRMENT: A SURVEY OF NORTH AMERICAN CLINICAL NEUROPSYCHOLOGISTS

A Dissertation

Submitted to the School of Graduate Studies and Research

in Partial Fulfillment of the

Requirements for the Degree

Doctor of Psychology

Kristina Lynne Talbert

Indiana University of Pennsylvania

August 2013

Indiana University of Pennsylvania
School of Graduate Studies and Research
Department of Psychology

We hereby approve the dissertation of

Kristina Lynne Talbert

Candidate for the degree of Doctor of Psychology

_____ _*Signature on file*_____
David J. LaPorte, Ph.D.
Professor of Psychology, Advisor

_____ _*Signature on file*_____
Susan Zimny, Ph.D.
Professor of Psychology

_____ _*Signature on file*_____
Beverly Goodwin, Ph.D
Professor of Psychology

ACCEPTED

_*Signature on file*_____ _____
Timothy P. Mack, Ph.D.
Dean
School of Graduate Studies and Research

Title: Battery Construction, Age Bias, and the Detection of Neurological Impairment: A Survey of North American Clinical Neuropsychologists

Author: Kristina Lynne Talbert

Dissertation Chair: Dr. David J. LaPorte

Dissertation Committee Members: Dr. Susan Zimny
Dr. Beverly Goodwin

The present study had three major objectives: (1) to evaluate the similarity of flexible assessment batteries used to evaluate dementia, (2) to investigate a possible age-bias in neuropsychological diagnosis, and (3) to gain insight into the perceived necessity of common clinical information. These objectives were accomplished using a survey mechanism to conduct the experiment. The online survey was completed by 125 INS members who are clinical psychologists currently offering neuropsychological assessment services. Demographic information was collected to assess the generalizability of study results.

After viewing a standard referral request respondents were asked to list the tests commonly used to evaluate a client with subjective memory complaints. Then, each respondent was presented with two clinical vignettes: a reference vignette, which was invariant across respondents, and a test vignette that varied by age (young or old) and test performance (average, borderline, impaired). For each vignette, respondents made two diagnostic ratings (for the presence any impairment and dementia) and associated confidence ratings. Finally, respondents rated the necessity of clinical information. The primary analyses involved a between-subjects, factorial multivariate analysis of variance (MANOVA) to investigate the effects of age and test performance on diagnostic and

confidence ratings. Follow-up univariate ANOVA analyses were also conducted as were post-hoc pairwise comparisons.

This project resulted in several important findings. The first section of this expanded on previous literature through exploration of the process of battery selection. The results of this project indicate that neuropsychologists differ in the specific tests selected for inclusion in an dementia evaluation battery, but tend to assess similar cognitive domains. Another important finding in this study was the presence of an age bias in neuropsychological diagnosis as demonstrated by the differential accuracy of the diagnostic ratings. Finally, the results of this study suggest that a lower threshold is used for some clinical decisions, such as the diagnosis of neurological impairment. Typically, neuropsychological diagnosis relies on a standard impairment classification of two standard deviations below expected performance. However, this study found a tendency to diagnosis neurological impairment and/or dementia at lower threshold (1-1.5 standard deviations below expected performance), especially in older individuals.

ACKNOWLEDGEMENTS

I would like to thank Dr. David J. LaPorte, for his careful criticism, his encouragement, and his resolute dedication to helping me navigate my way through this process. I am also indebted to my (past and present) committee members, Dr. Susan Zimny, Dr. Donald Robertson, and Dr. Beverly Goodwin, whose passion for psychological education and training has inspired me to take my own passions seriously.

I would also like to thank my family—my mother, Kim, my stepfather, David, my two loving siblings, Kayla and Jacob, my grandparents, Alan and Mary, my parents-in-law, Dave and Sylvia, and my brother-in-law, Jonathan—for their love and support during my odyssey in Pennsylvania. From nearby and far away, they have never stopped encouraging me towards excellence.

Finally, I am eternally obliged to Mr. Marcus Pickett, tireless proof-reader and dinner-preparer, whose patience and humor have rescued me from peril more times than I can recall.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The present study has three major objectives: (1) to evaluate the similarity of flexible assessment batteries used to evaluate dementia, (2) to investigate a possible age-bias in neuropsychological diagnosis, and (3) to gain insight into the perceived necessity of common clinical information. These objectives will be accomplished using a survey mechanism to conduct the present experiment. Survey research, by identifying the primary characteristics of the field and its practitioners, has documented the evolution of clinical neuropsychology as a specialty. This research has focused on a broad array of topics, including salaries, payment sources, geographic distribution of clinical practices, and use of assistants and psychometricians (e.g., Guilmette, Faust, Hart, & Arkes, 1990; Putnam, Deluca, & Anderson, 1994), as well as a variety of clinical practices and beliefs (Sweet, Moberg, & Suchy, 2000a).The value of survey research lies in its ability to reveal trends, answer important questions related to professional practice, and identify issues impacting the continued success of the field of clinical neuropsychology (Rabin, Barr, & Burton, 2005).

Study Aims

Assessment was initially the core defining feature of neuropsychology and continues to play an integral role in clinical practice. Recent practice trends reveal that neuropsychologists prefer the use of flexible batteries, yet no studies have investigated the degree of similarity between the flexible batteries used to assess specific types of clients (e.g., head injury, the elderly). Standardized assessment procedures are crucial to the continued advancement of the field and, in previous decades, the wide-spread use of a

few standardized batteries had largely ensured that all neuropsychologists were using consistent assessment techniques. However, with the advent of new neuropsychological tests, clinicians have abandoned standardized batteries in favor of a more flexible approach. Thus, one aim of this study is to investigate the assessment techniques used to assess for the presence of dementia. Individual neuropsychologists will benefit from an awareness of the common tests fellow clinicians use and this information will help ensure that the necessary cognitive areas are being consistently evaluated in everyday clinical practice.

In addition to survey research aimed at discovering *what* clinicians believe and do, there has also been research into *how* clinicians think and make decisions (Garb & Schramke, 1996). Most clinical decision-making research in neuropsychology has focused on the area of diagnosis, including the ability to detect neurological impairment and malingering. There is evidence that neuropsychologists sometimes overdiagnosis neurological impairment. In addition, there is evidence of age bias in the diagnosis of dementia. However, the findings in this area are contradictory, indicating that further clarification is necessary. A second aim of this study, then, is to assess whether there is evidence of an age bias among clinical neuropsychologists in the diagnosis of dementia.

A final aim of this study is to investigate the perceived necessity of clinical information. Although clinical decision-making studies have investigated the validity and reliability of neuropsychological diagnoses, it is unclear whether individual neuropsychologists use the same information when making a diagnosis. The perceived necessity of clinical information is one method that can shed light on the information clinicians feel is relevant to diagnostic decisions. Consistent standards as to the necessity

of clinical information would allow the field to advocate for itself as a unit, establishing and policing its own unique clinical standards.

CHAPTER II

BACKGROUND AND LITERATURE REVIEW

This literature review is divided into three broad sections. First, the development of clinical neuropsychology as a specialty of professional psychology will be briefed summarized. The second section will focus on the findings of survey research in psychological assessment, with an emphasis on surveys within the field of neuropsychology. The final section will begin with a brief overview of research in the general field of clinical judgment followed by a review of clinical judgment research specifically related to neuropsychological assessment and diagnosis.

Brief History of Clinical Neuropsychology

Clinical neuropsychology is an applied science occupied with the study of behavioral expressions of brain dysfunction (Lezak, Howieson, & Loring, 2004).This discipline evolved from experimental findings in neurology about brain-behavior relationships (Long, 1996) and has strong ties to a variety of scientific disciplines, including clinical psychology, neuroscience and medicine (Puente & Marcotte, 2000). Contributing to the growth of clinical neuropsychology as a specialty were parallel developments in cognitive psychology, the systematic analysis of functional impairment due to localized brain lesions, and the development of standardized assessment procedures (Meier, 1992).

Since neuropsychology's inception as a specialty, numerous researchers and clinicians have helped in defining and shaping the discipline into its current form. Among these, psychologists Arthur Benton and Hans-Lukas Teuber and physicians Norman Geschwind and A.R. Luria, are celebrated for their role in helping bridge the gap between

psychology and medicine. Other professionals have been recognized for their specific contributions to the field; for example, Brenda Milner, for describing temporal and frontal lobe functions, Muriel Lezak, for compiling neuropsychological assessment procedures and Edith Kaplan, for her comprehensive analysis of neuropsychological test performance (Meier, 1992). This list is by no means exhaustive; many other clinicians and researchers have provided crucial insights as well. However, these individuals help highlight the interdisciplinary nature of clinical neuropsychology.

The beginning of clinical neuropsychology as a distinct discipline can be traced to the establishment of the International Neuropsychological Society (INS) in 1966. This non-profit organization promotes research, education and service in neuropsychology in addition to facilitating communication among other disciplines in the scientific community that research brain-behavior relationships. During the 1960's and 1970's, before neuropsychology was recognized by the American Psychological Association (APA), the INS provided structure for the field and, by 1980, membership had expanded to over 2,000 professionals. Through the development of the Task Force on Education, Accreditation and Credentialing, INS strove to establish guidelines for the training and credentialing of practicing neuropsychologists. In 1980, APA Division 40, clinical neuropsychology, was established and this task force became a joint effort, publishing its first report in 1981. The aim of this report was to set requirements for competence and establish five training models, ranging from special coursework within a general clinical psychology program to specialized neuropsychology tracks (McCaffrey, Malloy, & Brief, 1985). This task force has continued to set guidelines meant to ensure the quality of neuropsychological services and adherence to professional standards.

In 1976 another professional neuropsychology organization was established, the National Academy of Neuropsychology (NAN), to fill a void left by the scientist-practioner model of INS and Division 40. NAN is more clinically focused; for example, one of its primary activities is to sponsor educational workshops in order to disseminate new neuropsychology techniques and principles (Puente, 1989). Efforts by all three professional organizations have resulted in collaborations with governmental and medical agencies to establish billing codes and reimbursement rates, the creation of honors for excellence in the field, and the organization of conventions and conferences devoted to neuropsychology theory, research and practice. In addition, numerous books and professional journals focused on neuropsychological topics have helped distinguish the field as a unique discipline. These efforts to establish a distinct specialty came to fruition in 1996, when the APA officially recognized clinical neuropsychology as clinical psychology's first specialty.

As the field of clinical neuropsychology has continued to expand, the number of practicing clinicians holding specialty diplomas has increased, reflecting the distinct training and education goals of neuropsychological professional organizations. Specialty diplomas in neuropsychology are awarded by the American Board of Professional Psychology (ABPP) and the American Board of Professional Neuropsychology (ABPN). These diplomas are awarded to professionals who have attained a high level of education, training and experience in neuropsychology and have passed an examination of these competencies (Meier, 1992). As of May 2009, the ABPP had awarded 700 diplomas and the ABPN had awarded 1700. Most practitioners in the field of neuropsychology, including those without specialty certification, posses a background in clinical

psychology with advanced coursework in neuroscience and additional training working with neurologically impaired populations.

An early drive in the field of clinical neuropsychology was to develop instruments that could detect and localize brain lesions. Early research focused on investigating fixed relationships between documented lesions and assessment measures (Benton, 1992). However, with the advent of sophisticated neuroimaging technology, the emphasis of neuropsychological evaluation shifted from localization of brain injury to describing functional impairments. Neuropsychological testing remains a non-invasive way to obtain detailed information regarding functional impairments due to neurological damage (Long, 1996). The continued growth and viability of neuropsychology is dependent on the ability of the field to adapt to changing consumer demands and scientific knowledge. Thus, recent trends in the field have included a focus on rehabilitation and the assessment of cognitive strengths and weakness that can be used to aid individuals in future treatment.

In order to meet future challenges, researchers have suggested that neuropsychologists need to understand their areas of competence and the limitations of their expertise, operate from a foundation of research, preform outcome and efficacy studies to establish the economic value of their services, and develop tests that not only diagnosis impairment but aid in rehabilitation (Heinrichs, 1990; Prigatano & Morrone-Strupinsky, 2010). It is clear that neuropsychology is continuing to develop as a specialty and new directions for the discipline are continuously emerging. The development of neuropsychology has been well-documented through the use of survey research to track

various aspects of the profession. One particular focus of this research has been tracking the development and use of different neuropsychological assessments techniques.

Survey Research in Neuropsychology and Assessment

Psychological Assessment Test Usage Surveys

Surveys have been used to track various aspects of psychological assessment since the 1930's. The first survey of test usage was published in 1935, in the Report of Committee of Clinical Section of APA. This survey noted the plethora of verbal and performance tests and a relative absence of personality measures. The next survey in this area was published a decade later in 1947 and demonstrated marked changes in the usage patterns of practicing psychologists, including the rise of intelligence, reading and vocational measures and the expanded use of projective personality tests (Louttit & Browne, 1947). Similar surveys were published in the decades following these initial investigations of assessment practices. Of note is Sundberg's (1961) survey which demonstrated that from 1935 to 1961 the use of intelligence tests decreased while the use of projective personality tests increased dramatically (Sundberg, 1961). In fact, the Rorschach (Beck, 1944) was the most widely used tests at the time this survey was completed. However, by 1971 the Wechsler Adult Intelligence Scale (WAIS; (Wechsler, 1955) had surpassed the Rorschach as the most frequently used assessment measure (Lubin, Wallis, & Paine, 1971).

Relevant to the current study is that some early test usage surveys in psychological assessment included neuropsychological assessment instruments. These instruments were added to surveys as they were developed; examples (along with the year they first appeared in survey research) include the Halstead Sorting Test (1947), the

Wechsler Memory Scale (1947), the Benton Visual Retention Test (1971) and the Halstead-Reitan Neuropsychological Battery (1971; (Lubin et al., 1971). The presence of these tests in general psychological assessment surveys indicates that they were being widely used and gained popularity before the establishment of clinical neuropsychology as distinct specialty. Beginning in the 1980's the number of test usage surveys increased dramatically and these surveys became more focused on test usage within specific disciplines, populations or settings. Still, the popularity of neuropsychological assessment is evident in these early tests usage surveys; by the mid 1980's the WMS was ranked the 12th most frequently used test in a survey of varied practice settings (Lubin, Larsen, & Matarazzo, 1984).

Surveys of Neuropsychological Test Usage and Practice

Although the surveys previously reviewed often included neuropsychological instruments, they rarely inquired about neuropsychological assessment directly. Beginning in the 1980's, researchers began to investigate the vast growth of clinical neuropsychology. A variety of issues related to the practice of clinical neuropsychology have been addressed through the use of surveys: graduate training programs in neuropsychology (McCaffrey et al., 1985); education and training of neuropsychology instructors (McCaffrey & Isaac, 1984; McCaffrey & Lynch, 1996); use of neuropsychological technicians (DeLuca & Putman, 1993); salary ranges (Putnam et al., 1994); practices and training among ABPP and non-ABPP neuropsychologists (Sweet, Moberg, & Suchy, 2000a); qualifications of neuropsychology internship supervisors (Ryan & Paolo, 1990); ethical beliefs of neuropsychologists (Brown, Gfeller, Ross, & Heise, 1999); use of the HRNB versus use of the Luria-Nebraska Neuropsychological

Battery (Guilmette & Faust, 1991); opinions regarding postconcussion syndrome (McMordie, 1988); neuropsychological test usage in forensic settings (Lees-Haley, Smith, Williams, & Dunn, 1996); characteristics of report-writing and content (Donders, 2001); techniques used to assess effort/malingering (Sharland & Gfeller, 2007); use of ecologically valid measures (Rabin, Burton, & Barr, 2007); and tests used to assess judgment (Rabin, Borgos, & Saykin, 2008). Thus, surveys have been used to track and chronicle most major areas of clinical neuropsychological practice.

Sweet et al. (2000a) investigated major trends and practices in neuropsychology over a ten-year time span using repeated surveys. Their results revealed a number of interesting trends, including the impact of managed care on the practice of neuropsychologists. From 1990 to 2000, more neuropsychologists were employed in private practice and worked more hours each week, but were reimbursed at lower rates. Most respondents to this survey indicated that managed care had been a major contributor to these changes. In addition, a preference for using a flexible battery approach was noted in the responses to this survey. Respondents were asked to characterize their battery approach as one of three options: the standardized battery approach, which was defined as a routine grouping of tests that is uniform across patients (e.g., HRNB, LBNB, Benton); a flexible approach, defined as a battery based on the needs of a individual case that is not uniform across patients; or a flexible battery approach, defined as variable by routine groupings of tests for different types of clients (e.g., head injury, elderly, substance abuse; Sweet et al., 2000a). Seventy percent of neuropsychologists who responded to this survey in 2000 endorsed a preference for the flexible battery approach

whereas 15% of respondents preferred the flexible approach and another 15% preferred to use a standardized battery.

Other recent surveys have investigated test usage patterns among both clinical neuropsychologists and clinical psychologists. In one such survey, 933 APA clinical psychologists (56% response rate) and 566 NAN neuropsychologists (47% response rate) responded to a survey investigating both test usage and assessment practices among practicing clinicians (Camara, Nathan, & Puente, 2000). Not surprisingly, the results revealed that neuropsychologists on average spend much more time per week preforming assessments than do clinical psychologists (10-20 hours per week vs. less than 5). Additionally, neuropsychologists reported using more tests on average than clinical psychologists (17.6 vs. 13.4). The following tests were reported as the top 15 most frequently used by neuropsychologists: the MMPI-2; WAIS-R; WMS-R; Trails; Controlled Oral Word Association Test (COWAT); Finger Tapping; HRNB; Boston Naming Test (BNT); Category Test; Wide Range Achievement Test-Revised/3rd Edition; Beck Depression Inventory (BDI); Rey-Osterrieth Complex Figure Test (ROCFT); Wisconsin Card Sorting Test (WCST); California Verbal Learning Test (CVLT); and the Grooved Pegboard Test.

A survey the following year asked neuropsychologists to list the tests they would use to assess memory, attention and executive functioning for an individual with a mild brain injury (Rabin, 2001). In the area of memory, the following tests were the five most frequently listed: WMS-R/WMS-III; CVLT; ROCFT; BNT; and WAIS-R/WAIS-III. The top five tests for the assessment of attention were: Trail Making Test; WAIS/WMS Digit Span Subtests; Paced Auditory Serial Addition Task; Stroop Test; and Continuous

Performance Test. Finally, in the area of the executive functioning, the following tests were most commonly used: WCST; ROCFT; Halstead Category Test; Trail Making Test; and COWAT. The survey also looked at the instruments used by neuropsychologists to assess a patient's ability to return to work, a common referral question in neuropsychological evaluations. The MMPI-2, WAIS-R/WAIS-III, a driving evaluation, BDI and clinical interview were the most commonly reported techniques used to assess the capacity to return to work; however, the authors noted that these instruments have questionable ecological validity (Rabin, 2001).

Overall, survey research in neuropsychology reveals several interesting trends. It is clear that the WAIS is by far the most frequently used instrument and has been for several decades. Also, neuropsychologists prefer the use of a flexible battery approach, which may help explain the decline of the standardized battery. This preference is likely due to the increased number of reliable neuropsychological assessment instruments available. Another trend revealed is the expanding roles and work settings of neuropsychologists. Most are employed in medical hospitals and private practices; however, an increasing number are employed in rehabilitation facilities (Rabin, 2001). Additionally, there is a great deal of diversity in the training and experience of neuropsychologists. Practicing clinicians, contrary to popular belief, are not engaged solely in assessment activities and also participate in in research, teaching and consultation. However, neuropsychologists' role in teaching and research has declined in recent years due to managed care policies that have resulted in increased case loads and lower reimbursement rates for many clinicians.

Two limitations of previous survey research should be noted. First, many of the surveys reviewed were conducted over a decade ago. Thus, there is a need for updated information regarding current clinical practices of neuropsychologists. Second, a preference for a flexible battery approach was found among neuropsychologists, yet no surveys investigated what tests neuropsychologists routinely use to assess different types of patients. Although surveys have reported the most popular tests used, it is unknown whether clinicians use similar combinations of tests to asses particular types of patients, such as those with a head injury or the elderly. The only survey in which clinicians were asked to select tests they would use based on a simulated client, provided clinicians with four broad areas (i.e., memory, attention, executive functioning, capacity to work) and asked what tests they would use to assess those specific areas. Thus, it is possible that neuropsychologists would have listed different tests designed to tap other cognitive domains had these prompts not been provided. The selection of assessment techniques characterizes the first stage of clinical judgment in neuropsychological assessment, a process that ends with test interpretation and diagnosis. Thus, the next section provides an overview of clinical judgment research with an emphasis on clinical decision-making in neuropsychology.

Clinical Judgment and Prediction

Clinical judgment, in both the psychological and medical literature, commonly refers to any artful or intuitive means used by clinicians in reaching a diagnosis or other decision that is based on previous experiences with patients. In contrast, statistical (or actuarial) refers to the use of any formal quantitative techniques or formulas, such as regression equations, for the same clinical tasks (Garb, 1998). In nearly every area

statistical and clinical predictions have been compared to one another, statistical prediction models have been shown to be more accurate and consistent with relatively few exceptions. As the field of professional psychology has grown, "clinical judgment" as a general term has been defined and redefined. Presently, the term "clinical judgment" refers to a variety of activities, from treatment decisions to outcome analysis, many of which are beyond the scope of this paper. Although clinical judgment is not necessarily synonymous with clinical prediction, the literature reviewed in this section will focus on prediction and diagnosis, and thus, the terms "clinical judgment" and "clinical prediction" will be used interchangeably.

Any discussion of clinical prediction must begin with a definition of what the term "clinical judgment" has come to mean in the empirical psychological literature. In some ways, clinical judgment is defined by what it is not. Professional psychologists claim that they are uniquely suited to making predictions about individuals and that these individualized predictions transcend predictions about people in general. That is, clinical judgment is seen as being the opposite of actuarial prediction, in which predictions are made about individuals based on their membership to a general class, about which much is known (Sawyer, 1966). In its extreme form, the clinical approach is an attempt to understand the inner workings of a particular individual view in isolation, rather than as a member of an aggregate group.

A variety of psychological fields use, in part, the actuarial approach. Industrial/Organization psychologists use training procedures in trying to improve worker productivity and satisfaction. For example, companies have instituted objective testing standards to their hiring procedures to ensure workers are suited for a given position

(Shadish, Cook, & Campbell, 2002). Social psychologists attempt to describe how people in general behave in certain social situations and how behavior can be altered through modification of the situation. In both cases, the individual is seen as representative of the general population being studied; thus, predictions are made about the behavior of a class of people rather than about a specific individual (Dawes, 1996; Hoshmand & Polkinghorn, 1992). These predictions are actuarial, or statistical, in nature. However, in clinical practice, the clinician is faced with an individual client and is asked to make predictions and diagnoses about this particular client. Proponents of the more intuitive clinical approach point out the difficulty in understanding a client in his or her complexity if one relies on statistical formulas (Garb, 1998). The expert clinician relies on knowledge that involves adapting previous learning and applying it to the uniqueness of a particular clinical situation. This process is not simply based on facts or sets of rules, instead it is described as a "dynamic and contextualized understanding that is the result of the interaction of cognitive patterns or meaning gestalts with environmental cues" (Hoshmand & Polkinghorn, 1992).

This intuitive approach is usually justified by referencing work on medical diagnosis and even chess expertise (Dawes, 1996). It is not at all clear, however, that clinical psychologists possess expertise similar to those in the medical field. Certainly there is a limited similarity between clinical prediction and chess playing, in which the player can be judged empirically by the number of games won. Medical diagnosticians use a great deal of explicit knowledge gained through diagnostic tests, and their performance can also be judged based on what they accomplish. In other words, their performance can be judged by how often they make the correct diagnosis when compared

to other medical professionals or mathematical models meant to represent expert judgment. Quantification of performance is often more difficult in the field of mental health. Still, psychologists using the clinical, intuitive approach to prediction and diagnosis can be compared to statistical prediction models like those in medicine. In a number of fields, including medicine, actuarial prediction models have consistently outperformed judges who rely exclusively on clinical prediction (Garb, 1998; Garb & Schramke, 1996).

Statistical vs. Clinical Prediction

The most accurate decision-making model has been long sought by psychological researchers. The debate began over 50 years ago with Meehl's (1954) book *Clinical Versus Statistical Prediction*, which summarized the findings from the existing decision-making literature. The research presented in this book found that in all but 1 of 20 studies, statistical models were more accurate than, or as equally accurate as, the clinician. Meehl concluded, therefore, that clinicians should focus their time on treatment and research and leave diagnostic judgments to statistical models (Meehl, 1954). Holt (1958) criticized Meehl's (1954) conclusion, citing two major issues: the identification and assessment of predictive variables and how they should be integrated.

Holt argued that clinicians, through extensive training, have the unique ability to identify the criterion they are predicting, what variables should be used in prediction, and the strength of the relationship between the predictors and criteria (Ægisdóttir et al., 2006). In addition, it was argued that the assessment of relevant factors was as much qualitative as it was quantitative (Holt, 1958). He also argued that Meehl (1954) unfairly compared statistical models with naïve clinical integration rather than focusing on

sophisticated clinical decision-making, which involves combining both qualitative and quantitative data. This data is gathered in a standardized manner and, furthermore, the data collected have known relationships with what is being predicted. Thus, the clinician is the primary diagnostic instrument and is able to make tailored predictions to each individual client. Holt (1958) presented data suggesting that his "sophisticated clinical approach" was superior to statistical procedures in predicting success in clinical training. Based on these findings, Holt (1958) argued for the combination of clinical and statistical methods (the sophisticated clinical approach) that would be systematic, controlled, and sensitive to individual cases (Aegisdottir et al., 2006). Thus, although Holt was criticizing Meehl's (1954) findings, he was still arguing against the use of "pure" clinical intuition.

Since the 1950's, numerous narrative and meta-analytic reviews of the literature on the differential accuracy of clinical and statistical prediction methods have been published (e.g., (Dawes, Faust, & Meehl, 1989; Garb, 1998; Grove & Meehl, 1996; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Kleinmuntz, 1990). In nearly every case, these review have supported Meehl's (1954) conclusion that statistical methods are more accurate than or are as equally accurate as clinical prediction methods. For example, Grove et al. (2000) found a consistent advantage (d =.12) for statistical prediction over clinical prediction across a variety of mental health predictors and criterion. Studies of clinical decision-making over the last 60 years have sought to both describe the process of clinical judgment and prediction and explain the near universal superiority of statistical models.

Methods of Modeling Clinical Judgment

In general, studies of clinical judgment are helpful to practicing clinicians in that they explicitly describe and prescribe utilization of readily available clinical information. In addition to comparing statistical models of prediction to clinical methods, studies of clinical judgment have attempted to describe the judgment process. Most studies in this area have used two general approaches: formal models, which are comprised of sets of if-then rules (i.e., regression equations), and cognitive models, which include heuristics, biases, and knowledge structures (Garb, 1998). Formal models include process-tracing, in which judges verbalize the steps they take in reaching decisions, and then computer-based models of these steps are created. These models are meant to recreate the judgment process rather than describe the judgment process. Heuristics are simple rules for making decisions and can describe how judgments are made, but are not meant to reproduce them. Biases describe errors in judgment, while knowledge structures include beliefs, theories, and information that is stored in memory that comprises implicit theories used to make judgments (Kahneman, Slovic, & Tversky, 1982).

Cognitive models: Information-processing approach. The information-processing approach is a valuable method for modeling how judges make decisions and can be used to identify the relevant factors that aid in diagnosis. In general, this approach looks at decision-making as occurring in a sequence, and it maintains that judges' verbalizations of what they do are valid (Dowie & Elstein, 1988). Inherent in this approach is the concept of bounded rationality (Newell & Simon, 1972), which states that there are limits on the human capacity for rational thought that are not due to unconscious motives but, rather, reflect limitations of cognitive capacities, such as the limited size of

working and short-term memory. Thus, cognitive heuristics and biases have developed as a way of coping with our innate cognitive limitations, and the simplification of complex situations is a necessary adaptation (Dowie & Elstein, 1988).

Information-processing approaches have been criticized for relying on highly circumscribed conditions in evaluating judgments, which reflects that fact that these methods are time-consuming and, thus, allow only a small sample of the individual's performance to be gathered. Thus, these models may match our own intuitive understanding of the decision-making process—in other words, what people think they do when making a decision—but are most useful when compared with statistical (actuarial) approaches to decision-making (Dowie & Elstein, 1988). However, information-processing studies have helped generate a model of the judgment process, at least the way in which people *think* they arrive at a decision.

Apropos to this paper is early research into medical-problem solving, which is most applicable to the diagnostic decisions asked of neuropsychologists, due to its reliance on explicit knowledge derived from objective test results. Elstein, Shulman and Sprafka (1978), studied the decision-making of internists deciding among three treatments for overactive thyroid. Their analysis revealed a four-step process in making medical decisions. 1. Cue acquisition, 2. Hypothesis generation, 3. Cue interpretation, and 4. Hypothesis evaluation.

In general, clinicians tended to generate a small number of hypothesis (4 or 5) early on in the clinical encounter based on a relatively small amount of information compared to the information that would be collected eventually. They then asked themselves what findings would be observed if a particular hypothesis was true. Thus,

data collection is tailored based on the hypotheses generated and is used to gradually reduce the difference between the clinician's state of knowledge and the knowledge that is needed to reach a particular conclusion. As previously stated, the number of hypotheses generated tends to be small, around 4 or 5, and are usually ideas that "pop" into the clinician's mind based on salient cues or combinations of cues. This relates to the availability heuristic (Tversky & Kahneman, 1974), which, in this case, states that hypotheses based upon vivid cues will seem more probable simply due to the ease with which they come to mind. Cue acquisition, then, begins to take place before hypothesis generation and is subsequently tailored depending on the hypotheses under consideration.

The third step of the decision-making process involves cue-interpretation, where more biases exert their influence, including the availability heuristic (Garb, 1998). In this study (Elstein, Shulman & Sprafka 1978), most clinicians rated cues as either confirmatory, disconfirmatory, or noncontributory. This weighting scheme is roughly equivalent to a regression equation in which only the signs of the coefficients, and not the magnitudes, are important.

Diagnostic accuracy, which is part of the hypothesis evaluation stage, was found to be correlated with both thoroughness of cue acquisition and accuracy of cue interpretation; however, cue acquisition and accuracy of cue interpretation were uncorrelated (Elstein et al., 1978). Therefore, inaccuracy can be due to either incomplete data collection or misinterpretation, but errors in interpretation are not easily ameliorated by more data collection, as would be expected based on the principle of bounded rationality. Simply put, the more data there is, the less likely it will all be used in reaching a decision (Oskamp, 1965). Thus, greater thoroughness alone will not solve problems in

diagnostic accuracy; simplifying strategies are needed to reduce cognitive burden (Garb, 1998). During this stage of decision-making, if a satisfactory hypothesis is found, the process ends and any additional evidence will likely be subjected to confirmatory hypothesis testing, or a biased interpretation of clinical data in favor of an already chosen diagnosis (Dowie & Elstein, 1988; Tversky & Kahneman, 1974). If no hypothesis fits with available data, the process begins again with additional data collection (cue acquisition; Elstein, et al., 1978).

Another commonly found cognitive bias that impacts decision-making is the overemphasis of positive findings, which states that data is remembered better if it fits with a favored hypothesis (Meehl, 1973). In addition to leading to a failure to consider disconfirmatory data, this bias will also lead judges to assign positive weights to noncontributory findings. This phenomenon is facilitated by the fact that data are often related probabilistically to the criterion (diagnosis), reflecting the fact that a symptom can be caused by a variety of underlying conditions (Dowie & Elstein, 1988). Because individual pieces of evidence can be related to a variety of underlying causes, it would appear that the collection of more data would aid in determining the cause of this cluster of findings. Perhaps additional data collection would increase the accuracy of diagnosis, but only if all the additional data collected are relevant, a factor that is not always easy to determine (Dawes, 1996).

Despite the fact that most clinical data are correlated due to a common underlying cause, it is still more efficient to use fewer cues and properly weight them than it is to collect more data, especially in light of limited working memory capacity. In addition, redundant data is sometimes used to erroneously bolster confidence, but more

information does not necessary improve accuracy (Garb, 1998). However, in most clinical settings, the reliability of data sources is low, thus arguing for redundancy to protect against overreliance on unreliable data (Meehl, 1973). This begs the question: Can clinicians discriminate between reliable and unreliable data? In some settings this question may be controversial; however, it is presumed that neuropsychologists are trained to investigate the reliability of assessment techniques prior to using them. Therefore, neuropsychologists should be in the unique position of knowing exactly the reliability for each individual test administered. It remains an open question whether more information improves accuracy in neuropsychological assessment or only increases confidence as in other areas of clinical judgment (Garb, 1998).

Cognitive heuristics, biases, and knowledge structures. The information-processing approach produces a model of clinical judgment that more reflects how judges *think* they arrive at a decision and is useful only to the extent that judges are actually aware of their cognitive processes. Numerous studies have demonstrated that often people are unaware of their cognitive process and, thus, have a limited ability to describe their actual decision-making process (e.g. Nisbett & Wilson, 1977; Smith & Miller, 1978; White, 1980). In addition, cognitive heuristics and biases can affect the decision-making process at every stage.

Examples of heuristics include the following: The *representativeness* heuristic describes judgments that are made based on how similar an object or person is to a category or class. For example, if making a diagnosis of anxiety, a clinician may compare the client to what is characterized as the typical client with an anxiety disorder. The *availability* heuristic describes judgments that are impacted by the ease with which

objects and events can be recalled from memory. For example, a clinician may be more likely to make a diagnosis of bipolar disorder than schizophrenia if the patients with bipolar disorder are recalled more easily. The *anchoring-and-adjustment* heuristic describes judgments that vary as a function of the order in which information was presented. Finally, the *past behavior* heuristic describes predictions of future behavior that are based upon knowledge of previous behavior (Tversky & Kahneman, 1974). The classic work on these heuristics has been extensively reevaluated and expanded since the 1970's and the relationship between cognitive heuristics and clinical judgment will be discussed below.

Biases include the confirmatory bias, the hindsight bias, the misestimation of covariance, and ignoring base rates or norms (Garb, 1998). Confirmatory bias occurs when judges seek or recall only that information that confirms their hypothesis. In addition, the confirmatory bias can lead judges to ignore information that does not support their hypothesis, interpret ambiguous information as supporting their hypothesis, or fail to consider the possibility that information may support an alternative hypothesis. Hindsight bias is said to occur when knowledge of an outcome increases the perceived likelihood of the outcome. Misestimation of covariance is defined as judgments that are made when clinicians do not correctly describe the relation between two events—remembering occasions when a test score and certain trait co-occurred, for example, but failing to recall instances in which the test score occurred in the absence of the trait (Garb, 1998).

In addition to heuristics and biases, clinicians' schematic processing can impact the judgment process through the reliance on personal stereotypes, prototypes, and scripts

(Nisbett & Wilson, 1977). A stereotype is defined as a clinician's beliefs about a typical client (e.g. typical depressed client), while a prototype is defined as a clinician's views of a prototypical client (e.g. client with all the features of depression). A script describes a person's beliefs about how events are likely to unfold (Schank & Abelson, 1977).

Collectively, cognitive heuristics, biases, and schematic processing affect all stages of the judgment process. Studies have demonstrated the impact of information primacy (Ambady & Rosenthal, 1992; Kendell, 1973), the anchoring and adjustment heuristic (Ellis, Robbins, Schult, Ladany, & Banker, 1990; Pain & Sharpley, 1989), and confirmatory hypothesis testing (Dallas & Baron, 1985) on data collection. The hindsight bias has been used to explain why clinicians tend to be overly deterministic when trying to understand the causes of a client's behavior (Einhorn, 1988; Hawkins & Hastie, 1990). Judgments can also be inaccurate due to the way in which clinicians remember information, as highlighted by research that demonstrates clinicians consistently fail to accurately describe the relation between two co-occurring events by remembering only those instances in which both events occurred (Arkes, 1981). As an example of the availability heuristic, the strength of verbal associative connections has been shown to impact how clinicians interpret the Rorschach (Chapman & Chapman, 1969). The impact of these cognitive heuristics and biases helps explain the limited accuracy of clinical prediction and diagnosis. In addition, the limited awareness of cognitive processes makes it difficult to model clinical judgment simply by asking clinicians to verbalize their decision-making process (Garb, 1998).

The representativeness heuristic, along with stereotypes and prototypes, has been shown to describe how clinicians integrate information to arrive at a judgment, thereby

introducing inaccuracies. In addition, the representativeness heuristic appears to describe the way in which most clinicians arrive at a diagnosis (Garb, 1996). Many researchers have agreed with Garb's (1996) stance that the representativeness heuristic seems to model clinical judgment quite well (Nilsson, Juslin & Olsson, 2008). In addition, the representativeness heuristic has been used to explain two common biases in clinical judgment: the conjunction fallacy and base-rate neglect (Tversky & Kahneman, 1983). However, as a general concept, the representativeness heuristic has been criticized for being vague with regard to the cognitive processes and types of representations involved (Nilsson, Olsson & Juslin, 2005). As a result of this criticism, two theories emerged to explain how the representative impacts probability judgments: the prototype hypothesis (Kahneman & Frederick, 2002) and the exemplar hypothesis (Juslin & Persson, 2002).

The prototype hypothesis states that the representative heuristic is based on individual judges having an idea of what an 'average' member of a particular class looks like; that is, this hypothesis requires abstraction concept formation in memory such that a single representation emerges that contains all features of the class (Kahneman & Frederick, 2002; Nilsson, Juslin, & Olsen, 2008). For example, an individual with clinical depression that has all of the typical features of depression would be considered a single prototype. Then, when judging a new individual, the prototype hypothesis states that a judge retrieves the prototype for that class and all other relevant classes from memory and a decision is reached based on how similar the individual is to the prototypes retrieved (Kahneman & Frederick, 2002). Thus, if a clinician is judging a new client who may be depressed, this hypothesis states that the judge would retrieve the prototype for depression as well as the prototype for other similar disorders, such as anxiety and

Bipolar. The diagnosis would be based on how similar the new client is to one of the prototypes retrieved during the decision-making process.

The exemplar hypothesis defines exemplars as previously encountered objects belonging to a certain class; thus, this model does not require abstraction in memory as judges are believed to hold multiple exemplars for a certain class in memory to use when making probability decisions (Juslin & Persson, 2002; Nilsson, Juslin & Olsson, 2008). When making a judgment, all exemplars that are similar to the new object are retrieved from memory and a decision as to the category the new object belongs is made based on how similar it is to all retrieved exemplars that belong in a particular category as well as how similar the new object is to all retrieved exemplars in general. Thus, this model takes into account both similarity to a category and the frequency with which this category is encountered (Nilsson, Juslin & Olsson, 2008). As an example, imagine a clinician who is again evaluating a new client with possible depression. The exemplar hypothesis states that the clinician will retrieve from memory all exemplars that are similar to the new client. The diagnosis of this client would be based on how similar the client is to exemplars belonging to the category of 'depressed individuals' relative to how similar the new client is to other exemplars retrieved, for instance a similar client who was instead diagnosed with an anxiety disorder.

A recent study comparing these two hypotheses through the use of a computer simulation appeared to support the exemplar hypothesis as underlying the impact of the representativeness heuristic on probability judgments (Nilsson, Juslin & Olsson, 2008). The authors of the study suggested that the representativeness heuristic is not so much a cognitive 'tool' as it is a side effect of exemplar memorization. These findings suggest

that the representativeness heuristic is most likely to impact probability estimates when judges have access to many exemplars; in other words, when judges are engaged in a familiar task (Nilsson, Juslin & Olsson, 2008). This argument supports other findings in clinical judgment research, such as research demonstrating that experts are not more accurate judges than novices (e.g., Faust, Guilmette, et al., 1988; Gaudette, 1992; Wedding, 1983). In addition, this research could also be used to support Garb's (1996) argument that the representativeness heuristic underlies much of diagnostic decision making as making a diagnosis is a very common clinical task.

Formal models: Process-tracing and statistical models. Process-tracing models provide a bridge between the more qualitative descriptions of clinical judgment provided by information-processing methods and the quantitative method of statistical modeling. Process-tracing also involves clinicians verbalizing the steps taken in arriving at a particular decision; however, these verbal descriptions are then used to create algorithms that model how the decision was made. Although limited to the extent that judges can verbalize their cognitive process, researchers have argued that process-tracing models can be valuable for describing clinicians who make valid judgments (Garb, 1984). Thus, this methodology can help represent how these judges are making valid decisions and can be used as the basis for linear regression models. Although not directly comparable, process-tracing models and liner regression equations can describe and predict the same judgments. However, researchers have found that regression models are usually more accurate (Einhorn, Kleinmuntz & Kleinmuntz, 1979).

Many types of statistical analyses have been used to model clinical judgment, although the most frequently used has been multiple regression (Garb, 1998). Numerous

studies of statistical models for clinical judgment have used a data set collected by Meehl (1959). In this classic study, 13 clinical psychologists and 16 clinical psychology graduate students were given MMPI profiles for 861 psychiatric patients. The MMPI profiles had been obtained from seven different sites, such as hospitals and outpatient clinics. All participants were asked to sort the profiles on an 11-point, forced normal distribution ranging from neurotic to psychotic. Although the participant clinicians were given less information than they would typically receive in an average clinical encounter, the task is considered to tap complex information-processing abilities (Garb, 1998). The results of this study demonstrated that statistical models outperformed clinicians in predicting actual diagnoses (Meehl, 1959); however, the results could not describe how well statistical rules can model clinical judgment.

Subsequent studies, however, have demonstrated the ability of statistical models to accurately model clinical judgment. Wiggins and Hoffman (1968), tested several statistical models and found that a linear-regression model was superior to all other models in capturing the judgment process of the clinicians. Linear models have been shown in other studies to model clinical judgment quite well, despite clinicians' insistence that they use cues configurally and integrate information in a more complex manner (Dawes, 1996; Garb, 1998; Wiggins & Hoffman, 1968).

Statistical models of clinical judgment have also been generated through the information-processing approach discussed earlier. In these studies, judges are asked to both make decisions and describe the relevant factors used in reaching their decision (Kahneman, Slovic & Tversky, 1982). Using the relevant factors identified, a statistical prediction model can be generated and compared to the judges' decisions. Early research

has demonstrated the ability of expert judges to identify the relevant factors crucial to making a valid decision. Unfortunately, the judges tend to use these factors inconsistently (Dudycha & Naylor, 1966). However, Gauron & Dickinson (1966) asked clinicians to rank important information in making diagnosis and found that clinicians often had little awareness of what information is actually important. This study showed that not only were individual judges using information inconsistently across different patients, but that clinicians often rated different information as being important when judging the same psychiatric case (Gauron & Dickinson, 1966).

However, later research suggests that expert opinion can still be used to make accurate statistical models. Einhorn (1972) studied how severity of biopsy results was related to predicted survival time in cancer patients. Expert medical doctors were asked to rate nine characteristics in terms of how related they were to severity of disease and subsequent survival time. Einhorn (1972) then produced actuarial tables from the combined ratings of the judges and made a regression model to predict survival time. He demonstrated that although the doctor's overall severity ratings were unrelated to survival time, the formulas generated from these ratings were effective in predicting survival time. Thus, statistical models can not only describe the clinical judgment process, they can improve upon it.

Why is Statistical Prediction Better?

One reason for the consistent superiority of statistical prediction is that these models are designed to discover patterns in variability—to detect signal amongst noise. They achieve this by combining available information optimally to detect a pattern (Dawes, Faust & Meehl, 1989). Furthermore, the predictions of these models are so

robust that small differences in the weights of different variables do not significantly alter predictions when compared to the optimal weights (Garb, 1998). In addition, these models are able to automatically compare variables that human judges find difficult to compare directly, for example college GPA and GRE scores. In order to compare these variables, one would need to know distributions of these predictors and their predictability. This information is not readily available to human judges, but it does form the basis of the statistical model (Dowie & Elstein, 1988).

Another reason that statistical methods are superior to intuitive methods is that intuitive methods are based on cognitive heuristics, such as availability and representativeness (Grove et al., 2000). Heuristics have some validity, which is why human judges usually do better than chance. Some researchers have argued that concept discrimination, or the ability to distinguish important characteristics that define a concept from the unimportant ones, is an accurate description of what a clinician does when making a diagnosis (Einhorn, 1972). In addition, human experts are useful at picking out potentially useful predictors, so consensus among practitioners can be used as identification of useful predictors, which can then be used to establish a statistical formula. This formula can then be tested to see how predictive it is and, if successful, distributed for general use. It is important to remember that the accuracy of any judgment is limited by the accuracy of the techniques employed. One study showed that clinical neuropsychologists were not more accurate with expertise (Faust, Guilmette, Hart, & Arkes, 1988); thus, there is a clear need for a general decision aid for neuropsychology.

Objections to the Use of Statistical Models

Despite the limited accuracy of clinical prediction and the superiority of statistical models, expert judges do demonstrate a unique ability to pick out the relevant factors important to diagnosis. If experts are able to provide useful information that can be used to create accurate statistical prediction models, then why are the models not used more frequently in clinical practice? Some critics of the statistical method point to the idea of case uniqueness, which refers to an unexpected cue that is salient but not universal and, therefore, is not included in the regression-equation prediction model (Garb, 1998). Dawes (1971) argues that despite these instances statistical models are still best because they are fair and consistent, save time, and make clinicians accountable for their decisions by forcing them to show how they arrived at the decision. Thus, even when there is an unexpected cue, it is likely that the more common cues included in the model are still present, and, thus, the statistical prediction model will still function adequately (Wainer, 1978). In addition it is not known to what extent "experts" agree on which relevant factors should be used in statistical prediction models, as most studies have used a small number of experts and cannot be seen as representative.

Another issue that surrounds the use of statistical models is whether or not linear models are accurate representations of the decision-making process, despite research that has shown the superiority of linear models in making predictions (Dowie & Elstein, 1988). An alternative to a linear model is a configural model, which includes interaction terms to account for the possibility that a particular judge may interpret an item of information as being contingent upon a second. However, research has shown that these models usually fail to improve accuracy and that no more of the variance in judgments is

accounted for by an interaction term (i.e. r_{x1x2}; Dawes & Corrigan, 1974). Thus, simple linear models continue to be accurate despite judges' insistence that they are using cues configurally (Dowie & Elstein, 1988). Thus, forcing judges to rank order the cues they used in making their decision in order of importance will accurately capture their decision-making process, even if they felt they used two cues in combination with one another (and thus, they have equal importance). It is important to remember that simple linear models are not meant to be seen as identical to the judgment process happening inside the judge's head, but rather is a close approximation of the process (paramorphic representation; Dowie & Elstein,1988).

Another objection is that the individuals studied were not true experts. Dawes (1996) argues that this objection relies on a definition of "expertise" that is so extreme only a small minority of professional clinicians would qualify. Yet, the majority of clinical psychologists are seen by the public and state licensing boards as being qualified to make valid clinical predictions. An additional criticism of previous clinical judgment research is that the techniques used, such as vignettes, are not ecologically valid (Bigler, 1990). Again Dawes (1996) provides an excellent analogy for answering this criticism. The argument for using vignettes is that, although they may not represent actual clinical work, they do represent components of clinical decision-making. Put anther way, if an individual claims he can play Beethoven's fifth on the piano, a rational request would be to ask that individual to play some scales. If the person performs poorly while playing scales, it would be safe to assume that he could not play Beethoven's fifth. However, the individual in question could always claim the test was not ecologically valid because there are no scales in Beethoven's fifth, and most rational people would dismiss the

criticism easily. The same can be said of vignettes. They may not be ecologically valid, but they are a good place to start. Any findings should, of course, be replicated with more ecologically valid measures following an initial, significant finding.

Finally, objections to statistical prediction often describe the statistical method as "dehumanizing" and, thus, relying on statistical formulas is unethical because it reduces the complexity of a human being to mere numbers (Garb, 1998). However, responses to this objection have pointed out that there is nothing in the statistical approach that implies a judgment about what people *are*; the point is to make the best possible prediction, which is in the best interest of everyone involved in a particular clinical case. In addition, statistical models can be made public, open to scrutiny, and modified appropriately.

Moreover, given that the research showing the superiority of statistical models has been reaffirmed for decades, it is strange that these models have not been applied more widely to clinical practice. This phenomenon is often attributed to the objections described above or to the perception that clinical criterion are inherently unpredictable and do not lend themselves to easily useable regression equations (Dowie & Elstein, 1988). While this may be the case in some psychological clinical settings (although research has not shown this to be the case), statistical prediction models have been used successfully elsewhere. Clinicians in forensic settings, for example, have developed models for classifying juvenile and adult prison inmates, such as the Violence Risk Appraisal Guide (VRAG; Quinsey, Harris, Rice, & Cormier, 1998).

In summary, clinical judgment research has revealed that the accuracy of clinical judgment is often disappointing. Judges decisions are influenced by a variety of cognitive heuristics and biases in addition to suffering as a result of limited cognitive resources.

Statistical models of judgment have been found to be superior to clinical judgment in hundreds of studies across a variety of disciplines, yet there is continued resistance to their use. Thus, statistical models, despite their consistent superiority, have not achieved widespread use, although they have been used successfully in some areas, such a medical settings and college admissions. Finally, research has demonstrated that expert judges can identify successful decision making strategies and relevant factors necessary to make an accurate diagnosis, but they apply this information inconsistently when making decisions. Therefore, a combined 'sophisticated clinical' approach that integrates both clinical and statistical prediction has been recommended by clinical judgment researchers.

Clinical Judgment Research and Neuropsychological Assessment

Judgment research in clinical neuropsychology, as in other areas of professional psychology, has been fraught with controversy. Neuropsychologists have been criticized by a number of researchers for a variety of factors related to clinical decision-making, including questions regarding the validity of their diagnoses and for not using statistical prediction rules (Guilmette & Giuliano, 1991; Wedding, 1991; Wedding & Faust, 1989). There have also been questions regarding the role of neuropsychologists in court proceedings and their ability to detect malingering (Faust, Hart, Guilmette, & Arkes, 1988; Garb & Schramke, 1996). Research on neuropsychological judgment has covered a variety of areas encompassing most aspect of clinical practice. This section will briefly cover the relevant research with a focus on the validity of diagnoses based on neuropsychological assessment.

Reliability of Judgments

Interrater reliability is in the good to excellent range for detecting impairment and the localization of brain impairment, and fair to excellent for describing cognitive strengths and weaknesses (Garb, 1998). However, these studies used standardized batteries, such as the Luria-Nebraska or the Halstead Reitan (HRB; Garb & Schramke, 1996). There are no studies on the reliability of impairment detection when neuropsychologists use flexible batteries, which is unfortunate as most neuropsychologists prefer the use of more flexible approaches as opposed to standardized batteries. The flexible battery approach is defined as a variable but routine combination of tests used to assess certain types of clients, such as those with an head injury or possible dementia. The flexible approach, however, does not involve routine groupings of tests but instead involves selecting tests based on each unique client presentation. Only 15% of surveyed neuropsychologists reported that they use standardized batteries frequently, whereas 70% reported using flexible batteries and 15% preferred the flexible approach (Sweet et al., 2000). It is likely that reliability for detecting impairment will be lower for neuropsychologists who use flexible batteries because they use a different combination of tests. However, in one study, three neuropsychologists were asked to describe the cognitive strengths and weaknesses of 41 6-year-old children using a variety of tests instead of a standard battery (Brown, del Dotto, Fisk, & Taylor, 1993). Interrater reliability was excellent for rating intellectual abilities, auditory language skills, and visuomotor skills; good for rating memory and overall functioning; and fair for rating attention (Garb & Schramke, 1996).

Validity of Judgments

In certain circumstances, research has shown that neuropsychologists make valid judgments, for example distinguishing impaired vs. normal patients. In a meta-analytic review, neuropsychologists obtained a hit rate of 84% (correct positives + correct negatives, divided by total number of judgments) in detecting whether any neurologic impairment existed. Most of these studies used standardized batteries, along with WAIS and WMS results, for a total of 2, 383 ratings (Garb & Schramke, 1996). In addition, most of these studies reported base rates, and neuropsychologists were found to be significantly more accurate than base rates (65% base rate vs. 84% hit rate; Garb & Schramke, 1996). Some researchers examine the ability to detect neurologic impairment by examining false positive, correct negative, false negative, and correct positive diagnostic decision rates. Validity estimates using this criteria vary widely, but there is evidence that neuropsychologists often overdiagnose neurologic impairment. Meta-analytic studies have reported that approximately 33% of nonimpaired individuals are misdiagnosed as having neurologic impairment after neuropsychological testing (Faust, Hart, & Guilmette, 1988; Faust et al., 1988). Garb & Schramke (1996) reexamined these results and found that misdiagnosis occurred in 21% of nonimpaired individuals.

The authors speculated that there is a variety of reasons that might explain the significant difference in misdiagnosis found in these studies. As with many meta-analyses, these reasons are usually related to differences in individual study methodology, such as the severity of impairment present in the protocols the judges were asked to rate or how nonimpaired patients were enrolled in the study. In addition, some studies failed to adequately describe the experience of the judges, although Faust et al.

(1988) reported that more experienced judges tended to misdiagnose nonimpaired cases more frequently than less experienced judges. The authors note that these results were of borderline significance, and it is difficult to compare this finding to other studies due to a lack of information regarding experience. Furthermore, when nonimpaired individuals were referred by a medical professional for testing as opposed to being recruited for the study, the false positive rates differ. The false positive rate was 25% for participants referred due to suspected neurologic impairment and 20% when participants were recruited as control subjects (Garb & Schramke, 1996).

Neuropsychologists are less accurate when making precise localization ratings, although it is difficult to directly compare studies as the judgment criterion often varied. As would be expected, validity was higher when asked to diagnosis right- vs. left-hemisphere impairment (hit rate= 89%) than when asked to differentially diagnosis normal, diffuse, left hemisphere, or right hemisphere impairment (hit rate=65%). Accuracy was also higher when the rating of "no impairment" was excluded, approximately 70% (Garb, 1998). As before, information on experience was limited, so the impact of this factor is not known. Surprisingly, accuracy was higher when nonimpaired participants were referred as opposed to recruited (73% vs. 63% accuracy) based on eight studies (Garb & Schramke, 1996). Unfortunately, the authors of this meta-analysis did not specify what "accuracy" referred to in this case. Thus, it may be that judges in the studies that used referred nonimpaired participants simply made more diagnoses of impairment overall, thereby increasing both their accuracy and false positive rate. Meta-analytic reviews have also compared neuropsychologists' ratings of normal functioning versus diffuse, left-hemisphere, or right-hemisphere impairment with a base

rate level of prediction based on eight studies. The base rate level of prediction overall was 38%, and the hit rate achieved by neuropsychologists was significantly more accurate at 67%, even though judges did not always have access to base rate information (Garb, 1998).

In two different studies, neuropsychologists were asked to make precise localization ratings. In one study, Reitan (1964), a single expert neuropsychologist made ratings of right-anterior, right-posterior, left-anterior, left-posterior, or diffuse impairment. The hit rate was 79%. In a second study (Faust et al., 1988), neuropsychologists made ratings of right-anterior, right-posterior, left-anterior, left-posterior, or no impairment. The overall hit rate was 56%. This may be a result of factors related to the studies themselves, which were done at different times (1970's-1990's) and looked at a variety of cases. Thus, one study could be looking at gross impairment, while another looked at more subtle impairment. In addition, the Reitan study used one expert judge, whereas the Faust et al. (1988) study sampled a range of judges; therefore, the difference between the hit rates between the two studies could be a reflection of the inconsistency in accuracy amongst the Faust et al. judges (1988) (Garb & Schramke, 1996).

There are several studies in which neuropsychologists were asked to diagnose the etiology of neurologic impairment (e.g., Alzheimer's disease, head injury, seizure disorder). Validity results varied widely. Two studies reported hit rates of 84 and 85% (Filskov & Goldstein, 1974; Reitan, 1964; respectively); however, Faust et al., (1988) reported a hit rate of only 23%. The Faust study did not attempt the explain why their result differed from previous research, possibly because they were unaware of this

discrepancy (neither the Reitan article nor the Filskov & Goldstein article was cited by the authors). Additionally, in all three studies the judges had access to similar information, and the etiology was comparable across studies. Although the Reitan (1964) study used an expert neuropsychologist, the judges in the Filskov & Goldstein study were not reported to be expert judges. Garb & Schramke (1996) posited that the ability of neuroradiology procedures in the 1980s to detect subtle impairment was superior to techniques used in the 1960s and 1970s, and this difference may help explain the difference in hit rate.

Other studies have looked at neuropsychologists' ability to detect malingering or "faking bad." Neuropsychologists were given test results from the Halstead-Reitan Battery (HRB) and the WAIS-R or WISC-R. Results from some studies have been disappointing, but, again, hit rates vary widely. Faust, Hart & Guilmette (1988) and Faust, Hart et al. (1988) both reported a hit rate of 0%, while other researchers have reported a hit rate of 50-69% (Heaton, Smith, Lehman, & Vogt, 1978). Garb (1998) points out that these hit rates may not be indicative of how well neuropsychologists detect malingering in clinical practice because most neuropsychologists also collect accurate history information. Furthermore, if they are unable to collect accurate information from the patient due to inconsistent reporting, they may suspect that he or she is lying. In the Faust, Hart et al. (1988) studies, experimenters provided the judges with fabricated history information, which the judges likely treated as fact given that they had no information regarding the veracity of this information.

To answer these criticisms, Faust and Guilmette (1990) explained that they did not think the 100% error rate obtained was an accurate reflection of how well

neuropsychologists can detect malingering in everyday practice, but rather see this result as a reason to doubt neuropsychologists' presumed ability to detect malingering easily. Still, although their study illustrates that neuropsychologists can be misled, it does not indicate how often this occurs. In addition, Garb & Schramke (1996), conclude that the Heaton et al. (1978) study may also underestimate the ability of neuropsychologists to detect malingering for three reasons. Four of the 20 participants who were told to malinger were eliminated from the study because they scored in the average range on the test battery. In all likelihood, these individuals would have been correctly identified as nonimpaired had they been included. Also, technicians who administered the test battery questioned the behaviors of seven participants during testing, but these observations were not reported to the judges. In actual clinical practice, it is expected that the technicians would have made the neuropsychologist aware of these observations. Finally, it is widely recommended that verified history information be used when detecting malingering, but this information was not available to the judges in this study.

Experience and Expertise

Numerous researchers have shown that experience is often unrelated to accuracy and validity. Ratings by experienced neuropsychologists are, in general, no more valid that ratings made by less experienced neuropsychologists (Faust, Guilmette, et al., 1988; Gaudette, 1992; Heaton et al., 1978; Nadler, Mittenberg, DePiano, & Schneider, 1994; Wedding, 1983). Results are mixed in terms of whether or not neuropsychological expertise leads to more accurate judgments. Neuropsychologists with the ABPP diploma were not more accurate than non-board certified neuropsychologists with few years of clinical experience when asked to describe the localization of brain impairment

(Gaudette, 1992). However, Wedding (1983) demonstrated that one presumed expert neuropsychologist was actually more accurate than other neuropsychologists (hit rates 63% and 54%, respectively) and completed the protocol rating in significantly less time. However, research has shown that there is a trend for experience being positively related to the accuracy of confidence ratings (Garb & Schramke, 1996).

Confidence Ratings and Judgments

Two studies looked at the relationship between validity of judgments and confidence in ratings. Wedding (1983) had 14 neuropsychologists categorize protocols as left or right hemisphere brain damage, diffuse damage, schizophrenia, or non-impaired. In another study, six neuropsychologists were asked to diagnose the presence of brain impairment and decide if the impairment was diffuse or lateralized to the right or left hemisphere (Gaudette, 1992). Both studies also had judges rate the likelihood that their judgments were correct. The correlation between the validity and confidence was .29, $t(24)=1.48$, $.05 < p < .10$ (Garb, 1998). This correlation implies that confident judges make only slightly more valid judgments than less confident judges. In another study (Trueblood & Binder, 1997), neuropsychologists made diagnosis of malingering, functional impairment, or neurologic impairment. Some of the neuropsychologists were given more test data, especially the results from forced-choice tests used to detect malingering. Neuropsychologists who received forced-choice test results made significantly more valid judgments and were more confident in their ratings than neuropsychologists who diagnosed malingering without force-choice test data as well as those who diagnosed functional or neurological impairment.

Although validity and confidence can be positively correlated, there is evidence that neuropsychologists are often overconfident. A *well calibrated rater* is defined as an individual who provides an estimate of making a correct judgment that is close to the actual probability of being correct (Garb, 1998). Thus, if the chances of being correct are 50% and the judge subjectively estimates being correct half the time, then they are said to be well calibrated. Therefore, it has been argued that a positive correlation between confidence and validity does not necessarily indicate good calibration. Research in this area indicates that neuropsychologists tend to be overconfident, but these results need to be clarified and the conditions in which overconfidence is likely to occur need to be clarified. Karekan & Williams (1994), asked neuropsychologists to estimate premorbid intelligence and set 95% confidence intervals around their estimates. The researchers found evidence that neuropsychologists were too confident in their IQ estimates as the confidence intervals were often too narrow (Kareken & Williams, 1994).

Consistent with this result, the Gaudette (1992) study demonstrated that neuropsychologists estimated their diagnostic hit rate to be 77.5% but achieved an actual hit rate of 62%. However, in the Wedding (1983) study, neuropsychologists were actually underconfident, as they estimate their hit rate to be 47% but achieved a hit rate of 55%. Garb (1998) suggests that these results may be due to the task used in the study. Neuropsychologists were asked to make a differential diagnosis of schizophrenia, left or right hemisphere impairment, diffuse impairment, or normal functioning; however, distinguishing between schizophrenia and brain impairment is no longer considered an important clinical task in neuropsychological practice.

Use of Decision Aids

The term 'decision aid' includes the use of empirical norms and base rates as well as the implementation of automated assessment programs and statistical prediction rules. Although empirical norms are widely used (e.g. Garb & Boyle, 2003; Heaton, Grant, & Matthews, 1991), attending to the base rates for the different conditions and behaviors seen in clinical settings is an area which could benefit neuropsychological practice (Garb & Schramke, 1996).

Statistical prediction models could also improve neuropsychologists diagnostic accuracy (Dawes, 1996; Dawes et al., 1989). These studies have found that statistical prediction models are helpful in other areas of psychological assessment. For example, one study found that actuarial prediction models were correct 70% of the time when diagnosing an individual as psychotic or neurotic based on MMPI profiles. Clinicians, even when using this formula, could not get close to this level of accuracy (Dawes et al., 1989). This finding has been shown in areas of neuropsychological assessment as well (Gaudette, 1992; Heaton et al., 1978; Wedding, 1983). In psychological evaluation and prediction, research has shown that clinical judgment is never superior, despite having access to more information than is included in statistical model. Thus, several researchers have concluded that only two or three relevant variables are needed for accurate diagnosis (Dawes, 1996; Garb & Schramke, 1996). However, clinical judgment can be improved through the use of formulas. Indeed, a study that looked at predicting intellectual disability due to brain damage demonstrated that the accuracy of clinical judgments was better when clinicians used diagnostic rules based on statistics (Leli & Filskov, 1981).

Still, research has shown that neuropsychologists prefer intuitive (clinical) methods over statistical formulas to detect intellectual deficit (Guilmette et al., 1990). While expert clinical judgment studies are almost universally disappointing, clinicians can offer useful information regarding potentially useful variables that can create accurate statistical models (Einhorn, 1972).

In an unusual finding, one study reported that neuropsychologists outperformed predictions made by the Halstead Reitan Brain Impairment Index (75% model, 86% clinicians; Russell, 1995). Several factors have been suggested to explain this result. First, the clinicians had more information available to them as the Trail Making Test and Aphasia Screening Test, although part of the HRB, are not scored as part of the Halstead Index. Second, it may be possible that neuropsychologists are able to recognize patterns within and across tests that may be related to the presence of brain dysfunction. However, interactions are not included in the Halstead Index. Finally, clinicians may regard some results as being more pathognomonic than other tests, while the Halstead Index equally weights all test results (Garb, 2000). Also, researchers have noted that the neuropsychologists in studies of statistical vs. clinical prediction often receive only psychometric and demographic information and, therefore, statistical prediction models may not be as accurate as neuropsychologists who have access to commonly collected clinical information (e.g. history information collected during a clinical interview; Garb & Schramke, 1996). Until the incremental validity of this additional information is studied, this comparative accuracy will remain an open question.

Overperception of Impairment

There is evidence that neuropsychologists sometimes overdiagnose neurological impairment. As previously reported, a meta-analytic review found that neuropsychologists diagnosed 21% of 555 nonimpaired participants as having neurological impairment (Garb & Schramke, 1996). Two studies (Nadler et al., 1994, replicated by Garb & Florio, 1997) showed that neuropsychologists do not always use norms by asking clinicians to rate the likelihood of dementia for a 38-year-old and a 75-year-old. Both client profiles were created using average scores (40^{th} – 65^{th} percentile) taken from the normative table for their respective age groups; however, neuropsychologists were much more likely to rate the 74-year-old as having dementia (58% diagnosed the elderly individual as having neurological impairment and 23% diagnosed dementia).

However, Garb and Boyle (2003) failed to replicate this finding. This study attempted to overcome some of the limitations of the Nadler et al., (1994) and Garb and Florio (1997) studies by assessing whether or not judges used normative data when making their diagnoses and using a web-based survey to ascertain how much time clinicians spent making their ratings. In addition, the test data provided was changed. The initial studies gave results from the HRB, the WAIS-R, the Controlled Oral Word Association Test, and the Rey Auditory-Verbal Learning Test. However, less than half of the clinicians in the Garb and Florio (1997) study reported using the HRB in the last year (the Nadler et al. (1994) study did not ask clinicians if they normally used the tests included in the study protocol). Also in both studies, clinicians were not given WAIS-R verbal IQ, performance IQ, or full-scale IQ; instead, they were given subscale raw scores.

To compensate for these short-comings, the authors provided neuropsychologists with comparable data from popular neuropsychological tests rather than from a standardized battery (see rationale section for a complete list of tests used). In addition, verbal, performance, and full-scale IQ scores were included. The researchers asked judges to rate the likelihood of any neurological impairment (0 to 10 scale) as well as rate the likelihood of a specific diagnosis of dementia (0 to 10 scale).

The results of this study found that, although the effect of age was statistically significant, there is no evidence for age bias (Garb & Boyle, 2003). Age bias occurs when the accuracy of judgments varies as a function of age, which was found by the Nadler et al. (1994) study. However, Garb and Boyle (2003) demonstrated that clinicians were unlikely to make a diagnosis of dementia for either client, although they did rate the elderly patient as being somewhat more likely to have neurological impairment (rating of 1.6 vs. 2.72). Nevertheless, this rating indicated that an actual diagnosis of any type of neurological impairment was unlikely for either client; thus, the neuropsychologists in this study appear to be correctly attending to base rates, in which an elderly individual is more likely to be experiencing neurological impairment than a middle-aged individual (Garb & Boyle, 2003).

However, several shortcomings were noted in this study. The authors admit that their results may not be representative of actual clinical judgments for two reasons. One, they sampled only neuropsychologists who had ABCN certification, which represents a small minority of practicing neuropsychologists. Thus, the neuropsychologists who participated in this study were, most likely, highly competent to complete this task. In addition, out of 187 professionals sampled, about half of ABCN certified

neuropsychologists, only 25 participated in this study, which further limits the generalizability of these findings. Two, the judgment task used in this study was much simpler and easier than what is typically encountered in clinical practice. The test data lacked any variability as all scores fell in the average range, around the 50th percentile. In typical neuropsychological test data, there is frequently more variability in scores, with at least a few individual scores falling below normal expectations. The authors conclude that, although the results of their study are encouraging, there is still a question as to how often neuropsychologists overdiagnose neurological disorders and under what conditions neurological disorders are likely to be overdiagnosed (Garb & Boyle, 2003).

Research indicates the neuropsychologists frequently make reliable and moderately valid judgments. These judgments are consistently more valid that chance levels and base rates, with the exception of malingering in which the results were not clear. However, the levels of validity were only moderately high, which some researchers feel may be due to the fact that judges were not given access to all information commonly available in clinical practice (Garb & Schramke, 1996). Still, the incremental validity of nonpsychometric data in neuropsychological assessment has not been established, although this information (history data and clinical interview) has been shown to increase accuracy in personality assessment (Garb, 1994). In addition, the reliability of ratings when neuropsychologists use a flexible battery or flexible approach has not been studied. Experience and presumed expertise was often not related to validity and studies employing the use of confidence ratings have demonstrated those neuropsychologists are often overconfident. Neuropsychologists frequently use norms when evaluating the results of neuropsychological tests which increase diagnostic accuracy. Despite research

suggesting that statistical prediction rules and decision aids can improve the accuracy of diagnosis, they are not widely used in neuropsychological practice (Sweet et al., 2000). Finally, there is evidence that neuropsychologists frequently overdiagnose neurological impairment, although the conditions in which this is likely to occur are unclear (Garb & Boyle, 2003).

CHAPTER III

RATIONALE FOR CURRENT STUDY

Clinical neuropsychology was described as an "emerging discipline" as early as 1970, and researchers have noted that all indicators point to its continued growth and importance (Benton, 1987). Advances in clinical practice and empirical research have supported these early observations, enabling neuropsychology to emerge as one of psychology's fast growing specializations. Various journals catering to neuropsychological topics, specialized clinical training programs, professional organizations for practicing clinicians, and the development of specialty diplomas are a testament to the rapid growth and influence of the discipline. As most of clinical neuropsychology involves identifying and describing the cognitive and behavioral deficits associated with brain dysfunction, the development of standardized assessment procedures has been critical to the growth of neuropsychology.

For over 30 years, researchers have used survey research to track and characterize the rapid growth within the field of clinical neuropsychology. Survey research has focused on a broad array of topics, including professional activities (e.g., work setting, referral sources, patient characteristics), education and training, journal preferences, and views on professional issues. In addition, a recent trend has been the identification of instruments used to assess specific cognitive and functional domains (e.g., broad cognitive areas such as attention and memory, malingering, judgment ability; Rabin et al., 2005, Rabin et al., 2008; Slick, Tan, Strauss, & Hultsch, 2004). Previous surveys have noted the preference by most neuropsychologists toward the use of a flexible battery approach and a departure from a past preference for standardized batteries. Other

researchers have documented the tests most commonly used by practicing professionals in the field. The use of a flexible battery approach implies the use of similar test combinations for similar client problems. Battery construction has largely been inferred from the popular tests used to assess broad cognitive domains and the domains neuropsychologists say they regularly assess. However, in one survey that investigated battery construction for the assessment of mild brain injury, important differences in battery construction were found (Rabin et al., 2007). Thus, one aim of this study is to investigate the similarity of batteries used by neuropsychologists to assess another common area of clinical practice—dementia.

The survey design has also been used in the area of clinical judgment within neuropsychology to investigate the reliability and validity of judgments and cognitive biases that may affect diagnoses (e.g., Garb & Boyle, 2003, Gaudette, 1992, Nadler et al., 1994). One explanation offered for why clinical judgment is sometimes less than optimal is that clinicians' cognitive processes are negatively impacted by heuristics and biases. The impact of cognitive heuristics and biases has been extensively studied in the areas of psychiatric diagnosis and personality assessment, but has rarely been studied in the context of neuropsychological assessment (Garb & Schramke, 1996). One study has been conducted on the hindsight bias (Arkes, Faust, Guilmette, & Hart, 1988), and three studies have assessed for the presence of an age bias (Garb & Boyle, 2003; Nadler et al., 1994). If cognitive heuristics and biases describe how laypeople and professional psychologists make judgments, then one would expect that they would also play a role in neuropsychological judgment.

Thus, a second aim of this study is to investigate the impact of cognitive biases, specifically age bias, on neuropsychological judgment. This will be achieved in two ways. One, an attempt will be made to replicate the finding of Garb and Boyle (2003) regarding the lack of an age bias in the diagnosis of neurological impairment and dementia with a more representative sample of neuropsychologists. The present study will also address one limitation of this study by adding variability in a few of the test results, thus more closely approximating everyday clinical conditions in which one or more test results usually falls below expectations. Two, it has been demonstrated that cognitive biases are most likely to effect judgment in situations where information is ambiguous (Tversky & Kahneman, 1974). Although ambiguity has been defined in a variety of ways by different researchers, in this study 'ambiguity' will refer to test results that are equivocal, and, therefore, do not explicitly denote the presence or absence of neurological impairment (e.g., borderline scores, z = -1 to -1.33). Thus, neuropsychologists in the present study will be presented with vignettes containing test data from the normative tables representing average, borderline, or impaired performance. It is hypothesized that age bias will not be present when the results are unequivocal (i.e., the average and impaired conditions) but will affect judgments in the more ambiguous condition (i.e., borderline scores) when compared to base rates.

Although numerous clinical decision-making studies have investigated the accuracy and reliability of neuropsychological diagnosis, these studies did not assess whether neuropsychologists used all of the information provided to them in making a diagnosis. Therefore, it is unknown whether neuropsychologists use the same clinical information when arriving at a decision. A final aim of this study is evaluate the degree

of consistency in clinical decision-making among practicing neuropsychologists by examining both the accuracy of diagnostic decisions and the perceived necessity of clinical information. This addresses a limitation in the current literature as previous studies in this area have focused on accuracy and have not assessed whether neuropsychologists view the same information as relevant or necessary in making a diagnosis. It may be that differences in the information perceived as relevant can help explain variability in the accuracy of diagnosis. In addition, consensus is largely regarded as a sign of maturity within an area of specialty (Donders, 2001) and can help inform training and clinical standards field-wide.

Consistency in the perceived necessity of clinical information for diagnosis is important for a variety of reasons. Neuropsychological assessment is an area in which actuarial prediction is presumed to be more frequent (Dawes, 1996). Furthermore, the discipline prides itself on the use of a convergence of evidence; that is, the use of multiple, overlapping indicators that produce somewhat redundant information, thereby allowing the clinician to evaluate patterns that converge on a single diagnosis (Lezak et al., 2004). Thus, neuropsychologists maintain that their form of clinical judgment is much closer to actuarial predication than other branches of clinical psychology, and this research will evaluate this implicit claim. It is hypothesized that the rankings of necessity of clinical information in making a diagnostic decision will be similar across neuropsychologists, reflecting this underlying actuarial model.

This study will present two case vignettes and ask neuropsychologists to make diagnostic ratings and then rate the information presented in terms of how necessary it was in arriving at a diagnosis. The use of vignettes is a common way of evaluating

clinical decision-making and has been used in a variety of areas (Hughes & Huby, 2004). Clinical-decision making research, however, has also revealed that human beings are not very adept at explaining how they arrive at decisions. Rather, they report how they *think* they make decisions, instead of how they actually arrive at a judgment (Garb, 1998). Still, the way people *think* they make a decision provides some insight into how the decision-making process occurs (Garb, 1998). Furthermore, neuropsychologists are expected to explicate their decision-making process in every comprehensive report, thereby justifying their diagnosis. Considering the lack of field-wide information concerning how neuropsychologists arrive at a diagnosis, the use of survey-based vignettes is a good place to start. At the very least, this methodology can help shed light on whether neuropsychologists *think* they are making decisions in a consistent manner across the field.

It is expected that the results of this study will inform the practice of neuropsychology by providing information as to how professional neuropsychologists construct batteries to assess for the presence of dementia and whether there is a consensus as to the necessity of various clinical information in making a diagnosis of dementia. Dementia was chosen in order to replicate the findings of Garb and Boyle (2003) and also represents a common differential diagnosis that most neuropsychologists are routinely asked to make. Consensus regarding the relevant information necessary to accurately diagnose dementia would further inform clinical practice. A major issue facing clinical neuropsychologists is receiving reimbursement for services, which is difficult, in part, due to variable standards set by individual institutions, professional organizations, and insurance companies (Prigatano & Morrone-Strupinsky, 2010). Thus, enabling the field

as a whole to explicitly list the information necessary to make a diagnosis may result in more consistent diagnostic standards and reimbursement rates.

CHAPTER IV

METHOD

This study utilized a survey mechanism to conduct an experiment designed to assess test usage practices for the evaluation of dementia, the presence of age bias in diagnosis, and the extent to which certain test data are perceived as necessary by clinical neuropsychologists when making a diagnosis. The following section will describe the participants, questionnaire design, and procedures.

Participants

Participants were randomly selected members of the International Neuropsychological Society (INS). The INS includes members from various disciplines who possess an interest in neuropsychology, so it was made clear in the initial contact email that practicing clinicians involved in neuropsychological assessment were the focus of the study. Consistent with previous survey methodology in this area, only members who possessed a doctoral degree (i.e., Ph.D., Psy.D., M.D., or Ed.D.), resided in the United States or Canada, and provided an email address were selected for inclusion. The INS provides a membership directory in which members can include contact information that will be shared with other INS members for the purposes of communication regarding research, best practices, and other clinical or academic information. The initial email list was pulled from this membership directory and included only participants who included their email address and listed their residency as the United States or Canada. Information regarding professional activities was also collected in the demographic section, so that unsuitable participants could be screened out.

Prior to the beginning of this study, statistical power was determined according to procedures outlined by Cohen (1992). In order to achieve a medium effect size (.25) with a statistical significance level of 0.05 and a power of 0.80, approximately 26 participants were needed for each of the six groups included in the statistical analysis for this study, thus a minimum of 156 participants are necessary. Survey invitations were sent to 1,000 members of the INS who met selection criteria; however, the desired number of responses was not achieved. Thus, the current study may be underpowered. In total, 125 INS members completed the survey, representing a response rate of 12.5%, which is consistent with previous response rates in survey research of practicing clinicians (Dillman, 2000).

Questionnaire

To participate in this study, participants signed on to a password protected Web site called Qualtrics, which hosted the survey created and monitored by the primary researcher (see Appendix A-F for text version of this survey). This questionnaire was composed of three main sections: test usage for a hypothetical client; diagnostic impressions for two case vignettes; and demographic/practice-related information. In addition, questionnaires varied along two dimensions (described in greater detail below): (a) age of simulated client (*38-year-old or 74-year*-old) and (b) level of performance reflected in test results (*average, borderline, or impaired* scores).

Typical Tests Used to Assess Dementia

In the first part of the questionnaire, participants were presented with a vignette consisting only of the following referral information: "Client is a 67-year-old, right-handed male, with 16 years of education who has been referred for testing due to

suspected memory impairments." They were then asked what techniques they would use in evaluating this simulated client. Techniques were defined as both specific tests used and any additional information gathered. A checklist of common, popular tests and techniques used by neuropsychologists was compiled from previous surveys to facilitate easy response (see Appendix A for text version). In addition, an "other" box provided write-in space for any additional tests not listed. This question was included for two reasons. One, previous surveys have indicated the majority of neuropsychologists use a flexible battery approach, while other surveys have asked neuropsychologists what tests they use most frequently. The flexible battery approach indicates using a specific combination of tests for certain common complaints, such as possible dementia or brain injury; however, no survey has asked neuropsychologists what tests they typically include in their flexible batteries. Thus, the question will address whether neuropsychologists who use a similar testing approach also use similar techniques.

Clinical Vignettes

Each questionnaire was composed of 2 clinical vignettes, the first of which was a reference vignette describing a simulated client, who is a 59-year-old, right-handed male with 14 years of education. Consistent with previous methodology (Garb & Boyle, 2003), the following referral information was provided: "Client referred by primary-care physician due to memory complaints. Initial interview with client and spouse indicates client is not depressed." Test results for this vignette were taken from the normative table representing average performance (i.e., 30^{th} to 60^{th} percentile range). All participants were presented with the reference vignette.

The second vignette portrayed either a 48-year-old or 74-year-old client. Both clients were described as male, right-handed and having 12 years of education. In addition, the following referral information was presented for both clients: "Client complains of memory problems. Client denied any symptoms of anxiety or depression during intake interview. There is no reported history of a head injury." Previous researchers have pointed out that this referral information is not misleading because people may complain of memory problems without having a neurological disorder (Garb & Boyle, 2003).

Participants either viewed the vignette portraying the 48-year-old or the 74-year-old. In addition, the test results presented for each client were taken from the normative table for their ages and represented either average, borderline, or impaired performance. Borderline performance in this study referred to scores that fall 1 to 1 ½ standard deviations below the mean for that age group. Impaired referred to test scores that fall 2 or more standard deviations below the mean for age. Handedness and level of education were invariant across all three possible vignettes to control for any possible confounding effects due to these variables. Thus, only age and level of performance was varied across all three vignettes. Participants were presented with either the young or old client at one of three possible levels of impairment; thus, there were six total versions of the survey.

After reviewing the simulated case data, participants were asked to make two separate diagnostic ratings for each vignette in the same manner as they would in routine clinical practice. First, they rated the likelihood of the client having any type of neurological impairment. Then, they rated the likelihood that the client has dementia. These ratings were made on a scale of 0 to 10; participants were told that a rating of 0

indicates that the client definitely does not have an impairment or disorder and that a rating of 10 indicates that the client definitely has an impairment or disorder. Additionally, it was explained that a rating of 5 indicates the client has a 50-50 chance of having an impairment or disorder. Thus, if the participant believes the client meets criteria for a diagnosis, the participant should make a rating of 8, 9, or 10. After making each rating, participants were asked to rate their level of confidence in their diagnosis on a 5-point Likert scale: (1) very low, (2) low, (3) moderate, (4) high, and (5) very high (Trueblood & Binder, 1997).

After making these ratings for either the young or old client, participants were asked to rate how necessary each piece of information provided in the vignette was for making their diagnostic ratings on a 4-point Likert scale ranging from 'absolutely unnecessary' to 'absolutely necessary' (Groenier, Pieters, Hulshof, Wilhelm, & Witteman, 2008). A scale with no neutral category, such as 'equally necessary and unnecessary' or 'no opinion' option, was deliberately chosen in order to force respondents to at least rate the information as either necessary or unnecessary. This decision was based on the idea that respondents may be more likely to overuse a neutral category when unsure of how necessary they found a particular piece of information as well as the desire to gather information that would shed light on the importance and usefulness of individual pieces of clinical information.

Demographic and Practice-Related Information

The final section of the questionnaire asked participants to provide basic demographic and practice-related information including: gender, age, degree type and

field, board certification status, percentage of time devoted to various professional activities, primary work settings, patient characteristics, and battery approach.

Protocols

Results from the following tests were presented for all vignettes: Wechsler Adult Intelligence Scale, 4th Edition (WAIS-IV; Wechsler, 2008); Wechsler Memory Scale, 4th Edition (WMS-IV, Wechsler, 2009); Boston Naming Test (Kaplan, Goodglass, & Weintraub, 1983; Farmer, 1991); Controlled Oral Word Fluency Test (Axelrod & Henry, 1991); Trail-Making Test (Part A), Trail-Making Test (Part B) (Heaton, Grant, & Matthews, 1991; Reitan, 1958); Hooper Visual Organization Test (Hooper, 1983); Brief Visual Memory Tests (Benedict, Schretlen, Groninger, Dobraski, & Spritz, 1996); Hopkins Verbal Learning Test-Revised (Benedict, Schretlen, Groninger, & Brandt, 1998); Wisconsin Card Sorting Test (Axelrod & Henry, 1992); and Finger Tapping Test and Grooved Pegboard (Heaton et al., 1991). Test scores for each vignette were based on the normative data for each level of performance (see Appendix B-D for copies of protocols), although participants were told that the results were obtained from the client in each vignette. These tests were chosen in order to replicate previous findings (Garb & Boyle, 2003) and are also commonly used neuropsychological tests with which most clinicians should be familiar (Rabin et al., 2007).

With the exception of scores from the WAIS-IV, all scores were varied according to level of performance in the vignette (i.e., average, borderline, or impaired range). WAIS-IV scores fell in the average range for all vignettes for two reasons. One, general intellectual functioning is often used as a basis for comparison with other neuropsychological tests results, thus deviations from both normative data and

expectations based upon intellectual functioning may be used to justify a diagnosis. In this case, borderline scores on all test measures may have lead some participants to conclude that there is no impairment in functioning, especially when given limited background information. Two, WAIS-IV scores in the extremely low (impaired) range may have lead to similar confusion, especially in light of research demonstrating that individuals with IQs lower than 80 often fall in the impaired range on other neuropsychological tests (Gilleard, 1997).

For each of the three levels of performance in the young/old client vignette, scores were somewhat variable rather than all falling in a given range. This was done in order to address a limitation in previous research and to more closely approximate tests results that are encountered in everyday clinical practice. Because the diagnostic decision in this case was about the presence of dementia, general memory scores always fell in the specified level of performance (e.g., average, borderline, impaired). In the average performance protocols, COWAT scores fell in the low average range as did Spatial Span scores on the WMS, while scores on the Trail Making Test (Past A) and finger-tapping (both hands) fell in the borderline range. In the borderline performance protocols, COWAT and Spatial Span scores also fell in the low average range, representing better than expected performance, while scores on the Trail-Making Test (Part A) and finger-tapping (both hands) fell in the impaired range. For the impaired performance protocols, facial-recognition and finger-tapping scores fell in the low average range. These scores introduced some of the variability typically found in neuropsychological assessment but were not indicative of the presence of any formal neurological impairment or dementia.

Procedure

A recent meta-analysis of web- or internet-based surveys suggested several steps to maximize response rates that were used in this study (Cook, Heath, & Thompson, 2000). In order to maximize response rates, all potential participants were first emailed a letter (see Appendix G) explaining they had been selected to participate and the purposes of the study. Potential participants were given an opportunity to opt-out of the study at this time. Then, approximately a week after the initial letter, an email was sent with a link to the survey along with explanations for its completion. Any potential participants who had not responded were sent a single reminder email approximately two weeks later (see Appendix H).

Informed Consent

The initial letter served as informed consent (see Appendix G). Participants were informed that the survey should take approximately 15-20 minutes to complete. Additionally, participants were informed that the foreseeable risks (e.g., loss of time) of participating in the study were minimal and that the survey was voluntary in nature. Further, responses were kept confidential and are released only as summary findings. After questionnaire responses were recorded, any identifying information was separated from the questionnaire to maintain anonymity. Respondents were directed to contact the principal investigator by email if any questions arose regarding participation in the study.

CHAPTER V

RESULTS

Participants: Demographics

189 individuals began the survey, but only 125 surveys were completed in their entirety. Only the completed surveys were included in the analyses. Table 1 summarizes the demographics of respondents.

Of the included participants, 62% were male with an average age of 49 years. Most participants held doctoral degrees in clinical psychology (84%), although the fields of counseling psychology (8%) and school psychology (5%) were also represented in the sample. The remaining participants held medical degrees, with the exception of one participant who wrote in a response specifying a doctorate degree in clinical neuropsychology. Fifteen participants (12%) carried a certification from the American Board of Professional Psychology (ABPP) and 3 (2.4%) carried a certification from the American Board of Clinical Neuropsychology (ABCN). A majority of the participants indicated that they worked primary in either a medical/psychiatric hospital (56%) or private/group practice (34%). Another 12% of respondents worked in a VA hospital setting, while the remaining respondents worked either in community mental health settings or college counseling centers.

Table 1

Demographic Information

Category		Count	% of respondents
Age:	Under 30	3	2.4%
	30-39	26	20.8%
	40-49	28	22.4%
	50-59	42	33.6%
	60+	22	17.6%
Gender:	Male	78	62.4%
	Female	47	37.6%
Degree:	Ph.D	102	81.6%
	Psy.D	18	14.4%
	Ed.D	4	3.2%
	Other	1	0.8%
Degree Field:	Clinical Psychology	104	83.2%
	Counseling Psychology	10	8%
	School Psychology	6	4.8%
	Other	4	3.2%
Board Certificate:	ABPP	15	12%
	ABCN	3	2.4%
Work Setting:	Medical/Psych Hospital	70	56%
	Private/Group Practice	43	34.4%
	VA Hospital	14	11.2%
	Community Mental Health	15	12%
	College Counseling Center	3	2.4%
Years of Practice:	Less than 5	19	15.2%
	5-10	28	22.4%
	11-20	35	28%
	21-30	34	27.2%
	30+	9	7.2%
# of Assessments per month:	1-15	91	72.8%
	15-30	21	16.8%
	30+	7	5.6%

Assessment Approach:	Flexible Battery	107	85.6%
	Flexible	17	13.6%
	Standardized	0	0%

The respondents had been engaged in neuropsychological practice for an average of approximately 16 years (range 1 year to 40 years). Most participants, 73%, indicated that they performed 1-15 neuropsychological assessments per month; this question was included to screen out participants who did not perform at least one neuropsychological assessment per month. Of the remaining respondents, 16% performed between 15 and 30 neuropsychological assessments per month, and only 5% of respondents performed more than 30 assessments per month.

Overall, the included participants indicated that they spend 53% of their professional time devoted to neuropsychological assessment and 33% of their time devoted to other forms of psychological assessment, such as psychodiagnostic or school-based evaluations. Although individual participants varied dramatically in terms of how much of their professional time was devoted to neuropsychological assessment (range 1%-100%), the overall trend suggests that this sample of clinicians, on average, spend a great deal of their professional time engaged in neuropsychological work. In addition, participants frequently reported being engaged in other professional activities, such as research, teaching, rehabilitation, and psychotherapy. This would suggest a rather diverse sample of participants in terms of professional settings and activities. Participants were also asked the age of the population they work with most frequently; overall, the included participants reported working equally with all age ranges including children, adolescents, adults, and older adults.

Finally, respondents were asked about the type of approach they usually use to select assessment procedures: a standard battery of fixed tests for all clients, a flexible battery of tests that is uniform across clinical presentations, or a flexible approach that varies based on individual client needs. Consistent with previous research, the participants in this study indicated using a flexible battery approach most often, with 86% of participants endorsing this method of battery construction. Furthermore, the flexible approach was endorsed by 14% of the sample participants with no one in the sample endorsing the standardized approach. The clinicians included in the sample also indicated that they use, on average, between 8 and 20 tests per evaluation, though the numbers varied significantly. Several respondents wrote that they found the question difficult to answer as the number of tests selected varies according to client needs, time allotted, and other unforeseen circumstances, such as the speed with which a client completes each test.

Dementia Battery Selection

In the first section of the survey, respondents were presented with a vignette of a 67 year-old man who complained of subjective memory complaints and were asked to choose the tests they would include to assess for the presence of dementia. Table 2 summaries the tests selected for inclusion in a dementia battery.

Table 2

Dementia Battery Test Selection

Name of Measure (alphabetical)	Number of Respondents
21 Item Test	5 (4%)
Aphasia Screening Exam	14 (11.2%)
Beck Depression Inventory	64 (51.2%)
Bender-Gestalt	0 (0%)
Booklet Category Test	2 (2.4%)
Boston Naming Test	85 (68%)
Brief Visual Memory Test	30 (24%)
California Verbal Learning Test	42 (33.6%)
CERAD Neuropsychological Battery	0 (0%)
Clock Drawing Test	60 (48%)
Cognistat/Neurobehavioral Status Exam	4 (3.2%)
Connors Continuous Performance Test	5 (4%)
Controlled Oral Word Association	93 (74.4%)
Delis-Kaplan Tests of Executive Function	14 (11.2%)
Facial Recognition Test	22 (17.6%)
Finger Tapping	40 (32%)
Fuld Object Memory Evaluation	2 (1.6%)
Geriatric Depression Scale	38 (30.4%)
Grooved Pegboard Test	58 (46.4%)
Halstead Category Test	1 (0.8%)
Halstead-Reitan Neuropsychological Battery	0 (0%)
Hooper Visual Organization Test	25 (20%)
Hopkins Verbal Learning Test	23 (18.4%)
Judgment of Line Orientation	34 (27.2)
Kauffman Brief Intelligence Test	6 (4.8%)
Luria-Nebraska Neuropsychological Battery	0 (0%)
Mini-Mental Status Exam	29 (23.2%)
Minnesota Multiphasic Personality Inventory	6 (4.8%)
Paced Auditory Serial Addition	1 (0.8%)
Personality Assessment Inventory	8 (6.4%)
Purdue Pegboard Test	2 (1.6%)
Raven's Progressive Matrices	2 (1.6%)
Repeatable Battery for the Assessment of Neuropsychological Status	7 (5.6%)
Rey 15 Item Memory Test	9 (7.2%)
Rey Auditory Verbal Learning Test	23 (18.4%)
Rey-Osterrieth Complex Figure Test	83 (66.4%)
Rorschach	1 (0.8%)
Sentence Repetition	35 (28%)
Stroop Test	88 (70.4%)
Tactual Performance Test	1 (0.8%)
Test of Memory and Malingering	45 (36%)

Token Test	5 (4%)
Trail Making Test (Part A & B)	103 (82.4%)
Validity Indicator Profile	1 (0.8%)
Visual Form Discrimination Test	4 (3.2%)
WAIS (Wechsler Adult Intelligence Scale-III/IV)	19 (15.2%)
WAIS Letter-Number Sequencing	8 (6.4%)
WAIS Vocabulary	57 (45.6%)
WAIS/WMS Digit Span	50 (40%)
WASI (Wechsler Abbreviated Scale of Intelligence)	25 (20%)
Warrington Recognition Memory	1 (0.8%)
Wechsler Individual Achievement Test	1 (0.8%)
Wide Range Achievement Test	6 (4.8%)
Wisconsin Card Sorting Test	41 (32.8%)
WMS (Wechsler Memory Scale- III/IV-full)	40 (32%)
WMS Information and Orientation Subtest	5 (4%)
WMS Logical Memory Subtest	72 (57.6%)
WMS Mental Control Subtest	32 (25.6%)
WMS Spatial Addition Subtest	3 (2.4%)
WMS Verbal Paired Associates Subtest	5 (4%)
WMS Visual Reproduction Subtest	19 (15.2%)
WMS Word List Subtest	2 (1.6%)
Woodcock-Johnson Tests of Achievement	2 (1.6%)

As table 2 demonstrates, there was a high degree of variability in terms of tests selected to evaluate the hypothetical client. The ten most popular tests were the Trail Making Test, the Controlled Oral Word Association Test, the Stroop Test, the Boston-Naming Test, the Rey-Osterrieth Complex Figure Test, the WMS Logical Memory subtest, the Beck Depression Inventory, the Clock Drawing Test, Grooved Pegboard, and the WAIS Vocabulary subtest. The popularity of these tests and others included in the table is consistent with previous surveys of neuropsychologists (Rabin et al., 2007) and demonstrates the continued dominance of the Wechsler family of tests.

Overall, there was little consistency in the tests selected by neuropsychologists in terms of total battery, although some tests were clearly more popular than others. This was partly due to the number of tests selected for inclusion in the battery, which ranged

from 5 to 25 tests. The magnitude of this variation was somewhat unexpected as the vignette stipulated only 3-4 hours of face-to-face testing time, which was included in order to mimic real-world testing conditions. In addition, respondents varied in their use of standard sets of tests, such as the full WAIS or WMS, versus selected subtests from these larger test collections, such as only the Vocabulary or Logical Memory subtest. However, the domains assessed by respondents did show substantial overlap, as one would expect. In general, respondents included tests that tap the following areas of cognitive performance: memory, intellectual ability, executive functioning, emotion regulation, and motor skills. Thus, the number and type of test used to assess these areas varied significantly from one respondent to another, but virtually all respondents chose tests that would evaluate the same five areas of cognitive and emotional functioning. About half of the respondents included a malingering measure in their battery as well.

A follow-up analysis was run to ascertain whether the inability to find a prototypical battery that was similar across individual neuropsychologists was due to individual preferences and training or due to the clinical population with which respondents commonly worked. To conduct this analysis, participants who indicated that they worked primarily with children and adolescents were excluded as these individuals are unlikely to be familiar with the assessment of dementia because it not commonly seen in younger age groups. This excluded 27 of the 125 respondents. The frequency of individual test selections and of batteries constructed was then reanalyzed to determine if a prototypical battery emerged when looking only at the 98 individuals who commonly assess adult and older adult populations. However, very little difference was found when the data were split according to patient population. Overall, the ten most popular tests

stayed the same when respondents who work primarily with children were excluded (in order of greatest to least popularity: Trail Making Test, the Controlled Oral Word Association Test, the Stroop Test, the Boston-Naming Test, the Rey-Osterrieth Complex Figure Test, the WMS Logical Memory subtest, the Clock Drawing Test, the WAIS Vocabulary subtest; Grooved Pegboard, and the Beck Depression Inventory). In addition, there was still little overlap in the testing batteries constructed by respondents as individual test batteries varied greatly in number of tests selected (range: 7-19) and the specific tests selected. The only difference observed was a tendency to favor selecting individual subtests from the WAIS-IV and WMS-IV rather than using the omnibus test as a whole. Therefore, it is likely that the difference in battery selection among neuropsychologist is due to some combination of test availability and individual clinician preferences and training. However, it should be noted that this analysis still demonstrated the common assessment of the five clinical domains of cognitive performance found in initial analysis of all respondents (i.e., memory, intellectual ability, executive functioning, emotion regulation, and motor skills) throughout the batteries selected to assess for possible dementia.

The checklist of tests included in the survey was taken from previous survey research and represented the most popular tests chosen. However, it was not feasible to include all possible tests; therefore, respondents were offered an opportunity to write in any test not included in the checklist. Five respondents used this section to explain their selection on the checklist and provide rationale for their selection. Another 49 respondents used this section to write-in a test not included on the checklist. The following tests were written-in by more than one respondent: Green's Word Memory

Test (3 respondents), Animal or Categorical Fluency (10 respondents), the Rivermead (3 respondents), WRAT word reading subtest only (4 respondents), and the North American Adult Reading Test (9 respondents). In addition, three respondents indicated in this section that they use an individually created measure, such as a specialized aphasia-screening tool.

Finally, a section containing non-psychometric clinical data routinely collected during neuropsychological evaluations was included, and respondents were asked to select any additional data they would collect in this case. Table 3 summarizes this checklist.

Table 3

Non-psychometric Data Collected During Dementia Evaluation

Type of Clinical Data	Number of Respondents
Behavior During Testing	123 (98.4%)
Family History	123 (98.4%)
Historical Data	125 (100%)
Interview with family/caretakers	121 (96.8%)
Clinical Interview	122 (97.6%)
Medical Record Review	105 (84%)

It is clear from this table that almost all respondents were in agreement as to the necessity of non-psychometric clinical data. This is consistent with the results of the psychometric data collected in that most respondents routinely collected information about the same domains of cognitive functioning.

Clinical Vignettes

The second section of the survey presented each respondent with two clinical vignettes: a reference vignette, which was invariant across respondents, and a test vignette that varied by age (young or old) and test performance (average, borderline, impaired). For this stage of analyses a between-subjects, factorial multivariate analysis of variance (MANOVA) was used to investigate the effects of age and test performance on diagnostic and confidence ratings. The dependent variables used in this analysis were deviation scores, which were calculated by subtracting the diagnostic and confidence ratings made for the reference vignette from the ratings made for the second test vignette (young or old at average, borderline, impaired range) for each participant. Thus, for each of the four ratings (dependent variables), the deviation scores were calculated using this formula: test vignette ratings-reference vignette ratings. Five responses were randomly eliminated in order to equalize the sample size in each group for the 6 clinical vignettes, leaving this stage of analyses to be completed using a sample of 120 respondents. This was done in response to a significant result of Levene's test and Box's test of equality of covariance, which represented a violation of the assumption of homogeneity of variance. However, according to the work of Tabachnick and Fidell (2007), a violation of this assumption can be overcome by equalizing group sample sizes. In addition, Pilli-Bartlett's trace was used as the test statistic for the MANOVA as it is most robust in the face of assumption violations (Bray & Maxwell, 1985).

Deviation scores were chosen as the unit measurement for the dependent variables in these analyses to help control within subject variance. All respondents received the same reference vignette, which portrayed a 59-year-old male with average test

performance. Their scores on this reference vignette were subtracted from their scores on the test vignette, which portrayed either a 48-year-old male or a 74 year-old male at one of 3 levels of performance (average, borderline, or impaired). Thus, positive deviation scores indicated higher ratings on the test vignette than the reference vignette, and negative scores indicate lower ratings on the test vignette. Deviation scores were calculated for both sets of diagnostic ratings and confidence ratings. Through the use of deviation scores, subject response style (e.g., a tendency to use or to avoid using the extreme ends of the rating scale) can be adequately controlled. Again, respondents were asked to make two diagnostic ratings: first, the likelihood of any neurological impairment and, second, the likelihood of a dementia. Then, respondents were asked to rate their confidence level for each diagnostic rating. Given that the reference vignette portrays a middle-aged male with average scores, both sets of diagnostic ratings were expected to be low. Analyses of the reference vignette indicated this to be the case, with an average rating of 2.9 (out of 10; standard deviation of 1.7) for the likelihood of any neurological impairment and an average rating of 2.2 (standard deviation of 1.2). This rating would correspond to little or no chance of impairment according to the diagnostic scale. In addition, confidence ratings were expected to be high for both ratings, which was the case. Confidence ratings for the likelihood of any neurological impairment had an average rating of 3.76 (out of 5, standard deviation of 0.76) and the confidence ratings for the likelihood of dementia averaged 3.93 (standard deviation of 0.7). This would correspond to a 'high' level of confidence in the subjective diagnostic ratings. Thus, the reference category appears to be an adequate basis for comparison. Table 4 summarizes the results of the primary MANOVA analysis. Using Pillai's trace, there was a significant

main effect of age on the deviation scores of diagnostic and confidence ratings, V = 0.31, $F(4, 111) = 12.26$, $p < 0.01$; a significant main effect of performance level, V = 0.68, $F(8, 224) = 14.41$, $p < 0.01$; and a significant effect of the interaction between age and performance level, V = 0.24, $F(8, 224) = 3.81$, $p < .05$.

Table 4

MANOVA Source of Variance Summary

Effect		Value	F	df	Error df	Sig.
Performance	Pillai's Trace	0.679	14.406	8	224	0.000
	Wilks' Lambda	0.361	18.466[a]	8	222	0.000
	Hotelling's Trace	1.663	22.865	8	220	0.000
	Roy's Largest Root	1.593	44.614[b]	8	112	0.000
Age	Pillai's Trace	0.306	12.258[a]	4	111	0.000
	Wilks' Lambda	0.694	12.258[a]	4	111	0.000
	Hotelling's Trace	0.442	12.258[a]	4	111	0.000
	Roy's Largest Root	0.442	12.258[a]	4	111	0.000
Performance * Age	Pillai's Trace	0.239	3.308	8	224	0.000
	Wilks' Lambda	0.765	3.981[a]	8	222	0.000
	Hotelling's Trace	0.302	4.152	8	220	0.000
	Roy's Largest Root	0.282	7.901[b]	8	112	0.000

a. Exact statistic
b. Upper bound on F that yields a lower bound on the significance level

However, separate univariate ANOVAs on the outcome variables revealed non-significant effects of age on both sets of confidence ratings, $F(1, 114) = 2.67$, $p > .05$ and $F(1, 114) = 3.58$, $p > .05$. In addition, univariate ANOVAs revealed a non-significant interaction between age and performance level on the diagnostic ratings, $F(2, 114) = 2.52$, $p > .05$ and $F(2, 114) = 0.67$, $p > .05$. Turning to the significant main effects in the univariate ANOVAs, there was a significant main effect for age on both sets of

diagnostic ratings: any impairment $F(1,114) = 4.53$, $p < .05$ and dementia $F(1,114) = 39.55$, $p < .05$. This finding suggests that the age of the client in the clinical vignette impacted the diagnostic ratings, but not the confidence in these ratings. The main effect of performance level was found to be significant across all four dependent variables: the two diagnostic ratings, any impairment $F(2, 114) = 44.32$, $p < .05$, and dementia $F(2,114) = 74.53$, $p < .05$; as well as the two confidence ratings, $F(2,114) = 3.55$, $p < .05$ and $F(2,114) = 3.31$, $p < .05$. Thus, the performance level in the clinical vignette impacted both the diagnostic ratings and the confidence in those ratings. Table 5 summarizes the results of the univariate ANOVAs.

Table 5

Univariate ANOVAs: Source of Variance Summary

Source	Dependent Variable	Sum of Squares	df	Mean Square	F	Sig.	Part. Eta Sqrd.	Obs. Power
Perform Level	1. Any Impair	549.017	2	274.5	44.319	0.000**	0.437	1.000
	2. Any Impair Confidence	2.217	2	1.1	3.549	0.032*	0.059	0.650
	3. Dementia	583.800	2	291.9	74.528	0.000**	0.567	1.000
	4. Dementia Confidence	2.217	2	1.1	3.308	0.040*	0.055	0.617
Age	1. Any Impair	28.033	1	28.03	4.526	0.036*	0.038	0.559
	2. Any Impair Confidence	0.833	1	0.83	2.669	0.105	0.023	0.367
	3. Dementia	154.133	1	154.1	39.353	0.000**	0.257	1.000
	4. Dementia Confidence	1.200	1	1.2	3.581	0.061	0.030	0.467
Perform * Age	1. Any Impair	31.217	2	15.6	2.520	0.085	0.042	0.496
	2. Any Impair Confidence	3.317	2	1.6	5.310	0.006**	0.085	0.829
	3. Dementia	5.267	2	2.6	0.672	0.513	0.012	0.161
	4. Dementia Confidence	5.850	2	2.9	8.729	0.000**	0.133	0.967
Error	1. Any Impair	706.100	114	6.1				
	2. Any Impair Confidence	35.600	114	0.3				
	3. Dementia	446.500	114	3.9				
	4. Dementia Confidence	38.200	114	0.3				
Total (Corrected Total)	1. Any Impair	3348.000 (1314.367)	120 (119)					
	2. Any Impair Confidence	42.000 (41.467)	120 (119)					
	3. Dementia	3158.000 (1189.700)	120 (119)					
	4. Dementia Confidence	54.000 (47.467)	120 (119)					

An analysis of the interaction should help clarify these statistical findings. The univariate ANOVAs show a significant interaction between age and performance level on the confidence ratings, $F(2,114) = 5.31$, $p < .05$ and $F(2, 114) = 8.73$, $p < .05$. However, as mentioned previously, there was a non-significant interaction of age and performance level on the diagnostic ratings. Figures 1-4 illustrate the interaction for each of the four dependent variables.

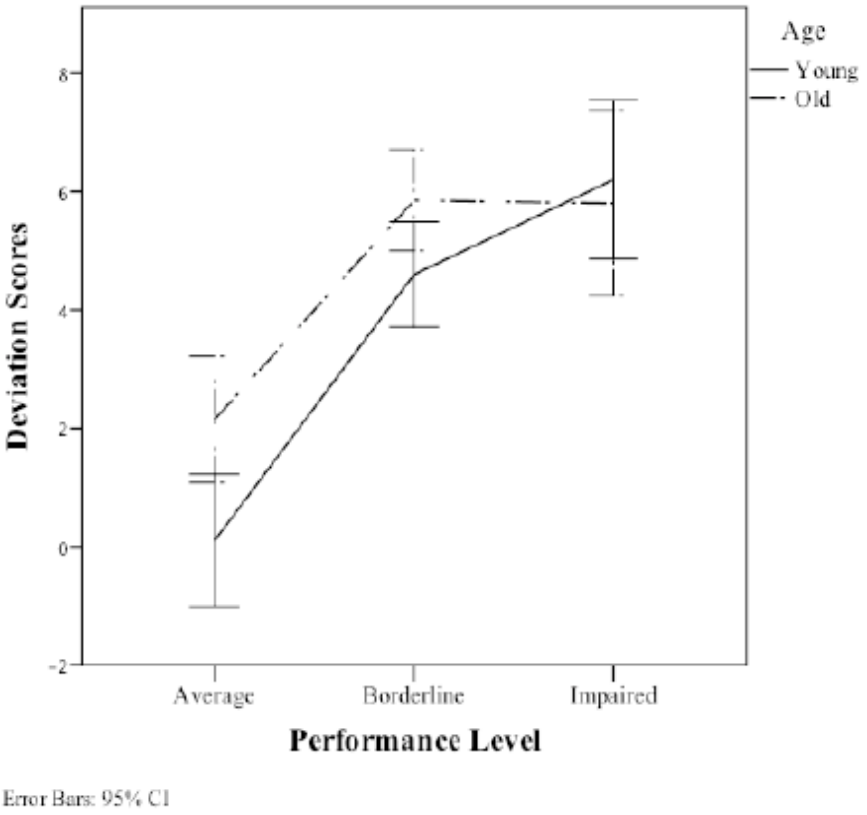

Figure 1. Diagnostic ratings for the likelihood of any neurological impairment.

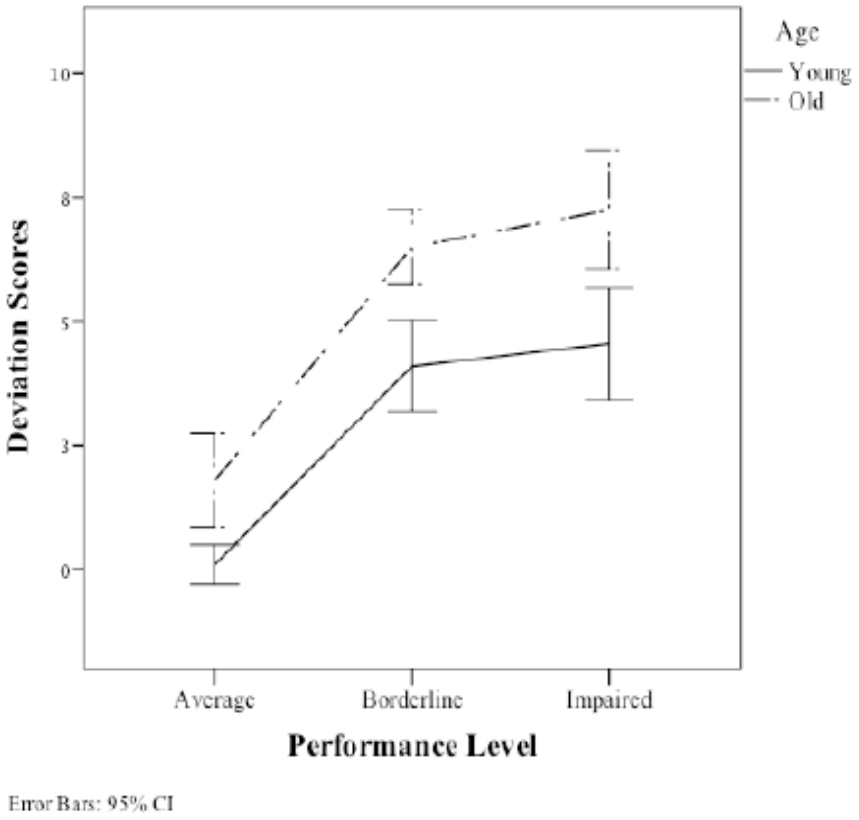

Figure 2. Diagnostic ratings for the likelihood of dementia.

As displayed by these two figures, the interaction between age and performance level was not significant for the two diagnostic ratings. Although the lines appear to cross in Figure 1, this does not indicate a significant interaction as evidenced by the almost entirely overlapping error bars. These figures also demonstrate the main effects of age and level of performance on the diagnostic ratings for each of the clinical vignettes. Figure 1 demonstrates that the older client received a slightly higher rating for the likelihood of any neurological impairment; however, overall the ratings for the older client versus the younger client did not significantly differ for this diagnostic rating category. Furthermore, the main effect of performance level is demonstrated in this figure, as diagnostic ratings tended to increase between the average and borderline levels of performance before leveling off between the borderline and impaired levels of

performance. As shown by Figure 2, the ratings for the likelihood of dementia followed a somewhat different pattern. Again, there is no interaction between age and level of performance in the diagnostic ratings for dementia. However, the main effect of age is shown to have a clear effect, with the older client receiving significantly higher ratings across all three levels of performance. Given base rates for cognitive impairment, which follows a linear pattern, these results are entirely expected. In addition, the main effect of performance level follows the expected pattern, with average levels of performance receiving the lowest ratings and impaired performance receiving the highest ratings.

These figures demonstrate both the expected main effect of performance level, with diagnostic ratings increasing as impairment level increases, and the more complex main effect of age. Although there was no main effect for age in the diagnostic ratings for the likelihood of any neurological impairment, this was not the case for the likelihood of dementia diagnostic rating. In this case, the age of the client presented in the vignette had a significant effect, with the older client receiving higher diagnostic ratings at each level of performance.

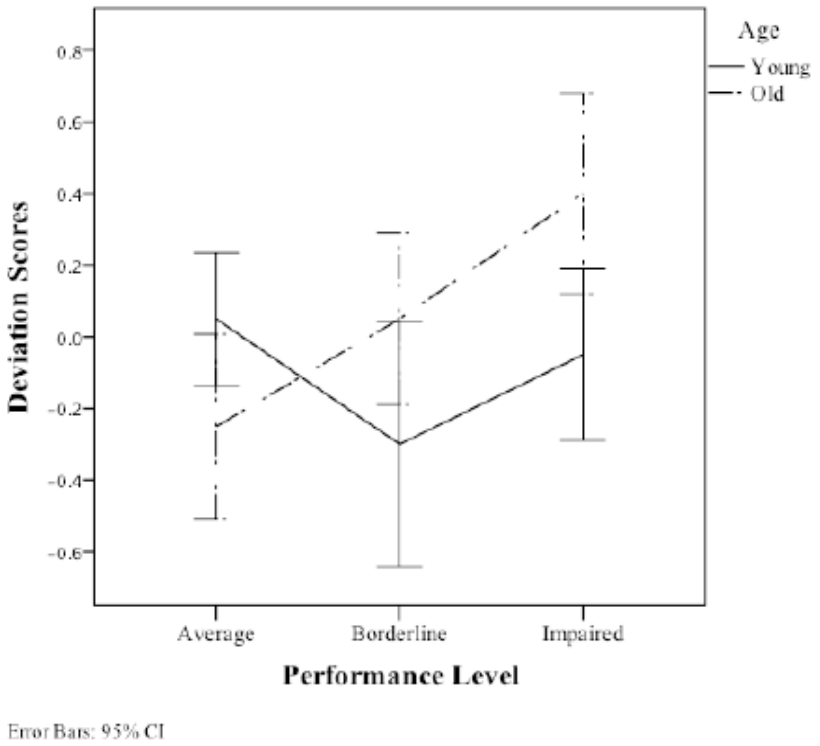

Figure 3. Confidence ratings for the likelihood of any neurological impairment.

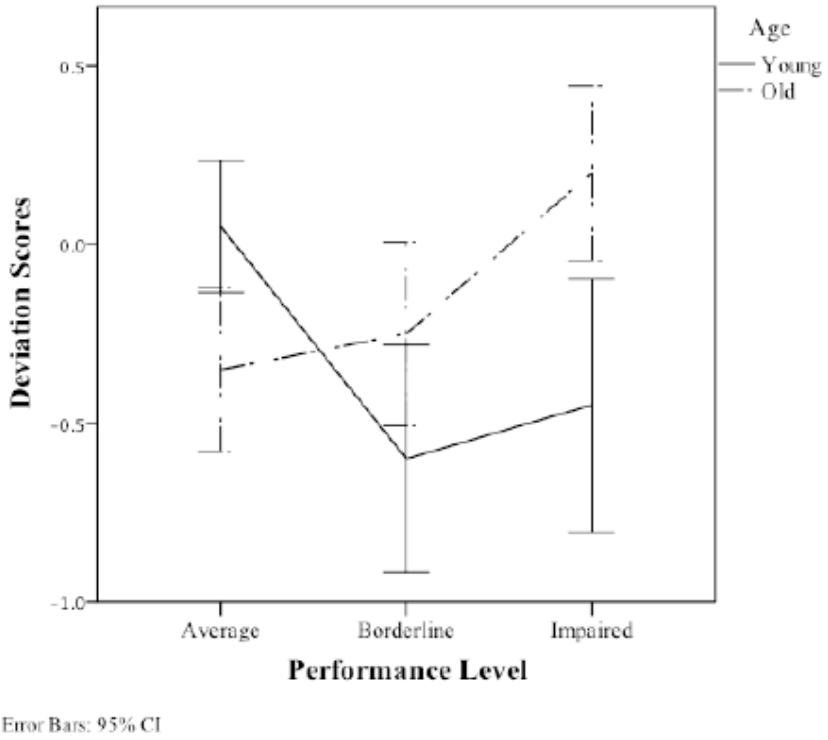

Figure 4. Confidence ratings for the likelihood of dementia.

Figures 3 and 4 display the interaction between age and level of performance for the confidence ratings. It was hypothesized that confidence ratings would follow a pattern in which the ratings were higher for the average and impaired categories and lower for the borderline category. As reported earlier in this section, there is a significant interaction between age and level of performance with regard to the confidence ratings. Thus, the main effect of performance impacted the confidence ratings differently for the old client versus the young client. As can be seen on Figures 3 and 4, the ratings for the young client followed the predicted pattern, higher for the average and impaired conditions—situations in which the clinical data are clear—and lower in the borderline condition when the clinical data do not have an easy interpretation. Although this pattern is less clear when looking at the ratings for the older client, the error bars overlap significantly in both Figures 3 and 4, indicating that the confidence ratings for the young client versus the old client were not substantially different, with the exception of the confidence ratings on the likelihood of dementia for the impaired level of performance vignettes. In this case, respondents were significantly more confident in their dementia diagnostic rating for the old client at the impaired level of performance than they were in their rating for the young client at the same impaired level of performance. In general, respondents became more confident in their diagnostic ratings for the older client as the level of impairment increased.

In addition to univariate ANOVAs, post-hoc Bonferroni pair-wise comparisons were run to analyze the simple main effects using the Bonferroni correction to control for the family-wise error rate. The follow-up analysis examined the deviation scores for each of the six groups (young (Y) or old (O) client, 3 levels of impairment: average, borderline

or impaired).Thus, the respondents were broken into six groups representing the test vignette they rated: the young client with average scores (Ya); the young client with borderline scores (Yb); the young client with impaired scores (Yi); the old client with average scores (Oa); the old client with borderline scores (Ob); and the old client with impaired scores (Oi). Table 6 summarizes these abbreviations.

Table 6

Vignette Abbreviations

Abbreviation	Description
Ya	Young client with average scores
Yb	Young client with borderline scores
Yi	Young client with impaired scores
Oa	Old client with average scores
Ob	Old client with borderline scores
Oi	Old client with impaired scores

For the diagnostic ratings, it was predicted that the following group comparisons would be found significant: Ya/Yb, Ya/Yi, Oa/Ob, Oa/Oi, Yb/Ob. As displayed in Table 7, these predictions were accurate for both diagnostic ratings, with the exception of the Yb/Ob comparison for the diagnostic rating of the likelihood of any neurological impairment.

Table 7

Summary of Bonferroni Pairwise Comparisons

Bonferroni Comparisons		Dependent Variables Mean Difference I-J (p-value)			
Group (I)	*Group (J)*	*Any Impairment Rating* Std. error=0.787	*Any Impairment Rating-Confidence* Std. error=0.177	*Dementia Rating* Std. error=0.626	*Dementia Rating-Confidence* Std. error=0.183
Ya	Yb	-4.60 (0.000)*	0.35 (0.751)	-4.00 (0.000)*	0.650 (008)*
	Yi	-6.10 (0.000)*	0.10 (1.000)	-4.45 (0.000)*	0.50 (0.110)
	Oa	-2.05 (0.156)	0.30 (1.000)	-1.70 (0.114)	0.40 (0.464)
	Ob	-5.75 (0.000)*	0.00 (1.000)	-6.40 (0.000)*	0.30 (1.000)
	Oi	-5.70 (0.000)*	-0.35 (0.751)	-7.15 (0.000)*	-0.15 (1.000)
Yb	Ya	4.50 (0.000)*	-0.35 (0.751)	4.00 (0.000)*	-0.65 (0.008)*
	Yi	-1.60 (0.667)	-0.25 (1.000)	-0.45 (1.000)	-0.15 (1.000)
	Oa	2.45 (0.035)*	-0.05 (1.000)	2.30 (0.005)*	-0.25 (1.000)
	Ob	-1.25 (1.000)	-0.35 (0.751)	-2.40 (0.003)*	-0.35 (0.876)
	Oi	-1.20 (1.000)	-0.70 (0.002)*	-3.15 (0.000)*	-0.80 (0.000)*
Yi	Ya	6.10 (0.000)*	-0.10 (1.000)	4.45 (0.000)*	-0.50 (0.110)
	Yb	1.60 (0.667)	0.25 (1.000)	0.45 (1.000)	0.15 (1.000)
	Oa	4.05 (0.000)*	0.20 (1.000)	2.75 (0.000)*	-0.10 (1.000)
	Ob	0.35 (1.000)	-0.10 (1.000)	-1.95 (0.035)*	-0.20 (1.000)
	Oi	0.40 (1.000)	-0.45 (0.183)	-2.70 (0.001)*	-0.65 (0.008)*
Oa	Ya	2.05 (0.156)	-0.30 (1.000)	**1.70 (0.114)**	-0.40 (0.464)
	Yb	-2.45 (0.035)*	0.05 (1.000)	-2.30 (0.005)*	0.25 (1.000)
	Yi	-4.05 (0.000)*	-0.20 (1.000)	-2.75 (0.000)*	0.10 (1.000)
	Ob	-3.70 (0.000)*	-0.30 (1.000)	-4.70 (0.000)*	-0.10 (1.000)
	Oi	-3.65 (0.000)*	-0.65 (0.005)*	-5.45 (0.000)*	-0.55 (0.049)*
Ob	Ya	5.75 (0.000)*	0.00 (1.000)	6.40 (0.000)*	-0.30 (1.000)
	Yb	1.25 (1.000)	0.35 (0.751)	2.40 (0.003)*	0.35 (0.876)
	Yi	-0.35 (1.000)	0.10 (1.000)	1.95 (0.035)*	0.20 (1.000)
	Oa	3.70 (0.000)*	0.30 (1.000)	4.70 (0.000)*	0.10 (1.000)
	Oi	0.05 (1.000)	-0.35 (0.751)	-0.75 (1.000)	-0.45 (0.232)
Oi	Ya	5.70 (0.000)*	0.35 (0.751)	7.15 (0.000)*	0.15 (1.000)
	Yb	1.20 (1.000)	0.70 (0.002)*	3.15 (0.000)*	0.80 (0.000)*
	Yi	-0.40 (1.000)	0.45 (0.183)	2.70 (0.001)*	0.65 (0.008)*
	Oa	3.65 (0.000)*	0.65 (0.005)*	5.45 (0.000)*	0.55 (0.049)*
	Ob	-0.05 (1.000)	0.35 (0.751)	0.75 (1.000)	0.45 (0.232)

* denotes significant difference ($p < 0.05$)

The comparison of Ya and Oa groups was expected to be non-significant as this would replicate previous findings in which there was an absence of a diagnostic age bias

when test results are within the average range (Garb & Boyle, 2003). As displayed in Table 7, this prediction was also found to be accurate; however, the graphed interaction shows these groups to be significantly different with regard to the likelihood of dementia diagnostic rating with the older client receiving higher ratings at each level of performance. The Bonferroni correction tends to be very conservative as it is primarily used to control for Type I error rate. Thus, the conservative statistic used in these post-hoc analyses can help explain this discrepancy as the Ya/Oa comparison approaches significance (which can be seen in Table 7). To clarify this finding, a second post-hoc analysis was run for only this comparison using a less conservative test, the Games-Howell. It should be noted that this second, less conservative, test was only run to clarify the discrepancy between the significant difference shown in Figure 2 and the non-significant Bonferroni pairwise comparison. If there had not been a discrepancy between these two test findings, a less conservative test would not have been run on this pairwise comparison; however, it seemed to prudent to run a final analysis in order to make sense of these contradictory findings. Using this test statistic, the comparison between the Ya group and the Oa group was significant (mean difference 1.20, $p = 0.020$), indicating that these groups were significantly different from one another. Despite presenting average scores for both groups, the client in the Oa vignette was rated to have a higher likelihood of dementia than the client in the Ya group. However, there was no significant difference in the ratings for the presence of any neurological impairments between the Ya client and the Oa client in either the Bonferroni comparison (as shown in Table 7) or the Games-Howell comparison (mean difference 2.05, $p= 0.085$).

The Bonferroni comparisons also help illuminate the differences in the confidence ratings for each test vignette. As previously discussed, there was a significant difference between the confidence ratings for the Yi client and Oi client with regard to the dementia diagnostic rating. In addition, the pair-wise comparisons show that respondents were significantly more confident in both of their diagnostic ratings for the old client at the impaired performance level than they were for the average performance level. This finding is contrary to both the hypothesized findings and the findings for the young client, in which the confidence ratings are nearly the same for the average and impaired levels of performance. It was hypothesized that the confidence ratings would be nearly the same in the average and impaired condition due to the unambiguous nature of the clinical data. Thus, it is interesting that respondents, when rating the older client, felt more confident diagnosing impairment when it was present than not diagnosing impairment when it was not present, as both judgments would be considered accurate. It is possible that knowledge of base rates contributed to this phenomenon.

Finally, Table 8 summarizes the average rating in each the 6 vignette groups and the percentage of respondents diagnosing a neurological impairment or dementia. As a reminder, respondents were told that a rating of 0, 1, or 2 indicated that a diagnosis would not be made; a rating of 5 indicated that there was a 50/50 chance of a diagnosable impairment; and that a rating of 8, 9, or 10 indicated that diagnostic criteria for impairment were met.

Table 8

Average Ratings and Percentage of Respondents Diagnosing Impairment

		Younger Client		Older Client	
Level of Performance	Diagnostic Rating	Mean Rating (SD)	% of respondents diagnosing	Mean Rating (SD)	% of respondents diagnosing
Average	Any Imp.	2.65 (1.46)	0%	5.52 (1.86)	4.8%
	Dementia	1.95 (0.76)	0%	4.52 (2.06)	4.8%
Borderline	Any Imp.	7.64 (1.62)	40.9%	8.52 (1.21)	52.4%
	Dementia	5.95 (1.84)	4.8%	8.57 (1.36)	66.7%
Impaired	Any Imp.	8.67 (2.13)	85.7%	9.25 (1.52)	75%
	Dementia	6.71 (2.17)	28.6%	9.65 (1.66)	80%

As shown in this table, respondents indicated, on average, that the older individual with average test scores had a 50% chance of both neurological impairment or dementia with an average rating of 5.52 and 4.52 respectively. However, the younger individual with average test scores had a mean rating of 2.65 for the presence of any neurological impairment and a mean rating of 1.95 for the presence of dementia, indicating that a diagnosis of impairment or dementia was very unlikely. For both the young and old client, a diagnosis (i.e rating of 8, 9, or 10) was infrequent; only one respondent made a diagnosis for the older client, and no respondents diagnosed the younger client with a cognitive impairment. Despite the infrequency of a diagnosis, the difference in the mean ratings for the young and old client suggests that ratings were not equally accurate for each client. Given that the test scores for both the young and old

client were taken from the average range of the standardized normative tables for the client's respective age, this difference in ratings indicates that respondents were less accurate in their diagnostic ratings for the older client, as the presence of a cognitive impairment is unlikely because each client performed in the average range for their age.

Furthermore, it was expected that the borderline test score vignettes would yield average ratings around five, indicating a 50/50 chance of a cognitive impairment, as these scores were taken from the normative table around one to one and half standard deviations below the mean. Although these scores indicate some level of impairment, they do not meet the typical standards for diagnosis, which indicate that a diagnosis should be made once impairment reaches two standard deviations below expected performance (i.e. the average range; Lezak, Howieson, & Loring, 2004). However, the mean rating for the older client at this level of performance was 8.52 for the presence of any neurological impairment and 8.57 for the presence of dementia. This indicates that, on average, respondents chose to diagnosis the older client with a cognitive impairment despite the borderline test scores. For the younger client, respondents were much more conservative in their ratings: the mean rating for any impairment was 7.64 and 5.95 for the presence of dementia. One half to two-thirds of respondents diagnosed the older client with a neurological impairment or dementia, respectively. Although 41% of respondents diagnosed the younger client with a neurological impairment, only one respondent diagnosed the younger client with dementia. These differential ratings suggest that respondents were using a lower threshold of cognitive impairment to diagnosis the older client with a cognitive impairment than was used to evaluate the younger client. In addition, respondents were more likely to diagnosis the older client with a *specific*

condition, such as dementia, than a *general* condition, such as any neurological impairment despite the fact that, statistically speaking, a general neurological impairment is much more common than a specific condition like dementia. This tendency was only seen in the diagnostic ratings for the older client.

For the impaired performance vignettes, it was expected that most respondents would diagnose the client, young or old, with a cognitive impairment and, thus, mean ratings would be around nine. This expectation held for both sets of diagnostic ratings for the older client, with mean ratings of 9.25 for the presence of any neurological impairment and mean ratings 9.65 for the presence of dementia. The ratings for the younger client, however, varied more than expected. The mean rating for the young client was 8.67 for the presence of any neurological impairment, but the mean rating for the presence of dementia was only 6.71. Less than a third of respondents diagnosed the younger client with dementia, despite 85.7% indicating that a diagnosable neurological impairment was present. This finding again indicates a difference in diagnostic thresholds in which performance falling two standard deviations below expected performance was not sufficient for the diagnosis of dementia in the younger client, but the same level of performance was seen as evidence of dementia in the older client. Again, respondents were slightly more likely to diagnosis the older client with the specific condition of dementia rather than the more general condition of a neurological impairment. This tendency was not observed in the ratings for the younger client.

Given this difference in diagnostic thresholds, an exploratory analysis was conducted to evaluate the possibility of experience-based cohort effects in diagnostic accuracy. The rationale for this analysis was based on the idea that differences in

diagnostic thresholds may reflect an evolving understanding in the field regarding the presence of mild cognitive impairment. This disorder is included in the ICD-10 taxonomy (World Health Organization, 2004) and has now been added to the DSM-5 (America Psychiatric Association, 2013). Therefore, younger clinicians may have more awareness of this condition, and thus, experience may explain the lower diagnostic threshold found in this study. A dummy coded variable was added to the analysis measuring 'high' or 'low' experience using a median distribution split. Thus, respondents with 15 or more years of experience providing neuropsychological services were coded in the 'high' experience group and respondents with fewer than 15 years of experience were coded in the 'low' experience group. Diagnostic frequencies for each group were compared to evaluate the effect of experience level on the diagnostic accuracy for both neurological impairment and dementia in each of the six clinical vignette groups. The results of this analysis were inconclusive as the level of experience did not produce a reliable effect. Overall, both the high-experience and low-experience respondents had similar rates of diagnosis. However, there were contradictory differences between vignette groups. For example, in the vignette group that rated the young client with borderline performance, respondents with less experience were more accurate; however, in the vignette that rated the older client with borderline performance, respondents with more experience were more accurate. In addition, the individual vignette groups are small (20 respondents per vignette), so splitting these groups in half by experience significantly limited the power of these analyses. Thus, it can be concluded that level of experience did not significantly impact diagnostic accuracy.

Utility of Clinical Information

After viewing the second test vignette, each respondent was asked to rate the perceived necessity of each piece of psychometric information in determining his or her diagnostic ratings for the previous vignette (the test vignette). In the final stage of analysis, the ratings of the necessity of clinical information were analyzed by comparing both the average ratings made by neuropsychologists who viewed the same vignette (i.e. average scores for the 48-year-old) and the average ratings made by neuropsychologists across the various levels of performance for a single client age (i.e., the average, borderline and impaired levels for the 48-year-old). First, survey responses were compared within each group (e.g., Ya or Yi) to investigate the perceived necessity of information when the same referral and psychometric information was viewed. In this stage of analysis, missing values were replaced with the arithmetic mean (Field, 2009); in total, 6 missing ratings were replaced by the mean for that piece of clinical information. Also, the same equalized sample used in the case vignette MANOVA analysis was used for this stage of analysis, resulting in a sample of 120 respondents. A between subjects MANOVA, this time using test vignette group membership as the independent variable (6 levels) and the perceived necessity ratings as the dependent variables (19 ratings), did not reveal any significant differences between the six groups in terms of the perceived necessity of clinical information (Pillai's Trace: $V = 0.756$, $F(95, 500) = 0.937$, $p > 0.05$). A follow-up analysis using univariate ANOVAs did not reveal any significant group differences for any of the 19 psychometric test results given in the clinical vignette.

Similarly, a between subjects, factorial MANOVA, using age and performance level as independent variables, did not find any significant differences between the

groups in terms of perceived necessity of clinical information. Thus, there was no main effect for age (Pillai's V = 0.157, $F(19, 96) = 0.943$, $p > 0.05$), no main effect of level of performance (Pillai's V = 0.291, $F(38, 194) = 0.870$, $p > 0.05$), and no interaction between age and level of performance (Pillai's V = 0.320, $F(38, 194) = 0.973$, $p > 0.05$). Follow-up analysis with univariate ANOVAs revealed only one significant difference for the level of impairment on one specific piece of clinical information, the Verbal Paired Associates subtest score. However, given the number of dependent variables used in this stage of analysis, it is possible that this is a spurious finding due to an inflated family wise error rate. A follow-up analysis using a Bonferroni correction to compensate for this inflated error rate revealed that this result did not meet significance at the corrected critical value ($\alpha = .00263$).

It was hypothesized that the necessity of clinical information would vary as a function of level of impairment, in which more severely impaired clinical findings would be rated as more necessary for making diagnostic decisions for both young and old clients. This hypothesis is not supported. Overall, frequency data reveals that the average rating for each piece of clinical information ranged from 2.5 (out of 4) to 3.75, indicating a tendency for respondents to rate each piece of clinical information as either "somewhat necessary" or "very necessary." Furthermore, statistical analysis reveals that these rankings did not vary as a function of the clinical test vignette presented. Table 9 displays the overall average rating for each piece of clinical information.

Table 9

Perceived Necessity of Clinical Information

Test Name	Mean	SD
WAIS-IV: Index Scores	3.68	0.550
WMS-IV: Index Scores	3.74	0.537
WMS-IV: Visual Reproduction	2.67	0.803
WMS-IV: Logical Memory	2.80	0.856
WMS-IV: Verbal Paired Assoc.	2.60	0.793
WMS-IV: Designs	2.51	0.830
WMS-IV: Spatial Addition	2.41	0.794
WMS-IV: Spatial Span	2.45	0.818
WMS-IV: Digit Span	2.74	0.835
Boston Naming Test	3.21	0.593
Controlled Oral Word Association	3.39	0.652
Trail-Making Test: Part A	3.43	0.560
Trail-Making Test: Part B	3.46	0.533
Hooper Visual Organization	2.48	0.850
Brief Visual Memory	2.66	0.835
Hopkins Verbal Learning	3.01	0.761
Wisconsin Card Sort	3.06	0.665
Finger Tapping	2.61	0.677
Grooved Pegboard	2.64	0.658

a. WAIS-IV: Wechsler Adult Intelligence Scale, 4th Edition
b. WMS-IV: Wechsler Memory Scale, 4th Edition
c. Ratings were made on a 4-point Likert scale (very unnecessary, somewhat unnecessary, somewhat necessary, and very necessary).

CHAPTER VI

DISCUSSION

The present study had three major objectives: (1) to evaluate the similarity of tests used by neuropsychologists to evaluate dementia, (2) to investigate a possible age-bias in neuropsychological diagnosis, and (3) to gain insight into the perceived necessity of common clinical information. These objectives were accomplished using a survey mechanism to deliver the experimentally manipulated vignettes. The online, four-part survey was completed by 125 INS members who are clinical psychologists currently offering neuropsychological assessment services. This chapter presents a discussion of the survey findings and is comprised of three sections. The first section will provide summary and interpretation of the study's findings. Included in this section is a brief description of respondent characteristics, followed by exploration of the results of hypothesis testing for each survey section with an emphasis on comparison with the empirical literature. The study's strengths will be highlighted throughout this section. In the second section, the study's perceived limitations will be explored and suggestions for further research will be offered. The final section will focus on the study's conclusions and implications for the practice of clinical psychology.

Summary and Interpretation of Study Findings

Respondent Characteristics

The respondents were 125 clinical psychologists currently practicing in the United States or Canada with an INS membership affiliation. The average respondent was 49 years old, held a doctoral degree in clinical psychology, and had been practicing neuropsychological assessment for 16 years. The percentage of women in this study

(38%) exceeds that of some earlier studies (Sweet et al., 2000), and is consistent with more recent surveys (Rabin, Barr & Burton, 2005). Previous researchers have attributed this increase in female response to clinical practice surveys to the growth in female participation in the practice of neuropsychology as a field (Rabin, 2001). However, the percentage of women in neuropsychology (46%; Hilsabeck & Martin, 2010) is still far less than the number of women receiving new doctoral degrees in clinical psychology (75%; APA, 2010). The respondents indicated working with a wide-variety of clinical populations, although the majority of respondents indicated working primarily with one clinical population (e.g. children, adults). However, no specific clinical population was more represented than another; thus, this study included a relatively equal number of professionals working with children, adolescents, adults, and older adults.

Overall, the respondents reported spending 53% of their professional time engaged in neuropsychological assessment, along with other professional activities. In addition, the majority of respondents (73%) indicated that they performed 1-15 neuropsychological assessments per month, with an average of 15 instruments per evaluation. Consistent with previous survey research, the majority of respondents also endorsed using a flexible battery approach to select evaluation procedures (e.g., Rabin, Barr & Burton, 2005). In terms of professional setting, 56% of participants reported working in either a medical or psychiatric hospital while 34% indicated that they primarily work in private or group practice. This finding was somewhat surprising given that previous studies (Rabin, Borgos & Saykin, 2008; Sweet et al., 2000) have found a trend toward more private sector employment among neuropsychologists, which was attributed to dwindling opportunities in hospital settings. Thus, the included respondents

in this study may not accurately represent the work setting distribution of neuropsychologists in general. This finding may impact the generalizability of study results as the respondents in this study may not be representative of neuropsychologists in general. Therefore study results should be interpreted cautiously in light of this fact.

Dementia Battery Construction

The first major goal of this study was to investigate the similarity amongst neuropsychological test batteries used to assess for the presence of dementia. Recent practice trends reveal that neuropsychologists prefer the use of flexible batteries (Sweet, Moberg, & Suchy, 2000a), yet no studies have investigated the degree of similarity between the flexible batteries used to assess specific types of clients (e.g., those with a head injury, the elderly). Standardized assessment procedures are crucial to the continued advancement of the field; however, with the advent of new neuropsychological tests, clinicians have abandoned standardized batteries in favor of a more flexible approach. Thus, the question posed by this study was whether or not this flexible battery approach used by individual neuropsychologists yielded a similar test battery.

Results of this survey expand upon previous survey research in several ways. First, there have been several surveys of practicing psychologists that have ascertained which tests are most commonly used in clinical practice (e.g. Slick et. al., 2004; Sweet et al, 2000;), but only one previous study (Rabin, Barr & Burton, 2005) looked at battery selection or the specific measures to assess for particular neuropsychological problem. That study differed from the current survey in that it specified the individual cognitive domains, such as memory or attention, and asked neuropsychologists which tests they would use to assess functioning in this domain. The current study strove to replicate

neuropsychological practice by presenting respondents with a typical referral and then asking them to specify the tests they would chose to assess this patient's overall functioning. Second, this survey verified the expanding popularity of flexible batteries. An overwhelming majority of participants in this study indicated that they use the flexible battery approach which, when coupled with previous research in this area (i.e. Sweet et al., 2000), suggests the near dominance of this approach to neuropsychological testing. Thus, it appears that the standardized batteries used during the early years of neuropsychological practice have been largely abandoned, with a few selected subtests being retained for use with new assessment combinations. However, despite the popularity of flexible batteries, a number of tests continue to be used by the vast majority of neuropsychologists, indicating that there is still a level of standardization in the field.

Third, this survey found that tests chosen for inclusion in each individual neuropsychologist's flexible battery varied widely in several categories. It was hypothesized that there would be substantial overlap among neuropsychologists in terms of the tests selected to assess for the presence of dementia, such that a prototypical battery would emerge from clinician consensus. However, this hypothesis was not supported. Neuropsychologists differed on the number of tests used, the specific measures selected, and whether they selected individual tests versus comprehensive scales, such as the WAIS or WMS. More specifically, it was hypothesized that the Wechsler Memory Scale (WMS), The Controlled Oral Word Association Test (COWAT), Trails, the Wisconsin Card Sort Test (WCST), finger-tapping, and grooved pegboard would be the most popular tests selected by neuropsychologists. This hypothesis was partially supported; the COWAT, Trails, and grooved pegboard were

among the top ten tests selected by neuropsychologists. However, most respondents indicated a preference for using selected subtests from the WMS, such as Logical Memory, rather than the entire scale.

Although there was little consistency between the batteries selected by respondents, there were general similarities in the domains assessed. In general, respondents included tests that tap the following areas of cognitive performance: memory, intellectual ability, executive functioning, emotion regulation and motor skills. They also tended to include multiple measures to evaluate certain critical cognitive areas, like memory and executive functioning, highlighting the importance of understanding a client's functioning in these areas when making a dementia diagnosis. This likely reflects the multifaceted nature of many cognitive areas. For instance, memory functioning is comprised of many constituent parts such as verbal memory, nonverbal memory, immediate recall, delayed recall, and recognition. Executive functioning often includes measurement of planning, organization, and inhibition abilities as part of the overall construct. Thus, although there does not appear to be consensus among neuropsychologists regarding the specific tests in include in a dementia battery, there is agreement regarding the critical functional domains that must be assessed. Additionally, multiple measurements using a variety of tasks in order to adequately capture the functioning of individual cognitive domains is recommended in the neuropsychological literature (Lezak, Howieson, & Loring, 2004) and this finding reflects that these recommendations have been incorporated into neuropsychological training and practice. Previous studies (Rabin, Barr & Burton, 2005; Sweet et al., 2000) have criticized neuropsychology for its failure to regularly assess malingering and the results of this

survey indicate that only about half of respondents included a malingering measurement in their battery.

Age Bias and the Accuracy of Diagnosis

The second section of this survey involved the use of clinical vignettes to study the accuracy of clinical diagnoses and explore the possibility of an age bias in the diagnosis of neurological impairment and dementia. This section of the survey was an expanded replication of Garb and Boyle's (2003) study, which found no age bias when neuropsychologists were presented with average scores for an old and young client. In that survey 25 board-certified neuropsychologists were asked to provide diagnostic ratings, on the same scale used in this study, for an old and young client using scores taken from the average range of the normative table. The authors concluded that, although there was as statistically significant effect for age, in which the old client was rated higher on the diagnostic scales, there was no age bias (Garb & Boyle, 2003). This conclusion is based on the idea that diagnoses are considered to be 'biased' when the accuracy of the judgments varies for different groups of people (Garb, 1997; Widiger & Spitzer, 1991). In this case, a diagnosis can be classified as demonstrating an age bias if the accuracy of diagnosis varies as a function of age. Thus, one would have to conclude that a bias is present if the ratings made for the younger client were more accurate than the ratings made for the older client. In the Garb and Boyle (2003) study, age bias did not occur because none of the participants made a diagnosis of impairment or dementia (a rating of 8, 9, or 10 on the diagnostic scale). In their study the mean rating for the older client was 2.72 for any neurological impairment and 2.04 for the presence of dementia. It might be argued that these ratings were inaccurate as they clearly indicate that a diagnosis

would not be made. The authors further argue that the statistically significant effect of age is an indication that neuropsychologists were adequately attending to base rates as neurological impairment and dementia are more likely with increased age (Garb & Boyle, 2003).

However, the results of the current study differ from these findings in several areas. First, this study not only sought to replicate the Garb and Boyle (2003) finding, but to expand the exploration of the conditions in which an age bias might occur. According to classic research in the area of clinical judgment and cognitive heuristics, biased decision-making is likely to occur in situations where two conditions are met: when the information provided is ambiguous and when the decision must be made quickly (Tversky & Kahneman, 1974). In these types of situations, people often rely on cognitive heuristics in order to make rapid decisions, thereby increasing the likelihood of inaccurate judgments. This study attempted to meet these conditions and increase the likelihood of biased decision-making in two ways. One, three levels of performance were used instead of only one: an average, borderline, and impaired level. The average and impaired performance categories provide unambiguous information as the individual is portrayed as either cognitively intact (scores taken from 0.5 standard deviations above or below the mean) or clearly impaired (scores taken from 2 or more standard deviations below the mean). Thus, the borderline performance category meets the first condition for biased decision-making by providing clinical data without a clear interpretation (scores taken from 1 to 1.5 standard deviations below the mean). In addition, in the Garb and Boyle (2003) study, participants were given a longer survey and were paid for their time in an effort to ensure that they spent adequate time making their ratings. In this study,

participants were asked to take a shorter survey (average completion time was between 10 and 15 minutes) and were not paid for participation, thus increasing the chances that they would feel the need to make decisions quickly. However, it should be noted that this was not an explicit manipulation of the study. The survey itself was untimed, thus respondents could take all the time they wanted to complete the diagnostic ratings. This incidental manipulation was based on the idea that respondents would display a tendency to complete the survey quickly as they did not receive incentives to participate.

These added conditions may account for the difference in findings between the current study and the Garb and Boyle (2003) survey. In addition to the expected statistically significant effect of age, there was evidence of an age bias. As can be seen in Table 8 (reproduced below), the ratings for the average level of performance differed in this study as compared to the findings of Garb and Boyle (2003). Again, respondents were told that a rating of 0, 1 or 2 indicated that a diagnosis would not be made; a rating of 5 indicated that there was a 50/50 chance of a diagnosable impairment; and that a rating of 8, 9 or 10 indicated that diagnostic criteria for impairment was met. The ratings for the clients in the Garb and Boyle (2003) study were 1.60 for any impairment and 0.84 for the presence of dementia for the young client versus 2.74 and 2.04 for the older client. Overall, the ratings in the current survey were higher: 2.65 and 1.95 for the younger client and 5.52 and 4.52 for the older client. Although respondents in this study were accurate in their diagnoses for the younger client (i.e., none of the respondents diagnosed the young client with average scores with a cognitive impairment), they were less accurate for the older client. Only one respondent diagnosed this older client with a cognitive impairment at this level of performance; however, the mean rating for this

vignette indicates that respondents rated the older client with average performance as having a 50% chance of cognitive impairment. This difference in accuracy meets the definition of bias as used in the Garb and Boyle (2003) study and indicates that older clients were more likely to be seen as impaired despite performing with the expected range for their age. This finding was surprising; it was hypothesized that there would not be an age bias at this level of performance based on previous research (Garb & Boyle, 2003.

Table 8

Average Ratings and Percentage of Respondents Diagnosing Impairment

		Younger Client		Older Client	
Level of Performance	Diagnostic Rating	Mean Rating (SD)	% of respondents diagnosing	Mean Rating (SD)	% of respondents diagnosing
Average	Any Imp.	2.65	0%	5.52	4.8%
	Dementia	(1.46)	0%	(1.86)	4.8%
		1.95		4.52	
		(0.76)		(2.06)	
Borderline	Any Imp.	7.64	40.9%	8.52	52.4%
	Dementia	(1.62)	4.8%	(1.21)	66.7%
		5.95		8.57	
		(1.84)		(1.36)	
Impaired	Any Imp.	8.67	85.7%	9.25	75%
	Dementia	(2.13)	28.6%	(1.52)	80%
		6.71		9.65	
		(2.17)		(1.66)	

Furthermore, the survey results indicate that the effect of age became more pronounced as the performance level decreased. It was hypothesized that the diagnostic ratings for the borderline performance level would be around 5 for the younger client, indicating an equal chance of impairment or no impairment, and somewhat higher for the older client due the influence of base rates on diagnostic ratings. This expectation was based on the idea that an age bias, if present, would be most obvious at this level of performance due to the ambiguous nature of the clinical data provided. However, this hypothesis was not supported. Instead, a majority of the respondents diagnosed the older client with neurological impairment (52%) and with dementia (67%). The mean ratings for the older client were 8.52 and 8.57, which places the client in the diagnosable range. Even the younger client was rated higher than expected at this level of performance; 41% of respondents diagnosed the younger client with neurological impairment, although only one respondent diagnosed the young client with dementia. Thus, the average rating for any neurological impairment was much higher than expected based on previous research using this same rating scale at 7.64 (Garb & Boyle, 2003). The rating for the likelihood of dementia, however, was in the expected range at 5.95. These results suggest that survey respondents were likely to make a diagnosis of dementia for the older client based on a lower threshold of impairment. Typically, neuropsychological diagnoses use a cutoff score of two standard deviations below expected performance as a basis for judging impairment (Lezak, Howieson, & Loring, 2004). In this study, the majority of respondents demonstrated a willingness to diagnosis the older client with neurological impairment and dementia at one to one-and-a-half standard deviations below expectations, indicating a lowered threshold for diagnoses.

Respondents also demonstrated a willingness to frequently diagnosis the younger client with neurological impairment at this performance level as well, which is perhaps a defensible position. Although this client does not meet the two standard deviation criteria typically used in neuropsychological diagnoses, this performance is still lower than expected and may indicate the presence of some kind of cognitive impairment despite not meeting criteria for a *specific* condition. Nevertheless, the most accurate rating at this level of performance is around five; therefore, it can be concluded that respondents were accurate in their rating of likelihood of dementia for the younger client with a mean rating of 5.95, but were less accurate in their rating of the older client with a mean rating of 8.57. This again indicates the presence of an age bias, but only in the dementia ratings. There was no age bias present in the ratings of the likelihood of any neurological impairment; however, there are indications of overdiagnosis in the rating of any neurological impairment as evidenced by the willingness to diagnosis the younger and older client at this borderline level of impairment.

In addition, there is evidence of the conjunction fallacy at this level of performance as respondents were more willing to diagnosis the older client with dementia, a specific condition, than with the more general condition of neurological impairment. The conjunction fallacy occurs when judges rate the likelihood of specific conditions as being higher than the likelihood of a single condition or a more general condition (Tversky & Kahneman, 1982). The base rate of *any* form of neurological impairment is always going to be higher than the base rate of a *specific* form of neurological impairment, such as a dementia. Thus, there is further indication that

cognitive biases impacted the respondent's diagnostic ratings in this performance level, but only for the older client.

In the final performance level, in which scores fell in the impaired range, it was hypothesized that nearly all neuropsychologists would diagnosis both the old and young client with both neurological impairment and dementia. Thus, it was expected that there would be no evidence of an age bias at this level of performance as the clinical data clearly shows evidence of impaired cognitive functioning, particularly in the area of memory. This hypothesis was supported. The older client was diagnosed with neurological impairment by 75% of respondents and with dementia by 80% of respondents, and the younger client was diagnosed with neurological impairment by 86% of respondents. However, only 29% of respondents diagnosed the younger client with dementia at the impaired level of performance. Although this finding seems to contradict the study predictions and suggest the presence of an age bias, the difference in diagnostic accuracy could be explained by a reliance on base rates when making dementia likelihood ratings.

The Garb and Boyle (2003) study concluded that the use of appropriate clinical base rates explained why the older client was consistently rated higher than the younger client; however they did not find differences in diagnostic accuracy between the young and old client vignettes in their study. Still, their study employed the use of only average scores and, although the older client was rated higher on the diagnostic scales, no participants diagnosed the vignette client as having a dementia. Thus, the authors state that the diagnostic ratings, overall, were accurate as respondents properly failed to diagnosis the client with any cognitive impairment and that the proper use of base rates

accounted for the statistically significant effect of age as risk of dementia increases with age (Garb & Boyle, 2003).

Although the present study found differences in the diagnostic accuracy when respondents rated the likelihood of a dementia for the old client versus the young client, this result can still be explained as an indication of the use of base rates by respondents. The base rate of dementia prior to age 60 is so low as to be almost non-existent (Petersen, 2003); thus, respondents would have found it extremely unlikely that the young client, age 48, would be experiencing a dementia. The vast majority of respondents accurately diagnosed the young client as having some form of neurological impairment, so it is clear that they did not misinterpret the clinical data presented. Furthermore, base rates are most influential on clinical decision making when data is ambiguous or limited (Garb, 1996). The clinical data provided in the vignette likely did not provide strong enough evidence of a dementia because it lacked information that is routinely collected during a comprehensive neuropsychological evaluations, specifically, information regarding a client's history, current functioning, and behavior during testing. Thus, respondents seem to have relied on clinical base rates when making these diagnostic decisions and rated the likelihood of a dementia for the young client as very unlikely because the base rate of dementia for a 48-year-old is so low.

It should be noted that although dementia is much less likely to be present at 48-years-old than at 74-years-old, it can and does occur at that age. However, the most typical dementia that presents at this age is a frontotemporal dementia, usually attributed to Pick's Disease, which does not typically present with primary memory impairments in the early stages. Instead, this form of dementia is usually accompanied by impairments in

behavior, personality, and language rather than memory (Welsh-Bohner, 2008). The vignettes in this study emphasized the presence of both subjective memory complaints and significant memory deficits in the impaired performance condition. Thus, an awareness and understanding of the base rates of various forms of dementia, including the rarer forms that can strike at a relatively young age, would likely have still led respondents to rate the likelihood of a dementia for the young client in the impaired condition as low because of the emphasis on memory deficits.

It is also possible that this finding demonstrates the use of another cognitive bias in diagnosis, the overreliance on prototypes and exemplars. 'Prototype' refers to the comparison of a clinical presentation to an 'ideal' presentation of a certain clinical diagnosis (Garb, 1996); while 'exemplar' refers to the comparison of a clinical presentation to previously encountered members belonging to a certain clinical diagnostic category (Juslin & Persson, 2002). Thus, when making a clinical judgment about a particular client, a clinician may compare this client's presentation with what a person would look like if they met all criteria for a diagnosis. The degree to which this real world client corresponds to the prototypical client with a particular disease would help the clinician make a diagnostic decision. It can be assumed that, for most clinicians, the prototypical client with dementia is an older individual who demonstrates a profound memory impairment along with other cognitive symptoms, such as functional impairments. Perhaps the respondents in this study were relying on the use of prototypes to make clinical decisions quickly and made a diagnostic error due to overreliance on this cognitive heuristic. Furthermore, exemplars are based on both similarity to a particular clinical diagnostic category, like prototypes, and the frequency with which a judge

encounters members of that clinical diagnostic category (Juslin & Persson, 2002). Thus, the use of exemplars would also lead respondents to perceive the older client as more likely to have dementia as most previously encountered individuals with dementia would be older adults.

Prototypes and exemplars are also examples of the representativeness heuristic, which describes judgments that are made based on how similar an object or person is to a category or class (Tversky & Kahneman, 1974; Nilsson, Juslin & Olsson, 2008). Furthermore, respondents were more likely to diagnosis the older client with dementia than with a general neurological impairment, which is an example of the conjunction fallacy (Tversky & Kahneman, 1982). However, they did not commit this fallacy when judging the younger client at any performance level. Researchers have argued that the conjunction fallacy is another special case of the representativeness heuristic (Tversky & Kahnman, 1983; Garb, 1998), which may explain this finding. Thus, through the use of prototypes, exemplars, and the representativeness heuristic, respondents would have quickly identified the older client with impaired performance as having a dementia, but would also be more likely to fail to diagnosis dementia in the younger client with the same level of performance.

In summary, the results of this survey did not support the hypothesis that there would be no age bias present in the average performance vignettes. There were indications of an age bias across the average level of performance, in which the older client was inaccurately rated as having a 50% chance of cognitive impaired and dementia. Furthermore, this age bias was also found in the borderline condition as was evidence of the conjunction fallacy. However, despite differences in diagnostic accuracy for the

likelihood of a dementia in the impaired condition, it does not appear that there was an age bias present in the ratings for this performance level. Rather, this finding may reflect a reliance on base rates due to the limited clinical data available and could also be interpreted as an example of misjudgment due to the reliance on cognitive biases, such as the representativeness heuristic and use of prototypes.

Confidence Ratings

After making each diagnostic rating, respondents were asked to rate their confidence in the rating. It was hypothesized that confidence ratings would follow a pattern in which the ratings were higher for the average and impaired categories and lower for the borderline category. This expectation was based upon the idea that respondents would feel more confident in their rating when clinical data were clear (i.e., the average and impaired performance conditions) and less confident when clinical data were ambiguous (i.e., the borderline performance condition). This hypothesis was partially supported. There was a significant interaction between age and level of performance with regard to the confidence ratings, meaning that the main effect of performance level impacted the confidence ratings differently for the old client versus the young client. The confidence ratings for the young client followed the predicted pattern, higher for the average and impaired conditions—situations in which the clinical data are clear—and lower in the borderline condition when the clinical data do not have an easy interpretation in both diagnostic conditions.

However, for the older client, a trend toward a different pattern was found with regard to the confidence ratings for the dementia diagnosis. Respondents were significantly more confident in their dementia diagnostic rating for the old client at the

impaired level of performance than they were in their rating for the young client at the same impaired level of performance. In addition, confidence ratings in the dementia diagnosis increased as the level of performance decreased for the older client. The confidence ratings for the diagnosis of any neurological impairment were not significantly different from the confidence ratings for the younger client, although they do follow the same general trend as the confidence ratings for the dementia diagnostic rating. Thus, overall, respondents became more confident in their diagnostic ratings for the older client as the level of impairment increased. It is interesting that respondents, when rating the older client, felt more confident diagnosing impairment when it was present than not diagnosing impairment when it was not present as both judgments would be considered accurate.

Due to the non-significant difference between the confidence ratings for the younger and the older client (i.e. there was no significant main effect for age with regard to the confidence ratings), it is impossible to make conclusions based on the trend observed. However, this is an interesting avenue for future research as it suggests that a knowledge of base rates and cognitive heuristics may have contributed to this trend. As mentioned previously in this section, clinicians often rely on the representativeness heuristic when making diagnostic decisions. In addition to impacting the nature of clinical judgments, cognitive heuristics also impact an individual's confidence in their judgments (Garb, 1996). Thus, the older individual in the impaired condition is consistent with the clinical prototype of a client with dementia and a diagnosis of this type of client would be supported through the use of clinical base rates. Therefore, respondents were not only more likely to accurately diagnosis the client in this vignette, but they also felt

most confident in their diagnostic ratings in this conditions due to the degree with which this vignette is consistent with clinical expectations based on knowledge of base rates and clinical prototypes. However, the older client with average performance is not consistent with the prototypical client of that age who also endorses subjective memory impairments. Thus, this discrepancy may have led respondents to doubt their decision more in this condition because it was not consistent with expectations, especially in light of the limited clinical information available. Furthermore, future research may be able to investigate this possibility and determine if the non-significant trend observed in this study would be found in more powerful research designs.

Research on eyewitness testimony may help shed some light on the common phenomenon in which judges are often inaccurate in their decision making, but also demonstrate a high degree of confidence in their decisions. Kassin (1985) argues that the lack of correlation between accuracy and confidence is due to discrepancy between one's self-perception and subjective self-report. In a series of experiments, this author demonstrated that adding a condition in which participants saw a video of themselves making a decision (i.e. picking a thief out of a lineup) before being asked to rate their confidence in their decision led to better calibrated confidence ratings (Kassin, 1985). Thus, the respondents were better able to rate the likelihood that their decision was accurate when they were able to watch themselves making the decision. Kassin (1985) concluded that this finding was due to the participant's ability to pick up on nonverbal cues, such as response time and facial expressions, which enabled them to more reliably assess their probability of making an accurate decision. This research suggests that

clinicians may be able to increase their ability to accurately rate their confidence if they take time to reflect on their decision-making process.

Furthermore, research on eyewitness testimony suggests reasons why individuals tend to be over-confident in the accuracy their judgments. Investigators have argued that the accuracy of judgments, such as eyewitnesses identifying criminal perpetrators, and the confidence in those judgments depends on the type information recalled and the route of recall, such as free recall versus recognition recall (Migueles & Garcia-Bajos, 1999). Research has shown that free recall is often better for central information (i.e. the gist and central details of an event) than for peripheral information (i.e. details immediately preceding or following the central event). However, when looking at recognition errors, judges were found to be very inaccurate regardless of information type (Migueles & Garcia-Bajos, 1999). This finding suggests that individuals are willing to accept false information, as evidenced by participants making recognition errors, yet feeling confident in the accuracy of their recognition, if it fits with previously learned cognitive scripts (Migueles & Garcia-Bajos, 1999). Thus, in the current study, which did not require free recall of information, respondents may have made diagnostic errors due to reliance on the representativeness heuristic, as argued earlier in this discussion, but would have still felt confident in these ratings due to the fact that their diagnostic ratings fit with previously learned cognitive scripts (i.e. older individuals are more likely to have dementia).

Perceived Necessity Ratings

The third section of the survey asked respondents to rate the perceived necessity of the psychometric data provided in each of the clinical vignettes. Respondents rated the 19 pieces of clinical data presented in terms of how necessary the information was in

making their diagnostic ratings. This section was included for two reasons. First, previous surveys of neuropsychological clinical practices noted that clinicians are routinely asked to spend less time testing clients due to limits on insurance reimbursements (Sweet et al., 2000). As a result, neuropsychologists are not able to give as many tests and sometimes make decisions based on limited clinical data. This survey sought to ascertain if some psychometric information was more useful to making diagnostic decisions as this would help neuropsychologists prioritize which measures to give first during an evaluation. Second, clinical judgment research has demonstrated that human judges believe their judgments to be more accurate when they have access to more information (Gauron & Dickinson, 1966). However, numerous studies have shown that judgements are no more accurate when additional information is provided beyond that which is necessary to make a decision (Grove et al., 2000). Indeed, some studies have shown that additional information can decrease the accuracy of judgments due to an increased burden on cognitive resources (Groenier et al., 2008). Thus, this section also sought to ascertain if some psychometric data was seen as superfluous, which would imply that clinical data collected during evaluation could be reduced thereby saving time and financial resources while also increasing the likelihood of accurate diagnoses.

It was hypothesized that necessity ratings would vary as a function of the vignette performance level with respondents who viewed the impaired vignettes (old or young client) rating most clinical information as "very necessary" due to the need to make a clinical diagnosis. However, this hypothesis was not supported. There was no difference between the necessity ratings across the six clinical vignettes. Instead, most respondents, regardless of the test vignette they viewed, rated each of the 19 pieces of psychometric

data as either ‘necessary’ or ‘very necessary.’ This finding was surprising, especially as some of the clinical data provided could be seen as redundant. For example, both the visual reproduction subtest on the WMS and the Brief Visual Memory Test (BVMT) involve presenting simple geometric figures for a brief period of time (i.e., 10 seconds) and then asking the examinee to draw the figures from memory. Although there are subtle differences in the information provided by each of these tests, they tap the same general cognitive domain. Still, there were no statistically significant differences between the six clinical vignette groups for any of the pieces of clinical data.

Neuropsychologists have long held that their methods of reaching diagnostic decisions are closer to statistical prediction models, and thus superior to other form of clinical prediction, due to the use of empirically validated, reliable and normed measures of cognitive functioning (Guilmette et al., 1990). Through the use of reliable psychometric tests, neuropsychologists can use objective data to justify their diagnosis, which is not possible in other areas of psychological diagnosis due to the use of subjective, unreliable measures such as clinical interview and client self-report (Gaudette, 1992). Although clinical judgment research has shown many of the claims to be inflated and has demonstrated various diagnostic inaccuracies in neuropsychological practice (e.g., Garb & Schramke, 1996; Wedding, 1983), this line of thinking can help to explain why all the psychometric data in this survey were judged to be necessary in making diagnostic ratings. Neuropsychological diagnosis prides itself on the use of objective, reliable measures and the more data available the more secure a clinician can feel in their decision making. Copious amounts of psychometric data allow a neuropsychologist to justify conclusions about the functioning of a wide variety of cognitive domains, thereby

bolstering the final diagnostic conclusion. It is also likely that this part of the survey was not properly designed to answer the research questions posited. This possibility will be further discussed in the following section.

Study Limitations and Suggestions for Future Research

Before discussing the implications of this study, it is necessary to first consider its limitations. Several of these limitations involve the survey design and the process of implementation. Survey research, in general, faces numerous obstacles leading to potential error by virtue of its design. These error sources include, but are not limited to, coverage, nonresponse, and measurement. A coverage error occurs when every unit in the survey population does not possess a known, non-zero chance of being included in the sample (Dillman, 2000). In the current study, only neuropsychologists who hold an INS membership and included their contact information in the voluntary member registry had a chance of being included in the sample. Previous surveys utilized wider sample frames, such as included members of the APA Division 40 along with members of the National Academy of Neuropsychologists (NAN).

Minimizing coverage error has become more difficult in recent years, as access to membership directories has tightened. Many APA divisions have changed bylaws to limit directories to only professional mailing lists (American Psychological Association, Division 40, http://www.div40.org/bylaws.html). NAN, The National Academy of Neuropsychology, has a similar policy of only releasing mailing addresses on a fee-per-name basis, despite recent trends in survey research that have highlighted the advantages of web-based survey design (Cook, Heath & Thompson, 2000). Web-based surveys have numerous advantages that led to their use in this study, including minimal financial cost,

speed of implementation and response, and compatibility with statistical software that limits initial data coding time. Thus, the INS sample frame was utilized despite the increased likelihood of coverage errors because it allowed for the use of web-based survey technology. Based on current trends, it is likely that survey research will continue to utilize new technology and mailed paper surveys will become obsolete. Therefore, professional organizations will need to structure new policies that protect their members' privacy, without creating undue barriers to valuable survey research that tracks important practice trends. Nevertheless, this survey likely overlooked a certain percentage of North American neuropsychological professionals.

Another source of error, nonresponse error, occurs when a significant number of non-responders included in the survey sample possess different characteristics from survey responders (Dillman, 2000). In the current study, the overall response rate was 12.5%. Although commensurate with previous research (e.g. Sweet et al., 2000), a higher return rate was desired and would have increased the power and generalizability of survey results. Past meta-analytic research has suggested that use of incentives do not reliably increase survey response (Cook, Heath & Thompson, 2000); however, it is possible that a sizable incentive, such as payment for time spent on the survey that approaches typical hourly wages, may have increased survey response. Given the relatively short duration of the survey, the financial cost of such a study might be manageable for future researchers. It is also possible that a financial incentive, combined with a similar survey design, may be superior in trying to approximate the cognitive processes used by neuropsychologists in a clinical setting. In addition, potential

respondents in this current study were sent one follow-up email, and it is possible that additional contact may have further enhanced the response rate.

The low response rate in this survey impacted the power of the statistical tests used and further limits the interpretation of the current study's results. Several of the statistical tests were underpowered due to several small effect sizes found in this study; effect sizes greater than 0.25 were adequately powered. Thus, the effect of performance level on both diagnostic ratings, the effect of age on the dementia ratings, and the interaction of performance and age on confidence ratings all achieved adequate power with the obtained survey sample size. However, the interaction of age and performance level on the diagnostic ratings, the effect of performance on the neurological impairment rating, and the effects of age and performance level on confidence ratings were underpowered due to the low response rate. A larger sample would have increased the power of statistical tests and an a priori power estimation predicted than an additional 55 respondents would result in adequate power for all statistical tests. This limitation in power reduces the confidence one can place in statistical test results as it introduces more error into the analysis.

Measurement error, a final source of error, refers to a situation in which a respondent's answer to a survey question is inaccurate, imprecise, or cannot be compared in a useful way to other survey responses. Problems with a questionnaire's wording and construction is one common way measurement error may occur (Dillman, 2000). The questionnaire used in this study was constructed by the principal researcher and combined elements from several previous studies. In addition, the vignettes were constructed based on the study by Garb and Boyle (2003) for the purpose of replication.

Should this questionnaire be used again in future research, several changes would increase clarity and utility. First, several survey respondents found it difficult to answer the question regarding the number of instruments used in a typical battery. The most common reasons for confusion were based on whether comprehensive batteries, such as the WAIS or WMS, should be counted as one test or whether each subtest should be tallied as an "instrument." Thus, clarification in the question's wording would be beneficial to avoid this confusion. Second, there were a number of questions regarding the tests included in the battery selection checklist that require clarification. The Controlled Oral Word Association Test was meant to include both semantic and phonemic portions of the task, but several respondents indicated that they were unsure if the test included the semantic portion. Thus, a brief clarification next to the test name would have reduced confusion. There was similar confusion regarding the Brief Visual Memory Test (BVMT); a few respondents indicated that this test could refer to two different measures, which was not known to the principal investigator at the time of the survey construction. Thus, future researchers should endeavor to research other common names for tests in order to avoid similar confusion.

Finally, this survey's necessity of clinical information section generated very little useful information. There are at least two possibilities that might explain this section's inability to generate the desired information. First, several studies conducted in the field of clinical judgment have suggested that human judges are generally unable to accurately record the steps in their decision-making process (e.g. Grove et al., 2000; Meehl, 1954). Thus, the fact that this section followed the section in which diagnostic ratings were made may have made it difficult for judges to accurately recall which pieces of clinical

information were more or less useful in making their ratings. Future researchers may want to consider redesigning this section and combining it with the diagnostic rating section. A rating scale could be included next to each piece of clinical information in the vignette and respondents would rank the necessity of each piece of psychometric data as they are making their diagnostic ratings. The concurrent nature of this task design may enable judges to more accurately record their decision-making process as it is occurring, instead of relying on retrospective inferences.

Second, the results obtained in the current study may simply reflect the limited usefulness in asking respondents to subjectively rate the necessity of information. Clinical judgment research has demonstrated that expert judges tend to feel that their diagnostic judgments are better when they have access to additional information; however, researchers have demonstrated that accuracy does not improve with additional information once basic clinical information has been provided (Garb, 1998). Thus, the rating of all clinical information as either 'necessary' or 'very necessary' may reflect this underlying human tendency to believe that all clinical information is valuable. Thus, in order to capture the information desired in this study, future researchers may want to consider an alternative research design. Instead of asking judges to rate the necessity of information, an additional manipulation could be added to the study design. This manipulation would vary the amount of clinical information provided in the vignettes, in order to see if a certain amount of information maximizes diagnostic accuracy. Alternatively, a rank order system could be used in which respondents are asked to rank the clinical information from most useful to least useful, thus forcing respondents to provide differential necessity rankings. In addition, respondents could have also been

asked which test score they would remove first if forced to limit the clinical data available. This would help provide a sense of which clinical data was seen as least useful.

Suggestions for Future Research

In addition to the aforementioned limitations and tactics to avoid these limitations, future researchers may be interested in expanding the current study's investigation. One of the purposes of the current study was to expand and clarify previous studies of age bias in neuropsychological diagnosis. However, age bias is not the only documented diagnostic bias in neuropsychological practice, and the over-perception of impairment has been discussed by previous studies (e.g. Garb & Schramke, 1996). Still, the conditions under which overdiagnosis is likely to occur are still largely unknown. Future research may want to expand the current study to include other clinical presentations, such as head injuries or learning disabilities, to determine the situations in which overdiagnosis or misdiagnosis may occur. The use of clinical vignettes similar to the ones used in this study could facilitate such investigations. In addition, the first section of the survey, which asked about battery construction, could be expanded to other clinical presentations as well. Given the relative dominance of the flexible battery, it would be beneficial to determine if neuropsychologists use similar tests to evaluate the relevant cognitive domains necessary to make a diagnosis.

Furthermore, the current study is limited in ecological validity due to the nature of the vignette task used. In typical neuropsychological evaluations, the clinician would have access to other sources of information, such as clinical interviews, medical records, and behavioral observations, prior to making a diagnostic decision. Thus, the vignette task used in the current study was only able to represent a fraction of this experience.

Still, the incremental validity of nonpsychometric data in neuropsychological assessment has not been established, although this information (history data and clinical interview) has been shown to increase accuracy in personality assessment (Garb, 1994). Future research could incorporate additional information, such as client background and behavior during testing, in order to determine how this information affects the accuracy of clinical information. However, this survey would be much longer in length, a factor that would impact response rates. Thus, future researchers would likely have to use some sort of incentive in order to generate a similar response rate to the current study.

Conclusions and Implications for the Practice of Neuropsychology

This doctoral project resulted in several important findings that have implications for the practice of neuropsychology. The first section of the current survey expanded on previous survey literature through exploration of the process of battery selection. Individual neuropsychologists benefit from an awareness of the common tests fellow clinicians use, and this information will help ensure that the necessary cognitive areas are being consistently evaluated in everyday clinical practice. The results of this project indicate that neuropsychologists may differ in the specific tests selected for inclusion in a dementia evaluation battery but tend to assess similar cognitive domains. This finding is encouraging and demonstrates consistent implementation of diagnostic criteria in selecting evaluation tasks in order to ensure that the necessary clinical information is collected.

Given the popularity of flexible battery, it is expected that neuropsychological training will begin to put more emphasis on understanding the critical diagnostic domains in different clinical presentations, in addition to an understanding of specific tests that can

be used to tap into the functioning of these cognitive domain areas. In the future, new practice parameters could focus on standardizing the domains that must be assessed, in addition to optional functional areas that can provide useful information. This would help ensure that the use of flexible batteries does not decrease the average quality of neuropsychological evaluations. A few professional organizations have sought to implement this strategy (e.g. the American Board of Clinical Neuropsychology); however, these organizations are only able to make suggestions regarding clinical standards. Thus, these standards are not consistently implemented or enforced. As there is already some consensus in the field regarding the crucial cognitive domains to assess in a dementia evaluation, it is likely that field-wide agreement could be obtained regarding standardized evaluation criteria. From a consumer protection perspective, enforceable evaluation standards would help increase consumer faith in the accuracy of neuropsychological evaluation and diagnosis.

Another important finding in this study was the presence of an age bias in neuropsychological diagnosis as demonstrated by the differential accuracy of the diagnostic ratings. This finding underscores the importance of continued reliance on clinical norms for the interpretation of psychometric data. It was presumed that neuropsychologists look at the age-based interpretations of clinical data and then base diagnostic decisions on this information. However, this study demonstrates that neuropsychologists do not always make accurate decisions, especially when under a time constraint or when clinical information is ambiguous. Therefore, neuropsychological training should emphasize the building of differential diagnosis skills and the ability to provide a diagnostic rationale to justify clinical decisions.

Finally, the results of this study suggest that a lower threshold is used for some clinical decisions, such as the diagnosis of a general neurological impairment. Typically, neuropsychological diagnosis relies on a standard impairment classification of two standard deviations below expected performance. However, this study found a tendency to diagnosis neurological impairment and/or dementia at a lower threshold (1-1.5 standard deviations below expected performance), especially in older individuals. This lower threshold for diagnosis may be defensible in some clinical situations, particularly when the clinician has access to additional information that suggests an impairment (i.e. evidence of discreet functional changes in prior abilities provided by family members or medical records; family history of a certain clinical condition). Although the vignette used in this study did not provide such information and, therefore, the inaccuracy demonstrated is likely indicative of biased decision making, it may also reflect the need for a diagnosis that reflects mild neuropsychological impairments. Indeed, the *DSM-V* (5th ed.; American Psychiatric Association, 2013), which was released earlier this year, supports this implication as it included a new diagnosis of Mild Neurocognitive Disorder. A similar diagnosis has been included in the ICD-10 (World Health Organization, 2004), but North American neuropsychologists have been unable to receive insurance reimbursement for this diagnosis until its inclusion in the *DSM-V* (2013).

For the past decade, researchers in the fields of neuropsychology and neurology have investigated the need for a diagnostic category that reflects cognitive changes that result in functional impairment, but do not yet reach the level of impairment seen in an individual with a dementia (Petersen & Morris, 2005). Although the definition and treatment of this so-called 'mild cognitive impairment' is still intensely debated, there is a

clear consensus regarding the need for a diagnostic category that draws attention to these issues (Gauthier & Touchon, 2005). Mild Cognitive Impairment is considered to be reflective of cognitive changes that are greater than expected with normal aging and may also reflect cognitive impairments resulting from traumatic brain injuries or psychiatric conditions, such as Major Depressive Disorder (Petersen & Morris, 2005). In addition, researchers have made great strides in recognizing the prodromal stages of Alzheimer's disease and continue to emphasize the need for a diagnostic category that allows for individuals to seek early intervention prior to the onset of a full-blown dementia (Backman, Jones, Berger, Laukka & Small, 2005). Although clear prodromal stages are more difficult to determine in other forms of dementia, such as Lewy Body Dementia or Frontotemporal Dementia, neuropsychological deficits are present in most individuals prior to a dementia diagnosis and can be measured by current testing measures (Welsh-Bohmer, 2008). Therefore, research over the past decade has continuously highlighted the need for early detection and diagnosis of cognitive impairments prior to the onset of full-blown dementia.

Thus, the tendency of respondents to diagnose a neurological impairment at a lower threshold than the standard rubric may reflect the acknowledgement that some sort of impairment is likely present and that this impairment is sufficient to warrant clinical attention. With the inclusion of the Mild Neurocognitive Impairment diagnostic category, neuropsychologists will be able to more accurately describe functional impairments while ensuring that the necessary services are available as most clinical settings require a diagnosis prior to beginning treatment.

REFERENCES

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S…Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction Stefanía. *The Counseling Psychologist*, *34*(3), 341-382. doi:10.1177/0011000006286696

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*, 256-274. doi:10.1037/0033-2909.111.2.256

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, D.C.: Author.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, D.C.: Author.

American Psychological Association, Division 40. (2005). Division 40 Bylaws. Retrieved from http://www.div40.org/bylaws.html

Arkes, H. R., Faust, D., Guilmette, T. J., & Hart, K. (1988). Eliminating the hindsight bias. *Journal of Applied Psychology*, *73*(2), 305-307.

Arkes, H. R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology*, *49*(3), 323-330. doi:10.1037/0022-006X.49.3.323

Backman, L., Jones, S., Berger, A., Laukka, E. J., & Small, B. J. (2005). Cognitive

impairment in preclinical Alzheimer's disease: A meta-analysis. *Neuropsychology, 19*: 520-531.

Beck, S. J. (1944). *Rorschach's test I: Basic processes.* Oxford, England: Grune & Stratton.

Benton, A. L. (1987). Evolution of a clinical specialty. *Clinical Neuropsychologist*, *1*(1), 5-8. doi:10.1080/13854048708520030

Benton, A. L. (1992). Clinical neuropsychology: 1960–1990. *Journal of Clinical and Experimental Neuropsychology*, *14*(3), 407-417. doi:10.1080/01688639208407616

Bigler, E. D. (1990). Neuropsychology and malingering: Comment on Faust, Hart, and Guilmette (1988). *Journal of Consulting and Clinical Psychology*, *58*(2), 244-247. doi:10.1037/0022-006X.58.2.244

Bray, J. H., & Maxwell, S. E. (1985). *Multivariate Analysis of Variance.* Newbury Park, CA: Sage.

Brown, G. G., del Dotto, J. E., Fisk, J. L., & Taylor, H. G. (1993). Analyzing clinical ratings of performance on pediatric neuropsychological tests. *Clinical Neuropsychologist*, *7*(2), 179-189. doi:10.1080/13854049308401520

Brown, L., Gfeller, J., Ross, M., & Heise, R. (1999). A survey of neuropsychologists' ethical beliefs and practices. *Archives of Clinical Neuropsychology*, *14*(8), 756-757.

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice*, *31*(2), 141-154.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, *74*(3), 271-280.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159.

Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in Web-or Internet-based surveys. *Educational and Psychological Measurement*, *60*(6), 821-836.

Dallas, M. E. W., & Baron, R. S. (1985). Do psychotherapists use a confirmatory strategy during interviewing? *Journal of Social and Clinical Psychology*, *3*(1), 106–122.

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist, 26,* 180-188.

Dawes, R. M. (1996). *House of Cards* (1st ed.). New York, NY: Free Press.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*(2), 95-106.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668-1674.

DeLuca, J. W., & Putman, S. H. (1993). The professional/technician model in clinical neuropsychology: Deployment characteristics and practice issues. *Professional Psychology: Research and Practice*, *24*(1), 100-106.

Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method* (2nd ed.). New York, NY: John Wiley.

Donders, J. (2001). A survey of report writing by neuropsychologists: General characteristics and content. *The Clinical Neuropsychologist*, *15*(2), 137-149. doi:10.1076/clin.15.2.137.1893

Dowie, J. E., & Elstein, A. S. (1988). *Professional judgment: A reader in clinical decision making.* London, England: Cambridge University Press.

Dudycha, L. W., & Naylor, J. C. (1966). Characteristics of the human inference process in complex choice behavior situations. *Organizational Behavior and Human Performance*, *1*(1), 110–128.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance, 7,* 86-106.

Einhorn, H. J. (1988). Diagnosis and causality in clinical statistical prediction. In D. C. Turk & P. Salovey (Eds.), *Reasoning, inference, and judgment in clinical psychology* (pp. 51-70). New York, NY: Free Press.

Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review*, *86*(5), 465-485. doi:10.1037/0033-295X.86.5.465

Ellis, M. V., Robbins, E. S., Schult, D., Ladany, N., & Banker, J. (1990). Anchoring errors in clinical judgments: Type I error, adjustment, or mitigation. *Journal of Counseling Psychology*, *37*(3), 343-351.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.

Faust, D., Guilmette, T. J., Hart, K. J., & Arkes, H. R. (1988). Neuropsychologists' training, experience, and judgment accuracy. *Archives of Clinical Neuropsychology*, *3*(2), 145-163. doi:10.1016/0887-6177(88)90060-1

Faust, D., Hart, K. J., & Guilmette, T. J. (1988). Pediatric malingering: The capacity of children to fake believable deficits on neuropsychological testing. *Journal of Consulting and Clinical Psychology*, *56*(4), 578-582. doi:10.1037/0022006X.56.4.578

Faust, D., Hart, K. J., Guilmette, T. J., & Arkes, H. R. (1988). Neuropsychologists' capacity to detect adolescent malingerers. *Professional Psychology: Research and Practice*, *19*(5), 508-515. doi:10.1037/0735-7028.19.5.508

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London, England: Sage.

Filskov, S. B., & Goldstein, S. G. (1974). Diagnostic validity of the Halstead-Reitan Neuropsychological Battery. *Journal of Consulting and Clinical Psychology*, *42*(3), 382-388.

Garb, H. N. (1998). *Studying the clinician* (1st ed.). Washington, D.C.: American Psychological Association.

Garb, H. N, & Schramke, C. J. (1996). Judgment research and neuropsychological assessment: A narrative review and meta-analyses. *Psychological Bulletin*, *120*(1), 140-153.

Garb, H. N., & Boyle, P. A. (2003). The diagnosis of neurological disorders in older adults. *Assessment*, *10*(2), 129-134.

Garb, H. N. (1984). The incremental validity of information used in personality assessment. *Clinical Psychology Review*, *4*(6), 641-655. doi:10.1016/02727358(84)90010-2

Garb, H. N. (1994). Toward a second generation of statistical prediction rules in psychodiagnosis and personality assessment. *Computers in Human Behavior*, *10*(3), 377-394. doi:10.1016/0747-5632(94)90063-9

Garb, H. N. (1996). The representativeness and past-behavior heuristics in clinical judgment. *Professional Psychology: Research and Practice*, *27*(3), 272-277. doi:10.1037/0735-7028.27.3.272

Garb, H. N. (2000). On empirically based decision making in clinical practice. *Prevention & Treatment*, *3*(1). doi:10.1037/1522-3736.3.1.329c

Gaudette, M. D. (1992). *Clinical decision-making in neuropsychology: Bootstrapping the neuropsychologist utilizing Brunswik's lens model.* Ann Arbor, MI: ProQuest.

Gauron, E. F., & Dickinson, J. K. (1966). Diagnostic decision making in psychiatry: Information usage. *Archives of General Psychiatry*, *14*(3), 225-237.

Gauthier, S., & Touchan, J. (2005). Mild cognitive impairment is not a clinical entity and should not be treated. *Archives of Neurology, 62*(7): 1164-1166. doi:10.1001/archneur.627.1164

Groenier, M., Pieters, J. M., Hulshof, C. D., Wilhelm, P., & Witteman, C. L. M. (2008). Psychologists' judgements of diagnostic activities: Deviations from a theoretical model. *Clinical Psychology & Psychotherapy*, *15*(4), 256–265.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293-323.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19-30. doi:10.1037/1040-3590.12.1.19

Guilmette, T. J., & Faust, D. (1991). Characteristics of neuropsychologists who prefer the Halstead-Reitan or the Luria-Nebraska Neuropsychological Battery. *Professional Psychology: Research and Practice*, *22*(1), 80-83.

Guilmette, T. J., Faust, D., Hart, K., & Arkes, H. R. (1990). A national survey of psychologists who offer neuropsychological services. *Archives of Clinical Neuropsychology*, *5*(4), 373–392.

Guilmette, T. J., & Giuliano, A. J. (1991). Taking the stand: Issues and strategies in forensic neuropsychology. *Clinical Neuropsychologist*, *5*, 197-219. doi:10.1080/13854049108404092

Hart, B. M., Wicherski, M., & Kohout, J. L. (2010). *2008-2009 Master's- and doctoral-level students in U.S. and Canadian graduate departments of psychology.* Washington, DC: American Psychological Association, Center for Workforce Studies.

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin*, *107*(3), 311-327.

Heaton, R. K., Smith, H. H., Lehman, R. A., & Vogt, A. T. (1978). Prospects for faking believable deficits on neuropsychological testing. *Journal of Consulting and Clinical Psychology*, *46*(5), 892-900. doi:10.1037/0022-006X.46.5.892

Heaton, R., Grant, I., & Matthews, C. (1991). *Comprehensive norms for an expanded Halstead-Reitan neuropsychological battery: Demographic corrections, research findings, and clinical applications.* Odessa, FL: Psychological Assessment Resources.

Heinrichs, R. W. (1990). Current and emergent applications of neuropsychological assessment: Problems of validity and utility. *Professional Psychology: Research and Practice*, *21*(3), 171-176. doi:10.1037/0735-7028.21.3.171

Hilsabeck, R. C., & Martin, E. M. (2010). Woman and advancement in neuropsychology: Real-life lessons learned. *The Clinical Neuropsychologist, 24*, 481-492.

Holt, R. R. (1958). Clinical and statistical prediction: A reformulation and some new data. *The Journal of Abnormal and Social Psychology*, *56*(1), 1-12. doi:10.1037/h0041045

Hoshmand, L. T., & Polkinghorne, D. E. (1992). Redefining the science-practice relationship and professional training. *American Psychologist*, *47*(1), 55-66.

Hughes, R., & Huby, M. (2004). The construction and interpretation of vignettes in social research. *Social Work and Social Sciences Review-An International Journal of Applied*, *11*(1), 36–51.

Juslin, P. & Persson, M. (2002). Probabilities from Exemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science, 26*, 563–607.

Kahneman, D. & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). New York, NY: Cambridge University Press.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. London, England: Cambridge University Press.

Kareken, D. A., & Williams, J. M. (1994). Human judgment and estimation of premorbid intellectual function. *Psychological Assessment*, *6*(2), 83-91.

Kassin, S. M. (1985). Eyewitness identification: Retrospective self awareness and the accuracy confidence correlation. *Journal of Personality and Social Psychology, 49,* 878-893.

Kendell, R. E. (1973). Psychiatric diagnoses: A study of how they are made. *British Journal of Psychiatry*, *122*(569), 437-445. doi:10.1192/bjp.122.4.437

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin*, *107*(3), 296-310. doi:10.1037/00332909.107.3.296

Lees-Haley, P. R., Smith, H. H., Williams, C. W., & Dunn, J. T. (1996). Forensic neuropsychological test usage: An empirical survey. *Archives of Clinical Neuropsychology*, *11*(1), 45-51. doi:10.1016/0887-6177(95)00011-9

Leli, D. A., & Filskov, S. B. (1981). Actuarial assessment of Wechsler Verbal-Performance Scale differences as signs of lateralized cerebral impairment. *Perceptual and Motor Skills, 53*, 491-496.

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological Assessment* (4th ed.). London, England: Oxford University Press.

Long, C. J. (1996). Neuropsychological tests: A look at our past and the impact that ecological issues may have on our future. In R. J. Sbordone & C. J. Long (Eds.), *Ecological validity of neuropsychological testing* (pp. 1-14). Salem, MA: CRC Press.

Louttit, C., & Browne, C. (1947). The use of psychometric instruments in psychological clinics. *Journal of Consulting Psychology*, *11*(1), 49-54.

Lubin, B., Larsen, R. M., & Matarazzo, J. D. (1984). Patterns of psychological test usage in the United States: 1935–1982. *American Psychologist*, *39*(4), 451-454.

Lubin, B., Wallis, R. R., & Paine, C. (1971). Patterns of psychological test usage in the United States: 1935-1969. *Professional Psychology*, *2*(1), 70-74.

McCaffrey, R. J., & Isaac, W. (1984). Survey of the educational backgrounds and specialty training of instructors of clinical neuropsychology in APA-approved graduate training programs. *Professional Psychology: Research and Practice*, *15*(1), 26-33.

McCaffrey, R. J., & Lynch, J. K. (1996). Survey of the educational backgrounds and specialty training of instructors of clinical neuropsychology in APA-Approved graduate training programs: A 10-year follow-up. *Archives of Clinical Neuropsychology*, *11*(1), 11–19.

McCaffrey, R. J., Malloy, P. F., & Brief, D. J. (1985). Internship opportunities in clinical neuropsychology emphasizing recent INS training guidelines. *Professional Psychology: Research and Practice*, *16*(2), 236-252.

McMordie, W. R. (1988). Twenty-year follow-up of the prevailing opinion on the posttraumatic or postconcussional syndrome. *Clinical Neuropsychologist*, *2*(3), 198-212.
doi:10.1080/13854048808520102

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

Meehl, P. E. (1959). A comparison of clinicians with five statistical methods of identifying psychotic MMPI profiles. *Journal of Counseling Psychology*, *6*(2), 102-109.
doi:10.1037/h0049190

Meehl, P. E. (1973). *Psychodiagnosis: Selected papers.* Minneapolis: University Minnesota Press.

Meier, M. J. (1992). Modern clinical neuropsychology in historical perspective. *American Psychologist*, *47*(4), 550-558.
doi:10.1037/0003-066X.47.4.550

Migueles, M., & Garcia-Bajos, E. (1999). Recall, recognition, and confidence patterns in eyewitness testimony. *Applied Cognitive Psychology, 13*, 257-268.

Nadler, J. D., Mittenberg, W., DePiano, F. A., & Schneider, B. A. (1994). Effects of patient age on neuropsychological test interpretation. *Professional Psychology: Research and Practice*, *25*(3), 288-295.

Newell, A., & Simon, H. A.. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nilsson, H., Juslin, P., & Olsson, H. (2008). Exemplars in the mist: The cognitive substrate of the representativeness heuristic. *Scandinavian Journal of Psychology, 49*, 201-212.
doi: 10.1111/j.1467-9450.2008.00646.x

Nilsson, H., Olsson, H., & Juslin, P. (2005). The cognitive substrate of subjective probability. *Journal of Experimental Psychology: Learning Memory and Cognition, 31*, 600–620.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231-259.

Oskamp, S. (1965). Overconfidence in case study judgments. *Journal of Consulting Psychology, 63,* 81-97.

Pain, M. D., & Sharpley, C. F. (1989). Varying the order in which positive and negative information is presented: Effects on counselors' judgments of clients' mental health. *Journal of Counseling Psychology*, *36*(1), 3-7.

Peterson, R. C. (2003). *Mild cognitive impairment: Aging to Alzheimer's disease*. London, England: Oxford University Press.

Petersen, R. C., & Morris, J. C. (2005). Mild cognitive impairment as a clinical entity and treatment target. *Archives of Neurology, 62*(7), 1160-1163. doi:10.1001/archneur.627.1160

Prigatano, G. P., & Morrone-Strupinsky, J. (2010). Advancing the profession of clinical neuropsychology with appropriate outcome studies and demonstrated clinical skills. *The Clinical Neuropsychologist*, *24*(3), 468-480.

Puente, A. E. (1989). Historical perspectives in the development of neuropsychology as a professional psychological speciality. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (pp. 3-16). New York, NY: Plenum Press.

Puente, A. E., & Marcotte, A. C. (2000). A history of Division 40 (Clinical Neuropsychology). In D. A. Dewsbury (ed.), *Unification through division: Histories of the divisions of the American Psychological Association* (pp. 137-160). Washington, D.C.: American Psychological Association.

Putnam, S. H., Deluca, J. W., & Anderson, C. (1994). The second TCN salary survey: A survey of neuropsychologists Part II. *The Clinical Neuropsychologist*, *8*(3), 245–282.

Quinsey, V., Harris, G., Rice, M., & Cormier, C. (1998). *Violent offenders: Appraising and managing risk.* Washington, D.C.: American Psychological Association.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology*, *20*(1), 33-65. doi:10.1016/j.acn.2004.02.005

Rabin, L. A., Borgos, M. J., & Saykin, A. J. (2008). A survey of neuropsychologists' practices and perspectives regarding the assessment of judgment ability. *Applied Neuropsychology*, *15*(4), 264-273.

Rabin, L. A., Burton, L. A., & Barr, W. B. (2007). Utilization rates of ecologically oriented instruments among clinical neuropsychologists. *The Clinical Neuropsychologist, 21*(5),727-743. doi:10.1080/13854040600888776

Rabin, L. A. (2001). *Test usage patterns and perceived ecological utility of neuropsychological assessment techniques: A survey of North American clinical neuropsychologists.* Ann Arbor, MI: ProQuest.

Reitan, R. (1964). Psychological deficits resulting from cerebral lesions in man. *The frontal granular cortex and behavior.* New York, NY: McGraw-Hill.

Russell, E. W. (1995). The accuracy of automated and clinical detection of brain damage and lateralization in neuropsychology. *Neuropsychology Review*, *5*(1), 1-68. doi:10.1007/BF02214929

Ryan, J. J., & Paolo, A. M. (1990). Credentials and clinical activities of internship supervisors in neuropsychology: A comparison of VA and non-VA training sites. *Archives of Clinical Neuropsychology*, *5*(1), 69-75. doi:10.1016/08876177(90)90008-D

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66,* 178-200.

Schank, R. C., & Abelson, R. P. (1977). *Scripts plans goals and understanding An inquiry into human knowledge structures*. Hillsdale, NJ: Erlbaum.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Sharland, M. J., & Gfeller, J. D. (2007). A survey of neuropsychologists' beliefs and practices with respect to the assessment of effort. *Archives of Clinical Neuropsychology*, *22*(2), 213–223.

Slick, D. J., Tan, J. E., Strauss, E. H., & Hultsch, D. F. (2004). Detecting malingering: a survey of experts' practices. *Archives of Clinical Neuropsychology*, *19*(4), 465–473.

Smith, E. R., & Miller, F. D. (1978). Limits on perception of cognitive processes: A reply to Nisbett and Wilson. *Psychological Review, 85*, 355-362.

Sundberg, N. D. (1961). The practice of psychological testing in clinical services in the United States. *American Psychologist*, *16*(2), 79-83. doi:10.1037/h0040647

Sweet, J. J., Moberg, P. J., & Suchy, Y. (2000a). Ten-year follow-up survey of clinical neuropsychologists: Part I: Practices and beliefs. *The Clinical Neuropsychologist*, *14*(1), 18-37. doi:10.1076/1385-4046(200002)14:1;1-8;FT018

Sweet, J. J., Moberg, P. J., & Suchy, Y. (2000b). Ten-year follow-up survey of clinical neuropsychologists: Part II: Private practice and economics. *The Clinical Neuropsychologist*, *14*(4), 479-495. doi:10.1076/clin.14.4.479.7201

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science san improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*(1), 1-26.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn & Bacon.

Trueblood, W., & Binder, L. M. (1997). Psychologists' accuracy in identifying neuropsychological test protocols of clinical malingerers. *Archives of Clinical Neuropsychology*, *12*(1), 13–27.

Tversky, A., & Kahneman, D. (1974). Judgments under uncertainty: Heuristics and biases. *Science*, *185*, 1124-1131.

Tversky, A., & Kahneman, D. (1982). Judgements of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky, (Eds.), *Judgments under uncertainty: Heuristics and biases* (pp. 422-444). London, England: Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgments. *Psychological Review, 90*, 293–315.

Wainer, H. (1978). On the sensitivity of regression and regressors. *Psychological Bulletin, 85,* 267-273.

Wechsler, D. (1955). *Manual for the Wechsler Adult Intelligence Scale.* Oxford, England: Psychological.

Wedding, D. (1983). Comparison of statistical and actuarial models for predicting lateralization of brain damage. *Clinical Neuropsychology*, *5*(1), 15-20.

Wedding, D. (1991). Clinical judgment in forensic neuropsychology: A comment on the risks of claiming more than can be delivered. *Neuropsychology Review, 2*(3), 233-239.
doi:10.1007/BF01109046

Wedding, D., & Faust, D. (1989). Clinical judgment and decision making in neuropsychology. *Archives of Clinical Neuropsychology*, *4*(3), 233-265.
doi:16/0887-6177(89)90016-4

Welsh-Bohner, K. A. (2008). Defining "prodromal" Alzheimer's disease, frontotemporal dementia, and Lewy body dementia: Are we there yet? *Neuropsychological Review, 18*(1): 70-72.
doi:10.1007/s 11065-008-9057-y

White, P. (1980). Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bem. *Psychological Review*, *87*(1), 105-112.
doi:10.1037/0033-295X.87.1.105

Wiggins, N., & Hoffman, P. J. (1968). Three models of clinical judgment. *Journal of Abnormal Psychology*, *73*(1), 70-77.

World Health Organization. (2004). *International Statistical Classification of Diseases and Health Related Problems*. Geneva, Switzerland: Author.

APPENDICES

Appendix A

Assessment Techniques Typically Used to Assess for the Presence of Dementia

Part I: The following client is referred to you for neuropsychological testing:

Referral Information: Client is a 67-year-old, right-handed male, with 16 years of education who has been referred for testing due to suspected memory impairments.

Please check all assessment techniques you would typically use to evaluate this client if you had 3-4 hours for face to face testing:

Assessment Instruments (in alphabetical order):

__ 21-item Test
__ Aphasia Screening Exam
__ Beck Depression Inventory
__ Bender-Gestalt
__ Benton Visual Retention Test
__ Booklet Category Test
__ Boston Naming Test
__ Brief Visual Memory Tests
__ California Verbal Learning Test
__ CERAD Neuropsychological Assessment Battery
__ Clock Drawing Test
__ Cognistat/ Neurobehavioral Cognitive Status Exam
__ Connors' Continuous Performance Test
__ Controlled Oral Word Association Test
__ Delis-Kaplan Tests of Executive Functioning
__ Facial Recognition Test
__ Finger Tapping Test
__ Fuld Object Memory Evaluation
__ Galveston Orientation and Amnesia Test
__ Geriatric Depression Scale
__ Grooved Pegboard Test
__ Halstead Category Test
__ Halstead-Reitan Neuropsychological Battery
__ Hand Dynamometer/ Grip Strength
__ Hooper Visual Organization Test
__ Hopkins Verbal Learning Test
__ Judgment of Line Orientation
__ Kauffman Brief Intelligence Test
__ Luria-Nebraska Neuropsychological Battery
__ Memory Assessment Scales
__ Mini-Mental Status Exam
__ Minnesota Multiphasic Personality Inventory-II

__ Paced Auditory Serial Addition Test
__ partial LNNB
__ Personality Assessment Inventory
__ Porteus Mazes
__ Purdue Pegboard Test
__ Raven's Progressive Matrices
__ Repeatable Battery for the Assessment of Neuropsychological Status
__ Rey 15 Item Memory Test
__ Rey Auditory Verbal Learning Test
__ Rey-Osterrieth Complex Figure Test
__ Rorschach
__ Ruff-Light Trail Learning Test
__ Seashore Rhythm Test
__ Sentence Repetition
__ Speech Sounds Perception Test
__ Stroop Test
__ Tactual Performance Test
__ Test of Variables of Attention
__ Tests of Memory and Malingering
__ Token Test
__ Tower Test
__ Trail Making Test (Part A and B)
__ Validity Indicator Profile
__ Visual Form Discrimination
__ WAIS Letter-Number Sequencing
__ WAIS Vocabulary
__ WAIS/WMS Digit Span
__ Warrington Recognition Memory Test
__ Wechsler Abbreviated Scale of Intelligence
__ Wechsler Adult Intelligence Scale (WAIS III or IV) (full)
__ Wechsler Individual Achievement Test (WIAT II or III)
__ Wechsler Memory Scale (WMS III or IV) (full)
__ Wide Range Achievement Test-III
__ Wisconsin Card Sorting Test
__ WMS Family Pictures
__ WMS Information and Orientation
__ WMS Logical Memory
__ WMS Mental Control
__ WMS Spatial Addition
__ WMS Verbal Paired Associates
__ WMS Visual Reproduction
__ WMS Word Lists
__ Woodcock Johnson Tests of Achievement-III
__ other (please specify)

Nonpsychometric clinical data:

___Behavior during testing
___Family history
___Historical data
___Interview with significant family members/caretakers
___Results of clinical interview
___Review of medical records

Appendix B

Reference Vignette

Part II: Please review the following information obtained during a neuropsychological evaluation

Referral Information:

Client is a 59-year-old, right-handed Caucasian male with 14 years of education. Client referred by primary-care physician due to memory complaints. Initial interview with client and spouse indicates client is not depressed.

Test Results:

The following age-adjusted test scores were obtained during face to face testing with this client.

WAIS-IV (standard scores)	
Verbal Comprehension Index:	103
Perceptual Reasoning Index:	98
Working Memory Index:	102
Processing Speed Index:	105
Digit Span Total Score:	26
WMS-IV (raw scores)	
Visual Reproduction I:	31
Logical Memory I:	22
Spatial Addition:	10
Verbal Paired Associates I:	26
Designs I:	58
Visual Reproduction II:	19
Logical Memory II:	21
Spatial Span:	21
Verbal Paired Associates II:	10
Designs II:	51
WMS-IV summary scores (standard scores)	
Auditory Memory Index:	100
Visual Memory Index:	90
Immediate Memory Index:	95
Delayed Memory Index:	93
Visual Working Memory Index:	102
Boston Naming Test (raw score):	50
Controlled Oral Word Association Test (FAS raw score):	39
Trail Making Test (Part A):	32
Trail Making Test (Part B):	68

Hooper Visual Organization Test:	27
Brief Visual Memory Test (raw scores)	
Trial 1:	5
Trial 2:	8
Trial 3:	9
Delayed Recall:	9
Recognition (number correct):	6
Recognition—false positives:	0
Hopkins Verbal Learning Test (raw scores)	
Trial 1:	6
Trial 2:	7
Trial 3:	9
Delayed Recall:	8
Recognition (number correct):	10
Recognition—false positives:	0
Wisconsin Card Sorting Test	
Perseverative Responses	16
Categories Achieved	4
Finger Tapping Test	
Dominant hand (averaged across 5 trials):	54
Nondominant hand (averaged across 5 trials):	48
Grooved Pegboard (in seconds)	
Dominant hand:	75
Nondominant hand:	84

Ratings: Please complete the following diagnostic ratings based on the information presented above.

Instructions for ratings: Ratings are made on a 0 to 10 scale. 0 indicates that neurological impairment is definitely absent; 10 indicates that neurological impairment is definitely present; 5 indicates there is a 50/50 chance that neurological impairment is present. A rating of 8, 9 or 10 indicates that the client meets criteria for a diagnosis of neurological impairment.

1a. Rate the likelihood that any type of neurological impairment is present:

0 1 2 3 4 5 6 7 8 9 10

1b. Rate your level of confidence in your rating:

___ very low ____ low ____ moderate ____ high ____ very high

2a. Rate the likelihood that a dementia is present:

0 1 2 3 4 5 6 7 8 9 10

2b. Rate your level of confidence in your rating:

___ very low ____ low ____ moderate ____ high ____ very high

Appendix C

Young Client Vignettes

Please review the following information obtained during a neuropsychological evaluation

Referral Information:

Client is a 48-year-old, right-handed Caucasian male with 12 years of education. Client complains of memory problems. Client denied any symptoms of anxiety or depression during intake interview. There was no reported history of a head injury.

Test Results: (note: average version scores appear first, borderline version scores appear next in parenthesis, impaired version scores appear last in **bold.** Participants will view only one set of scores).

The following age-adjusted test data were obtained during face to face testing with this client.

WAIS-IV (standard scores)	
Verbal Comprehension Index:	105 (for all three levels)
Perceptual Reasoning Index:	102 (for all three levels)
Working Memory Index:	97 (for all three levels)
Processing Speed Index:	96 (for all three levels)
Digit Span Total Score:	30 (22) **18**
WMS-IV (raw scores)	
Visual Reproduction I:	35 (32) **27**
Logical Memory I:	24 (19) **15**
Spatial Addition:	14 (10) **6**
Verbal Paired Associates I:	35 (24) **17**
Designs I:	71 (57) **46**
Visual Reproduction II:	30 (14) **9**
Logical Memory II:	22 (13) **9**
Spatial Span:	24 (16) **11**
Verbal Paired Associates II:	11 (8) **6**
Designs II:	59 (43) **23**
WMS-IV summary scores (standard scores)	
Auditory Memory Index:	102 (79) **68**
Visual Memory Index:	90 (77) **67**

Immediate Memory Index:	99 (75) **69**
Delayed Memory Index:	101 (79) **65**
Visual Working Memory Index:	94 (71) **61**
Boston Naming Test (raw score):	57 (51) **48**
Controlled Oral Word Association Test (FAS raw score):	39 (32) **25**
Trail Making Test (Part A):	26 (36) **48**
Trail Making Test (Part B):	58 (70) **88**
Hooper Visual Organization Test:	28 (16) **9**
Brief Visual Memory Test (raw scores)	
Trial 1:	7 (5) **3**
Trial 2:	10 (8) **6**
Trial 3:	11 (9) **7**
Delayed Recall:	10 (8) **7**
Recognition (number correct):	6 (5) **4**
Recognition—false positives:	0 (0) **1**
Hopkins Verbal Learning Test (raw scores)	
Trial 1:	7 (4) **2**
Trial 2:	9 (6) **4**
Trial 3:	11 (8) **6**
Delayed Recall:	10 (5) **3**
Recognition (number correct):	11 (8) **8**
Recognition—false positives:	0 (1) **2**
Wisconsin Card Sorting Test	
Perseverative Errors	7 (14) **19**
Categories Achieved	5 (4) **3**
Finger Tapping Test	
Dominant hand (averaged across 5 trials):	52 (43) **35**
Nondominant hand (averaged across 5 trials):	48 (41) **34**
Grooved Pegboard	
Dominant hand:	65 (75) **85**
Nondominant hand:	70 (89) **95**

Ratings: Please complete the following diagnostic ratings based on the information presented above.

Instructions for ratings: Ratings are made on a 0 to 10 scale. 0 indicates that neurological impairment is definitely absent; 10 indicates that neurological impairment is definitely present; 5 indicates there is a 50/50 chance that neurological impairment is present. A rating of 8, 9 or 10 indicates that the client meets criteria for a diagnosis of neurological impairment.

1a. Rate the likelihood that any type of neurological impairment is present:

0 1 2 3 4 5 6 7 8 9 10

1b. Rate your level of confidence in your rating:

___ very low ____ low ____ moderate ____ high ____ very high

2a. Rate the likelihood that a dementia is present:

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

2b. Rate your level of confidence in your rating:

___ very low ____ low ____ moderate ____ high ____ very high

Appendix D

Older Client Vignettes

Please review the following information obtained during a neuropsychological evaluation

Referral Information:

Client is a 74-year-old, right-handed Caucasian male with 12 years of education. Client complains of memory problems. Client denied any symptoms of anxiety or depression during intake interview. There is no reported history of a head injury.

Test Results: (note: average version scores appear first, borderline version scores appear next in parenthesis, impaired version scores appear last in **bold.** Participants will view only one set of scores).

The following age-adjusted test scores were obtained during face to face testing with this client.

WAIS-IV (standard scores)	
Verbal Comprehension Index:	105 (for all three levels)
Perceptual Reasoning Index:	102 (for all three levels)
Working Memory Index:	97 (for all three levels)
Processing Speed Index:	96 (for all three levels)
Digit Span Total Score:	26 (19) **16**
WMS-IV (raw scores)	
Visual Reproduction I:	31 (24) **20**
Logical Memory I:	29 (22) **13**
Verbal Paired Associates I:	19 (11) **6**
Visual Reproduction II:	16 (10) **5**
Logical Memory II:	14 (11) **7**
Spatial Span:	15 (10) **6**
Verbal Paired Associates II:	9 (6) **4**
WMS-IV summary scores (standard scores)	
Auditory Memory Index:	102 (79) **68**
Visual Memory Index:	90 (77) **67**
Immediate Memory Index:	99 (75) **69**
Delayed Memory Index:	101 (79) **65**
Visual Working Memory Index:	94 (71) **61**
Boston Naming Test (raw score):	49 (45) **40**
Controlled Oral Word Association Test (FAS raw score):	32 (24) **19**

Trail Making Test (Part A): (in seconds)	38 (48) **60**
Trail Making Test (Part B): (in seconds)	95 (120) **175**
Hooper Visual Organization Test:	22 (14) **3**
Brief Visual Memory Test (raw scores)	
Trial 1:	5 (3) **1**
Trial 2:	8 (5) **4**
Trial 3:	9 (6) **5**
Delayed Recall:	9 (5) **4**
Recognition (number correct):	5 (4) **3**
Recognition—false positives:	0 (1) **1**
Hopkins Verbal Learning Test (raw scores)	
Trial 1:	5 (2) **0**
Trial 2:	7 (3) **2**
Trial 3:	9 (5) **3**
Delayed Recall:	6 (4) **2**
Recognition (number correct):	11 (9) **8**
Recognition—false positives:	2 (2) **3**
Wisconsin Card Sorting Test	
Perseverative Errors	15 (28) **39**
Categories Achieved	4 (3) **1**
Finger Tapping Test	
Dominant hand (averaged across 5 trials):	53 (45) **38**
Nondominant hand (averaged across 5 trials):	47 (41) **34**
Grooved Pegboard (in seconds)	
Dominant hand:	83 (94) **105**
Nondominant hand:	89 (103) **115**

Ratings: Please complete the following diagnostic ratings based on the information presented above.

Instructions for ratings: Ratings are made on a 0 to 10 scale. 0 indicates that neurological impairment is definitely absent; 10 indicates that neurological impairment is definitely present; 5 indicates there is a 50/50 chance that neurological impairment is present. A rating of 8, 9 or 10 indicates that the client meets criteria for a diagnosis of neurological impairment.

1a. Rate the likelihood that any type of neurological impairment is present:

0 1 2 3 4 5 6 7 8 9 10

1b. Rate your level of confidence in your rating:

___ very low ____ low ____ moderate ____ high ____ very high

2a. Rate the likelihood that a dementia is present:

0 1 2 3 4 5 6 7 8 9 10

2b. Rate your level of confidence in your rating:

___ very low ____ low ____ moderate ____ high ____ very high

Appendix E

Ratings of Information Necessity

Instructions: Please rate how necessary the clinical information presented in the previous case study was for making your diagnostic ratings.

WAIS-IV (Index standard scores)

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

WMS-IV (raw scores)

Visual Reproduction I & II:

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Logical Memory I & II:

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Verbal Paired Associates I & II:

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Designs I & II:

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Spatial Addition:

__ absolutely unnecessary

__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Spatial Span:

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Digit Span Total Score (optional):

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

WMS-IV summary scores (standard scores)

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Boston Naming Test (raw score):

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Controlled Oral Word Association Test (FAS raw score):

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Trail Making Test (Part A):

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Trail Making Test (Part B):

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Hooper Visual Organization Test:

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Brief Visual Memory Test (raw scores)

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Hopkins Verbal Learning Test (raw scores)

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Wisconsin Card Sorting Test
Perseverative Responses

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Categories Achieved

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Finger Tapping Test

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Grooved Pegboard

__ absolutely unnecessary
__ somewhat unnecessary
__somewhat necessary
__absolutely necessary

Appendix F

Demographics

Part III: Please take a few minutes to answer the following questions.

1. Gender: ___ Male ___ Female

2. Age: ___ years

3. Highest degree earned: ___ Ph.D. ___ Psy.D. ___ Ed.D. ___Other (Specify)

4. Field in which highest degree was awarded (choose one):

___ Clinical Psychology ___ Counseling Psychology
___ School Psychology ___ Other (Specify)

5. Board Certification Status (check all that apply)

___ ABPP certification
___ ABPN certification

6. How many years have you been offering neuropsychological services? _____ years

7. Approximately what percentage of your professional activity is devoted to: (totals should sum to 100%)

neuropsychological assessment	_____%
rehabilitation and/or cognitive rehabilitation	_____%
psychotherapy	_____%
research and/or teaching	_____%
other (please specify) _____	_____%

8. On average, how many neuropsychological assessment do you preform each month?

____ less than 1 ____1-15 _____16-30 ____ more than 30

9. On average, how many instruments do you administer in a typical neuropsychological assessment battery?

_____ instruments

10. Indicate in what setting you primarily preform your neuropsychological work (choose all that apply)

____ business/industry
____ college/university counseling center
____ community mental health center
____ private or group practice
____ rehabilitation facility
____ medical hospital
____ psychiatric hospital
____ VA hospital
____ other (specify)

11. Indicate your PRIMARY philosophical approach toward test selection in neuropsychological assessment: (choose ONE)

___ Flexible (based upon the needs of an individual case, not uniform across patients)
___ Flexible battery (variable but routine groupings of tests for different types of patients)
___ Standardized battery (e.g., Hallstead-Reitan, Luria-Nebraska)
___ Other (please specify) _____

12. What percentage of your professional time is spent with the following populations? (totals should sum to 100%)

children (age < 12)	_____ %
adolescents (age 12-18)	_____ %
young adults (age 19-39)	_____ %
adults (age 40-65)	_____ %
older adults (age > 65)	_____ %

Appendix G

Initial Letter to Potential Participants

Dear _____ (name)

As part of my doctoral dissertation, I am surveying the perspectives and decision-making practices of clinical neuropsychologists. Respondents will include randomly chosen professional members of the International Neuropsychological Society (INS). As the practice of clinical neuropsychology continues to evolve, it is important to track various aspects of the subspecialty.

You are receiving this letter because you have been randomly selected to participate in this survey. The survey itself is web-based to facilitate easy responding. In approximately five days, you will receive an email with a link to the survey. The foreseeable risks (e.g., loss of time) for participating in this study are minimal. The estimated completion time of the survey is approximately 30 minutes.

All responses are confidential and will released only as summary findings; individual responses will not be identified. After you complete the survey, you name will be deleted from the mailing list and will not be linked to your responses. Participation in this survey is voluntary. If you prefer not to respond please reply to this email with "Opt-Out" in the subject line. Final copies of the report will be made available upon request. Please direct such inquires to the following email address: hxzp@iup.edu.

Questions regarding your participation in this study can be answered by Kristina Talbert, by email (k.l.talbert@iup.edu) or telephone (512-923-7712). This project has been approved by the Indiana University of Pennsylvania Institutional Review Board for the Protection of Human Subjects (telephone: 724-357-7730).

Thank you for your time and participation. Your cooperation is greatly appreciated!

Sincerely,

Kristina L. Talbert, M.A
Principal Investigator

David J. LaPorte, Ph.D
Professor and Director of Clinical Training
Indiana University of Pennsylvania

Appendix H

Follow-up Letter for Nonresponse

Dear Colleague,

Two weeks ago a request to participate in a research study and a link to a web-based survey were sent to you. The survey dealt with clinical decision-making and other practice issues as part of a doctoral research project. Your name was randomly drawn randomly from the online membership directory of the International Neuropsychological Society (INS).

Our records indicate that you have not yet responded to the survey. If this is an error, please let me know by replying to this email. If not, I would greatly appreciate your participation. The link to the survey is provided again in this email. The estimate completion time of the survey is 30 minutes. Please feel free to contact me if you have any questions about the study.

Again, thank you for your time and participation. Your cooperation is greatly appreciated!

Sincerely,

Kristina L. Talbert
Principal Investigator