2020

IMPACT OF BRAND RECOGNITION IN PSYCHOLOGICAL TEST SELECTION: IDENTIFICATION AND CORRECTIVE PROCEDURE

Aaron M. Baker
University of Rhode Island, aaron.m.baker@gmail.com

IMPACT OF BRAND RECOGNITION IN PSYCHOLOGICAL TEST SELECTION: IDENTIFICATION AND CORRECTIVE PROCEDURE

BY

AARON M. BAKER

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

PSYCHOLOGY

UNIVERSITY OF RHODE ISLAND

2020

DOCTOR OF PHILOSOPHY DISSERTATION

OF

AARON M. BAKER

APPROVED:

Dissertation Committee:

Major Professor David Faust

Ellen Flannery-Schroeder

Andrea Paiva

Hans Saint-Eloi Cadely

Nasser H. Zawia
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2020

ABSTRACT

The accuracy of psychological assessment may be determined largely by the quality of the test(s) selected; however, in clinical practice, tests may be selected impressionistically, and without sufficient consideration of test validity. Appreciating that not only psychometric standing, but also pragmatic considerations may be of import in test selection may help explain why surveys of test usage have not necessarily shown robust associations between frequency of test use and psychometric quality. The primary goal of this dissertation was to examine whether brand recognition (BR; presence/absence of the name of a well-known test) may sometimes diminish attention to psychometric qualities, and thus, when brand recognition exceeds test quality, impede optimal test selection. Participants (N = 123) were neuropsychologists and graduate students trained in neuropsychological assessment. This study explored the impact of BR in three primary areas: (1) appraising test-retest reliability; (2) estimating error in obtained scores; and (3) estimating the true discrepancy between two scores. Contrary to the hypothesized results, BR did not result in significant differences across any of the variables, an encouraging outcome suggesting that judgments were not swayed by a potential biasing factor. The null results, however, may have been due to focusing too heavily on judgment tasks (e.g., rating psychometric quality) that were assumed to be inherent to test selection, but instead may be partially independent. Certain interpretive practices based on configural relationships may be particularly vulnerable to test selection that places limited emphasis on psychometric adequacy. For example, study results suggested that participants markedly overperceived normal levels of scatter as rare or aberrant, and

that some neuropsychologists may not sufficiently account for measurement error. Although this study yielded positive or encouraging findings, given the frequent discordance between psychometric standing and frequency of test use found in survey research, concerns remain that BR or other variables can impede test selection and warrant further examination.

ACKNOWLEDGMENTS

I want to recognize and send my sincerest gratitude to my mentors, David Faust and Ellen Flannery-Schroeder, for their endless patience and support. They have demonstrated the most profound commitment to my professional and personal development. I also thank the other members of my dissertation committee, Andrea Paiva and Hans Saint-Eloi Cadely. This project is successful thanks only to the breadth of expertise bestowed generously by my committee. My committee chair, Joan Peckham, deserves special thanks. Her insight and support are greatly appreciated.

I am most grateful to be honored with a family that offers unwavering support across every avenue of my life. My wife, Eva Nieto-Baker, and children, Seven, Evangelina, and Kahlo, offer unmatched joy, laughter, adventure, wonder, authenticity, and love. The love we choose to share with each other provides an intrinsic complement to the balance of my life. My parents and three siblings offer unparalleled genuineness. They have granted me through lessons, modeling, and sincere conversations the experiences to appreciate the intrinsic value of ethics, virtue, integrity, resilience, humor, and love. The collection of each of their individual accomplishments will always bolster my own pursuits.

I dedicate this dissertation to my brother, Noah Jonathan Baker. His early passing, during the late stages of his Ph.D. education, presented a void in my life that will always persist. I strive to recognize this void as a gift to be cherished; a reminder for myself to model his genuine nature of humility. Paramount to all his decisions was the impact it had on the happiness and well-being of others. I experience reminders of you, Noah, on a daily basis, and while difficult, these memories offer certain solace

that helps guide my days. While quoted from the film Harvey, you instilled into me the profound appreciation of "'In this world, Elwood, you must be oh so smart or oh so pleasant.' Well, for years I was smart. I recommend pleasant. You may quote me."

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The accuracy of psychological assessment may be largely determined by the quality of the test(s) selected; however, in clinical practice, tests may be selected impressionistically, and without sufficient consideration of test validity. Although often not formally established, practice guidelines are available for test selection, which indicate that tests should meet standards of psychometric adequacy related to such qualities as acceptable levels of reliability, validity, and normative standards. Appreciating that there are numerous pragmatic considerations beyond psychometric adequacy that are relevant in test selection (e.g., time and cost), it is perhaps not surprising that surveys of psychological test usage have not shown a robust associations between frequency of test usage and psychometric quality. It is posited that decision-making in test selection is influenced by suboptimal processes related to limitations in clinical judgment and bias. As one example, brand recognition may impact test selection significantly. A scientific basis is lacking to evaluate and, when needed, reduce or eliminate the impact of factors that can degrade test selection.

It is further concerning that certain interpretive practices (e.g., scatter analysis) may be particularly vulnerable to test selection that places limited emphasis on psychometric adequacy. The primary goal of this study was to examine whether brand recognition (presence/absence of the name of a well-known test) negatively impacts neuropsychologists' appraisals of tests' psychometric qualities. More specifically, the

present study examined: (1) the impact of brand recognition on clinical judgment tasks that rely on test-retest reliability, (2) the effectiveness of a corrective procedure for diminishing or eliminating potential negative influences of brand recognition, and (3) clinical interpretive practices most susceptible to such impact.

CHAPTER 2

REVIEW OF LITERATURE

Test Selection

Neuropsychological organizations, such as the American Academy of Clinical Neuropsychology (AACN), have provided practice guidelines related to test selection (Board of Directors, 2007). The AACN guidelines indicate that tests must meet standards of psychometric adequacy related to acceptable levels of reliability, validity, and normative standards. The range of tests available in neuropsychology is vast, as demonstrated in authoritative texts (e.g., Lezak, Howieson, Bigler, & Tranel, 2012; Mitrushina, Boone, Razani, & D'Elia, 2005; Strauss, Sherman, & Spreen, 2006) and surveys of neuropsychologists' practices (Rabin, Barr, & Burton, 2005; Rabin, Paolillo, & Barr, 2016). Test manuals and published articles often provide considerable information on the psychometric qualities of various tests, which can be used to appraise their properties and standing. However, such information often conflicts across sources. Therefore, concern arises as how to optimally combine this information when appraising test quality without formal guidance.

The complexity and challenges related to ideal test selection have gained considerable attention (Bilder, 2011; Board of Directors, 2007; Brooks, Strauss, Sherman, Iverson, & Slick, 2009; Bush, 2010; Bush, Sweet, Bianchini, Johnson-Greene, Dean, & Schoenberg, 2018; Wong, 2006). Unfortunately, there is limited formal guidance available. Ideally, test selection should be guided primarily by the

test's capacity to assess, appraise, or detect what it purports to measure, taking into account, where applicable, the comparative quality of other measures that might be available and that are designed to measure the same thing(s). There are, of course, pragmatic concerns not intrinsically related to psychometric quality that deserve consideration (e.g., length and time to administer, cost, screening versus comprehensive testing). Certain domains or purposes of testing may also vary in how competing variables are prioritized (e.g., differences in test characteristics would be expected among psychological screening, personality testing, intellectual testing, and various aspects of neuropsychological testing). However, when test selection is determined through inferior methods, the result may well be diminished accuracy.

In neuropsychological assessment, this problem could be made exponentially worse with each additional test selected. Neuropsychological assessment often involves a combination of tests that could range from two tests to three tests to upwards of 40 tests/measures. Subsequently, the combinations between tests and corresponding co-joint properties grow exponentially. The complexity of this issue has been illustrated in multiple studies demonstrating that frequency of discrepancy across test scores and occurrence of low test scores change dramatically as the number of tests and comparisons multiply (Binder, Iverson, & Brooks, 2009; Brooks, Iverson, Sherman, & Holdnack, 2009; Schretlen, Munro, Anthony, & Pearlson, 2003; Schretlen, Testa, Winicki, Pearlson, & Gordon, 2008).

Psychometric Quality: Test-retest Reliability

Addressing each relevant feature in test selection (including both psychometric and pragmatic variables) is beyond the scope of this dissertation. Instead, this study

focuses on a specific psychometric variable that is particularly germane to test selection — test-retest reliability (Calamia, Markon, & Tranel, 2013). Speaking in more general terms, Lareau and Ahern (2012) described reliability as "Consistency and stability. Assuming the characteristic in question has not changed, if a test demonstrates reliability, the same or similar score should be obtained if the test is administered in the same manner or if different people administer the test in the same manner" (p. 282). Reliability is traditionally measured along any of three dimensions (internal consistency, test-retest, and inter-rater), each of which has value in its own respect.

The most common metric for measuring reliability is the correlation coefficient, which ranges from +1.0 to -1.0. In this context, the extremes represent perfect correlation, whereas a correlation of .00 represents measurement that is pure error (or randomness/chance) (Faust, 2012). The reliability coefficient can be understood most basically as indicating the extent to which an observed score represents *true measurement* as opposed to *error.* For example, if a test has a reliability of .60, then 60% of the observed score can be understood as true measurement and 40% as error variance. Therefore, if a test is highly reliable, then changes in scores are likely to reflect, at least for the most part, true changes versus measurement error. Test reliability (at least conceptually) is a necessary, but not sufficient, condition for test validity. As such, a test with deficient reliability will have deficient validity, but a test with satisfactory reliability may or may not be valid (Lareau & Ahern, 2012).

Test-retest reliability reflects a test's stability over time. There is no clear consensus regarding qualitative ranges for appraising test-retest reliability coefficients.

However, Strauss, Sherman, and Spreen (2006) provide the following guide: <.60 = low; .60-.69 = marginal; .70-.79 = adequate; .80-.89 = high; and .90+ = very high. Although there are discrepancies across professionals regarding such ranges and designations, there is general acceptance that in most situations reliability <.60 is unacceptably low and >.80 is moderate to high (Mitrushina, Boone, Razani, & D'Elia, 2005). It should be recognized that test-retest reliability is not a fixed quality, and appraising the acceptable range of reliability may vary across such dimensions as the domain being assessed, the length of the test-retest interval, and the clinical population of interest (Calamia, Markon, & Tranel, 2013; Duff, 2012).

Clinical Judgment and Bias: Brand Recognition

It may well be, given differences in belief and variation in the amount of psychometric information available or accessed in making choices, that test selection is determined primarily by clinical or impressionistic judgment. However, decision-making literature has identified various factors and judgment practices that can reduce accuracy below that which might otherwise be achieved when one relies primarily on more formal procedures, and some circumstances in which the rate of error can be seriously concerning. Research shows that both cognitive limitations and biases can limit or compromise judgmental accuracy (Faust, 1984; Faust & Ahern, 2012; Wedding & Faust, 1989). Examples of bias that may impact test selection includes confirmation bias (Nickerson, 1998), hindsight bias (Arkes, 1981), and the judged validity effect (Arkes, Boehm, & Xu, 1991; Arkes, Hackett, & Boehm, 1989).

Confirmation bias is the tendency of giving an unfair advantage to one's initial/favored belief (Nickerson, 1998). Under this phenomenon, individuals seek out

or interpret evidence in ways that are partial to their existing beliefs or expectations. Hindsight bias refers to the tendency to perceive outcomes, once known, as more predictable (in hindsight) than they truly are in foresight. Of particular relevance, the judged validity effect (or truth effect) refers to the potentially robust association between the number of times one hears about or is exposed to something and its perceived level of validity or quality, even if little or no true association exists.

Consumer research has demonstrated the impact of brand recognition on consumer preferences (Hauser, 2011; Thoma & Williams, 2013), which may occur as the result of a recognition heuristic. The recognition heuristic (Gigerenzer and Goldstein, 2011) posits that an object that is recognized will be judged to have more value, which has similarities to the judged validity effect. It seems likely that this effect would translate to professional decision-making in appraisal of test quality. Practitioners may often select tests with consideration of familiarity or early training, as opposed to psychometric quality alone. Rabin and colleagues (2016) suggested that surveys on the frequency of test usage are used to inform and likely guide test selection. Therefore, it is concerning if a non-optimal predictor (brand recognition) exerts a significant impact on the appraisal of test quality, which could in turn degrade clinical inferences.

Is Concern Warranted?

The potential problem of brand recognition exerting a negative influence on test selection may be substantial and pervasive within psychological assessment. For example, it may result in the selection of tests with unacceptably low reliability coefficients (Camara, Nathan, & Puente, 2000; Lees-Haley, Smith, Williams, & Dunn,

1996; Rabin, Barr, & Burton, 2005; Rabin, Paolillo, & Barr, 2016). Among other things, poor reliability increases the margin of error, sometimes to extreme levels, erodes the analysis of differences in scores across tests and possible test score patterns, and impacts the determination of expected scores on retesting (Faust, 2012).

Test-retest reliability coefficients can be used to adjust an observed score to determine the individual's most likely true score, or the score that is most likely to be obtained upon re-administration. This adjusted score might be thought of as the individual's *expected true score*. Putting aside the possibility of practice effects for the moment, the expected true score can be calculated by multiplying the observed score's difference from the mean (e.g., the z-score) by the test-retest reliability coefficient. To the extent obtained scores deviate from the mean, tests with poor reliability generate dramatic shifts when estimating the true score. For example, if a measure has a test-retest reliability of $r = .50$, then a score converted to $z = -2.0$ (or two standard deviations below the mean) would have an *expected true score* of $z = -1.0$; and a score converted to $z = -1.0$ would have an *expected true score* of $z = -0.5$. If such a substantial predicted shift is not recognized, it may have a highly detrimental impact on clinical inference. For example, this potentially impacts inferences made through the use of cut-scores (Charter, 2003; Charter & Feldt, 2001), or through the analysis of inter-test variability (i.e., examining the interrelations between scores, scatter analysis). Even if the potential impact on estimates of the true scores is recognized, the use of scores with such large error components or regression effects are often of little true value, or can easily lead to erroneous conclusions.

Clinical interpretive practices often rely on scatter – relative variability between high and low scores. Interpretation of scatter attends to the relationship between patterns of high and low test scores and comparison of such to expectations about *normal* versus *abnormal* test performance. Although limitations in scatter analysis have been recognized for well over half a century (e.g., Schofield, 1952), the appraisal of intra- and inter-test scatter[1] remains one of the most common approaches to the psychological evaluation of cognitive function and brain disorders (Lezak, et al., 2012). However, clinicians frequently underestimate normal level of scatter (Schretlen, Munro, Anthony, & Pearlson, 2003), leading to overidentification of pathology. The variability between test scores and measures is often altered by various factors that introduce artificial scatter into a profile, for example, the number of tests administered (Binder, Iverson, & Brooks, 2009; Brooks, Strauss, Sherman, Iverson, & Slick, 2009), and also low reliability, or the magnitude of error in the scores. Using a test with problematically low reliability may also significantly alter the interrelations among other test scores. Therefore, suboptimal test selection has the potential to worsen an already common, problematic judgment practice, and the impact may be pervasive. This dissertation explored whether the evaluation of a critical psychometric quality (i.e., test-retest reliability) would be compromised by a secondary, potentially irrelevant factor (i.e., brand recognition) when appraising test quality.

Hypotheses

H1: Participants provided with the name of a well-known test (i.e., CVLT-II)

[1] Scatter may refer to inter-test variability, which relates to the variability in scores across multiple tests, or intra-test variability, which relates to variability in scores within a single test (e.g., the subtests/indices within a single test).

would assign higher ratings of psychometric quality overall than participants kept blind to the name of the test. It was further hypothesized that an interaction would occur between ratings of psychometric quality and level of reliability, and more specifically that brand recognition would exert a significantly greater impact on ratings of psychometric quality when level of reliability was relatively ambiguous (i.e., falls between extremes of acceptable versus unacceptable quality, or "mediocre") versus more extreme (i.e., "good-excellent" and "poor").

H2: Presence (versus absence) of brand recognition would lead participants to underestimate the level of uncertainty or error in obtained scores. Here again, an interaction was hypothesized between underestimating error in obtained scores and level of reliability, with brand recognition expected to exert a significantly greater impact when level of reliability was relatively ambiguous versus more extreme.

H3: A corrective procedure (i.e., graphically displaying error variance) at the "mediocre" level of reliability and presence of brand would reduce the negative impact of brand recognition at that level. This impact would occur for both rating of psychometric quality and the probability that the examinee's estimated true score indicates deficit.

H4: Presence (versus absence) of brand recognition would lead participants to assign larger discrepancy figures when estimating the true variability between two scores. Here again, an interaction was hypothesized between estimating the true discrepancy between two scores and level of reliability, with brand

recognition exerting a significantly greater impact when level of reliability is relatively ambiguous versus more extreme.

H5 (exploratory): Participants would underestimate the level of scatter found in healthy individuals.

CHAPTER 3

METHODOLOGY

The primary goal of this dissertation was to examine whether brand recognition (i.e., presence/absence of the name of a well-known test) negatively impacts neuropsychologists' appraisals of tests' psychometric quality. The psychometric index of interest was test-retest reliability. This study primarily examined: (1) the impact of brand recognition on clinical judgment tasks that rely on test-retest reliability, (2) the effectiveness of a corrective procedure for diminishing or eliminating potential negative influences of brand recognition, and (3) clinical interpretive practices most susceptible to such impact. The study followed American Psychological Association ethical guidelines and was approved by the University of Rhode Island Institutional Review Board on Human Subjects.

Participants

Participants included professional psychologists with a focus in neuropsychology. A small percentage of participants were graduate students with a focus in neuropsychology, which is addressed below. Participants were initially recruited from the NPSYCH Listserv (approximate number of listserv members = 3500), an e-mail discussion list devoted to practice and research in adult neuropsychology. It is one of the more active neuropsychology listservs and is only open to neuropsychologists and other related specialist and researchers. Almost all members are expected to have had specialized training in neuropsychological assessment.

During the recruitment phase that was initially planned, it became apparent that the sample size would be suboptimal. Therefore, recruitment entered two subsequent phases. Following the initial postings to the NPSYCH Listserv, cross-posted recruitment e-mails were forwarded to the American Academy of Clinical Neuropsychology (AACN) Listserv (approximate number of members = 1700). Recruitment e-mails were also forwarded directly to members of AACN, which consists of board-certified neuropsychologists (approximate number of members = 1100). The sample pool therefore consisted of subscribers to the NPSYCH and/or AACN Listservs or neuropsychologists who are members of AACN but may not be subscribers to the Listserv. Recruitment e-mails acknowledged that the study was cross-posted. A brief description of the study with a link to the survey (hosted by www.qualtrics.com) was posted to the NPSYCH Listserv on three occasions and to the AACN Listserv on two occasions. Recruitment e-mails sent directly to AACN members were delivered on only one occasion. The Qualtrics program randomly provided participants with one of the seven possible stimuli (i.e., vignettes), as detailed below.

It became apparent that data collection was suboptimal and feasible recruitment strategies were becoming exhausted. Therefore, the last solicitation for participation sent to the AACN Listserv (i.e., the second recruitment posting to that listserv) narrowed the recruitment to include only two potential cells (i.e., poor level of reliability + absence of brand recognition vs. poor level of reliability + presence of brand recognition, which are detailed later). If overall data collection remained

suboptimal, it was believed that these groups would provide the most meaningful information based on trends in the preliminary data.

The current study had 236 individuals initiate participation. However, 113 participants discontinued immediately following the demographic portion of the study. It was unclear as to the reason why so many participants discontinued participation. Perhaps following the demographic portion of the survey, participation may have appeared more cognitive demanding or cumbersome than participants initially preferred. These potential participants did not provide any meaningful data regarding the dependent variables, and, therefore, were excluded from any analysis, as their responses only provided demographic information. The remaining participants (N = 123) were included within the study. There were no exclusionary criteria based on demographic features.

Out of the 123 participants who completed the study, some participants did not provide responses to all demographic variables and, on rare occasions, did not respond to one of the dependent variables. For example, only 122 participants responded to the question regarding gender, 122 participants responded to the first dependent variable, and 117 participants responded to the second dependent variable. Pairwise deletion of missing data occurred during statistical analyses.

Demographic features of the sample are summarized in Table 1 (see Appendix A for the demographic questionnaire). The sample was predominantly White (88.6%). Seventy-four participants (60.7%) identified as female and the remaining 48 participants (39.3%) identified as male, with the exception of missing data for one participant. Eleven participants (8.9%) reported their level of education as M.A./M.S.

These participants were included in the final analysis as their membership within the sample pool suggests they have had specialized training in neuropsychological assessment, or at least reasonable familiarity. Inclusion of predoctoral level participants introduces limitations to the generalizability of this study. All other participants (n = 112, 91.1%) reported having either a Ph.D. or Psy.D. Years since highest degree was categorized in one of four categories (<5 years, 5-10 years, 11-20 years, or >21 years), which was split about evenly across participants. The majority of participants were licensed psychologists (n = 105, 85.4%) and 53 participants (43.1%) were also board certified in clinical neuropsychology. Eighty-two participants (66.7%) indicated spending 76-100% of their professional time commitment on neuropsychological evaluations.

Procedure

Participants were provided with a brief vignette that included test-retest reliability coefficients at one of three levels (good-excellent, mediocre, and poor) from the California Verbal Learning Test-II (*CVLT-II*: Delis, Kramer, Kaplan, & Ober, 2000). The CVLT-II was selected because it is one of the most frequently used assessment instruments in clinical neuropsychology, especially within the domain of memory (Rabin, Paolillo, & Barr, 2016). The resultant measures from the CVLT-II include multiple subtest scores, indices, and process scores, which yield widely varying levels of test-retest reliability (i.e., ranging from r = .27 to .88), as reported in the test manual (Delis et al., 2000).

The CVLT-II is an individually administered test assessing learning and recall of verbal information. It comprises a 16-word list (List A) that is presented five times,

and then an interference 16-word list (List B) that is presented once. Each list is comprised of four words from four semantic categories (e.g., animals). Following presentation of List B, the examinee is immediately asked to recall List A spontaneously and then is provided with recognition cues (i.e., prompts for the semantic categories included in List A). Following a 20-minute delay, the examinee is again asked to recall List A spontaneously, and then again after cueing. This process is followed by a recognition trial (i.e., 48 items are presented and, for each word, the individual is asked whether or not it came from List A), and then an optional forced-choice recognition trial (i.e., 16 word pairs are presented and, for each pair, the individual is asked to identify the word included in List A), which was designed to serve as an embedded measure of performance validity or effort.

Each participant received one of seven stimuli and was asked to: (1) rate the overall psychometric quality of the reliability figures that are provided, (2) engage in two decision tasks, (3) indicate a dividing point of scatter commonly used to distinguish between normal and abnormal levels of variability, (4) rate the importance of measurement error in neuropsychological assessment, and (5) indicate level of familiarity with and frequency of use of the CVLT-II. In order to examine the potential impact of branding, half of the participants were informed that the reliability coefficients come from the CVLT-II, and the other half informed that the figures come from a test that is generically described as a memory measure.

This study utilized a partially crossed, independent, between-groups design with three independent variables, four dependent variables, and three potential covariates (see Appendix B). The primary design consisted of six groups (2 [presence/absence of

brand] X 3 [reliability level]). A seventh group, which served to test a corrective procedure, was partially crossed (1 X 1).

Independent Variables, Covariates, and Corrective Procedure

The first independent variable (brand recognition) had two conditions: presence vs. absence of the name of the assessment tool (i.e., the CVLT-II). The second independent variable (level of reliability) had three levels: good-excellent, mediocre, and poor. To obtain significant separation between groups, test-retest reliability data were selected from the CVLT-II at the higher, middle, and lower ranges. The groups were denoted as *good-excellent* (r = .88, .86, & .82), *mediocre* (r = .61, .57, & .56), and *poor* (r = .36, .30, & .27). This comparison allowed for an evaluation of the degree to which brand recognition had an impact across varying levels of test-retest reliability.

Participants were provided with a vignette with either the presence/absence of the test name and one of three levels of test-retest reliability coefficients (good-excellent, mediocre, poor)[2]. Thus, there were six different vignettes, given the 2 X 3 study design (see Appendix C-H for the vignettes, which also includes details of the dependent variables and covariates listed below). The only differences among the vignettes was: (1) the portion of the vignette that does or does not specifically name the test from which the reliability figures originate and (2) the level of test-retest reliability coefficients.

[2] Judgment tasks that participants engage in require one to three reliability coefficients from the respective levels, as opposed to considering all 3 for each question (e.g., participants receiving the "poor" test-retest reliability figures were provided r = .36, .30, & .27 for one question and just .27 for two questions).

The CVLT-II is published by one of the most prolific and well known psychological test companies, The Psychological Corporation. Therefore, brand recognition may well exert an influence, even if the participant does not have much familiarity with the CVLT-II (or other editions of the measure). In the event that a participant's responses are influenced by his/her prior knowledge of the CVLT-II, controlling for this potential influence will be important. Therefore, the present study attempted to control for three specific covariates. The first covariate addressed the participant's familiarity with the CVLT-II. The question addressing the first covariate, which appeared at the end of the vignette provided to participants, was: "How familiar are you with the CVLT-II?" The second covariate addressed the participant's frequency of CVLT-II use, as assessed by the following question: "When assessing memory abilities, how frequently do you use the CVLT-II?" A third potential covariate measured participants' perception of the degree to which measurement error is impactful during the interpretation of neuropsychological tests.

If brand recognition were to lead participants to underestimate psychometric problems, intervening to reduce this negative impact would be beneficial. As described earlier, a test score consists of true measurement and an error component. Graphically displaying the error variance with respect to the measure's test-retest reliability was expected to attenuate the impact of brand recognition. For example, if a test-retest reliability coefficient is .50, then the graphical display would be a pie chart that consists of 50% shaded blue (true measurement) and 50% shaded red (error). This third independent variable (graphical display of error variance) was crossed with presence of brand recognition at the mediocre level of reliability. Therefore, this

partially-crossed independent variable had two conditions: mediocre reliability + presence of brand recognition vs. mediocre reliability + presence of brand recognition + graphical display of error variance (see Appendix I for the corrective procedure vignette).

Dependent Variables

This study included four dependent variables: (1) rating of psychometric quality (DV1), (2) estimating the probability that the estimated true score indicates a deficit (DV2), (3) estimating the true discrepancy between two scores (DV3), and (4) judgments regarding level of scatter commonly used to distinguish between normal and abnormal levels of variability (DV4). Participants were first provided with a brief vignette, which provided a basis for addressing questions used to assess the dependent variables. The vignette included the presence or absence of brand recognition, along with test-retest reliability coefficients. Following the vignette, participants responded to questions related to the four dependent variables, and then the three covariates.

The question for the first dependent variable was: "How would you rate the psychometric quality of the overall test-retest reliabilities?" Next, a brief illustration of a hypothetical clinical case was provided, followed by presentation of a question addressing the second dependent variable: "…what is the probability that the examinee's estimated true score would fall below the -1 *SD* cut-off point?" Then, another brief illustration of a hypothetical clinical case was provided, followed by presentation of a question addressing the third dependent variable: "Which figure below best fits your estimation of the true discrepancy between the scores?" See

Appendices J-K for the method used to determine the accuracy of the responses to DV2 and DV3.

The fourth dependent variable evaluated clinical interpretive practices related to scatter analysis. Participants were asked: "…which of the following dividing points for the maximum discrepancy between highest and lowest score most closely matches the one you would use in distinguishing between normal and abnormal levels of variability?" Responses were measured on a continuous scale using standard deviation units.

Analyses

To achieve 80% statistical power, an overall sample of 245 (35 per cell) was needed. A medium effect size ($f = .25$) was anticipated. An *a priori* power analysis (based on calculations using G*Power 3.1.3) was conducted on the brand recognition groups at all three levels of reliability to calculate an adequate cell size. The power analysis was designed utilizing an initial plan of two separate ANOVAs serving as primary focus in this study. Therefore, a Bonferroni correction was employed to maintain an overall type I error rate of 5%. Thus, the α level for each analysis was set at .025. The statistical analytic techniques assumed in the power analysis were ultimately modified, but the above procedure still served to guide initial recruitment and study design. Given the significantly lowered sample size, if the null hypotheses were rejected (and demonstrated a medium effect size), the statistical power would subsequently be dramatically lowered.

The original plan considered utilizing a Multivariate Analysis of Variance (MANOVA) with subsequent follow up Analysis of Variance (ANOVA) and Tukey

Tests as indicated. However, it was determined instead to use three separate two-way ANOVAs to evaluate the first three dependent variables (respective to Hypotheses 1, 2, and 4) and two separate one-way ANOVAs regarding the corrective procedure of Hypothesis three. Given that the overall sample size was suboptimal and there was unequal sized groups due to narrowed recruitment during the last phase of recruitment that was aimed at increasing *n* in select cells, consideration of whether it was necessary to conduct non-parametric statistical analysis occurred. Levene's test also indicated unequal variances for the first dependent variable of rating of psychometric quality ($F = 4.901$, $p = .000$). However, Levene's test did not indicate unequal variance for the other dependent variables including estimating error ($F = .740$, $p = .596$) or estimating discrepancy ($F = 1.184$, $p = .322$). Given that the groups were independent and ANOVA is robust to violations of unequal variance, it was determined, with reservation, to continue with the planned parametric statistical analyses.

H1: A two-way ANOVA was conducted to compare the main effect of brand recognition and the interaction between brand recognition and level of reliability on ratings of psychometric quality.

H2: A two-way ANOVA was conducted to compare the main effect of brand recognition and the interaction between brand recognition and level of reliability on estimating error in obtained scores.

H3: Two separate one-way ANOVAs were conducted. These were segregated for dependent variables: rating of psychometric quality and estimating error. Means were compared at the mediocre level of reliability for presence of brand and the corrective

procedure. Therefore, each analysis had only two groups and a *t*-test may have been appropriate, but the initial proposal planned for an ANOVA. In any event, post-hoc tests were not appropriate.

H4: A two-way ANOVA was conducted to compare the main effect of brand recognition and the interaction between brand recognition and level of reliability on estimating the true discrepancy between two scores.

H5 (Exploratory): Descriptive statistics explored judgments regarding level of scatter commonly used to distinguish between normal and abnormal levels of variability.

CHAPTER 4

FINDINGS

H1: Brand Recognition and Rating of Psychometric Quality

It was hypothesized that participants provided with the name of a well-known test (i.e., CVLT-II) would assign higher ratings of psychometric quality overall than participants kept blind to the name of the test. It was further hypothesized that an interaction would occur between ratings of psychometric quality and level of reliability, and more specifically that brand recognition would exert a significantly greater impact on ratings of psychometric quality when level of reliability was relatively ambiguous (i.e., falls between extremes of acceptable versus unacceptable quality, or "mediocre") versus more extreme (i.e., "good-excellent" and "poor").

A two-way ANOVA was conducted to compare the main effect of brand recognition and the interaction between brand recognition and level of reliability on ratings of psychometric quality. As shown in Table 2, the main effect for brand recognition on ratings of psychometric quality was not significant, $F(1,103) = 1.048$, $p = .308$. The interaction effect was also not significant, $F(2,103) = .233$, $p = .792$. Respective mean ratings for no-brand versus brand at the varying levels of reliability were as follows: *good-excellent* reliability, $M = 5.94$, $SD = .574$ vs. $M = 6.06$, $SD = .539$; *mediocre* reliability, $M = 3.71$, $SD = .825$ vs. $M = 3.82$, $SD = 1.074$; and *poor* reliability, $M = 2.22$, $SD = 1.263$ vs. $M = 2.62$, $SD = 1.359$. Descriptive data are summarized in Table 3. For the purpose of simplicity, descriptive data for the

corrective procedure (as addressed in Hypothesis 3) are provided within this table, as will also be done in subsequent tables containing descriptive data.

H2: Brand Recognition and Estimating Error in Obtained Scores

It was hypothesized that presence (versus absence) of brand recognition would lead participants to underestimate the level of uncertainty or error in obtained scores. Here again, an interaction was hypothesized between underestimating error in obtained scores and level of reliability, with brand recognition expected to exert a significantly greater impact when level of reliability was relatively ambiguous versus more extreme.

A two-way ANOVA was conducted to compare the main effect of brand recognition and the interaction between brand recognition and level of reliability on estimating error in obtained scores. As shown in Table 4, the main effect for brand recognition on estimating error was not significant, $F(1,98) = .918$, $p = .340$. The interaction effect was also not significant, $F(2,98) = 2.425$, $p = .094$. Respective mean scores for estimating error in obtained scores for no-brand versus brand at the varying levels of reliability were as follows: *good-excellent* reliability, $M = 8.31$, $SD = 2.243$ vs. $M = 8.12$, $SD = 2.205$; *mediocre* reliability, $M = 7.29$, $SD = 1.637$ vs. $M = 5.56$, $SD = 2.065$; and *poor* reliability, $M = 4.65$, $SD = 2.572$. Descriptive data are summarized in Table 5.

H3: Corrective Procedure

It was hypothesized that a corrective procedure (i.e., graphically displaying error variance) at the "mediocre" level of reliability, combined with presence of brand, would reduce the negative impact of brand recognition at that level. It was further

predicted that such impact would occur for rating of (1) psychometric quality and (2) the probability that the examinee's estimated true score indicates deficit. As noted above, an impact from brand recognition was not found. However, two separate one-way ANOVAs were still conducted to examine possible differences between brand and brand/corrective on: (1) rating of psychometric quality and (2) estimating error in obtained scores. Means were compared at the mediocre level of reliability for presence of brand and the corrective procedure.

Given that two separate ANOVAs were conducted, a Bonferroni correction was used to maintain an overall type I error rate of 5%, i.e., the α level for each analysis was set at .025. In regards to rating of psychometric quality, an ANOVA indicated a non-significant result, $F(1,28) = .251$, $p = .620$. As such, there was no significant difference between *brand* ($M = 3.82$, $SD = 1.074$) and *brand/corrective* ($M = 3.62$, $SD = 1.193$) when using a graphic display of error variance at the mediocre level of reliability.

In regards to estimating error in obtained scores (i.e., that the hypothetical examinee's estimated true score would fall below the -1 *SD* cut-off point), a separate ANOVA with a Bonferroni correction also yielded a non-significant outcome, $F(1,27) = 4.316$, $p = .047$. As such, there was no significant difference between *brand* ($M = 5.56$, $SD = 2.065$; indicating the 50-59% probability) and *brand/corrective* ($M = 7.15$, $SD = 2.035$; indicating the 60-69% probability) when using a graphic display of error variance at the mediocre level of reliability. Results of these ANOVAs are summarized in Table 6.

H4: Brand Recognition and Estimating True Discrepancy Between Two Scores

It was hypothesized that presence (versus absence) of brand recognition would lead participants to assign larger discrepancy figures when estimating the true variability between two scores. Here again, an interaction was hypothesized between estimating the true discrepancy between two scores and level of reliability, with brand recognition exerting a significantly greater impact when level of reliability was relatively ambiguous versus more extreme.

A two-way ANOVA was conducted to compare the main effect of brand recognition and the interaction between brand recognition and level of reliability on estimating the true discrepancy between two scores. As shown in Table 7, the main effect for brand recognition on estimating the true discrepancy was not significant, $F(1,98) = .1.727$, $p = .192$. The interaction effect was also not significant, $F(2,98) = 1.250$, $p = .291$. Respective mean scores for estimating true discrepancy between two scores for no-brand versus brand at the varying levels of reliability were as follows: *good-excellent* reliability, $M = 2.44$, $SD = .512$ vs. $M = 2.65$, $SD = .606$; *mediocre* reliability, $M = 2.29$, $SD = .469$ vs. $M = 2.19$, $SD = .544$; and *poor* reliability, $M = 1.75$, $SD = .775$. Descriptive data are summarized in Table 8.

H5: Level of Scatter for Distinguishing Normal vs. Abnormal Levels of Variability

It was hypothesized that participants would underestimate the level of scatter found in healthy individuals. Results were analyzed by means of descriptive statistics. Of the 116 participants who responded to this item, 79.3% ($n = 92$) indicated a dividing point for distinguishing between normal and abnormal levels of variability at somewhere between 1.5 SD to 3.0 SD. The mean dividing point was 2.0 *SD* to 2.5 *SD*. Slightly less than 7% of participants ($n = 8$) indicated a dividing point at, or above, 4.0

SD. Table 9 provides cumulative percentages for level of scatter judged to distinguish between normal and abnormal levels of variability. Thus, as hypothesized, a majority of participants dramatically underestimated a cutoff for determining abnormal levels of scatter when the criterion for normal scatter was Schretlen and colleagues (i.e., scatter of maximum discrepancy in standard deviations of $M = 3.4$, $SD = 0.8$) or comparable findings (Binder, et al., 2009; Brooks et al., 2009; Schretlen et al., 2003; & Schretlen, et al., 2008)..

Presence versus absence of board certification in clinical neuropsychology did not alter the above mentioned findings for distinguishing between normal and abnormal levels of variability. Participants who are board certified in clinical neuropsychology ($n = 52$) had a mean dividing point of 2.0 *SD* to 2.5 *SD* ($M = 5.65$, $SD = 1.856$). Participants who were not board certified in clinical neuropsychology ($n = 64$) also had a mean dividing point of 2.0 *SD* to 2.5 *SD* ($M = 5.33$, $SD = 1.861$). (See Table 9 for anchor points indicating maximum discrepancy in standard deviation units, i.e., 1 = 0.0 *SD*, 2 = 0.5 *SD*, 3 = 1.0 *SD*…11 = >5.0 *SD*.)

Co-variates: Appreciation of Measurement Error and Familiarity/Usage of CVLT

Participants were asked to rate the degree to which they endorsed a statement suggesting that concerns about measurement error may be overstated, selecting among options on a Likert-scale (1 = Strongly Disagree; 4 = Moderately Agree; 7 = Strongly Agree). Strong disagreement with the statement (e.g., a response of "1" or possibly even "2") would seemingly have been the expected or proper response, which would affirm the importance of appreciating measurement error. However, of the 119 participants who responded to this item, only 24.2% ($n = 30$) endorsed a rating of 1 or

2. Instead, 64.7% (*n* = 77) endorsed a rating of 3, 4, or 5 (indicating moderate agreement), and 10% (*n* = 12) provided a rating of 6 to 7 (indicating strong agreement). The mean rating indicated moderate agreement (*M* = 3.5, *SD* = 1.455). Table 10 provides frequency of responses regarding appreciation of measurement error. Similar to the findings on scatter, those with board certification in clinical neuropsychology (*n* = 53) and those without this credential (*n* = 66) showed near equivalence in ratings, with respective means of 3.49 (*SD* = 1.368) and 3.52 (*SD* = 1.532).

Participants also separately rated their familiarity with the CLVT-II and the frequency with which they use the measure on a seven-point Likert-scale, ranging from "1" (not familiar at all with the CVLT-II on the first item and never use the CVLT-II on the second item) to "7" (extremely familiar; always use). Out of the 122 participants who answered these items, the respective mean ratings indicated a moderate to extreme level of familiarity with the CVLT-II (*M* = 5.56, *SD* = 1.373) and a rate of use of 50% (when assessing memory abilities) (*M* = 3.89, *SD* = 2.009). Table 11 and 12 provides frequency of responses on the first and second items, respectively.

Presence versus absence of board certification in clinical neuropsychology, again, did not provide meaningful differences on familiarity with or use of CVLT-II. Participants who are board certified in clinical neuropsychology (*n* = 53) had a mean familiarity rating of 5.25 (*SD* = 1.385) and a mean frequency of use of 3.62 (*SD* = 1.963). Participants who were not board certified in clinical neuropsychology (*n* = 69) had a mean familiarity rating of 5.80 (*SD* = 1.324) and a mean frequency of use of 4.09 (*SD* = 2.035).

CHAPTER 5

CONCLUSION

The primary goal of this study was to examine whether brand recognition (presence/absence of the name of a well-known test) negatively impacts neuropsychologists' appraisals of tests' psychometric qualities. More specifically, the present study examined: (1) the impact of brand recognition on clinical judgment tasks that rely on test-retest reliability, (2) the effectiveness of a corrective procedure for diminishing or eliminating potential negative influences of brand recognition, and (3) clinical interpretive practices that might be most susceptible to such impact.

The accuracy of psychological assessment may be largely determined by the quality of the test(s) selected; however, in clinical practice, tests may be selected impressionistically, and without sufficient consideration of test validity. Although often not formally established, practice guidelines are available for test selection, which indicate that tests should meet standards of psychometric adequacy related to such qualities as reliability, validity, and normative standards. Appreciating that not only psychometric standing, but also pragmatic considerations may be of import in test selection (e.g., time and cost) may help explain why surveys of psychological test usage have not necessarily shown robust associations between frequency of test usage and psychometric quality. However, decision-making in test selection may also be influenced by suboptimal processes related to limitations in clinical judgment and

bias. As one potential example, this study examined how brand recognition may impact test selection.

It is further concerning that certain interpretive practices based on configural relationships (e.g., scatter analysis) may be particularly vulnerable to test selection that places limited emphasis on psychometric adequacy, given the attention directed towards patterns and interrelations among test scores. The variability between test scores and measures is often altered by various factors that introduce artificial scatter into a profile. For example, as number of tests and comparisons multiply, frequency of discrepancies across test scores change dramatically. Using a test with problematically low reliability may significantly alter the interrelations among other test scores (Brooks, et al., 2009). Research has consistently demonstrated that clinicians frequently underestimate normal level of scatter; or, inversely, clinicians frequently overinterpret scatter leading to overidentification of pathology (Binder, et al., 2009; Brooks et al., 2009; Schretlen et al., 2003; & Schretlen, et al., 2008.). Although researchers have detailed limitations in scatter analysis for well over a half a century (Schofield, 1952), the appraisal of test scatter remains one of the most common approaches to evaluation of cognitive function and brain disorders (Lezak et al. 2012). The already problematic practice of overinterpreting scatter may be worsened when psychological tests are selected that have suboptimal, or deficient, psychometric qualities. A scientific basis is lacking to evaluate and, when needed, reduce or eliminate the impact of factors that can degrade test selection.

Impact of Brand Recognition

This study explored three primary areas regarding impact from brand recognition: (1) rating psychometric quality of test-retest reliability, (2) a judgment task of estimating error in obtained scores, and (3) a judgment task of estimating the true discrepancy between two scores. Under the assumption that an impact from brand recognition would be identified, a corrective procedure that graphically displayed error variance and was designed to reduce the impact from brand recognition was also examined. The impact (or lack thereof) from brand recognition on rating psychometric quality and judgment tasks was similar for each variable. Contrary to the hypothesized results, brand recognition did not result in significant differences across any of the variables. That is, brand recognition did not influence rating psychometric quality, estimating error in obtained scores, or estimating the true discrepancy between two scores, an encouraging outcome suggesting that judgments were not swayed by a potential biasing factor. The corrective procedure was examined as part of this study; however, given the lack of significant findings on brand recognition, it would have limited generalizability if significant. Similar to the findings mentioned above, the corrective procedure of graphically displaying error variance also did not result in significant differences for any of the variables.

The lack of a measurable effect from brand recognition should be appreciated within the appropriate context. Caution is advised on dismissing brand recognition outright as it relates to test selection. That is, the results do not suggest that brand recognition is independent from test selection. Instead, the null results suggest that brand recognition may be independent from rating psychometric quality and specific judgment tasks that rely upon test-retest reliability.

It could be considered that the current study design failed to identify a brand recognition influence that *truly* exists on appraisals of psychometric quality. However, trends within the data did not provide evidence to argue that a true effect is present, at least regarding the psychometric qualities of test-retest reliability. Alternatively, it may be surmised that psychometric quality (or at least test-retest reliability) is partially independent of test selection. If so, brand recognition may continue to be pertinent and warrant further examination.

It is possible that test selection places suboptimal emphasis on psychometric adequacy, and therefore, the current study design was overly narrow as it relates to the potential relationship between brand recognition and test selection. As part of one of the most comprehensive surveys on test-usage practices of clinical neuropsychologists, Rabin and colleagues (2016) revealed that there is extensive overlap among neuropsychologists in test selection and utilization of instruments, which was a pattern also observed in their initial 2001 assessment survey (Rabin et al., 2005). The authors state:

> Neuropsychologists may choose instruments based on psychometric considerations – the subset of highly used instruments could possess the strongest psychometric properties (e.g., reliability, validity, norms, and research base). However, this is likely not the case as serious concerns have been raised about several of the most commonly endorsed measures in terms of adequate reliability and validity, standardization, normative data, and/or patient classification…Another possibility is that neuropsychologists are drawn to instruments on which they were trained during graduate school, internship, or

> postdoctoral fellowship. This small group of highly utilized instruments may have earned popularity, in part, by virtue of its long history as training tools. In addition to being used in training and practice, these instruments tend to be among those cited frequently in assessment texts and journal articles. Consequently, when designing batteries, they are among the instruments that first come to mind. (Rabin et al., 2016, p. 223)

Adequate test selection is arguably the cornerstone to achieving accuracy in psychological assessment. Using a test with problematically low reliability may significantly alter the interrelations among test scores, a problem that is worsened with each additional test selected and combined within the interpretation. When test selection is overly impressionistic, there is significant concern that artifacts are included into the overall data used to make clinical inferences. Therefore, while the current study did not identify an impact from brand recognition on appraising psychometric quality, evaluating the decision-making process of test selection still warrants further attention. There is a clear need to better understand the factors involved in test selection and potentially how to make the process more ideal, as poor test selection will, in nearly all cases, degrade (or worsen) the accuracy of clinical interpretive practices.

Interpretive Practices: Measurement Error and Scatter

Underappreciation of Measurement Error

In this study, neuropsychologists were provided a statement regarding whether concerns about measurement error may be overstated and that building redundancy into psychological assessment and applying expert professional judgment can

circumvent problems with measurement error. Strong disagreement to the statement would have been the ideal response in that it affirms the paramount importance of appreciating measurement error and its subsequent impact on the accuracy of clinical inferences (Brooks et al., 2009; Faust, 2012; Lareau & Ahern, 2012). A substantial proportion of the neuropsychologists (64.7%) endorsed only moderate agreement and a smaller portion (10%) endorsed strong agreement, which indicated that concerns about measurement error are overstated or that building redundancy into psychological assessment or applying expert professional judgment circumvents problems with measurement error. Only 24.2% of the respondents endorsed strong disagreement, which emphasizes the paramount importance of appreciating measurement error. There was no meaningful difference in the findings between board certified clinical neuropsychologists and non-board certified. The study indicated that neuropsychologists, as a whole, may not sufficiently account for measurement error.

Appreciating measurement error is paramount for making determinations based on test scores and also safeguards clinicians from attaching meaning to scores that are not truly present (Brooks et al., 2009). Alongside other psychometric variables, it is necessary to consider reliability when interpreting test scores. Measurement error is inversely related to reliability and, as a simple rule of thumb, the greater the reliability the lower the measurement error and the lower the reliability the greater the measurement error. Within the current study, participants provided with *good-excellent* reliability coefficients rated the figures as approximately good-excellent in quality; participants provided with *mediocre* reliability coefficients rated the figures as

approximately mediocre[3]; and participants provided with *poor* reliability coefficients rated the figures as extremely poor. However, when participants were asked to make clinical judgments with respect to the reliability coefficients, judgements became less accurate when based upon worse reliability levels. Judgements utilizing reliability coefficients that were *good-excellent*, which indicates the estimated true score is comparable to the observed score, were most accurate. However, judgments utilizing reliabilities that were *poor*, which indicates that the estimated true score would be adjusted significantly from the observed score, were largely inaccurate.

It may be the case that as psychometric quality declines, psychologists rely more heavily on intuitive judgment and, therefore, disregard the mathematical properties when making inferences. Alternatively, psychologists may underappreciate how profound the impact is when interpreting a score with very poor reliability (e.g., if reliability was 0.0, they may not estimate the true score to be equal to the mean). Therefore, when tests are selected with poor psychometric qualities and measurement error is underappreciated, then there is concern that accuracy of psychological assessment would be degraded accordingly.

Neuropsychologists' Perception of Normal Scatter

In this study, neuropsychologists were asked to specify a cutoff (or dividing point) for maximum discrepancy between highest and lowest score that distinguishes between normal and abnormal levels of variability. Participants were provided a hypothetical situation where 15 tests were administered, which generated 32 scores,

[3] It is noteworthy that this study used the term "mediocre" to represent test-retest reliability coefficients of r = .61, .57, & .56, as these coefficients were generally in the middle between high and low coefficients. However, while it should be recognized that test-retest reliability is not a fixed quality, and appraising the acceptable range of reliability may vary across dimensions, guidelines would generally indicate such coefficients as having low to unacceptably low psychometric quality.

and all scores were adjusted for age, gender, and education. It was also noted that the norms for these tests had been derived from the same sample (i.e., co-normed). The hypothetical illustration was based on Schretlen and colleagues' (2003) Aging, Brain Imaging, and Cognition study (*ABC* study). The authors studied 197 healthy adults, age 20 to 92 with a mean age of about 55 years and a mean education of about 14 years. Each participant completed a neuropsychological battery of 15 tests that resulted in 32 measures or scores. The study revealed substantial intra-individual variability in the performance of presumably healthy, normal adults. For example, only 2% of the sample obtained a range of scatter of less than two standard deviations (*SD*), whereas 65% demonstrated a range of at least three *SD* and 20% a range of at least four *SD*. The mean level of intra-individual variability was about 3.4 standard deviations (*SD* = 0.8).

In the current study, nearly every respondent underestimated normal levels of scatter, many by a large margin. For example, 79.3% indicated a cutoff between 1.5 *SD* to 3.0 *SD*, levels well below those expected for *normal individuals* and very often exceeded by such groups. The mean dividing point indicated 2.0 *SD* to 2.5 *SD*. To provide a striking comparison, the *ABC* study revealed that no participant had a maximum discrepancy of less than 1.6 SD and only four participants (2%) were less than 2.0 *SD*. Within the current study, 31.1% of the participants rated a maximum discrepancy of ≤1.5 *SD* and 63% of the participants rated a maximum discrepancy ranging between 0 *SD* to 2.0 *SD* as an abnormal level of variability. Slightly less than 7% of the participants indicated a dividing point at, or above, 4.0 *SD*. There was no

meaningful difference in the findings between board certified clinical neuropsychologists vs. non-board certified.

Although the Schretlen data provides only a single source of information on scatter, the level of scatter found in that work is consistent with a considerable body of literature on the topic (Binder et al., 2009, Brooks et al., 2009). Consider further studies involving even a single general measure with about 10 or so subtests, such as the Wechsler Intelligence Scales, demonstrate levels of scatter among normal groups that equal or exceeds the cutoff levels that many respondents in the current study identified under the assumption that about triple the number of measures that were used. For example, the 11 primary subtests from the WAIS-III and the 10 primary subtests from the WAIS-IV both have a mean of about 2.2 SD between the highest and lowest scores (Wechsler, 1997; 2008). It is also a mathematical truism that increasing the number of tests or subtests within a neuropsychological battery that already includes such an intelligence test will produce a level of scatter that must at least equal, and will often exceed, the level of scatter produced by the intelligence test alone (Binder, et al., 2009). Furthermore, neuropsychological batteries are often comprised of various measures that are not co-normed, which is likely to accentuate scatter. Variability between test scores and measures may also be magnified by various artifacts, such as number of tests administered (Binder et al., 2009), scoring errors (Allard & Faust, 2000; Simons, Goddard, & Patton, 2002) and inadequate normative selection (Brooks, et al., 2009).

Considering whether variability (or scatter) is normal depends on many features, for example, number of tests administered (i.e., as number of tests increase, so does

the level of normal variability) and examinee characteristics (e.g., age, education, sex, ethnicity, and intellectual functioning). The degree of scatter in test batteries increases as test reliability decreases because there is more measurement error in scores with low reliability than in scores with high reliability Therefore, appraising normal variability depends on multiple variables. As a loose rule of thumb, scatter, across a comprehensive neuropsychological evaluation, typically does not become uncommon until you approach more than 4 *SD*. However, this still depends on the factors mentioned above and simultaneously on the criterion used to determine aberrance (e.g., observed in only 5% vs. 15% of the healthy population). The current results regarding neuropsychologists' perception of normal scatter argues that a common interpretive practice, which emphasizes scatter analysis and grossly underestimates normal levels of scatter, may well lead to the overidentification of pathology. This problem is worsened when psychologists select tests with inadequate psychometric qualities, in particular, those with poor reliability.

Limitations

One limitation of this study was the suboptimal sample size. The study aimed to recruit 245 participants (i.e., 35 participants per cell), but fell about 122 participants short (obtained n's = 13 – 26 participants per cell). As noted earlier, 236 participants initiated the study; however, 113 participants discontinued immediately following the demographic portion of the study and, thus, were not included in any analyses. This decreased sample size, along with the negligible effect size, reduced the study's overall statistical power. Another statistical limitation included the unequal sized groups, which was partially due to the narrowed recruitment during the last phase of

participant recruitment that was aimed at increasing *n* in select cells. Additionally, certain dependent variables indicated unequal variances, which argued for consideration of non-parametric statistical analysis. With reservation, as mentioned earlier, parametric statistical analyses were utilized. However, given the non-significant findings across the analysis and the absence of trends in the data regarding the primary dependent variables, an increased sample size would not likely have significantly altered the data.

Another study limitation involves the restricted data provided to participants. In standard clinical practice, a neuropsychologist will likely have access to detailed records, interview data, and other corroborating information, all of which might provide useful information. Efforts were taken to provide basic information and test data that would be sufficient to answer the interpretive questions. Participants may have preferred to have more detailed information regarding the hypothetical patient or specifics about the actual measures. However, decades of research suggest that clinicians reach more accurate conclusions overall if they disregard interview results and base their interpretations on test results alone (Faust & Ahern, 2012). Wording of select questions were nuanced and may have been determined to lack sufficient clarity. This may have been a reason why nearly 50% of the participants discontinued participation following the demographics section, which would have been when participants were asked to engage in the more cognitively demanding tasks of the study. Participants were provided an option to provide comments, and four individuals expressed confusion in the question wording. Alternatively, this may also have been

partially related to an inadequate appreciation (or knowledge-base) of specific psychometric factors.

Summary

In summary, participants in this study were primarily licensed, clinical neuropsychologists[4]. Participants were asked to engage in judgment tasks regarding rating test-retest reliability and make clinical inferences. The underlying assumption was that brand recognition would negatively impact participants judgments. Contrary to the hypothesized results, brand recognition did not influence rating psychometric quality or clinical judgments, an encouraging outcome suggesting that judgments were not swayed by a potential biasing factor. Caution is advised, however, on dismissing brand recognition outright as it relates to test selection. Psychometric quality and test selection may be partially independent.

Perhaps the impact of brand recognition was reduced (or simply undetected) because psychometric quality is not intrinsically associated with the specific judgements measured in the current study (i.e., clinical judgements may largely ignore, or place limited emphasis on, psychometric qualities, e.g., test-retest reliability). The current study may have been too narrow and did not specifically address the concern that brand recognition has a potential negative impact on test selection. Instead, the study may have focused too heavily on judgment tasks (e.g., rating psychometric quality) that were assumed to be inherent to test selection, but instead are partially independent from test selection.

In the current study, neuropsychologists' ratings suggested that there may be an underappreciation of measurement error within the field and/or a belief that building

[4] Graduate students trained in neuropsychological assessment made up 8.9% of the sample.

redundancy into psychological assessment or applying expert professional judgment can circumvent such concerns with measurement error. Neuropsychologists also misperceived normal levels of scatter as rare or aberrant when the criterion for normal scatter was Schretlen and colleagues (i.e., scatter of maximum discrepancy in standard deviations of $M = 3.4$, $SD = 0.8$) or comparable findings (Binder, et al., 2009; Brooks et al., 2009; Schretlen et al., 2003; & Schretlen, et al., 2008). This poses a problem because normal level of scatter may frequently be perceived as abnormal and lead to overpathologizing. Suboptimal test selection has the potential to worsen an already common, problematic judgment practice, and the impact may be pervasive.

This study provides evidence that brand recognition did not have an impact on specific judgment tasks related to test-retest reliability, but these findings did not necessarily alleviate concerns that brand recognition may be relevant to the overall test selection process. Although this study yielded positive or encouraging findings, given the frequent discordance between psychometric standing and frequency of test use found in survey research, concerns remain that brand recognition or other variables can impede test selection and warrant further examination. Further identifying the basis of the concern regarding suboptimal test selection and, as necessary, offering corrective approaches could take many directions.

A programmatic approach may start by exploring psychologists' beliefs of their own test selection practices vs. their actual test selection practices (or using test usage surveys as a proxy). Similarly, this could be explored through analysis of actual decision-making practices regarding the adoption of revised versions of tests or selection of novel tests. Upon exploring more broadly whether brand recognition (or

other suboptimal processes) negatively influences test selection, designing corrective procedures and how to implement such could be warranted. While a scientific basis may identify a potential problem in test selection practices, awareness of such a problem may not be sufficient to correct the negative influence by itself. A corrective approach may offer recommendations toward test publisher marketing practices or training program test selection practices. Following an empirical database that may arise, corrective procedures would be directed toward reducing negative influence of salient information that should be independent from the selection process and improving the adherence to the most principle variables[5]. Future research should: (1) determine whether brand recognition (or other suboptimal process related to limitations in clinical judgment and bias) influences psychological test selection, (2) appraise the variables that go into test selection and how to optimally combine them, and (3) if necessary, aim to develop evidence-based standards toward formalizing test selection guidelines.

[5] It may also be determined that perhaps brand recognition has a positive association with test selection (e.g., it may be a valid predictive variable), and, therefore, serve as a non-optimal, but useful heuristic.

Table 1

Demographic Features

		n	Frequency*
Gender		122	
	Male	48	39.3%
	Female	74	60.7%
	Missing data	1	--
Ethnicity		123	
	African American/Black	1	0.8%
	Caucasian/White	109	88.6%
	Asian or Pacific Islander	5	4.1%
	Hispanic/Latino	4	3.3%
	Bi-racial	2	1.6%
	Choose not to disclose	1	0.8%
	Not Listed	1	0.8%
Highest Degree		123	
	M.A/M.S.	11	8.9%
	Ph.D.	91	74.0%
	Psy.D.	21	17.1%
Years Since Highest Degree		123	
	< 5 years	37	30.1%
	5 – 10 years	26	21.1%
	11 – 20 years	26	21.1%
	> 21 years	34	27.6%
Currently Licensed		123	
	Yes	105	85.4%
	No	18	14.6%
Board Certification in Clinical Neuropsychology		123	
	Yes	53	43.1%
	No	70	56.9%
Percentage of Time Spent on Neuropsychological Evaluations		123	
	0%	0	0%
	1-25%	7	5.7%
	26-50%	16	13.0%
	51-75%	18	14.6%
	76-100%	82	66.7%
Forensic Involvement		123	
	Yes	47	38.2%
	No	76	61.8%

* Missing data were excluded when calculating overall percentages.

Table 2

Tests of Between-Subjects Effects for DV1: Rating Psychometric Quality

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	244.076[a]	5	48.815	45.512	.000	.688
Intercept	1734.676	1	1734.676	1617.310	.000	.940
Reliability_level	242.649	2	121.324	113.116	.000	.687
Presence_of_brand	1.125	1	1.125	1.048	.308	.010
Reliability_level * Presence_of_brand	.501	2	.250	.233	.792	.005
Error	110.475	103	1.073			
Total	2043.000	109				
Corrected Total	354.550	108				

a. R Squared = .688 (Adjusted R Squared = .673)

Table 3

Descriptive Statistics for DV1: Rating Psychometric Quality (N = 122)

Overall, how would you rate the psychometric quality of the test-retest reliabilities?[a]					
	N	*M* (*SD*)	Median	Min.	Max
Good-Excellent (*r* = .82, .86, & .88)[b]					
No Brand	16	5.94 (.574)	6	5	7
Brand	18	6.06 (.539)	6	5	7
Mediocre (*r* = .56, .57, & .61)[c]					
No Brand	14	3.71 (.825)	4	2	5
Brand	17	3.82 (1.074)	4	2	5
Corrective	13	3.62 (1.193)	4	2	6
Poor (*r* = .27, .30, & .36)[d]					
No Brand	18	2.22 (1.263)	2	1	6
Brand	26	2.62 (1.359)	2	1	6

[a] Measured on a Likert scale: 1 = Extremely Poor; 4 = Mediocre; 7 = Excellent
[b] Expected response = 6-7
[c] Expected response = 3-5
[d] Expected response = 1-2

Table 4

Tests of Between-Subjects Effects for DV2: Estimating Error

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	212.547[a]	5	42.509	8.557	.000	.304
Intercept	4323.298	1	4323.298	870.276	.000	.899
Reliability_level	189.978	2	94.989	19.121	.000	.281
Presence_of_brand	4.559	1	4.559	.918	.340	.009
Reliability_level * Presence_of_brand	24.089	2	12.045	2.425	.094	.047
Error	486.838	98	4.968			
Total	4990.000	104				
Corrected Total	699.385	103				

a. R Squared = .304 (Adjusted R Squared = .268)

Table 5

Descriptive Statistics for DV2: Estimating Error (N = 117)

Based on the T-score of 33 and the *r* = .XX test-retest reliability, what is the probability that the examinee's estimated true score would fall below the -1 *SD* cut-off point?					
	N	*M* (*SD*)	Median	Min.	Max
Good-Excellent (*r* = .88)[b]					
No Brand	16	8.31 (2.243)	9	2	10
Brand	17	8.12 (2.205)	9	2	10
Mediocre (*r* = .57)[c]					
No Brand	14	7.29 (1.637)	7.50	5	10
Brand	16	5.56 (2.065)	6	2	8
Corrective	13	7.15 (2.035)	7	3	10
Poor (*r* = .27)[d]					
No Brand	17	4.65 (2.572)	4	1	10
Brand	24	5.29 (2.368)	6	1	9

[a] Measured on continuous scale from 1 - 10: 1=0-9%; 2=10-19%; 3=20-29%; 4=30-39%; 5=40-49%; 6=50-59%; 7=60-69%; 8=70-79%; 9=80-89%; 10=90-100%

[b] Expected response = 9 (84%)

[c] Expected response = 6 (50%)

[d] Expected response = 2 (16%)

Table 6

Corrective Procedure

		Sum of Squares	df	Mean Square	F	Sig.
Rating	Between Groups	.319	1	.319	.251	.620
	Within Groups	35.548	28	1.270		
	Total	35.867	29			
Estimating_error	Between Groups	18.163	1	18.163	4.316	.047
	Within Groups	113.630	27	4.209		
	Total	131.793	28			
Estimating_discrepancy	Between Groups	.836	1	.836	3.681	.066
	Within Groups	6.130	27	.227		
	Total	6.966	28			

Table 7

Tests of Between-Subjects Effects for DV3: Estimating Discrepancy

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	7.707[a]	5	1.541	4.110	.002	.173
Intercept	503.994	1	503.994	1343.821	.000	.932
Reliability_level	6.602	2	3.301	8.802	.000	.152
Presence_of_brand	.648	1	.648	1.727	.192	.017
Reliability_level * Presence_of_brand	.937	2	.469	1.250	.291	.025
Error	36.754	98	.375			
Total	562.000	104				
Corrected Total	44.462	103				

a. R Squared = .173 (Adjusted R Squared = .131)

Table 8

Descriptive Statistics for DV3: Estimating Discrepancy (N = 117)

Assume that the duration between administrations eliminates any practice effects, and that the two administrations yield an initial T-score of 30 (*M* = 50, *SD* = 10) and a subsequent T-score of 60. Finally, assume that the measure has a test-retest reliability of *r* = .XX. Which figure below best matches your estimation of the true discrepancy between the scores?[a]					
	N	*M* (*SD*)	Median	Min.	Max
Good-Excellent (*r* = .88)[b]					
No Brand	16	2.44 (.512)	2	2	3
Brand	17	2.65 (.606)	3	1	3
Mediocre (*r* = .57)[c]					
No Brand	14	2.29 (.469)	2	2	3
Brand	16	2.19 (.544)	2	1	3
Corrective	13	1.85 (.376)	2	1	2
Poor (*r* = .27)[d]					
No Brand	16	1.75 (.775)	2	1	3
Brand	25	2.12 (.666)	2	1	3

[a] Measured on continuous scale from 1 – 3: 1 = <1.0 *SD*; 2 = 1.0-2.0 *SD*; 3 = >2.0 *SD*
[b] Expected response = 3 (>2.0 *SD*)
[c] Expected response = 2 (1.0-2.0 *SD*)
[d] Expected response = 1 (<1.0 *SD*)

Table 9

Level of Scatter Judged to Distinguish Between Normal and Abnormal Performance (N = 116)

…which of the following dividing points for the maximum discrepancy between highest and lowest score most closely matches the one you would use in distinguishing between normal and abnormal levels of variability?[a]			
Scatter[b]	Frequency	Percentage	Cumulative Percentage
>5.0	3	2.6	100.0
4.5	0	0.0	100.0
4.0	5	4.3	97.4
3.5	5	4.3	93.1
3.0	23	19.8	88.8
2.5	7	6.0	69.0
2.0	37	31.9	62.9
1.5	25	21.6	31.0
1.0	8	6.9	9.5
0.5	2	1.7	2.6
0.0	1	0.9	0.9
Mean =	5.47 (2.0-2.5 *SD*)		
SD =	1.858		

[a] Measured on a Continuous scale from 1 – 11, which ranges from 0.0 *SD* to >5.0 *SD* at increments of 0.5 *SD* (i.e., 1 = 0.0 *SD;* 2 = 0.5 *SD;* 3 = 1.0 *SD*; 4 = 1.5 *SD*; 5 = 2.0 *SD*; 6 = 2.5 *SD*; 7 = 3.0 *SD*; 8 = 3.5 *SD*; 9 = 4.0 *SD*; 10 = 4.5 *SD*; 11= *>5 SD*)

[b] Expected response = 9+ (i.e., >4.0 *SD*)

Table 10

Measurement Error (N = 119)

Concerns about measurement error may be overstated. Building redundancy into psychological assessment and applying expert professional judgment can circumvent problems with measurement error that many writers of assessment texts describe as major shortcomings. Please indicate the degree to which you agree with the previous statement.[a]		
Response[b]	Frequency	Percentage
1	13	10.9
2	17	14.3
3	26	21.8
4	36	30.3
5	15	12.6
6	11	9.2
7	1	0.8
Mean =	3.50 (Moderately Agree)	
SD =	1.455	

[a] Measured on a Likert scale: 1 = Strongly Disagree; 4 = Moderately Agree; 7 = Strongly Agree

[b] Expected response = 1-2 (Strongly Disagree)

Table 11

Familiarity with CVLT-II (N = 122)

How familiar are you with the CVLT-II?[a]		
Response	Frequency	Percentage
1	1	0.8
2	2	1.6
3	6	4.9
4	22	18.0
5	16	13.1
6	38	31.1
7	37	30.3
Mean =	5.56	
SD =	1.373	

[a] Measured on a Likert scale: 1 = Not familiar at all; 4 = Moderately familiar; 7 = Extremely familiar

Table 12

Frequency of Use with CVLT-II (N = 122)

When assessing memory abilities, how frequently do you use the CVLT-II?[a]		
Response	Frequency	Percentage
1	21	17.2
2	19	15.6
3	14	11.5
4	12	9.8
5	21	17.2
6	25	20.5
7	10	8.2
Mean =	3.89	
SD =	2.009	

[a] Measured on a Likert scale: 1 = Never; 4 = 50% of the time; 7 = Always

APPENDICES

Appendix A: Demographic Questionnaire

Please respond to the following questions concerning demographic information and professional practice.

1. **What best describes your gender:**

Male	Female	
Trans-man	Trans-woman	Agender
Non-binary	Gender Fluid	Genderqueer
Prefer not to respond		Not Listed _____

2. **Ethnicity:**

African American/Black	Caucasian/White
American Indian/Alaskan Native	Bi-racial
Asian or Pacific Islander	Not Listed _____
Hispanic/Latino	Choose not to disclose

3. **Highest Degree:** M.A./M.S. Ph.D. Psy.D Ed.D Other

4. **Years since Highest Degree:** <5 5-10 11-20 >21

5. **Currently Licensed as a Psychologist:** Yes No

6. **Board Certification in Clinical Neuropsychology:** Yes No

7. **Board Certification in other specialty:** Yes No

8. **Over the last two years, about what percentage of your time has been spent on neuropsychological evaluations or related activities in neuropsychology:**

   0% 1-25% 26-50% 51-75% 76-100%

9. **Over the last two years, what percentage of your time is spent with the following populations:**

Children and Adolescents (≤18 years)	0%	1-25%	26-50%	51-75%	76-100%
Adults (19-65 years)	0%	1-25%	26-50%	51-75%	76-100%
Geriatric Adults (>65 years)	0%	1-25%	26-50%	51-75%	76-100%

10. **Are you involved in forensic evaluations:** Yes No

    If yes, over the last two years, about what percentage of your time has been spent on forensic evaluations:

    N/A 0% 1-25% 26-50% 51-75% 76-100%

Appendix B: Graphical Display of Methodological Design

	Brand Recognition$^{(I.V.1)}$		*Corrective Procedure*$^{(I.V.3)}$
Level of Reliability$^{(I.V.2)}$	**Without Test Name**	**With Test Name**	**With Test Name + Error Variance**
GOOD-EXCELLENT			
MEDIOCRE			
POOR			

Appendix C: Vignette 1: Good-Excellent level of reliability + Absence of brand recognition

The following hypothetical situation examines the impact of test-retest reliability on aspects of assessment. Although the data provided are certainly less complete than would be typical in clinical practice, the information that follows should be sufficient to answer the items presented below.

Assume someone has administered a neuropsychological battery that included a few measures of verbal memory that address important constructs, and that you are required to formulate a preliminary impression. Assume that the reliability with which these constructs are measured is an important concern here.

The test-retest reliability coefficients for the respective memory subtests/indices are: *r* = .82, .86, & .88.

1) **How would you rate the psychometric quality of the overall test-retest reliabilities?**

Extremely poor **Mediocre** **Excellent**

1----------2----------3----------4----------5----------6----------7

2) **Consider the following hypothetical situation. Assume that the cut-off for identifying a deficit or weakness is a score that falls more than one standard deviation (*SD*) below the mean (e.g., a T-score below 40). Also, assume the examinee obtained a T-score of 33 (*M* = 50, *SD* = 10) on one of the memory measures referenced above, which has a test-retest reliability coefficient of *r* = .88. Based on the T-score of 33 and the *r* = .88 test-retest reliability, what is the probability that the examinee's estimated true score would fall below the -1 *SD* cut-off point?**

0%----------100%

3) **Assume the same memory test, drawn from the measures referenced above, is administered twice, and you are trying to determine whether the score on the second administration reflects genuine change from the score obtained on the first administration. Assume that the duration between administrations eliminates any practice effects, and that the two administrations yield an initial T-score of 30 (*M* = 50, *SD* = 10) and a subsequent T-score of 60. Finally, assume that the test has a test-retest reliability of *r* = .88.**

Which figure below best fits your estimation of the true discrepancy between the scores? (For example, the discrepancy between the observed scores would be 3 *SD*; 60 – 30 = 30.)

< 1.0 *SD*	1.0 – 2.0 *SD*	> 2.0 *SD*

4) **Assume that 15 tests have been administered, which generate 32 scores, and that all scores are adjusted for age, gender, and education. The norms for these tests have been derived from the same sample, or are co-normed.**

   Under these conditions, which of the following dividing points for the maximum discrepancy between highest and lowest score most closely matches the one you would use in distinguishing between normal and abnormal levels of variability?

   Anchor points indicate the maximum discrepancy between the highest and lowest score across a neuropsychological battery in standard deviation units

0.0 SD	0.5 SD	1.0 SD	1.5 SD	2.0 SD	2.5 SD	3.0 SD	3.5 SD	4.0 SD	4.5 SD	>5.0 SD

5) **Concerns about measurement error may be overstated. Building redundancy into psychological assessment and applying expert professional judgment can circumvent problems with measurement error that many writers of assessment texts describe as major shortcomings. Please indicate the degree to which you agree with the previous statement.**

Strongly Disagree			Moderately Agree			Strongly Agree
1	2	3	4	5	6	7

6) **How familiar are you with the CVLT-II?**

Not familiar at all			Moderately familiar			Extremely familiar
1	2	3	4	5	6	7

7) **When assessing memory abilities, how frequently do you use the CVLT-II?**

Never			50% of the time			Always
1	2	3	4	5	6	7

If you prefer, feel free to provide any comments regarding your responses.

Appendix D: Vignette 2: Mediocre level of reliability + Absence of brand recognition

The following hypothetical situation examines the impact of test-retest reliability on aspects of assessment. Although the data provided are certainly less complete than would be typical in clinical practice, the information that follows should be sufficient to answer the items presented below.

Assume someone has administered a neuropsychological battery that included a few measures of verbal memory that address important constructs, and that you are required to formulate a preliminary impression. Assume that the reliability with which these constructs are measured is an important concern here.

The test-retest reliability coefficients for the respective memory subtests/indices are: *r* = .56, .57, & .61.

1) **How would you rate the psychometric quality of the overall test-retest reliabilities?**

Extremely poor **Mediocre** **Excellent**
1----------2----------3----------4----------5----------6----------7

2) **Consider the following hypothetical situation. Assume that the cut-off for identifying a deficit or weakness is a score that falls more than one standard deviation (*SD*) below the mean (e.g., a T-score below 40). Also, assume the examinee obtained a T-score of 33 (*M* = 50, *SD* = 10) on one of the memory measures referenced above, which has a test-retest reliability coefficient of *r* = .57. Based on the T-score of 33 and the *r* = .57 test-retest reliability, what is the probability that the examinee's estimated true score would fall below the -1 *SD* cut-off point?**

0%----------100%

3) **Assume the same memory test, drawn from the measures referenced above, is administered twice, and you are trying to determine whether the score on the second administration reflects genuine change from the score obtained on the first administration. Assume that the duration between administrations eliminates any practice effects, and that the two administrations yield an initial T-score of 30 (*M* = 50, *SD* = 10) and a subsequent T-score of 60. Finally, assume that the test has a test-retest reliability of *r* = .57.**

Which figure below best fits your estimation of the true discrepancy between the scores? (For example, the discrepancy between the observed scores would be 3 *SD*; 60 – 30 = 30.)

< 1.0 ***SD***	**1.0 – 2.0** ***SD***	**> 2.0** ***SD***

4) **Assume that 15 tests have been administered, which generate 32 scores, and that all scores are adjusted for age, gender, and education. The norms for these tests have been derived from the same sample, or are co-normed.**

   Under these conditions, which of the following dividing points for the maximum discrepancy between highest and lowest score most closely matches the one you would use in distinguishing between normal and abnormal levels of variability?

   Anchor points indicate the maximum discrepancy between the highest and lowest score across a neuropsychological battery in standard deviation units

0.0 SD	**0.5 SD**	**1.0 SD**	**1.5 SD**	**2.0 SD**	**2.5 SD**	**3.0 SD**	**3.5 SD**	**4.0 SD**	**4.5 SD**	**>5.0 SD**

5) **Concerns about measurement error may be overstated. Building redundancy into psychological assessment and applying expert professional judgment can circumvent problems with measurement error that many writers of assessment texts describe as major shortcomings. Please indicate the degree to which you agree with the previous statement.**

Strongly Disagree			**Moderately Agree**			**Strongly Agree**
1	**2**	**3**	**4**	**5**	**6**	**7**

6) **How familiar are you with the CVLT-II?**

Not familiar at all			**Moderately familiar**			**Extremely familiar**
1	**2**	**3**	**4**	**5**	**6**	**7**

7) **When assessing memory abilities, how frequently do you use the CVLT-II?**

Never			**50% of the time**			**Always**
1	**2**	**3**	**4**	**5**	**6**	**7**

If you prefer, feel free to provide any comments regarding your responses.

Appendix E: Vignette 3: Poor level of + Absence of brand recognition

The following hypothetical situation examines the impact of test-retest reliability on aspects of assessment. Although the data provided are certainly less complete than would be typical in clinical practice, the information that follows should be sufficient to answer the items presented below.

Assume someone has administered a neuropsychological battery that included a few measures of verbal memory that address important constructs, and that you are required to formulate a preliminary impression. Assume that the reliability with which these constructs are measured is an important concern here.

The test-retest reliability coefficients for the respective memory subtests/indices are: *r* = .27, .30, &.36.

1) **How would you rate the psychometric quality of the overall test-retest reliabilities?**

Extremely poor **Mediocre** **Excellent**
1----------2----------3----------4----------5----------6----------7

2) **Consider the following hypothetical situation. Assume that the cut-off for identifying a deficit or weakness is a score that falls more than one standard deviation (*SD*) below the mean (e.g., a T-score below 40). Also, assume the examinee obtained a T-score of 33 (*M* = 50, *SD* = 10) on one of the memory measures referenced above, which has a test-retest reliability coefficient of *r* = .27. Based on the T-score of 33 and the *r* = .27 test-retest reliability, what is the probability that the examinee's estimated true score would fall below the -1 *SD* cut-off point?**

0%----------100%

3) **Assume the same memory test, drawn from the measures referenced above, is administered twice, and you are trying to determine whether the score on the second administration reflects genuine change from the score obtained on the first administration. Assume that the duration between administrations eliminates any practice effects, and that the two administrations yield an initial T-score of 30 (*M* = 50, *SD* = 10) and a subsequent T-score of 60. Finally, assume that the test has a test-retest reliability of *r* = .27.**

Which figure below best fits your estimation of the true discrepancy between the scores? (For example, the discrepancy between the observed scores would be 3 *SD*; 60 – 30 = 30.)

< 1.0	**1.0 – 2.0**	**> 2.0**
SD	***SD***	***SD***

4) **Assume that 15 tests have been administered, which generate 32 scores, and that all scores are adjusted for age, gender, and education. The norms for these tests have been derived from the same sample, or are co-normed.**

   Under these conditions, which of the following dividing points for the maximum discrepancy between highest and lowest score most closely matches the one you would use in distinguishing between normal and abnormal levels of variability?

   Anchor points indicate the maximum discrepancy between the highest and lowest score across a neuropsychological battery in standard deviation units

0.0 SD	**0.5 SD**	**1.0 SD**	**1.5 SD**	**2.0 SD**	**2.5 SD**	**3.0 SD**	**3.5 SD**	**4.0 SD**	**4.5 SD**	**>5.0 SD**

5) **Concerns about measurement error may be overstated. Building redundancy into psychological assessment and applying expert professional judgment can circumvent problems with measurement error that many writers of assessment texts describe as major shortcomings. Please indicate the degree to which you agree with the previous statement.**

Strongly Disagree			**Moderately Agree**			**Strongly Agree**
1	**2**	**3**	**4**	**5**	**6**	**7**

6) **How familiar are you with the CVLT-II?**

Not familiar at all			**Moderately familiar**			**Extremely familiar**
1	**2**	**3**	**4**	**5**	**6**	**7**

7) **When assessing memory abilities, how frequently do you use the CVLT-II?**

Never			**50% of the time**			**Always**
1	**2**	**3**	**4**	**5**	**6**	**7**

If you prefer, feel free to provide any comments regarding your responses.

Appendix F: Vignette 4: Good-Excellent level of reliability + Presence of brand recognition

The following hypothetical situation examines the impact of test-retest reliability on aspects of assessment. Although the data provided are certainly less complete than would be typical in clinical practice, the information that follows should be sufficient to answer the items presented below.

Assume someone has administered a neuropsychological battery that included a few measures of verbal memory that address important constructs, and that you are required to formulate a preliminary impression. Assume that the reliability with which these constructs are measured is an important concern here.

The test-retest reliability coefficients for the respective memory subtests/indices from the California Verbal Learning Test, 2nd Edition (CVLT-II) are: *r* = .82, .86, & .88.

1) **How would you rate the psychometric quality of the overall test-retest reliabilities?**

Extremely poor **Mediocre** **Excellent**
1----------2----------3----------4----------5----------6----------7

2) **Consider the following hypothetical situation. Assume that the cut-off for identifying a deficit or weakness is a score that falls more than one standard deviation (*SD*) below the mean (e.g., a T-score below 40). Also, assume the examinee obtained a T-score of 33 (*M* = 50, *SD* = 10) on one of the memory measures referenced above, which has a test-retest reliability coefficient of *r* = .88. Based on the T-score of 33 and the *r* = .88 test-retest reliability, what is the probability that the examinee's estimated true score would fall below the -1 *SD* cut-off point?**

0%----------100%

3) **Assume the same memory test, drawn from the measures referenced above, is administered twice, and you are trying to determine whether the score on the second administration reflects genuine change from the score obtained on the first administration. Assume that the duration between administrations eliminates any practice effects, and that the two administrations yield an initial T-score of 30 (*M* = 50, *SD* = 10) and a subsequent T-score of 60. Finally, assume that the test has a test-retest reliability of *r* = .88.**
Which figure below best fits your estimation of the true discrepancy between the scores? (For example, the discrepancy between the observed scores would be 3 *SD*; 60 – 30 = 30.)

< 1.0 *SD*	1.0 – 2.0 *SD*	> 2.0 *SD*

4) **Assume that 15 tests have been administered, which generate 32 scores, and that all scores are adjusted for age, gender, and education. The norms for these tests have been derived from the same sample, or are co-normed.**

   Under these conditions, which of the following dividing points for the maximum discrepancy between highest and lowest score most closely matches the one you would use in distinguishing between normal and abnormal levels of variability?

   Anchor points indicate the maximum discrepancy between the highest and lowest score across a neuropsychological battery in standard deviation units

0.0 SD	0.5 SD	1.0 SD	1.5 SD	2.0 SD	2.5 SD	3.0 SD	3.5 SD	4.0 SD	4.5 SD	>5.0 SD

5) **Concerns about measurement error may be overstated. Building redundancy into psychological assessment and applying expert professional judgment can circumvent problems with measurement error that many writers of assessment texts describe as major shortcomings. Please indicate the degree to which you agree with the previous statement.**

Strongly Disagree		Moderately Agree		Strongly Agree
1	2	3	4	5	6	7

6) **How familiar are you with the CVLT-II?**

Not familiar at all		Moderately familiar		Extremely familiar
1	2	3	4	5	6	7

7) **When assessing memory abilities, how frequently do you use the CVLT-II?**

Never		50% of the time		Always
1	2	3	4	5	6	7

If you prefer, feel free to provide any comments regarding your responses.

Appendix G: Vignette 5: Mediocre level of reliability + Presence of brand recognition

The following hypothetical situation examines the impact of test-retest reliability on aspects of assessment. Although the data provided are certainly less complete than would be typical in clinical practice, the information that follows should be sufficient to answer the items presented below.

Assume someone has administered a neuropsychological battery that included a few measures of verbal memory that address important constructs, and that you are required to formulate a preliminary impression. Assume that the reliability with which these constructs are measured is an important concern here.

The test-retest reliability coefficients for the respective memory subtests/indices from the California Verbal Learning Test, 2nd Edition (CVLT-II) are: *r* = .56, .57, & .61.

1) **How would you rate the psychometric quality of the overall test-retest reliabilities?**

Extremely poor **Mediocre** **Excellent**
1----------2----------3----------4----------5----------6----------7

2) **Consider the following hypothetical situation. Assume that the cut-off for identifying a deficit or weakness is a score that falls more than one standard deviation (*SD*) below the mean (e.g., a T-score below 40). Also, assume the examinee obtained a T-score of 33 (*M* = 50, *SD* = 10) on one of the memory measures referenced above, which has a test-retest reliability coefficient of *r* = .57. Based on the T-score of 33 and the *r* = .57 test-retest reliability, what is the probability that the examinee's estimated true score would fall below the -1 *SD* cut-off point?**

0%----------100%

3) **Assume the same memory test, drawn from the measures referenced above, is administered twice, and you are trying to determine whether the score on the second administration reflects genuine change from the score obtained on the first administration. Assume that the duration between administrations eliminates any practice effects, and that the two administrations yield an initial T-score of 30 (*M* = 50, *SD* = 10) and a subsequent T-score of 60. Finally, assume that the test has a test-retest reliability of *r* = .57.**

Which figure below best fits your estimation of the true discrepancy between the scores? (For example, the discrepancy between the observed scores would be 3 *SD*; 60 – 30 = 30.)

< 1.0 *SD*	1.0 – 2.0 *SD*	> 2.0 *SD*

4) **Assume that 15 tests have been administered, which generate 32 scores, and that all scores are adjusted for age, gender, and education. The norms for these tests have been derived from the same sample, or are co-normed.**

   Under these conditions, which of the following dividing points for the maximum discrepancy between highest and lowest score most closely matches the one you would use in distinguishing between normal and abnormal levels of variability?

   Anchor points indicate the maximum discrepancy between the highest and lowest score across a neuropsychological battery in standard deviation units

0.0 SD	0.5 SD	1.0 SD	1.5 SD	2.0 SD	2.5 SD	3.0 SD	3.5 SD	4.0 SD	4.5 SD	>5.0 SD

5) **Concerns about measurement error may be overstated. Building redundancy into psychological assessment and applying expert professional judgment can circumvent problems with measurement error that many writers of assessment texts describe as major shortcomings. Please indicate the degree to which you agree with the previous statement.**

Strongly Disagree			Moderately Agree			Strongly Agree
1	2	3	4	5	6	7

6) **How familiar are you with the CVLT-II?**

Not familiar at all			Moderately familiar			Extremely familiar
1	2	3	4	5	6	7

7) **When assessing memory abilities, how frequently do you use the CVLT-II?**

Never			50% of the time			Always
1	2	3	4	5	6	7

If you prefer, feel free to provide any comments regarding your responses.

Appendix H: Vignette 6: Poor level of reliability + Presence of brand recognition

The following hypothetical situation examines the impact of test-retest reliability on aspects of assessment. Although the data provided are certainly less complete than would be typical in clinical practice, the information that follows should be sufficient to answer the items presented below.

Assume someone has administered a neuropsychological battery that included a few measures of verbal memory that address important constructs, and that you are required to formulate a preliminary impression. Assume that the reliability with which these constructs are measured is an important concern here.

The test-retest reliability coefficients for the respective memory subtests/indices from the California Verbal Learning Test, 2nd Edition (CVLT-II) are: *r* = .27, .30, &.36.

1) **How would you rate the psychometric quality of the overall test-retest reliabilities?**

Extremely poor **Mediocre** **Excellent**
1----------2----------3----------4----------5----------6----------7

2) **Consider the following hypothetical situation. Assume that the cut-off for identifying a deficit or weakness is a score that falls more than one standard deviation (*SD*) below the mean (e.g., a T-score below 40). Also, assume the examinee obtained a T-score of 33 (*M* = 50, *SD* = 10) on one of the memory measures referenced above, which has a test-retest reliability coefficient of *r* = .27. Based on the T-score of 33 and the *r* = .27 test-retest reliability, what is the probability that the examinee's estimated true score would fall below the -1 *SD* cut-off point?**

0%----------100%

3) **Assume the same memory test, drawn from the measures referenced above, is administered twice, and you are trying to determine whether the score on the second administration reflects genuine change from the score obtained on the first administration. Assume that the duration between administrations eliminates any practice effects, and that the two administrations yield an initial T-score of 30 (*M* = 50, *SD* = 10) and a subsequent T-score of 60. Finally, assume that the test has a test-retest reliability of *r* = .27.**

Which figure below best fits your estimation of the true discrepancy between the scores? (For example, the discrepancy between the observed scores would be 3 *SD*; 60 – 30 = 30.)

< 1.0 *SD*	1.0 – 2.0 *SD*	> 2.0 *SD*

4) **Assume that 15 tests have been administered, which generate 32 scores, and that all scores are adjusted for age, gender, and education. The norms for these tests have been derived from the same sample, or are co-normed.**

   Under these conditions, which of the following dividing points for the maximum discrepancy between highest and lowest score most closely matches the one you would use in distinguishing between normal and abnormal levels of variability?

   Anchor points indicate the maximum discrepancy between the highest and lowest score across a neuropsychological battery in standard deviation units

0.0 SD	0.5 SD	1.0 SD	1.5 SD	2.0 SD	2.5 SD	3.0 SD	3.5 SD	4.0 SD	4.5 SD	>5.0 SD

5) **Concerns about measurement error may be overstated. Building redundancy into psychological assessment and applying expert professional judgment can circumvent problems with measurement error that many writers of assessment texts describe as major shortcomings. Please indicate the degree to which you agree with the previous statement.**

Strongly Disagree			Moderately Agree			Strongly Agree
1	2	3	4	5	6	7

6) **How familiar are you with the CVLT-II?**

Not familiar at all			Moderately familiar			Extremely familiar
1	2	3	4	5	6	7

7) **When assessing memory abilities, how frequently do you use the CVLT-II?**

Never			50% of the time			Always
1	2	3	4	5	6	7

If you prefer, feel free to provide any comments regarding your responses.

Appendix I: Vignette 7: Mediocre level of reliability + Presence of brand recognition + Corrective Procedure

The following hypothetical situation examines the impact of test-retest reliability on aspects of assessment. Although the data provided are certainly less complete than would be typical in clinical practice, the information that follows should be sufficient to answer the items presented below.

Assume someone has administered a neuropsychological battery that included a few measures of verbal memory that address important constructs, and that you are required to formulate a preliminary impression. Assume that the reliability with which these constructs are measured is an important concern here.

The test-retest reliability coefficients for the respective memory subtests/indices from the California Verbal Learning Test, 2nd Edition (CVLT-II) are: *r* = .56, .57, & .61.

= error

= true measurement

CVLT-II measures		
Test-retest Reliability Coefficients		
.56	.57	.61
		

1) **How would you rate the psychometric quality of the overall test-retest reliabilities?**

Extremely poor Mediocre Excellent
1----------2----------3----------4----------5----------6----------7

2) **Consider the following hypothetical situation. Assume that the cut-off for identifying a deficit or weakness is a score that falls more than one standard deviation (*SD*) below the mean (e.g., a T-score below 40). Also, assume the examinee obtained a T-score of 33 (*M* = 50, *SD* = 10) on one of the memory measures referenced above, which has a test-retest reliability coefficient of *r* = .57. Based on the T-score of 33 and the *r* = .57 test-retest reliability, what is the probability that the examinee's estimated true score would fall below the -1 *SD* cut-off point?**

0%----------100%

3) **Assume the same memory test, drawn from the measures referenced above, is administered twice, and you are trying to determine whether the score on the second administration reflects genuine change from the score obtained on the first administration. Assume that the duration between administrations eliminates any practice effects, and that the two administrations yield an initial T-score of 30 (*M* = 50, *SD* = 10) and a subsequent T-score of 60. Finally, assume that the test has a test-retest reliability of *r* = .57.**

Which figure below best fits your estimation of the true discrepancy between the scores? (For example, the discrepancy between the observed scores would be 3 *SD*; 60 – 30 = 30.)

< 1.0 *SD*	**1.0 – 2.0 *SD***	**> 2.0 *SD***

4) **Assume that 15 tests have been administered, which generate 32 scores, and that all scores are adjusted for age, gender, and education. The norms for these tests have been derived from the same sample, or are co-normed.**

Under these conditions, which of the following dividing points for the maximum discrepancy between highest and lowest score most closely matches the one you would use in distinguishing between normal and abnormal levels of variability?

Anchor points indicate the maximum discrepancy between the highest and lowest score across a neuropsychological battery in standard deviation units

0.0 SD	**0.5 SD**	**1.0 SD**	**1.5 SD**	**2.0 SD**	**2.5 SD**	**3.0 SD**	**3.5 SD**	**4.0 SD**	**4.5 SD**	**>5.0 SD**

5) **Concerns about measurement error may be overstated. Building redundancy into psychological assessment and applying expert**

professional judgment can circumvent problems with measurement error that many writers of assessment texts describe as major shortcomings. Please indicate the degree to which you agree with the previous statement.

Strongly Disagree — Moderately Agree — Strongly Agree

1----------2----------3----------4----------5----------6----------7

6) How familiar are you with the CVLT-II?

Not familiar at all — Moderately familiar — Extremely familiar

1----------2----------3----------4----------5----------6----------7

7) When assessing memory abilities, how frequently do you use the CVLT-II?

Never — 50% of the time — Always

1----------2----------3----------4----------5----------6----------7

If you prefer, feel free to provide any comments regarding your responses.

Appendix J: Estimated Correct Responses for D.V. 2

The following formula is used to calculate the estimated true score.

$X+[r_{xx}(x-X)]$ or $r_{xx}(z)$

X=Mean
r_{xx}=reliability coefficient
x=observed score
z= z-score

The following formula is used to calculate a score band with upper and lower limits around the estimated true score at 68% C.I. (+/-1*SD*).

$+/-1SD(\sqrt{1}- r_{xx})(\sqrt{r_{xx}})$

Using the observed score of T = 33 (or, z = -1.7), the following scores are calculated at each level of reliability.

Good - Excellent: .82 (.86) [.88]
Estimated true score = 36 (35) [35]
Range = 32 – 40 (32 – 38) [32 – 38]
Therefore, using *r* = .88, there is an approximate 84% probability of estimated true scores in the impaired range.

Mediocre: .61, .57, & .56 [scores estimated the same at each coefficient]
Estimated true score = 40
Range = 35 – 45
Therefore, using r = .57, there is an approximate 50% probability of estimated true scores in the impaired range.

Poor: .36 (.30) [.27]
Estimated true score = 44 (45) [45]
Range = 39 – 49 (40 – 50) [41 – 49]
Therefore, using r = .27, there is an approximate 16% probability of estimated true scores in the impaired range.

Appendix K: Estimated Correct Responses for D.V. 3

The calculations used to appraise responses for D.V. 3 use the formula from Appendix J in calculating the estimated true score. However, the focus at hand is comparing the magnitude of the true discrepancy between the two scores, as measured in standard deviations. This is in contrast to the discrepancy between the observed scores.

Two observed scores will be compared: T = 30 (or, z = -2.0) and T = 60 (or, z = 1.0). The following expected true scores and magnitude of true discrepancy are calculated at each level of reliability.

Good - Excellent: *r* = .88

T = 30; estimated true score is T = 32.4
T = 60; estimated true score is T = 58.8
Therefore, the true discrepancy between the two scores is 2.6 *SD*.

Mediocre: *r* = .57

T = 30; estimated true score is T = 38.6
T = 60; estimated true score is T = 55.6
Therefore, the true discrepancy between the two scores is 1.7 *SD*.

Poor: *r* = .27

T = 30; estimated true score is T = 44.6
T = 60; estimated true score is T = 52.7
Therefore, the true discrepancy between the two scores is 0.8 *SD*.

BIBLIOGRAPHY

Allard G. & Faust, D. (2000). Errors in scoring objective personality tests. *Assessment, 7*, 119-129.

Arkes, H. R., Boehm, L., & Xu, G. (1991). Determinants of judged validity. *Journal of Experimental Social Psychology, 27,* 576-605.

Arkes, H. R., Hackett, C., & Boehm, L. (1989). The generality of the relation between familiarity and judged validity. *Journal of Behavioral Decision Making, 2,* 81-94.

Arkes, H. R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology, 49,* 323-330.

Bilder, R. M. (2011). Neuropsychology 3.0: Evidence-based science and practice. *Journal of the International Neuropsychological Society, 17,* 7-13.

Binder, L. M., Iverson, G. L., & Brooks. B. L. (2009). To err is human: "Abnormal" neuropsychological scores and variability are common in healthy adults. *Archives of Clinical Neuropsychology, 10,* 27-46.

Board of Directors (2007). American Academy of Clinical Neuropsychology (AACN) practice guidelines for neuropsychological assessment and consultation. *The Clinical Neuropsychologist, 21,* 209-231.

Brooks, B. L., Strauss, E., Sherman, E. M. S., Iverson, G. L., & Slick, D. J. (2009). Developments in neuropsychological assessment: Refining psychometric and clinical interpretive methods. *Canadian Psychology, 50,* 196-209.

Bush, S. S. (2010). Determining whether or when to adopt new versions of psychological and neuropsychological tests: Ethical and professional considerations. *The Clinical Neuropsychologist, 24,* 7-16.

Bush, S.S., Sweet, J. J., Bianchini, K. J., Johnson-Greene, D., Dean, P. M., & Schoenberg, M. R. (2018). Deciding to adopt revised and new psychological and neuropsychological tests: An inter-organizational position paper. *The Clinical Neuropsychologist, 32,* 319-325.

Calamia, M., Markon, K., & Tranel, D. (2013). The robust reliability of neuropsychological measures: Meta-analyses of test-retest correlations. *The Clinical Neuropsychologists, 27,* 1077-1105.

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

Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *The Journal of General Psychology, 103,* 290-304.

Charter, R. A. & Feldt, L. S. (2001). Meaning of reliability in terms of correct and incorrect clinical decisions: The art of decision making is still alive. *Journal of Clinical and Experimental Neuropsychology, 23,* 530-537.

Delis, D., Kramer, J., Kaplan, E., & Ober, B. A. (2000). *California Verbal Learning Test* (2nd ed.). San Antonio, TX: The Psychological Corporation.

Duff, K. (2012). Evidence-based indicators of neuropsychological change in the

individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology, 27,* 248-261.

Faust, D. (1984). *The limits of scientific reasoning. Minneapolis, MN: University of Minnesota Press.*

Faust, D. (2012). Criteria for appraising scientific status I: Daubert factors. In D. Faust (Ed.), *Coping with Psychiatric and Psychological Testimony*, *6th ed.* (pp.42-87). New York, NY: Oxford University Press.

Faust, D. & Ahern, D. C. (2012). Clinical judgment and prediction. In D. Faust (Ed.), *Coping with Psychiatric and Psychological Testimony*, *6th ed.* (pp.147-208). New York, NY: Oxford University Press.

Faust, D., Ahern, D. C., & Bridges, A. J. (2012). Neuropsychological (brain damage) assessment. In D. Faust (Ed.), *Coping with Psychiatric and Psychological Testimony, 6th ed.* (pp. 363-469). New York, NY: Oxford University Press.

Gigerenzer, G. & Goldstein, D. G. (2011) The recognition heuristic: A decade of research. *Judgment and Decision Making, 6,* 100-121.

Hauser, J. R. (2011). A marketing science perspective on recognition-based heuristics. *Judgment and Decision Making, 6,* 100-121.

Lareau, C. R. & Ahern, D. C. (2012). A primer on psychological, intelligence, cognitive, and neuropsychological testing. In D. Faust (Ed.), *Coping with Psychiatric and Psychological Testimony, 6th ed.* (pp. 281-301). New York, NY: Oxford University Press.

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,* 45-51.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). New York, NY: Oxford University Press.

Martin, P. K., Schroeder, R. W., & Odland, A. P. (2015). Neuropsychologists' validity testing beliefs and practices: A survey of North American professionals. *The Clinical Neuropsychologist, 29,* 741-776.

Mitrushina, M., Boone, k. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). New York, NY: Oxford University Press.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2,* 175-220.

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,* 30-65.

Rabin, L. A., Paolillo, E., Barr, W. B. (2016). Stability in test-usage practices of clinical neuropsychologists in the United States and Canada over a 10-year period: A follow-up survey of INS and NAN members. *Archives of Clinical Neuropsychology, 31,* 206-230.

Schofield, W. (1952). Critique of scatter and profile analysis of psychometric data. *Journal of Clinical Psychology, 8,* 16-22.

Schretlen, D. J., Munro, C. A., Anthony, J. C., & Pearlson, G. D. (2003). Examining the range of normal intraindividual variability in neuropsychological test performance. *Journal of the International Neuropsychological Society, 9,* 864-870.

Schretlen, D. J., Testa, S. M., Winicki, J. M., Pearlson, G.D., & Gordon, B. (2008). Frequency and bases of abnormal performance by healthy adults on neuropsychological testing. *Journal of the International Neuropsychological Society, 14,* 436-445.

Simons, R., Goddard, R., & Patton, W. (2002). Hand-scoring rates in psychological testing. *Assessment, 9,* 292-300.

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

Sweet, J. J., Benson, L. M., Nelson, N. W., & Moberg, P. J. (2015). The American Academy of Clinical Neuropsychology, National Academy of Neuropsychology, and Society for Clinical Neuropsychology (APA Division 40) 2015 TCN professional practice and 'salary survey': Professional practices, beliefs, and incomes of U.S. Neuropsychologists. *The Clinical Neuropsychologists, 29,* 1069-1162.

Thoma, V. & Williams, A. (2013). The devil you know: The effect of brand recognition and product ratings on consumer choice. *Judgment and Decision Making, 8,* 33-44.

Wedding, D. and Faust, D. (1989). Clinical judgment and decision making in

neuropsychology. *Archives of Clinical Neuropsychology, 4,* 233-265.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale – Third Edition: Administration and scoring manual.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale – Fourth Edition: Administration and scoring manual.* San Antonio, TX: The Psychological Corporation.

Wong, T. M. (2006). Ethical controversies in neuropsychological test selection, administration, and interpretation. *Applied Neuropsychology, 13,* 68-76.