2-7-2019

Performance Validity Testing for Individuals with Limited English Proficiency

Kelly An
University of Windsor

Performance Validity Testing for Individuals with Limited English Proficiency

By

Kelly An

A Dissertation
Submitted to the Faculty of Graduate Studies
through the Department of Psychology
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy
at the University of Windsor

Windsor, Ontario, Canada

2019

Performance Validity Testing for Individuals with Limited English Proficiency

by

Kelly An

APPROVED BY:

____
E. D. Bigler, External Examiner
Brigham Young University

____
W. Park
School of Social Work

____
A. D. Baird
Department of Psychology

____
C. A. Abeare
Department of Psychology

____
L.A. Erdodi, Advisor
Department of Psychology

January 24, 2019

DECLARATION OF ORIGINALITY

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

Background

Performance validity tests (PVTs) are an integral component of neuropsychological assessments. Despite the growing literature on PVTs, little research has focused on how these instruments perform in individuals with limited English proficiency (LEP). Indeed, the majority of PVTs have been developed and validated with individuals who are native speakers of English (NSE), and their psychometric properties have not yet been established for an LEP population.

Objectives

The current dissertation aimed to (1) determine the effect of LEP on PVT performance; (2) examine signal detection properties of current PVTs in individuals with LEP; and (3) develop new PVT cutoffs for this population.

Methods

To examine these objectives, a two-part prospective study was conducted. Part 1 consisted of using a case-control design to compare PVT performance between LEP and NSE groups. Part 2 consisted of using a single-blind, experimental-malingering design to establish classification accuracy across a battery of PVTs with an LEP sample.

Participants (N = 140) were randomly assigned to either a non-malingering control or experimental-malingering condition. Research assistants, who were blinded to the experimental condition of the participant, administered a battery of neuropsychological tests containing PVTs with high verbal mediation (PVT_{HVM}) and low verbal mediation (PVT_{LVM}). Both a liberal cutoff, maximizing sensitivity, and a conservative cutoff, emphasizing specificity, were chosen from the literature to calculate base rates of failure (BR_{Fail}).

Results

Part 1. Under normal conditions (i.e., not instructed to malinger), participants with LEP had a higher BR_{Fail} on and failed more PVT_{HVM} compared to NSE. In contrast, BR_{Fail} and number of PVTs failed were similar between groups on PVT_{LVM}. English proficiency was highly correlated with BR_{Fail} on PVT_{HVM} but not on PVT_{LVM}.

Part 2. Using published cutoffs, PVT_{LVM} demonstrated good classification accuracy, while the majority of PVT_{HVM} were not specific to malingering for the LEP sample. Adjusted cutoffs resulted in high sensitivity while maintaining adequate specificity on many PVT_{LVM}, but an optimal balance of sensitivity and specificity was unable to be obtained on some PVT_{HVM} regardless of how cutoffs were adjusted.

Conclusions & Future Directions

PVT_{HVM} increased false-positive errors for individuals with LEP, as both experimental malingering and LEP produce an elevated BR_{Fail} on these tests. Although there were instrument-specific exceptions to the overall findings, it is generally recommended that examiners preclude the use of PVT_{HVM} for individuals with LEP. The current study established new cutoffs on many PVTs that are both specific and sensitive for this population.

As a field, neuropsychological testing with cultural and linguistic minorities have been identified as a prominent issue, and the need for further studies in individuals with LEP is evident. Future investigations should focus on validating the new LEP cutoffs with different demographic samples in clinical and forensic settings.

ACKNOWLEDGEMENTS

This work would not have been possible without the support of many individuals throughout my doctoral training. To my research advisor, Dr. Laszlo Erdodi, thank you for accepting me under your wing and fostering my interest and knowledge in PVTs. Your mentorship, while not always conventional, challenged me to think critically and grow as a researcher. To Dr. Anne Baird, you have been instrumental to my development throughout graduate school, and I am sincerely grateful for your warm support and guidance, especially during my Master's thesis and comprehensive exams. To Drs. Chris Abeare and Wansoo Park, your insightful comments and suggestions have helped shape this document into its current version. To my External Examiner, Dr. Erin Bigler, I am honoured to have you be part of my defense. To my research assistants, Jordan Charles, Sami Ali, Anca Enache, and Jasmine Dhuga, I owe you my deepest appreciation for your time and dedication in collecting data for this project.

My personal and professional development is also the result of the continuous support and encouragement of my friends and family. Orrin-Porter, I would not have survived graduate school without you. Hio Tong, your presence made Windsor infinitely better. To my parents, my motivation to pursue doctoral studies is the direct result of the strong values of achievement you instilled in me. Thank you for supporting my career aspirations and your guidance throughout. To my baby sister, Ashley, thank you for never making me explain my work to you – seeing you is a welcome respite from my adult responsibilities. To Makoto, thank you keeping my lap warm and rewarding me with purrs during the writing of this document. Finally, to my husband, Alex, your decision to move to Windsor means the world to me. Thank you for your unwavering love, patience, and perspective – I truly could not have done this without you.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I: INTRODUCTION

Neuropsychological testing rests on the assumption that there is a relation between brain and behavior. At its foundation, the field assumes that performance on neuropsychological tests is a valid measure of cognitive abilities. While there is strong evidence supporting the brain-behavior relationship (Lezak, Howieson, Bigler, & Tranel, 2012), a direct association between test performance and brain integrity is rarely observed, as performance may be affected by various non-neurological factors.

These interfering factors may be detected on performance validity tests (PVTs), which assess the integrity of the test data. PVTs are most widely used in settings where secondary gain may be involved (e.g., insurance benefits after a motor vehicle collision, ADHD assessment for student disability accommodations). Their utility extends beyond these settings, however, as threats to performance validity may occur during any assessment due to any number of reasons. Recent studies have shown that both healthy undergraduate students (An, Kaploun, Erdodi, & Abeare, 2016) and children (Kirkwood & Kirk, 2010) fail PVTs. Regardless of the underlying cause, failure on PVTs signals that noise has been introduced that obscures interpretation of the tests and suggests that an inaccurate portrayal of an individual’s abilities may have been painted.

An increasing appreciation of the role of non-neurological factors influencing test performance has created a rapidly growing interest in performance validity over the past few decades (Heilbronner et al., 2009). Although issues of performance validity had historically been a concern, interest in developing formal measures of performance validity began only in the late 1970’s (Heaton, Smith, Lehman & Vogt, 1978), and have since resulted in hundreds of publications (Martin, Schroeder, & Odland, 2015).

Despite the exponential growth, relatively few studies have examined PVTs in culturally- and linguistically-diverse groups. The majority of tests were developed and normed on White-Anglo, native speakers of English (NSE). Individuals with limited English proficiency (LEP) are less privileged in this respect; those who assess them are left without appropriate methods to evaluate performance validity. This is a concern as individuals with LEP may differ in unknown ways from NSE. Thus, methods to detect invalid performance on neuropsychological testing may themselves be less valid and reliable for minority groups.

Although issues regarding construct equivalency of tests for individuals with LEP apply to the broader field of clinical neuropsychology, performance validity research is especially lagging in this area. To date, only a handful of studies have been identified on this topic. However, the limited number of extant papers is not representative of the actual demographics of Canada, which is becoming increasingly culturally and linguistically diverse. According to the 2011 census, 20% of Canadians speak a mother tongue other than English or French and 17.5% speak more than one language at home, a substantial growth since the 2006 census (14.2%; Statistics Canada, 2011). Hence, the demand for services for cultural and linguistic minorities overshadows the actual availability of services, appropriate assessment instruments, and multiculturally competent neuropsychologists.

The current dissertation examines the effect of LEP on performance validity testing. While available guidelines recommend referring individuals with LEP to neuropsychologists competent in their native tongue, in actuality this best-practice standard is difficult to attain due to a lack of available clinicians proficient in other languages. Furthermore, it may be impossible to develop new assessment measures for all ethnic and language groups. Canada is represented by more than 200 languages, with numerous dialects and regional differences (Statistics Canada, 2011). In the Philippines, for example, over 75 dialects (e.g., Tagalog, Visayan) are spoken (Wong & Fujii, 2004). Likewise, Spanish-speaking Mexicans living in the suburbs of the west coast likely have a different set of experiences, values, and customs than Spanish-speaking Dominicans living in Toronto. These within-group differences may limit the validity and usefulness of applying norms for broad racial and language categories.

Instead of developing increasingly more group-specific instruments, the current research strived to examine the properties of and calculate new norms on existing PVTs for individuals with LEP. Specifically, the present dissertation had three objectives: (1) to examine the effect of LEP on PVT performance; (2) to examine whether current PVT cutoffs are useful in detecting non-credible performance in individuals with LEP; and (3) to develop new PVT cutoffs for this population. To this end, a prospective study was conducted consisting of two parts: the first comparing LEP and NSE participants using a case-control design, and the second calculating classification accuracy and cutoffs using an experimental-malingering design.

CHAPTER II: LITERATURE REVIEW

Performance Validity Testing

Terminology & conceptualization. Performance validity testing refers to measuring validity of performance on ability tests. Research has established that performance on PVTs is related to performance on neuropsychological tests (Green, Rohling, Lees-Haley, & Allen III, 2001). Although the term PVT was previously used interchangeably with symptom validity test (SVT; Pankratz, 1983), PVTs are now differentiated from SVTs to increase conceptual clarity, such that SVTs refer only to validity of symptom complaints (e.g., self-report personality or symptom questionnaires) as opposed to performance-based ability measures (e.g., intellectual and cognitive tests; Larrabee, 2012). This distinction is supported by research showing that performance and symptom validity load on different factors on factor analysis (Van Dyke, Millis, Axelrod, & Hanks, 2013). Thus, it has been recommended that performance and symptom validity be assessed independently. For clarity, the terms performance validity and PVTs will be used for the current research as the focus is on cognitive ability tests. Additionally, the terms invalid and non-credible performance will be used to refer to failure on PVTs.

Likewise, various terms have been used over the past few decades to describe failure on PVTs, ranging from insufficient, suboptimal, or poor effort to negative response bias to malingering. While no widespread consensus exists on the single most appropriate term to use, some terms have fallen out of favor in the field. Specifically, many neuropsychologists veer from using descriptors with *effort* (e.g., poor effort, insufficient effort) and from using *malingering*. The term *effort* has been criticized due to its vagueness and potential implication that failure on PVTs was volitional (Bigler, 2012). Performance invalidity does not imply inferences about the underlying cause of failures.

Similarly, the term *malingering* has been criticized for its implications. Malingering refers to the “intentional production of false or grossly exaggerated physical or psychological symptoms, motivated by external incentives….” (American Psychiatric Association, 2013, p. 726). Slick, Sherman, and Iverson (1999) originally proposed three categories with criteria for Malingered Neurocognitive Dysfunction: (1) Definite: presence of external incentives and below-chance performance on at least one forced-choice PVT, (2) Probable: presence of external incentives and at least two PVT failures (not below-chance), and (3) Possible: presence of external incentives and discrepant evidence from self-report.

As mentioned, there are numerous reasons for non-credible performance aside from deliberate exaggeration or fabrication of symptoms. For malingering to be diagnosed, there must be secondary external gain, such as financial compensation from a lawsuit or disability benefits after a motor vehicle collision. While it is possible to diagnose malingering via the DSM-V (Heilbronner et al., 2009), it cannot be diagnosed based on PVT failure alone. PVT failures may be used as one piece of evidence for diagnosing malingering, but other information is necessary, such as considering the context, historical and injury information, and behavioral observations.

The term *malingering* is also problematic on a conceptual level, as the conceptualization of conscious versus unconscious processes underlying PVT failures is not clearly dichotomous or mutually exclusive (Nies & Sweet, 1994). An individual may have varying degrees of both deliberate feigning of symptoms and unconscious motivations to maintain a “sick role” (e.g., factitious disorder). Research suggests that PVTs cannot differentiate malingering from psychiatric disturbances, such as somatic symptom disorders, as both cases result in failures on PVTs (Boone, 2007). To date, there are no tests that detect only malingerers and not those with somatic symptoms disorders. Teasing apart the conscious and unconscious forces and whether poor performance on testing represents truly experienced or feigned deficits is nearly impossible in a typical clinical practice. Nevertheless, in many forensic settings, the purpose of administering PVTs is to determine whether test results are valid; it may be less important to infer the underlying reason for the invalid results (Boone, 2007).

Test-taking effort has been conceptualized as a dynamic process that exist on a continuum (Boone, 2009). Instead of an “all or none” dichotomy, it is viewed as varying in levels from maximum to no effort. This conceptualization recognizes that many variables affect PVT scores. For example, pain, fatigue, and boredom may affect one’s level of engagement on tasks and result in decreased scores on PVTs. Additionally, test-taking effort is not static but a dynamic process; it can fluctuate within any given test session and across different tests and behaviors and may change in response to interactions with the environment (Boone, 2009).

Despite this dynamic conceptualization, decisions regarding performance validity are often conducted in a categorical manner in clinical settings. As with any categorical systems (e.g., DSM-V), this method of classification provides efficiency in clinical decision making but imposes an artificial dichotomy that often results in losing information in the process (Millon, Krueger, & Simonsen, 2010). As will be subsequently discussed, criteria for determining non-credible performance (e.g., how many PVT failures are required) has been heavily debated and stems from the process in translating test-taking effort as a continuous process into a dichotomous decision regarding overall performance validity on tests.

Methods to detect non-credible performance. Currently, an abundance of well-validated PVTs are available for clinicians, with no single “gold-standard” PVT (Bigler, 2012). What may be the most appropriate test and cutoff to use in one population and setting may have poor classification accuracy in another. As will be discussed, inferences based on any single PVT is generally not recommended; combining information from multiple indicators provides the most accurate information for determining performance validity. With this in mind, the following is a review of the types of PVTs and signal detection terminology necessary to evaluate PVTs.

Freestanding versus embedded tests. Two broad categories of PVTs are available. Freestanding measures are those developed specifically for the purposes of assessing performance validity and are administered separately from other neuropsychological measures (Bigler, 2014). For example, the Test of Memory Malingering (TOMM; Tombaugh, 1996) and Word Memory Test (WMT; Green, 2003) are two of the most frequently used freestanding PVTs (Martin et al., 2015).

In contrast, embedded measures are validity indicators “retro-fitted” into other measures; the tests from which they were derived were not originally developed to assess for performance validity but later validated for this purpose (Heilbronner et al., 2009). Numerous indicators have been developed in all major neuropsychological domains (memory, attention, visuospatial-perceptual abilities, executive function; Boone, 2007). These include, for instance, the Reliable Digit Span (RDS; Greiffenstein, Baker, & Gola, 1994) from the Wechsler Adult Intelligence Scale – Revised Edition (WAIS-R) Digit

Span subtest and the forced-choice recognition score from the California Verbal Learning Test – Second Edition (CVLT-II; Moore & Donders, 2004), two of the most frequently used embedded indicators (Martin et al., 2015).

Embedded indicators have several advantages over freestanding PVTs. From a resource perspective, they do not take additional time to administer and are cost-effective (Heilbronner et al., 2009). Neuropsychologists can use embedded indicators already available in their usual battery to assess validity without increasing testing time or buying separate material. Additionally, embedded PVTs overcome the potential limitation of coaching frequently present for freestanding PVTs. Some freestanding PVTs, because of their ease and distinct features (e.g., two forced-choice options), are more easily recognizable as PVTs and individuals can be coached to pass these measures in forensic settings (Suhr & Gunstad, 2000). Embedded PVTs are less susceptible to coaching because they are embedded within a more realistic test.

However, freestanding PVTs should not be jettisoned in favor of exclusive use of embedded indicators. Although broad generalizations comparing freestanding and embedded PVTs warrants caution given their range of signal detection properties, several embedded indices have been found to have lower sensitivity compared to freestanding measures at comparable specificity levels (e.g., Armistead-Jehle & Buican, 2013; Armistead-Jehle & Hansen, 2011). In a direct comparison of 17 embedded indicators with freestanding PVTs, Miele, Gunner, Lynch, and McCaffrey (2012) found that many embedded indicators produced greater false positives and negatives than freestanding PVTs. Hence, the additional time to administer freestanding PVTs may be justified as these indices often provide useful, non-redundant information over embedded measures (Heilbronner et al., 2009).

Regardless of the type, all PVTs are based on the concept of maintaining good face validity as genuine cognitive ability tests, despite actually being insensitive to neurological, psychiatric, or medical disorders (Larrabee, 2012). The TOMM, for example, presents as a challenging test of visual memory (e.g., memorization of 50 items is emphasized). However, performance is near ceiling for most credible samples (Bigler, 2012). Likewise, the Rey-15 Item Test (FIT) emphasizes that 15 different items need to be recalled after viewing the stimulus for 10 seconds, but in actuality only 5 sets of items need to be memorized as items are redundant and can be easily chunked (e.g., A B C; Boone, 2007). Near-ceiling performance is observed on these measures as they rely on the use of overlearned skills that are typically not affected by brain injury. Additionally, many tests rely on recognition memory (e.g., TOMM, WMT), which is easier than recall memory and often resilient to brain damage (Huppert & Piercy, 1976).

Diagnostic statistics in PVT research. In addition to below-chance performance, falling below norms-referenced cutoffs is the most common method to determine performance validity. Signal detection terminology is required to understand the interpretation of PVT results and is discussed below (Bianchini, Mathias, & Greve, 2001; Boone, 2007; Heilbronner et al., 2009).

True positive rate (TP) refers to the correct identification of a condition (i.e., invalid performance) from a positive test result (i.e., PVT failure). True negative rate (TN) refers to the correct identification of valid performance from a negative test result (i.e., passing PVTs). False positive rate (FP) refers to the incorrect identification of invalid performance in individuals who fail PVTs. False negative rate (FN) refers to the incorrect identification of valid performance in individuals who passed PVTs. Sensitivity refers to the percentage of invalid cases correctly identified as true positives. It is calculated by dividing the number of detected invalid cases by all invalid cases that exist in the particular sample (TP/TP + FN; Baldessarini, Finklestein, & Arana, 1983). Specificity refers to the percentage of valid cases correctly identified as true negatives. It is calculated by dividing the number of valid cases by all valid cases that exist in the sample (TN/TN + FP; Baldessarini et al.). To illustrate, a particular PVT cutoff may have a sensitivity of .60 and specificity of .90. This means that at this cutoff, the PVT failed to identify 40% of invalid cases and 10% of valid cases. Because of the high proportion of FN, failing the PVT at this cutoff suggests non-credible performance but passing does not necessarily suggest credible performance as 40% of invalid cases may have been missed.

There is a trade-off between sensitivity and specificity (Boone, 2007). As with the example above, for tests and cutoffs with lower sensitivity and high specificity, failing suggests invalid performance but passing does not necessarily imply credible performance because the cutoff may not detect all cases of poor effort (Boone, 2007). In contrast, cutoffs with high sensitivity but low specificity will result in a greater FP rate as they are less discriminative between poor and adequate effort.

Aside from sensitivity and specificity, positive predictive power (PPP) and negative predictive power (NPP) also offer useful data. Describing classification accuracy using PPP and NPP is advantageous over sensitivity and specificity as these concepts take into account population base rates (Bianchini et al., 2001). Sensitivity and specificity are calculated by dividing TP or TN by all individuals with (TP + FN) and without (TN + FP) the condition, respectively (Bianchini et al.). In contrast, PPP and NPP are calculated by dividing TP or TN by all individuals who are identified by the test as positive (TP + FP) or negative (TN + FN), respectively. Because base rates are taken into consideration, PPP and NPP vary as a function of the condition of interest in a specific population (Bianchini et al.). A setting consisting of personal injury litigants sustaining mild traumatic brain injury (TBI), for example, will have a much higher base rate of performance invalidity than an outpatient hospital clinic assessing older adults for dementia. Hence, PPP and NPP have direct clinical relevance as these statistics consider the properties of various clinical groups.

Both PPP and NPP can also be calculated using sensitivity (SN), specificity (SP), and prevalence (*p*) values (Baldessarini et al., 1983). Specifically, PPP is calculated by (SN x *p*)/[(SN x *p*) + (1 – *p*)(1-SP)] and NPP is calculated by (1 – *p*)SP/[(1 – *p*)SP + *p*(1 – SN)]. These equations are useful in PVT research as classification accuracy properties with hypothetical base rates (e.g., 10%, 30%, 50%) can be calculated in order provide clinically relevant information for different populations.

Practice standards & guidelines.

Continuous assessment of effort. As discussed, performance validity cannot be assumed to be static throughout a testing session. Fluctuations in performance may occur due to any number of reasons. For example, an internal or external event (e.g., panic attack, pain) during an assessment may cause an individual to exert less than optimal effort during some portions of the assessment. As such, it has been recommended that performance validity be continuously monitored throughout an assessment (Boone, 2009). Additionally, because individuals may differ in their strategies to feign impairment (e.g., performing poorly on only verbal memory tests), it has been recommended to incorporate PVTs that assess a variety of cognitive domains (Cottingham, Victor, Boone, Ziegler, & Zeller, 2014; Heilbronner et al., 2009).

Sensitivity & specificity. In clinical practice, a more conservative stance (e.g., emphasizing high specificity) for interpreting PVT results is generally recommended. Because of the potentially grave consequences of falsely identifying an individual's performance as non-credible (e.g., financial loss, misdiagnosis, emotional distress), a specificity of .90 is the accepted standard (Bianchini et al., 2001). However, the consequences of FN may be equally as damaging in some circumstances (e.g., limiting access to resources for individuals with legitimate impairments; Bianchini et al.). Hence, the costs of FP and FN should be considered in accordance with the setting. In many cases, specificity and sensitivity need to be balanced such that sensitivity is not unreasonably lowered while trying to maintain high specificity.

Interpreting PVT failures. Overall, there is consensus that judgments regarding performance validity should be made based on multiple PVTs and domains of behavior (Bigler, 2012; Boone, 2007; Heilbronner et al., 2009; Larrabee, 2003; Victor, Boone, Serpa, Buehler, & Ziegler, 2009). Because effort may vary within a given assessment, and PVTs do not have perfect specificity and sensitivity, one may make serious clinical judgements if interpretations are based on one instrument given at one point of an assessment. Research has shown that using multiple PVTs versus a single PVT greatly improves specificity (e.g., Larrabee, 2012; Victor et al.). Indeed, this standard of administering multiple PVTs is reflected in actual practice. In a survey of neuropsychologists with expert knowledge of PVTs, participants reported that the average number of PVTs they typically administer is six in clinical settings and eight in forensic settings (Schroeder, Martin, & Odland, 2016).

Despite general recommendations to incorporate multiple measures of performance validity into an assessment, the question of *how* to interpret a combination of PVTs has been heavily debated in the literature. For example, how many failures should be required before one arrives at an impression of non-credible performance? Victor and colleagues (2009) found high sensitivity (.95) but low specificity (.53) when any one of four PVTs administered was used as the criterion for failure, whereas specificity (.94) greatly improved using a "pairwise model" (i.e., failure on two PVTs) while maintaining adequate sensitivity (.84). Similarly, Larrabee (2003) found that using a combination of any two of five PVTs resulted in .89 specificity and .88 sensitivity in classifying individuals with moderate-to-severe TBI.

Another consideration is how to interpret a "near pass" (i.e., failures hovering just below the cutoff; Bigler, 2012). Although below-chance performance signals unequivocal invalidity on tests, interpretation of performance in the "near pass" range is not as clear. Pass/Fail cutoff points may overlap with a range sensitive to genuine impairment for some conditions (e.g., dementia), resulting in false positives if these borderline cases are deemed invalid. The ambiguity of interpreting the "near pass" has been recently addressed by some researchers using composite models that recognize the continuum of performance validity and accounts for both the number and extent of PVTs failures (e.g., An et al., 2019; Erdodi, Sagar, et al., 2018; Zuccato, Tyson, & Erdodi, 2018). Consideration of other methods for determining performance validity (e.g., neuroimaging) has also been suggested in addition to PVT interpretation (Bigler, 2015).

Nevertheless, although more research is required on PVT interpretation, there is a general consensus that considering a multimodal approach (e.g., interview, PVTs, neuroimaging) and using multiple PVTs provides the best classification accuracy.

PVT research designs. Two types of research designs are commonly used to examine the characteristics of PVTs: (1) malingering simulation and (2) known-groups (Heilbronner et al., 2009). Experimental-malingering simulations (i.e., analogue studies) compare participants who are instructed to feign cognitive impairment on tests to a control group not instructed to feign impairment (Bianchini et al., 2001). In contrast, a known-groups design compares PVT performance between participants determined to be credible to participants determined to be non-credible based on a set criterion (e.g., presence of external incentives, performance on an established PVT; Rogers, 2008).

As with other types of research designs, a tradeoff between experimental rigor and clinical relevance is paralleled with these two types of designs. Known-groups studies use participants who may have real external incentives for performing non-credibly, thus increasing external validity. However, these are case-control studies and lack the experimental control of simulation designs. Furthermore, defining the criterion groups may present a challenge (Rogers, 2008). Individuals who are purposely performing non-credibly are not likely to readily disclose their intentions. Moreover, using litigation status as a criterion may not adequately differentiate groups, as not all individuals will feign impairment. Using performance on other well-established PVTs to define criterion groups depends on the signal detection properties of the criterion PVT in the population of interest and may result in also result in incorrect classification.

In contrast, experimental-malingering studies may be limited by their poor external validity (Heilbronner et al., 2009). Several factors threaten the external validity in simulation studies. Firstly, participants in simulation studies often differ in characteristics from genuine malingerers in real-world settings. For example, college students, who are generally young, educated, and healthy, are frequently used in these studies but may be unable to reproduce the inner reality of a TBI patient involved in a motor vehicle-related litigation (Haines & Norris, 2001). Secondly, the simulation scenario participants are given during experiments may not be believable or relatable and depend on the participant's ability and willingness to engage in the scenario (Rogers, 2008). Finally, simulation studies lack the external incentives and consequences contingent on performance that are present in clinical or forensic settings (Rogers, 2008). A $20 research incentive for participation, for instance, does not come close to the motivation and consequences of receiving thousands of dollars in monetary gains or disability benefits (Bianchini et al.).

Despite their limitations, experimental-malingering studies are one of the best available methods to establish PVT classification accuracy. The American Academy of Clinical Neuropsychology (AACN; 2007) guidelines recommend that both experimental-malingering and known-group studies be conducted to validate new PVTs, as they complement one another in internal and external validity (Heilbronner et al., 2009).

Assessment of English Language Proficiency

The literature on language proficiency assessment is vast and falls under the specialty of educational assessment, outside the scope of neuropsychology. The following discussion will not do justice to the enormous amount of research in this area but will provide an overview of the literature relevant to the current dissertation.

Terminology. Language proficiency is a multi-domain construct that has been conceptualized as part of the larger umbrella term of language competence (Marian, Blumenfeld, & Kaushanskaya, 2007). It is distinct from but related to the concepts of language dominance and language preference, two other components of language competence. Language preference refers to one's subjective feelings toward a language, and language dominance is a relative term comparing usage in two or more languages. While preference and dominance may be congruent with one's proficiency level (e.g., one prefers speaking Mandarin, uses Mandarin in most settings, and has greater proficiency in Mandarin than English), these do not necessarily align.

Moreover, language proficiency may vary depending on the domain. For example, individuals may have different levels of English proficiency in reading, writing, speaking, and listening. Some areas may be more developed than others due to an individual's exposure and experiences. University students immersed in an English-speaking university in Canada, for example, may have a higher level of reading and writing proficiency but more limited conversation skills in English.

Assessment methods. Language proficiency can be assessed through various methods: self-report, interview, and performance-based measures. Self-report ranges from an informal question regarding one's language proficiency to well-validated questionnaires (e.g., LEAP-Q). Interviews may consist of a standardized semi-structured format, such as the Oral Proficiency Interview based on the American Council of Teaching Foreign Languages guidelines (Liskin-Gasparro, 2003). Finally, standardized performance-based tests such as the Multilingual Naming Test (MINT) or the Boston Naming Test (BNT) have been examined as objective measures of English proficiency. Research suggests that adults have accurate self-reports of their language proficiency, and this is the most popular assessment method in the literature (Marian et al., 2007). Additionally, self-reported proficiency is highly correlated with proficiency determined through objective testing, often with robust and large effects (Marian et al.).

However, there are some limitations of self-report methods. Specifically, language proficiency self-ratings have been found to be better predictors of language dominance rather than proficiency level, per se (Sheng, Lu, & Gollan, 2014). Furthermore, some studies suggest that performance-based assessments of English proficiency tend to indicate greater English proficiency and dominance than self-report measures (Gollan, Weissberger, Runnqvist, Montoya, & Cera, 2012; Sheng et al.). Because of inherent weaknesses with any one method, a multi-measure approach to language proficiency assessment has been advocated (e.g., Gollan et al., Sheng et al.).

Neuropsychological Testing with Culturally & Linguistically Diverse Populations

Terminology. Culture and ethnicity are complex, multidimensional constructs. There is debate on how race, culture, and ethnicity are defined within the literature, and some studies fail to operationalize these terms altogether. For the purposes of the current dissertation, these terms will follow the definitions from the *Guidelines on Multicultural Education, Training, Research, Practice, and Organizational Change for Psychologist*s (American Psychological Association [APA], 2003). Specifically, race is defined as a socially constructed category that is assigned to individuals based on physical characteristics (e.g., skin color). Ethnicity, on the other hand, is defined as “the acceptance of the group mores and practices of one's culture of origin and the concomitant sense of belonging" (APA, 2003, p. 380). Thus, while racial groups are often assigned to individuals based on physical appearance, ethnic groups are chosen by individuals based on their sense of belonging in and acceptance by a group. Finally, culture is a fluid and dynamic category that refers to "belief systems and value orientations that influence customs, norms, practices, and social institutions" that embody a certain worldview and may be learned and passed down (APA, 2003, p. 380). Discussion of culture in the literature often refers to ethnicity or race, but in fact encompasses a very large number of constructs (e.g., sexual orientation, gender identity, religion). For the purposes and sake of clarity in the current dissertation, any discussion of culture will be narrowly limited to ethnicity.

Influence of LEP & cultural factors on neuropsychological assessment. Differences in neuropsychological test performance between individuals from a White-European background and other ethnicities in North America have been found in both non-clinical (e.g., Jacobs et al., 1997) and patient populations (e.g., Boone, Victor, Wen, Razani, & Pontón, 2007). Individuals from ethnically diverse groups in North America have been found to obtain lower scores across tests of attention, learning and memory, language, visual-constructional ability, processing speed, and executive functions compared to a White-European sample (Boone et al., 2007). These differences have been studied and found mainly in African American (e.g., Manly, Byrd, Touradji, & Stern, 2004; Schwartz et al., 2004) and Hispanic samples (e.g., Gasquoine, 1999; Jacobs et al.). Fewer studies have examined neuropsychological performance in Asians as a group, with most studies categorizing Asians with other groups due to small sample sizes (e.g., Razani, Murcia, Tabares, & Wong, 2007).

Differences in neuropsychological tests scores have also been found between NSE and individuals with LEP in North America. Boone and colleagues (2007), for example, found that LEP participants performed significantly worse on the Digit Span, Boston Naming, and FAS tests compared to NSE participants. These findings not only apply for tests with a verbal component, but differences in performance have also been found on non-verbal tests (Rosselli & Ardila, 2003). As will be subsequently discussed, between-group differences in test performance are observed not only because of linguistic barriers, but because of lack of construct and test equivalency, thus making the perception of non-verbal measures as “culture-free” a misconception (Rosselli & Ardila, 2003).

Race and ethnicity are demographic descriptors, akin to age or gender. Similar to examining age as a moderating variable, race and ethnicity do not infer a causal relationship to test scores but are correlated with other variables that explain between-group differences (Brickman, Cabo, & Manly, 2006). The mechanisms underlying these differences are complex and multifactorial, and it would be an error to attribute test performance differences to inherent differences in cognitive abilities between ethnic groups, as research has not supported this view (Ojeda, Aretouli, Peña, & Schretlen, 2016). Instead, multiple factors associated with culture (e.g., acculturation level; Manly, et al., 2004; Saez et al., 2014) or inherent in the situation (e.g., stereotype threat; Steele & Aronson, 1995) underlie observed differences in test scores. Other factors associated with culture that have been found to influence test performance include quality of education (Cavé & Grieve, 2009; Chin, Negash, Xie, Arnold, & Hamilton, 2012; Fyffe et la., 2011;

Manly, Jacobs, Touradji, Small, & Stern, 2002; Sisco et al., 2014), test-wiseness (Manly et al., 2002), and degree of literacy (Manly, Touradji, Tang, & Stern, 2003).

Construct validity and test equivalency are also compromised when instruments are used with individuals who are vastly different from the normative sample (Brickman et al., 2006; Rivera Mindt, Byrd, Saez, & Manly, 2010). The majority of neuropsychological tests are developed with White, middle-to-upper class NSE in North America. When used in other cultural groups, differences in values and experiences may result in performance differences. For example, Western worldviews emphasize individualism (e.g., independence and achievement) and verbal communication, both of which are reflected in the concept of testing. Additionally, several components in cognitive testing may be incongruent with the values of other cultures (Ardila, 2005). For instance, there may be differences in how one is expected to behave in a one-to-one relationship with an authority figure or the value in performing one’s best (which is less emphasized in less competitive cultures; Ardila, 2005).

Many abilities are also not innate and depend on experience. Skills such as copying a figure or mental arithmetic, for example, are associated with schooling and may not be relevant in some cultures (Rosselli & Ardila, 2003). Differences in response styles, such as prioritizing speed versus accuracy, have been found across cultures and influence test performance (Ojeda et al., 2016). Items on tests may also be interpreted differently between cultural groups. What is deemed an “intelligent” response to a Vocabulary item from the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV), for instance, is decided through the lens of the culture that it is developed from. Individuals from developing countries, for example, tend to focus on the function (e.g., a lion runs fast and may harm you) when defining Vocabulary items, whereas individuals from industrialized countries focus on taxonomy (e.g., a lion is an animal; Flynn, 2007), the latter of which is rewarded in the current WAIS-IV scoring system. Thus, neuropsychological tests can be seen as cultural devices that reflect one particular worldview or value system (Cole, 1999). Tests cannot be “free” of cultural bias as one is inferring brain integrity from behavior, which is not independent of experience.

Practice guidelines versus actual practice.

Cultural competency & ethics. To serve the growing culturally diverse population in Canada, the importance of cultural competency has been increasingly emphasized. The most widely accepted definition of cultural competence involves three overarching components: awareness, knowledge, and skills (Sue, Arredondo, & McDavis, 1992; Rivera Mindt et al., 2010). Specifically, competency in multicultural issues incorporates awareness of one’s own biases, values, and attitudes towards other cultural groups, culture-specific knowledge regarding other groups, and appropriate clinical skills to work with a diversity of individuals. In addition, providing culturally competence services is an ethical duty of psychologists. Specifically, the APA Ethical Principles of Psychologists and Code of Conduct state that psychologists have the duty to "take into account the purpose of the assessment as well as the various test factors, test-taking abilities, and other characteristics of the person being assessed, such as situational, personal, linguistic, and cultural differences, that might affect psychologists’ judgments or reduce the accuracy of their interpretations" (APA, 2002, p.13).

While no neuropsychologists would disagree with the importance of considering demographic factors during an assessment, few practice guidelines in neuropsychology exist that are targeted towards working with individuals of different cultures. Indeed, the Practice Guidelines of the AACN have only one page dedicated to discussing cultural issues, with little in the way of specific recommendations pertaining to clinical practice (Rivera Mindt et al., 2010). Furthermore, the existing Guidelines are not consistently enforced in the field, resulting in a disparity between the Guidelines and actual practice (Elbulok-Charcape, Rabin, Spadaccini, & Barr, 2014). For example, the Guidelines suggest that the best practice for assessing individuals with LEP is to refer to another neuropsychologist who is competent in the client's native language. Although reasonable, in actuality this standard is difficult to uphold. In a survey of 512 doctorate-level psychologists, only 15% of participants identified as being adequately fluent to administer tests in another language (Elbulok-Charcape et al., 2014). Thus, underrepresentation of cultural and linguistic minorities in the field create challenges in actually referring a client to a neuropsychologist of similar background.

Development and translation of tests. Because the majority of neuropsychological tests are developed for an English-speaking population, tests are often modified and validated for other languages post-development. Although there are best practice guidelines for translation of tests (e.g., Artiola i Fortuny & Mullaney, 1998), this standard may be too high for consistent adherence. In many cases where appropriate non-English versions and norms are unavailable, psychologists may turn to translating tests within their practice in an attempt to provide a valid assessment for clients with LEP (Elbulok-Charcape et al., 2014). However, this "in house" method is not recommended. Many psychologists, even with native fluency in another language, are not trained in the translation of tests (Rivera Mindt et al., 2010). There are subtle differences in languages and translations from an untrained professional may result in meaning being lost. Furthermore, even if a linguistically equivalent version is produced, this practice does not guarantee construct and test equivalency (Brickman et al., 2006).

Assessment and interpretation of tests. Psychologists have an ethical responsibility to refer a client elsewhere if the assessment is outside of one's competence. This competence has been interpreted by some psychologists to include linguistic competence. Specifically, an argument has been made that it is unethical to assess individuals in another language if one is not fluent in that language (Artiola i Fortuny & Mullaney, 1998). In practice, this standard of referring to other neuropsychologists with bilingual fluency is nearly impossible (Brickman et al., 2006). Finding neuropsychologists competent in both a certain language *and* a specific area of practice may be difficult, even in larger cities. For instance, although some neuropsychologists residing in Ontario will have native fluency in Mandarin, there will be a sparsity of Mandarin-speaking neuropsychologists who also have the in-depth knowledge, expertise, and insight to assess a young adult survivor of childhood acute-lymphoblastic leukemia for cognitive effects of chemo-radiation. Indeed, although the majority of neuropsychologists (68.6%) would ideally refer an individual with LEP to a neuropsychologist competent in the client's language, "difficulty finding a colleague to whom the patient can be referred or who can be consulted" was identified as one of the top three challenges in assessment with ethnic or linguistic minorities reported in the same survey (Elbulok-Charcape et al., 2014, p. 357).

In the absence of a qualified neuropsychologist to whom to refer, using an interpreter as been suggested as the next best solution (Romero et al., 2009), and the second most common practice (40%) when assessing a client with LEP amongst neuropsychologists (Elbulok-Charcape et al., 2014). However, this option also poses many issues. For example, interpreters may not have the necessary knowledge of the terminology required in a neuropsychological assessment. Much nuanced information may be lost in the back-and-forth live interpretation during an assessment. Many verbal tests, such as the WAIS-IV Vocabulary subtest or a list-learning task such as the CVLT-II, are invalid once interpreted. Even the use of nonverbal tests is of questionable validity, as standardization is compromised with the use of an interpreter such that the original standardization environment usually does not include an interpreter (Brickman et al., 2006). Interpreters also may not have received adequate training to know how to work with the psychiatric and neurological populations often seen by neuropsychologists and may be unequipped to respond in crisis situations or when clients reveal details regarding suicidality and other psychiatric symptoms. They may also inadvertently and inappropriately reveal information about tests. Finally, the inherent difficulty in using interpreters is that it is impossible to assess the interpreter’s fluency level and verify the accuracy of their translation.

In terms of interpretation of test scores when assessing ethnic or linguistic minorities, the most common approaches reported by neuropsychologists involve using norms matching the client’s race/ethnicity (82.4%), using education-corrected norms (45.4%), and adjusting cognitive test scores (22.8%; Elbulok-Charcape et al., 2014). However, the lack of appropriate norms and tests were also reported as two of the greatest challenges. Thus, the disparity between the reported approaches to working with LEP clients and the perceived problems highlight the unresolved challenges in the field.

Performance validity testing with cultural and linguistic minorities. Unlike the literature on PVTs in general, which has exponentially bloomed over the past few decades, studies on PVTs for individuals with LEP have taken a slower pace. Many studies (e.g., Kim et al., 2010) specifically excluded individuals who were not NSE. Thus, it is unclear whether findings from these studies can be generalized to individuals with LEP. A small pool of extant studies has examined the TOMM in a Hong Kong sample (Chang, 2006), Digit Span embedded indicators in a Taiwanese sample (Yang et al., 2012), the TOMM, Dot Counting Test (DCT), Victoria Symptom Validity Test (VSVT), B Test, and FIT in Spanish samples (Burton, Vilar-López, & Puente, 2012; Vilar-López et al., 2007; Vilar-López, Gomez-Rio, Caracuel-Romero, Llamas-Elvira, & Perez-Garcia, 2008a; Vilar-López, Gómez-Río, Santiago-Ramajo et al., 2008b), the FIT and RDS in a Japanese sample (Yamaguchi, 2005), the Hiscock's Forced-Choice Digit Memory Test with a Chinese sample (Liu, Gao, Li, & Sheng, 2001), and DCT with a rural Indian sample (Weiss & Rosenfeld, 2010). There is also a handful of studies examining symptom validity in cultural and linguistic minorities (e.g., DuAlba & Scott, 1993). However, in an effort to stay within the scope of the current dissertation, the following review will be limited to performance validity.

Overall, relevant studies have found adequate classification accuracy of PVTs in other cultural groups when cutoffs were adjusted to maintain specificity. Although some studies did not support using some PVTs for certain groups, these studies contained methodological flaws or insufficient data. For example, Weiss and Rosenfled (2010) found that performance on the DCT was lower for their rural Indian sample compared to published norms and that no cutoffs provided both adequate specificity and sensitivity for this group. This study, however, was limited by their unclear differentiation between credible and non-credible groups and restricted range of cutoffs examined. Similarly, Yang and colleagues (2012) found significant differences between the Chinese version of the Digit Span in a Taiwanese sample and the WAIS-III Digit Span US norms. Although no classification accuracy data were reported, the authors recommended that Digit Span embedded indicators not be used in this population.

In contrast, other studies not only supported the use of PVTs with other cultural groups but reported no performance differences between groups. Vilar-López and colleagues (2008a, 2008b), for example, found no differences on the TOMM, DCT, VSVT, B Test, and FIT between a European Spanish-speaking sample sustaining mild TBI and published North American norms. Furthermore, these researchers found that all tests were able to differentiate credible and non-credible groups. However, it is unclear whether PVTs in these studies were administered in English or Spanish and the level of English proficiency of participants if tests were administered in English.

The studies reviewed above examined PVT performance in cultural groups outside of North America. Only one published study examined the influence of cultural and linguistic factors in English-speaking ethnic minorities or individuals who speak English as a second language within North America. In an archival study comparing PVT performance across a large U.S. sample (*N* = 168), which consisted of approximately 50% non-White and 17% LEP participants, differences were found between racial groups on several PVTs (Salazar, Lu, Wen, and Boone, 2007). Specifically, differences were found after co-varying for age and education on the Digit Span Age-Corrected Scale Score (DS-ACSS), RDS, Rey Auditory Verbal Learning Test (RAVLT) Recognition, effort equation, and discriminant function, Rey-Osterrieth Complex Figure Test (RCFT) effort equation, and RCFT/RAVLT discriminant function. No differences were found on the FIT, DCT E-score, and Warrington Recognition Memory Test.

Salazar and colleagues (2007) also compared PVT performance between LEP and NSE participants. Aside from worse performance on the RDS and better on the RCFT effort equation for LEP participants, no between-group differences were found. DS-ACSS and RDS were also related to age at which English was learned. No effect was found for number of years lived or educated in the United States. Salazar and colleagues also examined cutoffs for the LEP group that would produce a specificity of .90. Adjustments were necessary on most measures to maintain this level of specificity: DS-ACSS (≤4), RDS (≤5), DCT E-scores (≤19), FIT (≤12), and RCFT effort equation (≤45).

Although the above study is the only published research examining PVT performance in individuals with LEP in North America, there are several limitations. Aside from their retrospective design and small sample size of comparison groups, the most notable limitation relates to the lack of a non-credible comparison group and the omission of necessary classification accuracy to interpret their proposed cutoffs. Specifically, the study only compared base rates of PVT failure between LEP and NSE groups, which precludes the calculation of sensitivity data. Although Salazar and colleagues (2007) reported that RDS cutoff of ≤5 for the LEP group produces specificity of .96, for example, this may not be a useful cutoff if sensitivity is low. The present dissertation addressed this limitation by not only comparing PVT performance between groups, but also by calculating classification accuracy using a prospective, experimental design to determine clinically useful cutoffs for individuals with LEP.

CHAPTER III: HYPOTHESES

The present dissertation had three broad objectives: (1) to examine the effect of LEP on PVT performance; (2) to examine whether current PVT cutoffs are useful in detecting non-credible performance in individuals with LEP; and (3) to develop new cutoffs for this population. To this end, the research consisted of two parts using a prospective data collection.

Part 1: How do Individuals with LEP Perform on PVTs Compared to NSE?

This portion of the dissertation consisted of a prospective case-control design comparing performance differences on PVTs between individuals with LEP and NSE in Canada. Several *a priori* hypotheses were examined:

Hypothesis 1: Overall PVT performance. It was hypothesized that the LEP group would have a higher base rate of failure on PVTs (BR_{Fail}) and a greater number of PVTs failed than the NSE group at commonly used cutoffs. Both individual instruments (e.g., TOMM and RDS) and instruments combined were compared. Additionally, scores were compared as continuous variables in addition to BR_{Fail}.

Hypothesis 2: Level of English proficiency. Because the LEP group consisted of a range of English proficiency levels, English proficiency was also examined as a continuous variable. Specifically, it was hypothesized that participants with lower levels of English proficiency on self-rated and objective language measures would have a higher BR_{Fail}.

Hypothesis 3: Level of verbal mediation of PVTs. It was hypothesized that BR_{Fail} will be greater on PVTs with high verbal mediation (PVT_{HVM}) for the LEP compared to NSE group. In contrast, it was predicted that LEP and NSE participants would have similar BR_{Fail} on PVTs with low verbal mediation (PVT_{LVM}).

Part 2: Can Current PVTs Detect Non-Credible Performance for Individuals with LEP? What Cutoffs Provide Adequate Classification Accuracy in this Population?

Regardless of whether the hypotheses in Part 1 were confirmed or rejected, the question still remains on the usefulness of PVTs in detecting non-credible performance for individuals with LEP. Indeed, cutoffs may simply need to be adjusted to produce a similar signal detection profile as NSE. However, an alternative possibility may be that no cutoffs will result in adequate sensitivity or specificity, suggesting that certain instruments cannot be used to detect non-credible performance for this population.

To this end, the second part of the dissertation focused on calculating classification accuracy on a battery of PVTs for the LEP group. The purpose of this portion was twofold: (1) to determine whether published PVT cutoffs can adequately detect non-credible performance for individuals with LEP and, (2) to determine cutoffs that provide a good balance of specificity and sensitivity for this population. To this end, Part 2 involved an experimental-malingering design and used experimental malingering as a criterion for calculating cutoffs.

Hypothesis 4: Classification accuracy as a function of level of verbal mediation. It was hypothesized that PVT_{HVM} may not be good detectors of non-credible performance for individuals with LEP whereas PVT_{LVM} would have better utility. This prediction was based on the rationale that both the experimental-malingering and non-malingering control conditions will perform poorly on PVT_{HVM} given the language demand of these tests, thus making discrimination between the conditions challenging.

CHAPTER IV: METHODS

Participants

Recruitment. Participants were prospectively recruited through the University of Windsor's Psychology Participant Pool, Centre for English Language Development (CELD), and Windsor International Student Email List (WISEL). The latter two recruitment methods were used solely to target individuals with LEP.

For the Participant Pool, two screening questions were included: (1) "Do you speak English as a second language (ESL)?" and (2) "Would you rate your English proficiency (in either speaking, understanding, or reading) as less than Very Good?". If both questions are answered YES, students viewed the posting recruiting for the LEP group. If both questions are answered NO, students viewed the posting recruiting for the NSE group. Participants were compensated 2.5 credits commensurate to 2.0 hours of in-lab participation in according to Participant Pool guidelines.

For participants recruited through CELD and WISEL, individuals received similar screening questions over email regarding their language background and English proficiency. Participants were compensated $20 for their participation. An email reminder was sent to participants signed-up for the study 48 hours prior to their time slot.

Inclusion criteria. Inclusion in the LEP group required having LEP in either speaking, understanding, or reading English, and greater proficiency in their native language than English (i.e., non-balanced bilingual). This was operationalized as a score <8/10 on at least one of the three English Language Proficiency rating scales (speaking, understanding, reading) of the Language Experience and Proficiency Questionnaire (LEAP-Q; described in the Measures section), as well as higher proficiency ratings of their native language than English.

Individuals were included in the NSE group if, in addition to answering "no" on both screening questions, they rated their English proficiency on the LEAP-Q as ≥8/10 for speaking, reading, and understanding domains.

Participants were excluded if they had a current diagnosis of a major psychiatric or neurological disorder, developmental disability, or serious medical illness that would affect cognitive functioning or their ability to engage in testing. This information was assessed via a self-report questionnaire administered prior to commencing the administration of cognitive tests. The principal investigator also corresponded with all participants and screened for noticeable psychiatric symptoms (e.g., psychotic behaviors and severe anxiety) to make a final judgement regarding inclusion in the study. Additionally, exclusion criteria for the study were described in the advertisement postings to ensure that non-eligible individuals did not sign-up for the study. A summary of inclusion and exclusion criteria is listed in Table 1.

Table 1

Inclusion & exclusion criteria

Inclusion criteria:	
General:	Age ≥ 18
LEP group:	English as second language Proficiency in speaking, understanding, OR reading English < 8/10 Proficiency ratings of their native language > English proficiency
NSE group:	English as first language Proficiency of ≥ 8/10 in speaking and understanding English
Exclusion criteria:	
Major psychiatric disorders	Current depressive or manic episode, psychosis, severe anxiety disorders
Neurological conditions	Cerebrovascular disorders, dementia, traumatic brain injury (moderate-to-severe), other neurological disorders
Developmental disabilities	Intellectual disability, autism spectrum disorder
Serious medical conditions	Cancer treated with spinal/brain radiation and chemotherapy (e.g., meningioma, acute lymphoblastic leukemia), pituitary diseases

Description of the sample. A total of 140 participants was included in the study (70 LEP, 70 NSE). The majority of participants were female (74.3%) and right-handed (90.7%). Average age was 23.7 years old (*SD* = 6.3, range = 17-59) and average years of education was 15.3 years (*SD* = 1.9). Highest parental education was used as a proxy for socioeconomic status (SES), and the majority of the sample stated that their parents completed a post-secondary degree (maternal = 57.1%; paternal = 50.7%). There were no differences between LEP and NSE groups on age, handedness, or parental-education level (Table 2). However, the LEP group had a significantly higher percentage of males (40.0%) than the NSE group (11.3%): χ^2 (1, *N* = 140) = 14.96, *p* < .01, Φ^2 = .11 (large effect). This preponderance of males in the LEP group was likely a result of differences in academic programs of between groups. Specifically, whereas the NSE group consisted of undergraduate Psychology students, the LEP group was represented by a greater diversity of academic programs, including many from STEM graduate programs, which is largely male-dominated (Vogt, Hocevar, & Hagedorn, 2007; Wang & Degol, 2017).

Furthermore, the LEP group had on average significantly greater number years of education (*M* = 15.96, *SD* = 1.88) than the NSE group (*M* = 14.67, *SD* = 1.75): *t*(138) = -4.18, *p* <.01, *d* = .71 (large effect). Again, this difference may be an artifact of divergent recruitment strategies: while NSE were recruited exclusively from the Psychology Participant Pool, which consisted of mainly first- and second-year undergraduate students, participants with LEP were additionally recruited from University of Windsor’s CELD and WISEL. Graduate students were overrepresented in this category, inflating mean education level of the LEP group.

Table 2

Demographic Background of Sample (N = 140)

	LEP (*n* = 70)		NSE (*n* = 70)			
	Control	EM	Control	EM		
	(*n* = 40)	(*n* = 30)	(*n* = 40)	(*n* = 30)	*p*[a]	*d*/ Φ^2
Age	24.2 (2.9)	24.8 (5.1)	24.1 (8.5)	21.4 (7.1)	.16	.24
Education (years)	16.3 (2.1)	15.5 (1.5)	15.0 (1.9)	14.3 (1.4)	<.01	.71
Gender (% Male)	42.5%	36.7%	15.0%	6.7%	<.01	.11
Handedness (% Right)	97.4%	86.7%	92.5%	86.7%	.69	<.01
Maternal/Paternal Education (%)					.06/.63	.03/.03
Less than high school	20.0/15.0	23.3/16.7	5.0/10.0	3.3/10.0		
High School	22.5/17.5	13.3/16.7	10.0/17.5	23.3/23.3		
College diploma	25/22.5	23.3/26.7	32.5/27.5	43.4/20.0		
Bachelor's degree	22.5/30.0	26.7/30.0	42.5/22.5	10.0/23.3		
Master's degree	7.5/10.0	10/3.3	7.5/7.5	13.3/23.3		
Doctoral degree	2.5/5.0	3.3/6.7	2.5/12.5	6.7/0.0		
Primary culture – Canadian (%)	0.0	0.0	72.5	83.3	<.01	.90
Canadian identification (0-10)	3.4 (2.2)	3.9 (2.3)	9.0 (1.5)	9.4 (.9)	<.01	3.09
Years immigrated to Canada	1.2 (1.2)	3.2 (4.7)	9 (8.4)*	13 (5.6)**	<.01	1.97

$^*n$ = 2; $^{**}n$ = 4 (only 2 and 4 participants in the NSE sample were born outside of Canada)

[a]Contrasts are between LEP and NSE groups. Contrasts with categorical variables were conducted using Chi-square test and phi effect size. Contrasts with continuous variables were conducted using *t*-test and Cohen's *d.*

Note. LEP: Limited English proficiency; NSE: Native speakers of English; EM: Experimental Malingering.

The majority of NSE participants identified Canadian as their primary culture (77.1%), with an average rating of 9.2 out of 10 with respect to Canadian identification (Table 2). In contrast, none of the LEP participants identified Canadian as their primary culture and rated their identification with this culture being on average 3.6 out of 10.

As expected, LEP and NSE groups also differed on language background. Participants with LEP spoke significantly more languages, had less exposure and preference to read and speak in English, and had poorer self-reported proficiency in speaking, understanding, and reading English than NSE participants (Table 3). LEP participants also scored lower than NSE on an objective measure of English proficiency (Boston Naming Test Short Form – see Measures). Within the LEP sample, participants reported significantly better proficiency in reading than speaking and understanding English: *t*(69) = 2.32 to 3.55, *p* =.02 to <.01, *d* = .24 to 45 (small to medium effect).

LEP participants reported being on average 10.7 years old when they started learning English. In contrast, the majority of the NSE group (91.4%) was born in Canada, and the rest moved to Canada at a significantly younger age (*M* = 11.7 years ago, *SD* = 6.12) than LEP participants, 98.6% of whom were immigrants, and on average moved to Canada within the past 2 years (*SD* = 3.29 years).

Table 3

Language Background of the Sample (N = 140)

	LEP (*n* = 70)		NSE (*n* = 70)			
	Control (*n* = 40)	EM (*n* = 30)	Control (*n* = 40)	EM (*n* = 30)	*p*	*d*
Number of fluent languages	2.6 (.75)	2.7 (.80)	1.8 (.95)	1.5 (.94)	<.01	1.13
Age learned English	9.6 (3.5)	12.1 (5.7)	0.2 (1.0)	0.0	<.01	3.14
Current exposure to English (%)	44.1 (18.7)	52.3 (21.3)	92.5 (13.0)	95.2 (10.7)	<.01	2.79
Preference reading English (%)	44.4 (29.5)	52.1 (27.8)	95.6 (10.0)	95.6 (11.0)	<.01	2.22
Preference speaking English (%)	31.9 (22.4)	37.0 (22.8)	93.1 (14.8)	92.8 (16.8)	<.01	3.05
Proficiency in English (0-10)						
Speaking	6.1 (1.3)	6.3 (1.3)	9.1 (.8)	9.3 (.7)	<.01	2.82
Understanding	6.4 (1.5)	6.5 (1.2)	9.3 (.8)	9.3 (.7)	<.01	2.57
Reading	6.8 (1.4)	6.9 (1.4)	9.1 (.8)	9.0 (.6)	<.01	2.06
BNT-15 (objective measure of English proficiency)	6.5 (3.1)	6.3 (3.4)	13.9 (1.2)	12.8 (3.1)	<.01	2.62

[a]Contrasts are between LEP and NSE groups using *t*-test.

Note. Language background collected from the Language Experience and Proficiency Questionnaire (Marian et al., 2007); LEP: Limited English proficiency; NSE: Native speakers of English; EM: Experimental Malingering; BNT-15: Boston Naming Test – Short Form Accuracy raw score.

Overall, the sample was represented by at least 23 cultures and 20 languages, with multiple dialects within some language groups (Table 4). Aside from Canadian culture and English language, the majority of the sample identified their primary culture as Chinese or Indian and spoke a Chinese (e.g., Mandarin, Cantonese) or Indian dialect (e.g., Hindi, Gujarati, Urdu).

Table 4

Primary Cultures and Languages of the Sample (N = 140)

Primary Culture Identified	% of Sample	Primary Language	% of Sample
Canadian	38.6	English	49.3
Chinese	25.7	Mandarin/Cantonese/Chinese dialect	25.0
Indian (Hindu, Sikh)	11.4	Hindi	4.3
African (Nigerian)/African Canadian	5.0	Gujarati	3.6
Arab, Syrian, or Middle Eastern	4.3	Arabic	2.9
Iranian	2.1	Telugu	2.1
Pakistani	1.4	Persian/Farsi	2.1
Ukrainian	0.7	Punjabi	1.4
Portuguese	0.7	Spanish	1.4
Mexican	0.7	Nepali	0.7
Liberian	0.7	Italian	0.7
Jamaican	0.7	Kannada	0.7
Indigenous	0.7	Urdu	0.7
Polish	0.7	Kinyarwanda	0.7
Swiss	0.7	Tamil	0.7
Lebanese	0.7	Portuguese	0.7
German	0.7	Dinka	0.7
Italian	0.7	Turkish	0.7
South Asian (not specified)	0.7	Vietnamese	0.7
Brazilian	0.7	Assyrian	0.7
Spanish	0.7		
Turkish	0.7		
Vietnamese	0.7		

Measures

English language proficiency. As the purpose of assessing English proficiency in the current research was to examine the relationship between English proficiency and PVT performance, a comprehensive assessment of participants' language history was not collected. To this end, a truncated version of a validated self-report questionnaire (consisting of only the relevant sections of the LEAP-Q) and a short performance-based measure of English proficiency (BNT 15-item) were administered to participants. Each of these measures are detailed below.

Language Experience and Proficiency Questionnaire (LEAP-Q). The LEAP-Q (Marian et al., 2007) was developed to provide a thorough assessment of an individual's language status. Unlike unstandardized "homemade" questionnaires or improvised questions that assess for global language proficiency, the LEAP-Q has been validated to show good internal and criterion-based validity (e.g., LEAP-Q items correlate with several Woodcock Johnson Test of Achievement subtests; Marian et al.). Distinct aspects of language competence (proficiency, dominance, preference, usage) are assessed in each of the languages across three domains (speaking, listening, reading). The LEAP-Q is publicly available in many languages on the research group’s website (http://www.bilingualism.northwestern.edu/leapq/).

For the current dissertation, relevant sections from the Canadian Research version were administered. Specifically, ratings regarding language dominance, proficiency, order of language acquisition, and cultural and education background were included, while some language history sections (e.g., contributors to learned languages, detailed current usage) were omitted. The truncated LEAP-Q is provided in Appendix A.

Boston Naming Test – Short Form (BNT-15). The BNT-15 is the condensed 15-item version of the full-length 60-item BNT, a measure of visual confrontation naming (Strauss, Sherman & Spreen, 2006). On this task, examinees are asked to provide the name of line drawings of objects of increasing difficulty. The short version significantly cuts down administration time but has been shown to maintain good psychometric properties (Strauss et al.). Similar to the full version, the BNT-15 short form is affected by age, education, ethnicity, and linguistic background (Strauss et al.). Although many short forms have been developed, the Mack 15-item version (Mack, Freed, Williams, & Henderson, 1992) found in the beginning of the BNT stimulus booklet was chosen for the current study.

Recent literature has examined the BNT-15 as a performance-based measure of English proficiency (Erdodi, Jongsma, & Issa, 2017). Specifically, the BNT-15 has been shown to have high sensitivity (89%) in discriminating between individuals whose dominant language is English and whose dominant language is Arabic (Erdodi et al.). As such, the BNT-15 was administered in conjunction with the LEAP-Q as an objective measure of English proficiency.

Neuropsychological & performance validity tests. The tests included, with information on estimated administration times, cognitive domains, and type (e.g., freestanding, embedded) are listed in Table 5. The battery was chosen to include a balance of freestanding and embedded, high and low verbally-mediated, and established and experimental PVTs. In addition, tests were chosen to sample across multiple cognitive domains (e.g., verbal and visual memory, executive function and attention, processing speed, visual-spatial). All tests were administered using standardized instructions. Demographically corrected norms were used in the current study for calculating T-scores for FAS, Animals, Trail Making, and Complex Ideational Material tests (Heaton, Miller, Taylor, & Grant, 2004). Other standardized scores were calculated using norms published in the manual unless otherwise indicated.

Table 5

Characteristics of Included Test Battery

PVT$_{HVM}$				**PVT$_{LVM}$**			
Test	Time	Domain	Type	Test	Time	Domain	Type
ACS WCT	4	Memory	FS	TOMM T1	5	Memory	FS
Rey WRT	5	Memory	FS	FIT	2	Memory	FS
WAIS-III Digit Span	5	Attention	Embed	DCT	5	Attention	FS
BDAE CIM	5	Language	Embed	RCFT	12	Memory	Embed
FAS, Animals	4	EF	Embed	TMT	4	EF	Embed
Emotion Fluency	2	EF	Embed	WAIS-III Coding	2.5	PS	Embed
D-KEFS Stroop 1-3	6	EF/PS	Embed	WAIS-IV SS	2.5	PS	Embed
WRAT-4 Reading	2	Reading	–	Clock drawing	1	Visual-spatial	–
Other measures							
LEAP-Q Abbrev	4	Language	Quest	V-5	2	Mood	Quest
BNT-15	3	Language	–	GAD-7	2	Mood	Quest
				PHQ-9	2	Mood	Quest
Testing time:	40				40		

Note. PVT$_{HVM}$: Performance validity tests with high verbal mediation; PVT$_{LVM}$: Performance validity tests with low verbal mediation; Time: Estimated administration time (minutes); FS: Freestanding test; Embed: Embedded validity indicator; Quest: Questionnaire; EF: Executive Function; PS: Processing Speed; ACS WCT: Advanced Clinical Solutions Word Choice Test (Wechsler, 2009); Rey WRT: Rey Word Recognition Test (Greiffenstein et al., 1994); WAIS-III Digit Span: Wechsler Adult Intelligence Scale Third-Edition Digit Span Subtest (Wechsler, 1997); BDAE CIM: Boston Diagnostic Aphasia Examination – Complex Ideational Material (Goodglass et al., 2001); FAS & Animals: Controlled Oral Word Association (Benton & Hamsher, 1978); Emotion Fluency: Emotion Word Fluency Test (Abeare et al., 2017); D-KEFS Stroop 1-3: Delis-Kaplan Executive Function System Color-Word Interference Test Conditions 1 to 3 (Delis et al., 2001); WRAT-4 Reading: Wide Range Achievement Test Fourth-Edition Reading Subtest (Wilkinson & Robertson, 2006); BNT-15: Boston Naming Test – Short Form (Strauss et al., 2006); LEAP-Q Abbrev: Language Experience and Proficiency Questionnaire Abbreviated version (Marian et al., 2007); TOMM T1: Test of Memory Malingering Trial 1 (Tombaugh, 1996); FIT: Rey 15-Item Test (Rey, 1964); DCT: Dot Counting Test (Boone et al., 2002b); RCFT: Rey-Osterrieth Complex Figure Test (Meyers & Meyers, 1995); TMT: Trail Making Test (Reitan, 1992); WAIS-III Coding: Wechsler Adult Intelligence Scale Third-Edition Digit Symbol Subtest (Wechsler, 1997); WAIS-IV SS: Wechsler Adult Intelligence Scale Fourth-Edition Symbol Search Subtest (Wechsler, 2008); Clock Drawing: Clock Drawing Test (Strauss et al., 2006), V-5: Visual Analog Scale; GAD-7: General Anxiety Disorder 7-Item Scale (Spitzer, Kroenke, Williams, & Löwe, 2006); PHQ-9: Patient Health Questionnaire 9-Item Scale (Kroenke, Spitzer, & Williams, 2001).

Two sets of published cutoffs were chosen: conservative cutoffs to optimize specificity and liberal cutoffs to optimize sensitivity. Specifically, conservative cutoffs were chosen to maintain a specificity of ≥.90 to minimize false-positive errors, while liberal cutoffs were chosen for the highest sensitivity while still maintaining an acceptable specificity of ≥.85. Table 6 lists the cutoffs that were used to determine BR$_{Fail}$ for the current study.

Table 6

PVT Liberal and Conservative Cutoffs

PVT$_{HVM}$			**PVT$_{LVM}$**		
Test	Liberal	Conservative	Test	Liberal	Conservative
WCT Accuracy	≤47	≤43	RCFT Copy	≤26	≤23
WCT Time	≥156	≥171	RCFT IR	≤10	≤9.5
RDS	≤7	≤6	RCFT Recognition	≤16	≤15
DS-ACSS	≤6	≤5	RCFT Equation I	≤47	≤45
WRT Recognition	≤7	≤5	TOMM T1	≤44	≤39
WRT Combination	≤10	≤8	DCT E-Score	≥15	≥17
FAS T-score	≤33	≤31	FIT Combined Score	<21	<20
Animals T-score	≤33	≤31	TMT-A Time	≤39	≤34
Verbal Fluency LRE	≥.45	≥.475	TMT-B Time	≤37	≤30
Letter Fluency LRE	≥.5	≥.6	TMT A + B	≥137	≥170
CIM Raw Score	≤9	≤8	Digit Symbol ACSS	≤5	≤4
CIM T-Score	≤29	≤23	Symbol Search ACSS	≤6	≤5
Stroop Condition 1	≤7	≤5			
Stroop Condition 2	≤7	≤5			
Stroop Condition 3	≤7	≤5			

Note. PVT: Performance validity test; PVT$_{HVM}$: Performance validity tests with high verbal mediation; PVT$_{LVM}$: Performance validity tests with low verbal mediation; Liberal: Cutoffs optimized for sensitivity (i.e., chosen for the highest sensitivity while maintaining specificity of ≥.85); Conservative: Cutoffs optimized for specificity (i.e., chosen to maintain a specificity of ≥.90 to minimize false-positive errors); WCT: Word Choice Test (Barhon et al., 2015; Davis, 2014; Erdodi et al., 2016); RDS: Reliable Digit Span (Greiffenstein et al., 1994; Schroeder et al., 2012); DS-ACSS: Digit Span Age-Corrected Scaled Score (Axelrod et al., 2006; Babikian et al., 2006; Jasinski et al., 2011; Spencer et al., 2013; Young et al., 2012); WRT Recognition: Word Recognition Test Recognition score (Bell-Sprinkel et al. 2013; Greiffenstein et al.; Nitch et al., 2006); WRT Combination Score: WRT Recognition – number of false positives + WRT Recognition hits from first 8 words (Nitch et al.); FAS T-score: Letter fluency test demographically corrected T-score (Curtis et al., 2008; Sugarman & Axelrod, 2015); Animals T-score: Category animal fluency test demographically corrected T-score (Sugarman & Axelrod); Verbal Fluency LRE: Logistical regression equation combining FAS and Animal Fluency T-scores (Sugarman & Axelrod); Letter Fluency LRE: Logistical regression equation combining overall letter fluency output and pattern of performance (Johnson et al., 2012); BDAE CIM: Boston Diagnostic Aphasia Examination – Complex Ideational Material (Erdodi & Roth, 2017; Erdodi et al., 2016); Stroop Conditions 1-3: Delis-Kaplan Executive Function System Color Word Interference Test Conditions 1 to 3 (Laszlo, Sagar, et al., 2018); RCFT: Rey-Osterrieth Complex Figure Test; Copy: Copy Trial raw score; IR: Immediate Recall raw score; RT: Recognition Trial raw score (Reedy et al., 2013; Sugarman et al., 2016; Whiteside et al., 2011); Equation: CT raw score + (true positive recognition – Atypical recognition errors) x 3; Lu et al., 2003)]; TOMM T1: Test of Memory Malingering Trial 1 (Jones, 2013; Denning, 2011; Greve et al., 2006); DCT E-Score: Dot Counting Test Effort-Score (Boone et al., 2002a); FIT Combined Score: Rey 15-Item Test recall + recognition combination score (free recall + [recognition hits – false positives]; Boone et al., 2002c); TMT-A, TMT-B Time: Trail Making Test Part A and Part B Time score (Busse & Whiteside, 2012; Iverson et al., 2002); TMT A + B: Trail Making Test Trial A & B Total Combined Time Score (Busse & Whiteside, 2012; Shura et al., 2016); TMT B/A: Trail Making Test Part B Time score/Part A Time score (Iverson et al.; Ruffolo et al., 2000; van Gorp et al., 1999); Digit Symbol ACSS: Digit Symbol age-corrected scaled score (Etherton et al., 2006; Kim et al., 2010); Symbol Search ACSS: Symbol Search age-corrected scale score (Erdodi, Abeare, et al., 2017).

The Test of Memory Malingering (TOMM). The TOMM (Tombaugh, 1996) is a 50-item forced-choice test of visual-recognition memory and is the most widely used PVT amongst neuropsychologists (Martin et al., 2015). The task consists of two learning trials and a Retention Trial. After presentation of 50 pictures during the learning trials, examinees are shown 50 two-choice recognition panels one at a time consisting of a previously presented picture (target) and a picture not previously seen (foil). Examinees are asked to discriminate between the target and foil items and are given immediate feedback regarding their response after each target-foil pair.

Tombaugh (1996) suggested using a cutoff of <45 on Trial 2. However, subsequent research suggested that using this cutoff is too conservative and results in poor sensitivity at acceptable specificity levels (e.g., within a mild TBI sample; Greve, Bianchini, & Doane, 2006). Other research comparing the TOMM to other widely used PVTs, such as the WMT (Green, 2003), have also suggested that the TOMM is comparably less sensitive (Gervais et al., 2004). Given these shortcomings, subsequent researchers have proposed alternate cutoffs for the TOMM that produce higher sensitivity at adequate specificity levels in various populations (e.g., active military duty in an outpatient clinic, Jones; 2013; mild TBI in a private practice for medicolegal purposes, Stenclik, Miele, Silk-Eglit, Lynch, & McCaffrey, 2013; veterans in a VA hospital outpatient clinic; Kulas, Axelrod, & Rinaldi, 2014). These include a cutoff of Trial 2 ≤49 (.96 specificity, .86 sensitivity; Jones, 2013), Trial 2 ≤48 (.92 specificity, .75 sensitivity; Stenclik et al., 2013), Trial 1 ≤44 (.93 specificity, .86 sensitivity, Jones, 2013), Trial 1 ≤42 (.91 specificity, .66 sensitivity; Greve et al., 2006), and Trial 1 ≤40 (.94 specificity, .72 sensitivity; Denning, 2012). The current study used these more recently published cutoffs for Trial 1.

The Dot Counting Test (DCT). The DCT (Boone, Lu & Herzberg, 2002b) uses a non-forced choice format and simply involves the presentation of grouped and ungrouped dots on a set of 12 stimulus cards which examinees are asked to count as quickly as possible. As counting is typically well preserved in most patients with brain injury, this task does not tap into real cognitive impairment but provides an estimate of the examinee’s effort levels. Scores on this task takes into account both speed (response latency) and accuracy (number of errors), which are amalgamated into a total E-score (mean ungrouped dot counting time + mean grouped dot counting time + number of errors). The DCT has been found to have good specificity and sensitivity across various populations (Boone et al., 2002a). An E-score ≤17 produced the best classification accuracy in a mixed clinical population (excluding dementia), with good specificity (.91) and sensitivity (non-forensic: .76; forensic: 1.00).

Rey 15-Item Test (FIT). The FIT (Rey, 1964) is a brief task of short-term visual memory. It is one of the oldest and most widely used PVTs. The task involves presentation of 15 meaningful symbols on a stimulus page for 10 seconds followed by free recall of the items by asking the examinee to draw all the stimuli remembered. Memorizing the 15 items, while seemingly challenging, is actually quite easy as items are presented in a 3 X 5 matrix with each of the 5 rows being automatically chunked.

Although studies have suggested that the recall score is highly under-powered, scores that incorporate both the recall and recognition trial seem to dramatically improve sensitivity (Boone, Salazar, Lu, Warner-Chacon, & Razani, 2002c). Specifically, the combined recall and recognition score (i.e., free recall + [recognition hits – false positives]) provides better sensitivity (.71) and specificity (>.92) using a cutoff of <20 than using the recall score alone, which has good specificity (.90-1.00) but lower sensitivity (.47) at a cutoff of <9 (Boone et al., 2002c).

Rey Word Recognition Test (WRT). The WRT, also developed by Rey, is a freestanding PVT measuring verbal-recognition memory (Boone, 2007). This task involves presentation of 15-unrelated words (presented orally) followed by immediate recognition of the words from a list of 30 words (15 targets, 15 foils). In the standard administration, participants are provided the list of all 30 words at once, although modified versions (e.g., reading the recognition list aloud; Greiffenstein et al., 1994) have been reported in the literature.

The WRT has been found to good signal-detection properties in various studies. In the earliest investigations, a cutoff of ≤5 identified 88% of the non-credible group and 59% of the credible group in a post-concussive sample (Greiffenstein et al., 1994), while a cutoff of ≤6 identified 93% of the non-credible group and 80% in the credible post-concussive group (Greiffenstein, Baker, & Gola, 1996). Subsequent research found that this cutoff can be raised to ≤7 for women while maintaining good sensitivity (.81) and specificity (≥.90), although a cutoff of ≤5 was required to maintain similar signal-detection properties for men (Nitch, Boone, Wen, Arnold, & Alfano, 2006). This gender difference has also been found most recently in a mild TBI sample, such that sensitivity was higher in detecting non-credible female participants with mild TBI compared to their male counterparts (.68 versus .48 at cutoff ≤6; Bell-Sprinkel et al., 2013). This gender difference on the WRT has been hypothesized to be attributed to performance differences on verbal-based tasks, with women outperforming men (Boone, 2007).

A combination score for the WRT has also been created. The combination score is based on the finding that credible examinees have better performance on the first half of the list and double-weighs recognition words from the first half of the list (recognition hits – FP + recognition hits from first 8 words; Nitch et al., 2006). Using this score has produced comparable or better sensitivity at .88 specificity (Women: cutoff: ≤11, .85 sensitivity, Men: cutoff: ≤8, .75 sensitivity) in a heterogeneous sample compared to using the recognition score alone. Subsequent validation of this equation also reported adequate classification accuracy (cutoff: ≤8, .47 sensitivity, .92 specificity; Bell-Sprinkel et al., 2013). Both recognition and combination scores were used in the current study.

Word Choice Test (WCT). The WCT from the Advanced Clinical Solutions (ACS; Wechsler, 2009), an add-on to the WAIS-IV, is a PVT that uses a 50-item dichotomous forced-choice paradigm. On this task, examinees are presented with a series of 50 words both visually on a stimulus book and orally by the examiner. Examinees are asked to state whether the word is “natural” or “man-made” to ensure adequate attention to the material. Following the learning phase, examinees complete a recognition task consisting of 50 target-foil pairs.

Compared to a similar forced-choice recognition memory test (Warrington Recognition Memory Test – Word Trial [RMT-W]; Warrington, 1984), the WCT has been found to be superior in detecting performance invalidity (Davis, 2014; Erdodi et al., 2014). Using the ACS manual suggested cutoff (≤43), the WCT has been found to have a low sensitivity (.38-.41) and high specificity (.84-.96) in the literature (Bashem et al., 2014; Davis, 2014; Erdodi et al., 2014). In contrast, using more liberal cutoffs (≤47) have been found to produce better sensitivity while maintaining specificity (e.g., Davis: .87 specificity, .75 sensitivity; Erdodi et al., 2014: .84 specificity, .54 sensitivity; Erdodi et al., 2016: .87 specificity, .57 sensitivity).

In addition to the accuracy cutoff, a time-to-completion (T2C) score has also been proposed as an embedded indicator. Specifically, Erdodi, Tyson, and colleagues (2017) found that using a completion T2C cutoff of ≥171 seconds produced good sensitivity (.49) and specificity (.91) and identified 6-10% additional invalid cases above using only the accuracy score. Most recently, critical items have been explored for their utility to increase the overall classification on the WCT. Several critical items on the WCT have been identified and aggregates of these items have been shown to produce superior signal detection properties over recognition scores alone (Erdodi, Tyson, et al., 2018).

Trail Making Test (TMT). The TMT, a measure of attention, processing speed, and executive functioning, is a widely used test first originating as part of the Army Individual Test Battery and later adapted into the Halstead-Reitan Battery (Reitan, 1992; Strauss et al., 2006). The test involves two parts. In TMT Part A (TMT-A), examinees are asked to draw a line connecting numbers in order on a sheet of paper as quickly as possible. In TMT Part B (TMT-B), examinees are asked to alternate between connecting numbers and letters as quickly as possible, thus measuring both mental flexibility and psychomotor speed. Several scores can be derived from this test, including time and error scores and a difference ratio score (TMT B/A).

Age, education, and IQ have been found to affect test scores, with lower education and IQ and increasing age associated with poorer performance (Strauss et al., 2006). Some research has found that cultural and linguistic variables affect performance on this test (Strauss et al.). Being one of the most commonly used neuropsychological tests (Rabin, Barr & Burton, 2005), the literature is rich with support for its reliability, validity, and sensitivity to detect brain injury (Strauss et al.). Because it taps into several cognitive domains (e.g., attention, processing speed, executive functioning), it is a generalized test of brain integrity, and poor performance on the TMT may signal impairment in any one of these domains.

In addition to its long history as a neuropsychological test, the TMT has more recently been examined as an embedded PVT. However, the literature on using TMT scores as validity indicators have generally found that this measure is not very sensitive in detecting non-credible performance, especially in individuals with moderate-to-severe cognitive impairment (Boone, 2007). One study found that, although TMT-A and TMT-B completion times were significantly longer in the non-credible group compared to a credible head-injury group, sensitivity to detect the non-credible group was very poor across all TBI severities when specificity was at adequate levels (.02-.19 using the following cutoffs: TMT-A ≥63, TMT-B ≥200, TMT B/A ≤1.49; Iverson, Lang, Green & Franzen, 2002).

Research on the TMT B/A ratio as an embedded validity indicator has also produced mixed results. While earlier studies showed that real-world and experimental malingerers showed larger discrepancies between TMT-A and TMT-B times (Ruffolo, Guilmette & Willis, 2000; van Gorp et al., 1999), other studies find no difference in the B/A ratio between non-credible and credible head injury groups (Iverson et al., 2002; O’Bryant, Hilsabeck, Fisher, & McCaffrey, 2003). Similarly, examination of the TMT A+B combination score as an embedded indicator has produced mixed results. A conservative cutoff of ≥170 produced sensitivity ranging from .11 (Shura, Miskey, Rowland, Yoash-Gantz, & Denning, 2016) to .48 (Busse & Whiteside, 2012), while a more liberal a cutoff of ≥137 produced a sensitivity of .21 (Shura et al.) when specificity was adequate. These findings seem sensible given that, as mentioned, the TMT is a sensitive test of global brain injury. Thus, the TMT, like many embedded indicators, should be interpreted in conjunction with other PVTs for any decisions regarding performance validity.

Rey-Osterrieth Complex Figure Test (RCFT). The RCFT is a commonly used test of visual-spatial construction ability, planning and organization, and visual memory (Strauss et al., 2006) The test consists of a copy trial (CT; copying a two-dimensional picture of a complex figure), immediate recall (IR; drawing the figure 3-minutes after copying), delayed recall (DR; drawing the figure after a 30-minute delay), and recognition trial (RT; identifying 12 target components from 12 foils). While multiple versions and norms are available, the scoring and norms from the manual will be used, which are stratified by age (Meyers & Meyers, 1995).

In terms of its use as a PVT, several indicators can be derived. The CT and RT score have shown the best classification accuracy (Blaskewitz, Merten & Brockhaus, 2009; Sugarman, Holcomb, Axelrod, Meyers & Liethen, 2016; Whiteside, Wald, & Busse, 2011), although some studies find that the CT raw score produced low sensitivity (Lu, Boone, Cozolino & Mitchell, 2003). In one of the earliest investigations of the RCFT RT, Meyers and Volbrecht (1999) found that their litigating sample showed a particular profile of atypical responses across the trial ("memory error patterns"), which differed from non-litigants. However, sensitivity using solely the memory error patterns was low (28%) and subsequent research confirmed its low sensitivity (Lu et al., 2003). While the IR and DR raw scores have also been found to have utility in detecting non-credible performance in some studies (e.g., IR cutoff: <10, .88 specificity, .45 sensitivity;

Reedy et al., 2013), other studies find that credible and non-credible groups are not adequately classified based on IR and DR scores (Blaskewitz et al., 2009) or the RCFT equation (Sugarman et al., 2016).

Research has found that a combination of scores from the CT and RT produce higher sensitivity at acceptable specificity levels than using either alone (Lu et al., 2003; Reedy et al., 2013; Sugarman et al., 2016). Specifically, using the combination score equation CT score + (true-positive recognition – atypical-recognition errors) x 3 and a cutoff of ≤47, Lu and colleagues were able to identify 91% of the non-malingering clinical group and 76% of the suspect effort group. The atypical-recognition errors are false-positive responses of incorrectly selecting certain items (1, 4, 6, 10, 11, 16, 18, 21) that are vastly different from the actual components of the figure. Research has shown that even brain injury patients rarely endorse these items (Lu et al.). Subsequent cross-validation of this equation corroborated its superior signal detection properties compared to using only CT or RT scores between credible and non-credible patients (cutoff ≤50: .90 specificity, .80 sensitivity; Reedy et al.).

However, this equation was based on an atypical administration of the RCFT comprising of the copy trial, 3-minute recall, and recognition trial immediately following the 3-minute recall, with no 30-minute delay recall trial. Thus, applicability to the standard administration that includes the 30-minute delay recall is unclear in these two studies. A subsequent study using the standard administration format found that while sensitivity was slightly lower than previously reported, the RCFT equation was still useful in differentiating between clinical patients and litigants (cutoff ≤45: .95 specificity, .52 sensitivity; Blaskewitz et al., 2009). Additionally, a recent study using standard administration format and a different multivariate model that aggregated CT and RT scores produced high specificity (.91) and adequate sensitivity (.55) using a cutoff of >.425, and moderate specificity (.86) and good sensitivity (.64) using a cutoff of >.35 in a large veteran sample (Sugarman et al., 2016).

WAIS-III Digit Span. The Digit Span subtest is a measure of attention and working memory. In the WAIS-R and WAIS-III version, the test involves repeating sequences of progressively longer digit strings forward and in reverse order (Strauss et al., 2006). This version is the most widely used and heavily researched, and thus included in the present study. A newer Digit Span subtest of the WAIS-IV with the addition of a number sequencing component has been subsequently shown to also maintain good signal detection properties (Reese, Suhr, & Riddle, 2012; Spencer et al., 2013; Young, Sawyer, Roper, & Baughman, 2012).

The Digit Span test contains several embedded validity indicators. The most widely used is the RDS (Greiffenstein et al., 1994), which consists of summing the longest forward and backward digit sequences of the trials where both items are completed successfully. Indeed, recent meta-analyses found over 20 (Jasinski, Berry, Shanera, & Clark, 2011) and 35 studies (Schroeder, Twumasi-Ankrah, Baade, & Marshall, 2012) on the RDS over the past few decades. Across studies, the RDS has been shown to successfully discriminate between credible and non-credible performance (Jasinski et al.; Schroeder et al.). A cutoff of ≤7 or ≤6 is most frequently used (Schroeder et al.). In a large meta-analysis, a cutoff of ≤7 produced overall specificity rates of .82-.85 across clinical groups, which is lower than the gold standard .90 specificity clearance, although sensitivity is adequate (.48-.58; Schroeder et al.). In contrast, using a cutoff of

≤6 produced high overall specificity across clinical groups (.96-.97) but unacceptably low sensitivity (.30-.35). Hence, lowering the cutoff from ≤7 to ≤6, while boosting specificity, results in a large decrease in sensitivity. It is important to remember that these signal-detection properties are sample-specific. For example, in examining the RDS in samples of TBI and chronic pain patients, excellent specificity (.92-.93) and sensitivity (.60-.67) were obtained at a cutoff of ≤7 (Etherton, Bianchini, Greve, & Heinly, 2005; Mathias, Greve, Bianchini, Houston, & Crouch, 2002). In contrast, specificity has been found to be lower than .90 even when a cutoff of ≤6 was used with individuals with severe memory impairment, LEP, low education attainment, and low IQ scores (Schroeder et al.).

The DS-ACSS is another score that has been used as an embedded validity indicator. The DS-ACSS has been found to have comparable to the RDS such that both produce large effect sizes (*d* = 1.08-1.34) in detecting non-credible performance and have similar signal-detection properties (Jasinski et al., 2011; Spencer et al., 2013). A cutoff on the DS-ACSS of ≤6 and ≤5 has been found to have adequate specificity but, like many other embedded indicators, suffers in sensitivity when used by itself (Axelrod, Fichtenberg, Millis, & Wertheimer, 2006; Babikian, Boone, Lu, & Arnold, 2006; Spencer et al.; Young et al., 2012).

Verbal Fluency. Verbal fluency tests typically consist of phonemic fluency (also called the Controlled Oral Word Association – COWAT; Benton & Hamsher, 1978) and semantic fluency. Both fluency tasks require the examinee to orally state as many words as possible in 60 seconds that either begin with a certain letter (phonemic fluency) or that belong to a certain category (semantic fluency). While many versions exist, the letters

FAS and category of animals are most commonly used for phonemic and semantic fluency, respectively (Strauss et al., 2006).

Research results on verbal fluency indicators to discriminate between individuals with credible and non-credible performance have been mixed. While some studies found good signal-detection properties using FAS scores in a mild TBI sample (Backhaus, Fichtenberg, & Hanks, 2004; Curtis, Thompson, Greve, & Bianchini, 2008), other research found that, at acceptable specificity rates, FAS and Animal fluency produced extremely low sensitivity in a moderate-severe TBI samples (Curtis et al., 2008: FAS: .15; Whiteside et al., 2015: FAS: .09, Animals: .25). This is not surprising, given that verbal fluency measures are sensitive to actual cognitive impairment (Strauss et al., 2006). Hence, FAS and Animal fluency may only be useful to detect non-credible performance in cases where there is an absence of neurological dysfunction.

Recent research has also examined the utility of equations combining phonemic and semantic fluency scores. Silverberg, Hanks, Buchanan, Fichtenberg, and Millis (2008) found that an equation using scores from an extended version (CFLJW) produced good classification accuracy. Additionally, using Bayesian Model Averaging, Johnson, Silverberg, Millis, and Hanks (2012) found that an equation comprising of CFL total score and a measure of the pattern-of-performance over time produced good signal-detection properties in an outpatient mixed neurological sample. However, both of these models used the CFL version and did not examine models with semantic fluency scores.

Most recently, Sugarman and Axelrod (2015) found that using a logistic regression equation (LRE) combining FAS and Animal scores resulted in good signal-detection properties in a veteran hospital outpatient sample (cutoff ≥.475: .91 specificity,

.46 sensitivity), which outperformed using FAS (cutoff <30: .90 specificity, .30 sensitivity) and Animal T-scores (cutoff <33: .91 specificity, .42 sensitivity) individually. As this model incorporating both FAS and Animals is most applicable to the current study, the Sugarman and Axelrod LRE was included along with the T-scores. The Johnson and colleagues (2012) LRE was also included as it added a unique component of pattern-of-performance over time. Although this equation was based on a different letter fluency task (CFL instead of FAS), the two versions have been found to be highly comparable (Lacy et al., 1996).

Aside from the established phonemic and semantic fluency tasks, an Emotion Word Fluency Test has been recently developed and shown to have good construct validity and reliability (Abeare, Freund, Kaploun, McAuley, & Dumitrescu, 2017). Parallel to other verbal fluency tasks, this version involves naming as many emotions as possible in one minute. The Emotion Word Fluency Test was included in the current study for exploratory purposes.

Boston Diagnostic Aphasia Examination – Complex Ideational Material (BDAE-CIM). The CIM is a subtest of the BDAE (Goodglass, Kaplan, & Barresi, 2001) that assesses auditory language comprehension abilities. Examinees are required to respond yes/no to questions that vary from simple factual statements (e.g., “Is a hammer good for cutting wood?) to answering more syntactically and semantically complex questions about short stories read to the examinee. Because of its simple forced-choice format, the CIM has recently been examined as a PVT. Specifically, Erdodi and Roth (2017) and Erdodi, Tyson, and colleagues (2016) found that in a mixed neurological and psychiatric sample (excluding patients with aphasia), a raw score cutoff of ≤8 and ≤9 and

T-score cutoff of ≤23 and ≤29 best detected invalid performance on the CIM when compared against other established PVTs. At these cutoffs, the CIM was more likely to identify invalid performance than receptive language deficits. Thus, the CIM has promising signal-detection properties in individuals without aphasia.

The CIM has also been examined in an LEP sample and preliminary results showed that this instrument was sensitive to English proficiency (Erdodi, Jongsma, et al., 2017). Hence, although the CIM has been found to have good classification accuracy as a PVT in a general clinical population, its ability to distinguish between credible and non-credible performance in individuals who have LEP is unclear.

WAIS-III/IV Processing Speed subtests. Two subtests make up the Processing Speed Index (PSI): Symbol Search and Coding. In the WAIS-IV Symbol Search subtest, examinees are asked to visually scan pages for matching symbols as quickly as possible (Strauss et al., 2006). The WAIS-III Digit Symbol subtest (now the WAIS-IV Coding subtest) consists of transcribing digit-symbols as quickly as possible. There is a time limit of 2-minutes on both tasks.

Aside from serving as useful measures of graphomotor processing speed, the two subtests also show promise as embedded validity indicators. Research has found that the PSI is able to discriminate between credible and non-credible groups with mild TBI (Curtis, Greve, & Bianchini, 2009) and clinical pain samples (Etherton, Bianchini, Heinly, & Greve, 2006), with the optimal cutoffs ranging between PSI ≤70 and ≤75.

Similarly, the Digit Symbol subtest has also shown promising signal detection properties. In a clinical pain sample, the Digit Symbol ACSS had the best classification accuracy at a cutoff of ≤4 (.66 sensitivity, .96 specificity) and ≤5 (.81 sensitivity, .87

specificity; Etherton et al., 2006). However, in a mixed clinical group, the Digit Symbol ACSS was found to have lower sensitivity (.18 and .40 respectively) at adequate specificity levels (Kim et al., 2010). The PSI and Digit Symbol were found to have poor signal-detection properties for some populations, namely individuals with moderate-severe TBI, cerebrovascular accidents and genuine memory impairment (Curtis et al.; Etherton et al.). This is not surprising, given that a dose-response relationship exists between injury severity and scores on PSI subtests (Curtis et al.).

The PSI subtests from the most recent version (i.e., WAIS-IV Coding) have been examined in a mixed clinical sample (excluding moderate-severe TBI) and results corroborated findings from previous studies (Erdodi, Abeare, et al., 2017). Specifically, the PSI (cutoff ≤79: .92-98 specificity, .23-56 sensitivity) and Symbol Search subtest (cutoff ≤6: .88-93 specificity, .38-64 sensitivity) have good classification accuracy when compared against combinations of other established PVTs. The Coding subtest and a Coding-Symbol Search ratio and difference score also produced good specificity but low sensitivity, while a composite based on these five indices had a good balance of specificity and sensitivity at a cutoff of ≥3 (.89-.94 specificity, .23-.53 sensitivity).

A recognition trial for the WAIS-III Digit Symbol has also been developed for the purposes of assessing performance validity. This incidental recognition memory task, which is administered immediately after the main test, requires examinees to discriminate target symbols from three foils for each of the nine symbols (Kim et al., 2010). Recognition raw scores were found to produce higher sensitivity (.59) at a cutoff of ≤5 compared to Digit Symbol ACSS and raw scores (Kim et al.).

Wide Range Achievement Test Fourth Edition Reading Subtest (WRAT-4 Reading). The WRAT-4 Reading is an achievement test of reading ability (Wilkinson & Robertson, 2006). This subtest consists of orally reading a list of 55 words of increasing difficulty. Few studies have examined the effects of performance validity on reading tests or have examined reading tests as a measure of performance validity. It has been previously assumed by some researchers that "hold tests" such as WRAT-4 Reading, which are relatively insensitive to brain injury, are also unaffected by suboptimal effort. However, a few studies have shown that reading scores are indeed lower in non-credible groups than credible groups on the WRAT-4 (Sawyer, Yong, Roper & Rach, 2014), Test of Premorbid Functioning (TOPF; Martin et al., 2018), and North American Adult Reading Test (NAART; Davis, McHugh, Axelrod, & Hanks, 2012). The exception is a study comparing performance on the Wechsler Test of Adult Reading (WTAR) in individuals passing and failing the TOMM, in which no differences were found (Whitney, Shepard, Mariner, Mossbarger, & Herman, 2010). Thus, performance on reading tests cannot be assumed to be immune to performance invalidity.

Clock Drawing Test. The Clock Drawing test is a measure of visual-spatial-construction ability, although it is commonly seen as a quick "bedside" measure of cognitive functioning given its sensitivity to global cognitive deficits (Strauss et al., 2006). Examinees are asked to produce a freehand drawing of the face of a clock with its numbers and hands set to a specific time. Some versions also include trials with a pre-drawn circle and copying for individuals with more severe impairment to differentiate the underlying difficulties. Similar to the Digit Symbol subtest, performance on the Clock Drawing Test has been found to be minimally affected by LEP (Erdodi, Jongsma, et al., 2017). For the current study, the free drawing trial was administered and the Rouleau, Salmon, Butters, Kennedy, and McGuire (1992) qualitative scoring system was used.

Delis-Kaplan Executive Function System Color-Word Interference Test (D-KEFS Stroop). The D-KEFS Stroop (Delis, Kaplan, & Kramer, 2001) is a measure of cognitive flexibility, inhibition, and selective attention and involves naming color names printed in a different colored ink (Strauss et al., 2006). The entire task consists of four conditions. The first two conditions (reading and color naming) measure oral processing speed as these conditions simply require word reading and color naming as quickly as possible. These two conditions provide a baseline to compare the more challenging inhibition and switching demands of Conditions 3 and 4. Condition 3 consists of incongruent color-word stimuli (e.g., the word "blue" printed in green ink) and requires examinees to inhibit their dominant response of word reading in the face of incongruent ink-color stimuli. Condition 4 further engages cognitive flexibility by requiring switching between the automatic word reading task and naming the incongruent colors.

The D-KEFS Stroop was recently examined as a measure of performance validity (Erdodi, Sagar, et al., 2018). Although Conditions 3 and 4 are cognitively demanding and sensitive to neurological impairment, Conditions 1 and 2 are simple tasks and have potential utility as PVTs. In their mixed clinical sample, a cutoff ≤6 on any of the 4 conditions produced adequate classification accuracy against criterion measures (.87–.94 specificity, .34 –.71 sensitivity), and a multivariate model aggregating indicators produced even better classification. The current study included Conditions 1 to 3 to further explore this instrument as a PVT.

Other questionnaires. A brief demographic questionnaire to collect relevant demographic information (e.g., age, gender, SES) was administered (Appendix B). Three mood screening questionnaires (Visual Analog Scale, Patient Health Questionnaire 9-Item Scale, Generalized Anxiety Disorder 7-Item Scale) were also administered as part of the battery, although are not central to the main hypotheses.

Procedures

Testing was conducted in a quiet, distraction-free environment. The primary investigator (PI) explained the testing process, risk and benefits of participation, and compensation to participants and obtain their consent to participate. All participants were informed that this study investigates their cognitive functioning on a variety of neuropsychological tests, and no information about the hypotheses was revealed. After consent was obtained, participants were asked to complete a battery of neuropsychological tests administered by a trained undergraduate research assistant (RA).

The four RAs received extensive training to ensure proper adherence to standardized instructions and study protocols. Testing sessions were audio-recorded. Both the audio-recordings and scoring of the RAs were regularly reviewed by the PI, and feedback was consistently provided to the RAs. The RAs were aware of the general topic of the research (e.g., LEP and PVTs) but were blinded to the study hypotheses so not to introduce bias or testing demands when administering tests.

Order of tests. The questionnaires were administered first by the PI to confirm whether participants met eligibility for the study. The order of the remaining measures was counterbalanced across participants to control for fatigue and order effects. Tests were administered in one of the following two orders: (1) TOMM, WRT, RCFT CT, BNT-15, RCFT IR, Verbal Fluency (FAS, Animals, Emotional Fluency), TMT, Digit Span, DCT, Clock Drawing, CIM, Digit Symbol, Symbol Search, WRAT-4 Reading, RCFT DR + Rec, D-KEFS Stroop, FIT, WCT or (2) WCT, FIT, D-KEFS Stroop, RCFT CT, BNT-15, RCFT IR, WRAT-4 Reading, Symbol Search, Digit Symbol, CIM, Clock Drawing, DCT, Digit Span, TMT, Verbal Fluency, RCFT DR + Rec, WRT, TOMM. Finally, a brief post-session questionnaire (described below) was administered. Table 7 details the complete study protocol.

Table 7

Description of the study protocol

Examiner	Task	Description	Time
PI	1. Consent	The PI completed the consent process with the participant.	10
	2. Questionnaires	The PI administered the demographic questionnaire, LEAP-Q, V-5, GAD-7, and PHQ-9 to the participant.	15
	3. Experimental condition instructions	The PI provided written instructions corresponding to the experimental condition of the participant. Oral explanation was provided to clarify the malingering task when necessary.	10
	4. Pre-session manipulation check	The PI administered a multiple-choice questionnaire to the participant to assess for comprehension of condition instructions.	1
RA	5. Cognitive testing	An RA administered the neuropsychological test battery to the participant.	80
	6. Post-session questionnaire	An RA administered a post-session questionnaire to assess compliance of condition instructions.	2
PI	7. Compensation & debrief	The PI answered questions of the participant and delivered Participant Pool points or monetary compensation.	2
Total Session Time:			120

Note. Time: Administration time in minutes; PI: Primary investigator; RA: Research assistant; LEAP-Q: Language Experience and Proficiency Questionnaire; V-5: Visual Analog Scale; GAD-7: General Anxiety Disorder 7-Item Scale; PHQ-9: Patient Health Questionnaire 9-Item Scale.

Experimental Malingering & Non-Malingering Control Conditions.

Participants were randomized into one of two conditions: Experimental Malingering (EM) or Non-Malingering Control (NC). Participants in the NC condition received instructions to put forth their best effort in completing the tests. Participants in the EM condition received instructions to feign cognitive deficits commonly observed after a moderate-to-severe TBI and were provided with a scenario modelled after those developed by DenBoer & Hall (2007) and Suhr & Gunstad (2000). The instructions and scenario that were given to participants are provided in Appendix D.

The recommendations for simulation studies provided by Rogers (2008) were followed. Specifically, Rogers outlined six elements that ideally should be considered when conducting simulation research. These include comprehensibility (i.e., instructions should be easily understood by participants), specificity (i.e., instructions should be explicit and clear), context (i.e., participants should be familiar with the context being simulated), relevance (i.e., participants should be able to relate to the scenarios), motivation (i.e., participants should be motivated to comply with task instructions), and believability (i.e. participants should be advised to make a realistic presentation).

Comprehensibility and specificity was satisfied by providing instructions at an easy reading level and written in simple sentences that explicitly state the task. Because participants in the LEP group had a range of English proficiency and reading ability, written instructions were read, clarified, and simplified by the examiner as necessary. Participants were provided an opportunity to ask questions regarding the instructions. A pre- and post-session questionnaire regarding the instructions was also administered to ensure comprehension.

Context and believability were addressed by utilizing a realistic scenario regarding a motor-vehicle collision. Participants were asked to complete neuropsychological testing for determination of insurance benefits and were provided information on the nature of cognitive deficits (e.g., memory, processing speed) following a TBI. Additionally, participants were warned that the battery may contain PVTs and asked to make their presentation as believable as possible to avoid detection.

The principles of relevance and motivation were more challenging to achieve. For example, it may have been difficult for some participants to relate to the scenario if they have not encountered such a situation. Furthermore, real-world incentives to feign impairment and the resulting consequences (e.g., payout in millions of dollars) were not present in this context and may be difficult to imagine for some participants. Nevertheless, the level of motivation to comply with the instructions and the relatability of the scenario were assessed in the post-session questionnaire as described below.

As a check for recall, comprehension, and compliance with the task instructions, a pre-session and post-session questionnaire were administered to participants, as per Rogers (2008). Participants in the EM condition received a pre-session questionnaire consisting of three multiple-choice questions and a post-session questionnaire consisting of four questions. Participants in the NC condition received one multiple-choice question pre-session and two questions post-session. The pre- and post-session questionnaires are provided in Appendix E.

Research assistants were blinded to the randomly assigned conditions of participants, so as not to introduce demand characteristics during testing. Blinding was completed by having the PI, who was not involved with test administration, provide the condition instructions and scenario to participants prior to the RA beginning the neuropsychological testing. Participants were asked not to reveal their condition to the RA completing testing. Participants were encouraged to ask questions and clarify the instructions with the PI to ensure they fully comprehended instructions before starting neuropsychological testing with the RA.

Statistical Analyses

Part 1: How do individuals with LEP perform on PVTs compared to NSE?

Hypothesis 1: BR$_{Fail}$ will be higher in the LEP than NSE group. BR$_{Fail}$ was calculated by summing the number of participants who failed ≥1, ≥2, ≥3, and ≥4 PVTs. A Chi-Square test of independence was used to compare BR$_{Fail}$ between LEP and NSE groups. Comparisons were made with both liberal and conservative cutoffs. The total number of PVTs failed at each failure level was also calculated and compared between LEP and NSE groups using a *t*-test. In addition to the overall BR$_{Fail}$ across all PVTs, BR$_{Fail}$ was also calculated and compared between groups at the instrument and indicator level. Between-group comparisons were also completed on PVT scores as continuous variables using *t*-tests.

Hypothesis 2: English proficiency will be associated with BR$_{Fail}$. Point-biserial correlations were calculated to examine whether BR$_{Fail}$ varies as a function of level of English proficiency. Specifically, correlation analyses were conducted between BR$_{Fail}$ and the Speaking, Comprehending, and Reading proficiency ratings on the LEAP-Q. Correlations were also conducted between BR$_{Fail}$ and the BNT-15.

Hypothesis 3: BR$_{Fail}$ will be greater for LEP than NSE participants on PVT$_{HVM}$ but not on PVT$_{LVM}$. The BR$_{Fail}$ was calculated for the combination of PVT$_{HVM}$ and PVT$_{LVM}$ to examine any differences between groups. A mixed-design ANOVA was conducted with English proficiency group (LEP versus NSE) as the between-group variable, level of verbal mediation of PVTs (low versus high) as the within-group variable, and number of PVTs failed as the dependent variable.

Part 2: Can Current PVTs Detect Non-Credible Performance for Individuals with LEP? What Cutoffs Provide Adequate Classification Accuracy in this Population?

Specificity, sensitivity, PPP, and NPP were calculated using standard formulas, as described in the Literature Review section. Area under the curve (AUC) was calculated for all PVTs as a measure of the overall accuracy of each PVT in predicting NC and EM group membership. Values of 1.0 represent a perfect discrimination while .5 represents chance discrimination based on PVT scores. EM and NC conditions within the LEP and NSE groups were compared to obtain signal-detection properties at different cutoffs across PVTs.

In order to find the optimal cutoff on each PVT, sensitivity and specificity were calculated for numerous potential cutoffs. Liberal and conservative cutoffs were determined as defined by a specificity of .84 and .90, respectively. Positive and negative predictive power were also calculated for hypothetical base rates representing settings with low (10%), medium (30%), and high (50%) base rates of invalid performance.

CHAPTER V: RESULTS

The current dissertation utilized 20 neuropsychological tests, resulting in over 200 scores. Such a broad-based assessment was instrumental in providing a thorough test of the main hypotheses, especially those focused on multivariate models. Consequently, the results contain a rich variety of analyses. Out of concerns that covering every single detail might attenuate the core investigation, reporting on the results was focused on the initial hypotheses, plus an additional one that is a pertinent extension of the *a priori* predictions (i.e., experimental-malingering profiles of LEP vs. NSE participants). While the data lend themselves to further exploratory analyses and clinically relevant *post hoc* hypotheses, in the interest of providing a succinct coverage of the original research questions, such temptations for follow-up analyses were actively resisted.

Part 1: How do Individuals with LEP Perform on PVTs Compared to NSE?

Hypothesis 1: BR_{Fail} will be higher in the LEP than NSE group. Participants in the LEP-NC group had a significantly higher overall BR_{Fail} than participants in the NSE-NC group across both liberal and conservative cutoffs (Table 8). Specifically, LEP-NC participants were more likely to fail ≥1 PVT (RR: 1.30-2.00), ≥2 PVTs (RR: 2.24-3.50), ≥3 PVTs (RR: 2.75-4.40), and ≥4 PVTs (RR: 5.00-6.50): χ^2 (1, N = 140) = 8.54-31.75, p < .01, Φ^2 = .11-.40 (large-very large effects). The total number of PVTs failed as a continuous variable was also greater for the LEP-NC (M = 2.8-4.0, SD = 1.3-1.6) compared to the NSE-NC group (M = 0.9-1.7, SD = 1.3-1.6) at both liberal and conservative cutoffs, d = 1.41–1.45 (large effect; Table 9). Overall, results support the hypothesis that examinees with LEP under normal conditions (i.e., instructed to perform to the best of their ability) would fail PVTs at a higher rate than NSE.

Table 8

Comparing Combined Base Rates of Failure (All Tests) as a Function of English Proficiency Group in the Control (i.e., Non-Malingering) Sample (n = 80)

		English Proficiency					
		LEP (*n* = 40)	NSE (*n* = 40)				
Score	Cutoff	BR_{Fail}	BR_{Fail}	RR	χ^2	*p*	Φ^2
Fail ≥1 PVT	LIB	97.5	75.0	1.30	8.54	<.01	.11
Fail ≥1 PVT	CON	95.0	47.5	2.00	22.03	<.01	.28
Fail ≥2 PVTs	LIB	95.0	42.5	2.24	25.66	<.01	.32
Fail ≥2 PVTs	CON	87.5	25.0	3.50	31.75	<.01	.40
Fail ≥3 PVTs	LIB	82.5	30.0	2.75	22.40	<.01	.28
Fail ≥3 PVTs	CON	55.0	12.5	4.40	16.16	<.01	.20
Fail ≥4 PVTs	LIB	62.5	12.5	5.00	21.33	<.01	.27
Fail ≥4 PVTs	CON	32.5	5.0	6.50	9.93	<.01	.12

Note: PVT: Performance validity test; LEP: Limited English proficiency; NSE: Native speakers of English; BR_{Fail}: Base rate of failure (Failure on each indicator was only counted once within each test to reduce inflation); LIB: Liberal cutoffs optimized for sensitivity; CON: Conservative cutoffs optimized for specificity.

Table 9

Comparing the Total Number of PVTs Failed as a Function of English Proficiency Group in the Control (Non-Malingering) Sample (n = 80)

	English Proficiency						
	LEP (*n* = 40)		NSE (*n* = 40)				
Cutoff	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
LIB	4.0	1.6	1.7	1.6	-6.30	<.01	1.41
CON	2.8	1.3	0.9	1.3	-6.52	<.01	1.45

Note. PVT: Performance validity test; LEP: Limited English proficiency; NSE: Native speakers of English; LIB: Liberal cutoffs optimized for sensitivity; CON: Conservative cutoffs optimized for specificity.

Hypothesis 2: English proficiency will be associated with BR_{Fail}. LEAP-Q was significantly correlated with BR_{Fail} for failing ≥1, ≥2, ≥3, and ≥4 PVT_{HVM} at both liberal and conservative cutoffs for Speaking (r_{pb} = -.22 to -.62, $p < .05$), Comprehending (r_{pb} = -.21 to -.53, $p < .05$), and Reading (r_{pb} = -.20 to -.53, $p < .05$), accounting for 4% to 38% of the variance of BR_{Fail} (Table 10). The BNT-15 was also significantly correlated with BR_{Fail} at all levels of cutoffs for PVT_{HVM} (r_{pb} = -.41 to -.72, $p < .01$), with 16% to 52% shared variance. None of the English proficiency measures were correlated with BR_{Fail} on PVT_{LVM}.

Table 10

Point-Biserial Correlations Between Base Rates of Failure at Various Cutoffs and Measures of English Proficiency in the Control (Non-Malingering) Condition (n = 80)

Level of verbal mediation	BR$_{Fail}$		LEAP-Q English Proficiency Rating						BNT-15	
	Level	Cutoff	Speaking		Comprehending		Reading		Accuracy	
			r_{pb}	r^2	r_{pb}	r^2	r_{pb}	r^2	r_{pb}	r^2
HIGH	≥1 PVT	LIB	-.43**	.18	-.38**	.14	-.37**	.14	-.45**	.20
		CON	-.54**	.29	-.49**	.24	-.47**	.22	-.61**	.37
	≥2 PVTs	LIB	-.53**	.28	-.49**	.24	-.45**	.21	-.62**	.38
		CON	-.62**	.38	-.53**	.28	-.53**	.29	-.72**	.52
	≥3 PVTs	LIB	-.41**	.17	-.40**	.16	-.41**	.17	-.65**	.43
		CON	-.30**	.09	-.26**	.07	-.29**	.09	-.52**	.27
	≥4 PVTs	LIB	-.36**	.13	-.36**	.13	-.32**	.10	-.53**	.28
		CON	-.22*	.05	-.21*	.04	-.20*	.04	-.41**	.17
LOW	≥1 PVT	LIB	-.11	.01	-.08	.01	-.10	.01	-.15	.02
		CON	-.13	.02	-.09	.01	-.05	<.01	-.16	.02
	≥2 PVTs	LIB	-.01	<.01	.01	<.01	.03	<.01	-.08	<.01
		CON	.07	.01	.14	.02	.08	<.01	-.01	<.01
	≥3 PVTs	LIB	-.10	.01	-.12	.01	-.04	<.01	-.12	.01
		CON	-	-	-	-	-	-	-	-
	≥4 PVTs	LIB	-	-	-	-	-	-	-	-
		CON	-	-	-	-	-	-	-	-
OVERALL	≥1 PVT	LIB	-.25*	.06	-.24*	.06	-.21*	.04	-.30**	.09
		CON	-.47**	.22	-.44**	.19	-.40**	.16	-.54**	.30
	≥2 PVTs	LIB	-.48**	.23	-.45**	.20	-.40**	.16	-.55**	.30
		CON	-.55**	.30	-.50**	.25	-.48**	.23	-.69**	.47
	≥3 PVTs	LIB	-.47**	.22	-.45**	.20	-.45**	.21	-.59**	.35
		CON	-.43**	.18	-.32**	.11	-.39**	.16	-.57**	.33
	≥4 PVTs	LIB	-.38**	.14	-.37**	.13	-.38**	.15	-.56**	.31
	≥1 PVT	CON	-.30**	.09	-.22*	.05	-.20*	.04	-.46**	.21
	Test	Score								
	BNT-15	Accuracy	.82**	.67	.75**	.57	.73**	.54	-	-

**p*(one-tail) < .05; ***p*(one-tail) < .01

Note: PVT: Performance validity test; BR$_{Fail}$: Base rate of failure; LIB: Liberal cutoffs optimized for sensitivity; CON: Conservative cutoffs optimized for specificity; LEAP-Q: Language Experience and Proficiency Questionnaire; BNT-15: Boston Naming Test 15-Item Short Form; Negative correlation = Higher BR$_{Fail}$ is correlated with lower score on English proficiency measures.

Hypothesis 3: BR$_{Fail}$ will be greater for LEP than NSE participants on PVT$_{HVM}$ but not on PVT$_{LVM}$. As predicted, the difference in BR$_{Fail}$ between LEP-NC and NSE-NC groups was observed only on PVT$_{HVM}$. Specifically, LEP-NC participants had a significantly higher overall BR$_{Fail}$ on PVT$_{HVM}$ than the NSE-NC group at both liberal and conservative cutoffs, across changing psychometric definitions of invalid performance (RR: 1.85-8.50), while no difference in overall BR_{Fail} was found for PVT_{LVM} (RR: 1.00-1.22; Table 11). These results provide an important context for the previous finding (i.e.., overall BR_{Fail}: LEP > NSE), suggesting that the higher combined BR_{Fail} is driven by higher failures on PVT_{HVM} in LEP participants.

Table 11

Comparing Combined Base Rates of Failure as a Function of English Proficiency Group in the Control (i.e., Non-Malingering) Sample (n = 80)

			English Proficiency					
			LEP	NSE				
Level of Verbal			(*n* = 40)	(*n* = 40)				
Mediation	Score	Cutoff	BR_{Fail}	BR_{Fail}	RR	χ^2	*p*	Φ^2
HIGH	Fail ≥1 PVT	LIB	92.5	50.0	1.85	17.64	<.01	.22
	Fail ≥1 PVT	CON	90.0	32.5	2.77	27.9	<.01	.35
	Fail ≥2 PVTs	LIB	90.0	27.5	3.27	32.24	<.01	.41
	Fail ≥2 PVTs	CON	80.0	10.0	8.00	39.6	<.01	.49
	Fail ≥3 PVTs	LIB	65.0	12.5	5.20	23.23	<.01	.29
	Fail ≥3 PVTs	CON	40.0	5.0	8.00	14.01	<.01	.18
	Fail ≥4 PVTs	LIB	42.5	5.0	8.50	15.53	<.01	.19
	Fail ≥4 PVTs	CON	17.5	0.0	-	7.67	<.01	.10
LOW	Fail ≥1 PVT	LIB	70.0	57.5	1.22	1.35	.25	.02
	Fail ≥1 PVT	CON	45.0	37.5	1.20	0.46	.50	.01
	Fail ≥2 PVTs	LIB	17.5	15.0	1.17	0.09	.76	.00
	Fail ≥2 PVTs	CON	2.5	7.5	0.33		.62	.01
	Fail ≥3 PVTs	LIB	2.5	2.5	1.00		1.00	.00
	Fail ≥3 PVTs	CON	0.0	0.0	-		-	-
	Fail ≥4 PVTs	LIB	0.0	0.0	-		-	-
	Fail ≥4 PVTs	CON	0.0	0.0	-		-	-

Note: PVT: Performance validity test; LEP: Limited English proficiency; NSE: Native speakers of English; BR_{Fail}: Base rate of failure (Failure on each indicator is only counted once within each test to reduce inflation); LIB: Liberal cutoffs optimized for sensitivity; CON: Conservative cutoffs optimized for specificity; HIGH: Seven tests of high verbal mediation (Word Choice Test, WAIS-III Digit Span subtest, Word Recognition Test, FAS, Animals, CIM, Stroop) contribute to BR_{Fail}; LOW: Seven tests of low verbal mediation (Test of Memory Malingering, Dot Counting Test, Rey 15-Item Test, Trail Making Test, WAIS-III Digit Symbol subtest, WAIS-IV Symbol Search subtest, Rey-Osterrieth Complex Figure Test) contribute to BR_{Fail}.

At the instrument level, the LEP-NC group had notably higher BR_{Fail} on the WRT Combination, FAS, Animals, $\text{LRE}_{Johnson}$, CIM, and Stroop Color and Interference conditions than the NSE group (RR: 2.33-12.00; Table 12). The only three PVT_{HVM} that did not show this pattern were the Digit Span (RDS, ACSS), WCT Accuracy, and Stroop

Word condition. Markedly, Digit Span indicators had a reversal in the expected BR*Fail* direction, with NSE having a higher failure rate (RR:1.33-5.00). In contrast, there was no significant difference in BR*Fail* between the LEP-NC and NSE-NC groups on any of the PVT*LVM* (Table 13).

The LEP-NC group also had lower mean scores on several PVT*HVM* as continuous variables compared to the NSE-NC group, including the WCT T2C, FAS, Animals, CIM, and Stroop Color and Interference conditions (Table 14), while no meaningful differences (i.e., at least a medium effect) in scores were found on PVT*LVM*, with the exception of the Clock Drawing Test (Table 15). Amongst the PVT*HVM*, LEP-NC and NSE-NC participants performed similarly on the WCT Accuracy, Digit Span, and Stroop Word condition.

Table 12

Comparing Instrument-Level BR$_{Fail}$ on Tests of High Verbal Mediation as a Function of English Proficiency Group in the Control (i.e., Non-Malingering) Sample (n = 80)

			English Proficiency					
			LEP (*n* = 40)	NSE (*n* = 40)				
PVT	Score	Cutoff^a	BR$_{Fail}$	BR$_{Fail}$	RR	χ^{2b}	*p*	Φ^2
WCT	Accuracy	≤47	7.5	2.5	3.00		.62	.01
		≤43	2.5	0.0	-		1.00	.01
	T2C	≥156	8.1	0.0	-		.12	.04
		≥171	8.1	0.0	-		.12	.04
Digit Span$_{WAIS\text{-}III}$	RDS	≤7	7.5	10.0	0.75 (1.33)		1.00	<.01
		≤6	0.0	2.5	0		1.00	.01
	ACSS	≤6	2.5	12.5	0.20 (5.00)		.20	.04
		≤5	0.0	5.0	0		.49	.03
WRT	Recognition	≤7	15.0	2.5	6.00		.11	.05
		≤5	0.0	0.0	-		-	
	Combination	≤10	25.0	5.0	5.00	6.28	.01	.08
		≤8	17.5	0.0	-		.01	.10
FAS	T-score	≤33	40.0	15.0	2.67	6.27	.01	.08
		≤31	30.0	12.5	2.40	3.66	.06	.04
Animals	T-score	≤33	62.5	15.0	4.17	19.0	<.01	.24
		≤31	57.5	12.5	4.60	17.80	<.01	.22
	LRE$_{Johnson}$	≥.45	35.0	15.0	2.33	4.3	.04	.05
		≥.475	20.0	5.0	4.00	4.11	.04	.05
	LRE$_{Sugarman}$	≥.5	10.0	2.5	4.00		.36	.03
		≥.6	7.5	2.5	3.00		.62	.01
CIM	Raw	≤9	82.1	7.5	10.95	44.48	<.01	.56
		≤8	56.4	7.5	7.52	21.8	<.01	.28
	T-score	≤29	82.1	10.0	8.21	41.33	<.01	.52
		≤23	79.5	7.5	10.60	41.74	<.01	.53
Stroop	Color	≤7	52.5	15.0	3.50	12.6	<.01	.16
		≤5	27.5	0.0	-	12.75	<.01	.16
	Word	≤7	12.5	10.0	1.25		1.00	<.01
		≤5	7.5	5.0	1.50		1.00	<.01
	INT	≤7	30.0	2.5	12.00	11.11	<.01	.14
		≤5	20.0	2.5	8.00		.03	.08

[a]First row = Liberal cutoff; Second row = Conservative cutoff; [b]Fisher's Exact Test calculated when Chi-Square assumptions were violated (e.g., expected frequencies > 5).

Note: PVT: Performance validity test; LEP: Limited English proficiency; NSE: Native speakers of English; BR$_{Fail}$: Base rate of failure; WCT: Word Choice Test; T2C: Time to completion (seconds); Digit Span$_{WAIS\text{-}III}$: WAIS-III Digit Span subtest; RDS: Reliable Digit Span; ACSS: Age-corrected scaled-score; WRT: Word Recognition Test; FAS: Letter fluency test; Animals: Category animal fluency test; LRE$_{Johnson}$: Logistical regression equation combining overall letter fluency output and pattern of performance (Johnson et al., 2012); LRE$_{Sugarman}$: Logistical regression equation combining FAS and Animal Fluency T-scores (Sugarman & Axelrod, 2015) ; CIM: Complex Ideational Material; Stroop: D-KEFS Color-Word Interference Test; Color: Color Condition ACSS; Word: Word Condition ACSS; INT: Interference Condition ACSS.

Table 13

Comparing Instrument-Level BR$_{Fail}$ on Tests of Low Verbal Mediation as a Function of English Proficiency Group in the Control (i.e., Non-Malingering) Sample (n = 80)

			English Proficiency					
			LEP (*n* = 40)	NSE (*n* = 40)				
PVT	Score	Cutoff[a]	BR$_{Fail}$	BR$_{Fail}$	RR	χ^{2b}	*p*	Φ^2
TOMM	T1	≤44	5.0	10.0	0.50		.68	.01
		≤39	0.0	5.0	0		.49	.03
DCT	E-score	≥15	7.5	5.0	1.50		1.00	<.01
		≥17	2.5	2.5	1.00		1.00	<.01
FIT	Recall	<10	0.0	0.0	-		-	
		<9	0.0	0.0	-		-	
	Recognition	<11	5.0	0.0	-		.49	.03
		<10	2.5	0.0	-		1.00	.01
	Combined	<21	2.5	0.0	-		1.00	.01
		<20	2.5	0.0	-		1.00	.01
TMT	A T-score	≤39	50.0	32.5	1.54	2.53	.11	.03
		≤34	32.5	25.0	1.30	0.55	.46	.01
	B T-score	≤37	40.0	22.5	1.78	2.85	.09	.04
		≤30	5.0	5.0	1.00		1.00	<.01
	A + B Raw	≥137	5.0	7.5	0.67		1.00	<.01
		≥170	2.5	2.5	1.00		1.00	<.01
CD$_{WAIS-III}$	ACSS	≤5	0.0	2.5	0		1.00	.01
		≤4	0.0	2.5	0		1.00	.01
SS$_{WAIS-IV}$	ACSS	≤6	2.5	2.5	1.00		1.00	<.01
		≤5	2.5	0.0	-		1.00	.01
RCFT	Copy	≤26	0.0	5.0	0		.49	.03
		≤23	0.0	0.0	-		-	
	IR	≤10	0.0	7.5	0		.24	.04
		≤9.5	0.0	7.5	0		.24	.04
	Recog	≤16	5.3	2.5	2.12		.61	<.01
		≤15	5.3	0.0	-		.23	.03
	Equation	≤47	2.6	2.5	1.04		1.00	<.01
		≤45	2.6	2.5	1.04		1.00	<.01

[a] First row = Liberal cutoff; Second row = Conservative cutoff; [b] Fisher's Exact Test calculated when Chi-Square assumptions violated (e.g., expected frequencies > 5).

Note: PVT: Performance validity test; LEP: Limited English proficiency; NSE: Native speakers of English; BR$_{Fail}$: Base rate of failure; TOMM T1: Test of Memory Malingering Trial 1; DCT E-Score: Dot Counting Test Effort-Score; FIT: Rey 15-Item Test; TMT A T-score, B T-score: Trail Making Test Part A and Part B T-score; A + B: Trail Making Test Trial A & B Total Combined Time Score; CD$_{WAIS-III}$: WAIS-III Digit Symbol subtest; SS$_{WAIS-IV}$: WAIS-IV Symbol Search subtest; RCFT: Rey-Osterrieth Complex Figure Test; Copy: Copy Trial raw score; IR: Immediate Recall raw score; Recog: Recognition Trial raw; Equation: CT raw score + (true positive recognition – Atypical recognition errors) x 3 (Lu et al., 2003).

Table 14

Descriptive Statistics of Tests of High Verbal Mediation as a Function of English Proficiency Sample in the Control (Non-Malingering) Sample (n = 80)

		English Proficiency						
		LEP (*n* = 40)		NSE (*n* = 40)				
Measure	Score	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
WCT	Accuracy	49.1	1.5	49.7	0.7	2.16	.04	0.51
	T2C	95.8	38.1	72.4	20.6	-3.29	<.01	0.76
Digit Span$_{\mathit{WAIS\text{-}III}}$	Total Raw	17.0	3.7	17.4	4.0	.46	.64	0.10
	RDS	9.8	2.1	10.0	2.1	.32	.75	0.10
	ACSS	9.7	2.6	10.0	2.7	.63	.53	0.11
WRT	Recognition	10.1	2.3	10.8	1.7	1.48	.14	0.35
	Combination	14.2	4.3	15.8	3.3	1.87	.07	0.42
FAS	Raw	30.5	6.5	39.7	10.8	4.61	<.01	1.03
	T-score	34.0	5.9	43.0	9.9	4.95	<.01	1.10
Animals	Raw	16.3	3.7	23.6	5.3	7.13	<.01	1.60
	T-score	30.0	9.9	46.9	10.9	7.27	<.01	1.62
Emotional Fluency	Raw	8.1	2.7	13.3	6.1	5.03	<.01	1.10
CIM	Raw	7.4	2.7	11.1	1.2	7.76	<.01	1.77
	T-score	16.3	15.4	44.2	13.8	8.51	<.01	1.91
Stroop	Color Raw	33.9	6.9	27.5	4.6	-4.91	<.01	1.09
	Colors ACSS	7.2	3.0	10.1	2.1	5.02	<.01	1.12
	Word Raw	22.1	3.8	20.9	4.1	-1.35	.18	0.30
	Word ACSS	10.0	2.3	10.7	2.4	1.34	.18	0.30
	INT Raw	57.1	17.1	43.8	8.6	-4.41	<.01	0.98
	INT ACSS	8.7	3.6	11.8	2.0	4.89	<.01	1.06
BNT-15	Accuracy	6.5	3.1	13.9	1.2	14.02	<.01	3.15
	T2C	185.8	57.2	43.4	27.8	-14.17	<.01	3.17
Reading$_{\text{WRAT-4}}$	SS	85.4	10.2	102.7	11.7	7.05	<.01	1.58

Note: LEP: Limited English proficiency; NSE: Native speakers of English; WCT: Word Choice Test; T2C: Time to completion (seconds); Digit Span$_{\mathit{WAIS\text{-}III}}$: WAIS-III Digit Span subtest; RDS: Reliable Digit Span; ACSS: Age-corrected scaled score (*M* = 10, *SD* = 3); WRT: Word Recognition Test; FAS: Letter fluency test; Animals: Category animal fluency test; T-score (*M* = 50, *SD* = 10); Emotional Fluency: Category emotional fluency test; CIM: Complex Ideational Material; Stroop: D-KEFS Color-Word Interference Test; INT: Interference Condition; BNT-15: Boston Naming Test 15-Item Short Form; Reading$_{\text{WRAT-4}}$: Wide Range Achievement Test 4th Edition Reading subtest; SS: Standard score (*M* = 100; *SD* = 15).

Table 15

Descriptive Statistics of Tests of Low Verbal Mediation as a Function of English Proficiency Sample in the Control (Non-Malingering) Sample (n = 80)

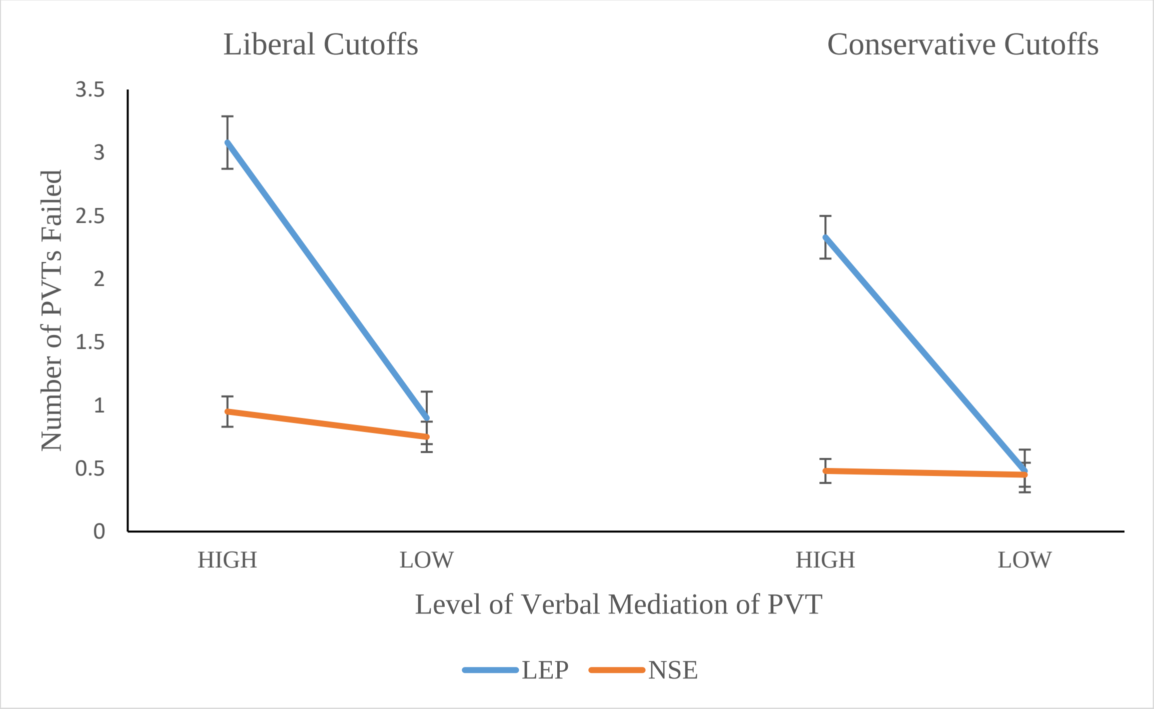
		English Proficiency						
		LEP (*n* = 40)		NSE (*n* = 40)				
Measure	Score	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
TOMM	T1	48.6	2.0	47.4	3.1	-2.13	.04	0.46
DCT	E-score	11.1	2.3	10.3	2.7	-1.34	.19	0.32
	Errors	1.0	0.9	0.9	1.0	-0.24	.81	0.11
FIT	Recall	14.6	1.1	14.9	.47	1.87	.07	0.35
	Recognition	14.1	2.0	14.5	1.0	1.18	.24	0.25
	Combined	28.7	2.7	29.5	1.4	1.62	.11	0.37
TMT	A Raw	32.5	18.8	32.4	18.8	-0.02	.98	0.01
	A T-Score	37.9	10.8	41.0	14.9	1.07	.29	0.24
	B Raw	62.9	19.3	58.4	23.6	-0.95	.35	0.21
	B T-Score	42.9	9.6	47.8	11.8	2.04	.04	0.46
CD*WAIS-III*	Raw	86.7	13.1	87.0	13.7	0.08	.93	0.02
	ACSS	11.5	2.6	11.6	2.4	0.27	.79	0.04
	Recognition	7.8	1.2	8.3	1.1	1.76	.08	0.43
SS*WAIS-IV*	Raw	36.3	6.3	37.0	7.1	0.43	.67	0.10
	ACSS	11.0	2.3	11.3	2.7	0.58	.56	0.12
RCFT	Copy	34.2	1.7	33.5	2.5	-1.39	.17	0.33
	T2C	171.4	113.8	146.3	54.2	-1.26	.21	0.28
	IR Raw	23.0	4.9	23.6	6.8	0.48	.63	0.10
	IR T-Score	45.2	12.4	48.6	14.8	1.12	.27	0.25
	DR Raw	22.6	4.6	23.1	6.7	0.41	.68	0.09
	DR T-Score	44.2	11.2	46.9	14.0	0.96	.34	0.21
	Recog Raw	20.1	2.5	21.3	2.0	2.21	.03	0.53
	Recog T-Score	41.5	14.3	48.1	12.9	2.13	.04	0.48
CDT	Raw	8.4	1.5	9.7	0.6	4.94	<.01	1.14

Note: LEP: Limited English proficiency; NSE: Native speakers of English; TOMM T1: Test of Memory Malingering Trial 1; DCT E-Score: Dot Counting Test Effort-Score; FIT: Rey 15-Item Test; TMT-A, TMT-B: Trail Making Test Part A and Part B; T-score (*M* = 50, *SD* = 10);CD*WAIS-III*: WAIS-III Digit Symbol subtest; ACSS: Age-corrected scaled score (*M* = 10, *SD* = 3); SS*WAIS-IV*: WAIS-IV Symbol Search subtest; RCFT: Rey-Osterrieth Complex Figure Test; T2C: Time to completion; IR: Immediate Recall; DR: Delayed Recall; Recog: Recognition; CDT: Clock Drawing Test.

A mixed-design ANOVA was conducted as a formal measure of the interaction between level of verbal mediation of PVTs (low vs. high) and English proficiency (LEP vs. NSE) on the number of PVTs failed. In addition to the univariate main effects presented above, results revealed a significant interaction using both liberal, *F* (1, 78) = 38.96, *p* <.01, $\eta^2_{partial}$ = .33 (very large effect), and conservative cutoffs, *F* (1, 78) = 49.92, *p* <.01, $\eta^2_{partial}$ = .39 (very large effect; Figure 1). The outcome of multivariate analyses confirms earlier conclusions that the difference in BR_{Fail} between groups are attributable to the level of verbal mediation of PVTs.

Figure 1

Interaction Between Level of Verbal Mediation and English Proficiency Sample on the Number of PVTs Failed in the Control (Non-Malingering) Condition (n = 80)



Note. PVT: Performance validity test; NSE: Native speakers of English; LEP: Limited English proficiency.

Specifically, LEP-NC participants failed on average more PVT_{HVM} than NSE-NC participants regardless if liberal, *t*(78) = -7.23, *p* < .01, *d* = 1.62 (large effect), or conservative cutoffs, *t*(78) = -7.76, *p* < .01, *d* = 1.73 (large effect), were used (Table 16). In contrast, there was no difference in the number of PVT_{LVM} failed between LEP-NC and NSE-NC participants at either the liberal, *t*(78) = -0.88, *p* = .38, or conservative cutoffs, *t*(78) = -0.19, *p* = .85. Taken together, the results suggest that the higher BR_{Fail} and greater number of PVT failures in the LEP-NC group was limited to PVT_{HVM}.

Table 16

Comparing Number of PVTs Failed as a Function of English Proficiency Group in the Control (Non-Malingering) Sample (n = 80)

		English Proficiency						
Level of verbal mediation		LEP (*n* = 40)		NSE (*n* = 40)				
	PVT Type	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
HIGH	LIB	3.1	1.4	1.0	1.2	-7.23	<.01	1.62
	CONS	2.3	1.3	0.5	0.8	-7.76	<.01	1.73
LOW	LIB	0.9	0.7	0.8	0.8	-0.88	.38	0.20
	CONS	0.5	0.6	0.5	0.6	-0.19	.85	0.05

Note. PVT: Performance validity test; LEP: Limited English proficiency; NSE: Native speakers of English; LIB: Liberal cutoffs optimized for sensitivity; CONS: Conservative cutoffs optimized for specificity.

Part 2: Can Current PVTs Detect Non-Credible Performance for Individuals with LEP? What Cutoffs Provide Adequate Classification Accuracy in this Population?

In the LEP group, AUC ranged from .55 to .88 for PVT_{HVM} and from .72 to .93 for PVT_{LVM}. In the NSE group, AUC ranged from .69 to .94 for PVT_{HVM} and from .74 to .93 for PVT_{LVM} (Tables 17-18). A closer examination of AUC values revealed that LEP participants had significantly lower AUC for at least one indicator on all PVT_{HVM} compared to NSE participants, while no significant between-group differences were found on PVT_{LVM}.

Table 17

Receiver Operating Characteristics for Tests of High Verbal Mediation as a Function of Language Group with Experimental Condition (Control vs. Malingering) as the Criterion Variable (N = 140)

		English Proficiency					
		LEP (*n* = 70)			NSE (*n* = 70)		
PVT	Score	AUC	*p*	95% CI	AUC	*p*	95% CI
WCT	Accuracy	.88	<.01	.78–.97	.94	<.01	.88-1.00
	T2C	.70	.01	.56-.84	.90	<.01	.83-.97
Digit Span$_{\textit{WAIS-III}}$	RDS	.87	<.01	.77-.96	.78	<.01	.67-.89
	ACSS	.86	<.01	.76-.96	.74	<.01	.63-.86
WRT	Recognition	.83	<.01	.73-.93	.78	<.01	.65-.90
	Combination	.85	<.01	.75-.95	.78	<.01	.67-.89
Verbal Fluency	FAS T-score	.55	.48	.40-.70	.70	.01	.57-.82
	Animals T-score	.66	.03	.52-.79	.80	<.01	.70-.91
	LRE$_{\textit{Johnson}}$	.56	.38	.42-.71	.69	.01	.56-.82
	LRE$_{\textit{Sugarman}}$	.64	.05	.50-.78	.83	<.01	.72-.93
CIM	Raw	.65	.04	.52-.79	.80	<.01	.70-.91
	T-score	.59	.21	.45-.73	.78	<.01	.67-.89
Stroop	Color	.74	<.01	.62-.86	.79	<.01	.68-.91
	Word	.78	<.01	.66-.91	.77	<.01	.65-.90
	INT	.64	.05	.50-.79	.87	<.01	.77-.96

Note: PVT: Performance validity test; LEP: Limited English proficiency; NSE: Native speakers of English; AUC: Area under the curve; WCT: Word Choice Test; T2C: Time to completion (seconds); Digit Span$_{\textit{WAIS-III}}$: WAIS-III Digit Span subtest; RDS: Reliable Digit Span; ACSS: Age-corrected scaled-score; WRT: Word Recognition Test; FAS: Letter fluency test; Animals: Category Animal fluency test; LRE$_{\textit{Johnson}}$: Logistical regression equation combining overall letter fluency output and pattern of performance (Johnson et al., 2012); LRE$_{\textit{Sugarman}}$: Logistical regression equation combining FAS and Animal Fluency T-scores (Sugarman & Axelrod, 2015); CIM: Complex Ideational Material; Stroop: D-KEFS Color-Word Interference Test; Color: Color Condition ACSS; Word: Word Condition ACSS; INT: Interference Condition ACSS.

Table 18

Receiver Operating Characteristics for Tests of Low Verbal Mediation as a Function of Language Group with Experimental Condition (Control vs. Malingering) as the Criterion Variable (N = 140)

		English Proficiency					
		LEP (*n* = 70)			NSE (*n* = 70)		
PVT	Score	AUC	*p*	95% CI	AUC	*p*	95% CI
TOMM	T1	.93	<.01	.85-1.00	.93	<.01	.86-1.00
DCT	E-score	.90	<.01	.83-.98	.89	<.01	.81-.97
FIT	Combined	.84	<.01	.73-.94	.83	<.01	.72-.94
TMT	A T-score	.78	<.01	.66-.89	.76	<.01	.64-.87
	B T-score	.72	<.01	.59-.86	.74	<.01	.61-.86
	A + B	.80	<.01	.69-.91	.77	<.01	.65-.89
CD$_{\mathit{WAIS-II}}$	ACSS	.89	<.01	.81-.97	.88	<.01	.80-.97
SS$_{\mathit{WAIS-IV}}$	ACSS	.84	<.01	.74-.94	.80	<.01	.69-.92
RCFT	Copy	.88	<.01	.78-.98	.82	<.01	.71-.92
	IR	.86	<.01	.76-.96	.77	<.01	.65-.88
	Recog	.75	<.01	.63-.87	.83	<.01	.73-.92
	Equation	.84	<.01	.74-.94	.90	<.01	.83-.97

Note: PVT: Performance validity test; LEP: Limited English proficiency; NSE: Native speakers of English; AUC: Area under the curve; TOMM T1: Test of Memory Malingering Trial 1; DCT E-Score: Dot Counting Test Effort-Score; FIT: Rey 15-Item Test; TMT A T-score, B T-score: Trail Making Test Part A and Part B T-score; A + B: Trail Making Test Trial A & B Total Combined Time Score; CD$_{\mathit{WAIS-III}}$: WAIS-III Digit Symbol subtest; SS$_{\mathit{WAIS-IV}}$: WAIS-IV Symbol Search subtest; RCFT: Rey-Osterrieth Complex Figure Test; Copy: Copy Trial raw score; IR: Immediate Recall raw score; Recog: Recognition Trial raw score; Equation: CT raw score + (true positive recognition – Atypical recognition errors) x 3 (Lu et al., 2003).

In terms of determining optimal cutoffs for the LEP group, at least one indicator on all PVT$_{LVM}$ was found to have a sensitivity of ≥.50 at specificity levels of .84 and .90 (Tables 19-20). In contrast, several PVT$_{HVM}$ had low sensitivity (sensitivity = .13 to .48) at these specificity levels, including the WCT T2C, FAS, Animals, Verbal Fluency LREs, CIM, and Stroop Color and Interference conditions (Tables 21-22). Furthermore, several cutoffs on PVT$_{HVM}$ had to be made so conservative (e.g., Stroop Color ACSS ≤2 or CIM T-score ≤3) that their clinical utility becomes questionable. These findings suggest that, while PVT$_{LVM}$ are useful for distinguishing between honest and feigned performance independent of English proficiency, many PVT$_{HVM}$ have compromised classification accuracy in examinees with LEP.

Table 19

Classification Accuracy of Embedded PVT$_{LVM}$ with Experimental Condition (Control vs. Malingering) as the Criterion Variable in LEP Sample (n = 70)

	Signal detection properties					Hypothetical base rates		
Test cutoff	SENS	SPEC	+LR	-LR		.10	.30	.50
Trail Making Test								
A T-score ≤ 24	**.50**	**.93**	**6.67**	**.54**	PPP	.44	.75	.88
					NPP	.94	.81	.65
A T-score ≤ 29	**.57**	**.83**	**3.24**	**.53**	PPP	.27	.59	.77
					NPP	.95	.82	.66
B T-score ≤ 30	.47	.95	9.33	.56	PPP	.51	.80	.90
					NPP	.94	.81	.64
B T-score ≤ 31	**.50**	**.90**	**5.00**	**.56**	PPP	.36	.68	.83
					NPP	.94	.81	.64
B T-score ≤ 33	**.57**	**.88**	**4.53**	**.50**	PPP	.35	.67	.83
					NPP	.95	.83	.67
A+B ≥ 118	**.60**	**.85**	**4.00**	**.47**	PPP	.31	.63	.80
					NPP	.95	.83	.68
A+B ≥ 120	.57	.90	5.67	.48	PPP	.39	.71	.85
					NPP	.95	.83	.68
A+B ≥ 123	**.57**	**.95**	**11.33**	**.46**	PPP	.56	.83	.92
					NPP	.95	.84	.69
CD$_{WAIS-III}$								
ACSS ≤ 6	.60	1.00	-	.40	PPP	1.00	1.00	1.00
					NPP	.96	.85	.71
ACSS ≤ 7	**.70**	**.93**	**9.33**	**.32**	PPP	.53	.81	.91
					NPP	.97	.88	.76
SS$_{WAIS-IV}$								
ACSS ≤ 6	.43	.98	17.33	.58	PPP	.70	.90	.96
					NPP	.94	.80	.63
ACSS ≤ 7	.47	.93	6.22	.58	PPP	.43	.74	.87
					NPP	.94	.80	.64
ACSS ≤ 8	**.60**	**.85**	**4.00**	**.47**	PPP	.31	.63	.80
					NPP	.95	.83	.68
RCFT								
Copy ≤ 30	.67	.98	26.67	.34	PPP	.79	.93	.97
					NPP	.96	.87	.75
Copy ≤ 31	**.73**	**.93**	**9.78**	**.29**	PPP	.54	.82	.91
					NPP	.97	.89	.78
IR ≤ 15.0	.43	.90	4.33	.63	PPP	.32	.65	.81
					NPP	.93	.79	.61
IR ≤ 16.5	**.67**	**.88**	**5.33**	**.38**	PPP	.38	.71	.85
					NPP	.96	.86	.73
Equation ≤ 51	**.53**	**.92**	**6.76**	**.51**	PPP	.42	.74	.87
					NPP	.95	.82	.66
Equation ≤ 52	**.57**	**.87**	**4.31**	**.50**	PPP	.33	.65	.81
					NPP	.95	.83	.67

Note: PVT: Performance validity test; LEP: Limited English proficiency; SENS: Sensitivity; SPEC: Specificity, +LR: Positive likelihood ratio; -LR: Negative likelihood ratio; PPP: Positive predictive power; NPP: Negative predictive power; TMT A T-score, B T-score: Trail Making Test Part A and Part B T-score; A + B: Trail Making Test Trial A & B Total Combined Time Score; CD$_{WAIS\text{-}III}$: WAIS-III Digit Symbol subtest; ACSS: Age-corrected scaled-score; SS$_{WAIS\text{-}IV}$: WAIS-IV Symbol Search subtest; RCFT: Rey-Osterrieth Complex Figure Test; Copy: Copy Trial raw score; IR: Immediate Recall raw score; Recog: Recognition Trial raw score; Equation: CT raw score + (true positive recognition – Atypical recognition errors) x 3 (Lu et al., 2003), **Bolded** = Best cutoffs when SENS ≥.50 and SPEC ≥.85 (liberal) and ≥.90 (conservative).

Table 20

Classification Accuracy of Freestanding PVT$_{LVM}$ with Experimental Condition (Control vs. Malingering) as the Criterion Variable in LEP Sample (n = 70)

	Signal detection properties					Hypothetical base rates		
Test cutoff	SENS	SPEC	+LR	-LR		.10	.30	.50
TOMM								
T1 ≤44	.83	.95	16.67	.18	PPP	.65	.88	.94
					NPP	.98	.93	.85
T1 ≤45	**.87**	**.93**	**11.56**	**.14**	PPP	.58	.84	.93
					NPP	.98	.94	.88
T1 ≤46	**.90**	**.88**	**7.20**	**.11**	PPP	.45	.76	.88
					NPP	.99	.95	.90
DCT								
E-score ≥ 13.4	**.80**	**.88**	**6.40**	**.23**	PPP	.43	.74	.87
					NPP	.98	.91	.81
E-score ≥ 14.6	.70	.90	7.00	.33	PPP	.44	.75	.88
					NPP	.96	.88	.75
E-score ≥ 15.2	**.70**	**.95**	**14.00**	**.32**	PPP	.61	.86	.93
					NPP	.97	.88	.76
FIT								
Recall ≤14	.48	.91	5.33	.57	PPP	.37	.70	.84
					NPP	.94	.80	.64
Recall ≤ 13	.42	.93	6.00	.62	PPP	.40	.72	.86
					NPP	.94	.79	.62
Combined ≤ 24	.52	.93	6.90	.52	PPP	.45	.76	.88
					NPP	.95	.82	.66
Combined ≤ 25	**.59**	**.90**	**5.86**	**.46**	PPP	.40	.72	.86
					NPP	.95	.84	.69
Combined ≤ 26	**.69**	**.88**	**5.52**	**.35**	PPP	.40	.72	.86
					NPP	.95	.84	.69

Note: PVT: Performance validity test; LEP: Limited English proficiency; SENS: Sensitivity; SPEC: Specificity, +LR: Positive likelihood ratio; -LR: Negative likelihood ratio; PPP: Positive predictive power; NPP: Negative predictive power; TOMM T1: Test of Memory Malingering Trial 1; DCT E-Score: Dot Counting Test Effort-Score; FIT: Rey 15-Item Test. **Bolded** = Best cutoffs when SENS ≥.50 and SPEC ≥.85 (liberal) and ≥.90 (conservative).

Table 21

Classification Accuracy of Embedded PVT$_{HVM}$ with Experimental Condition (Control vs. Malingering) as the Criterion Variable in LEP Sample (n = 70)

	Signal detection properties					Hypothetical base rates		
Test cutoff	SENS	SPEC	+LR	-LR		.10	.30	.50
Digit Span$_{WAIS-III}$								
ACSS ≤ 6	**.70**	**.98**	**28.00**	**.31**	PPP	.80	.94	.97
					NPP	.97	.88	.77
ACSS ≤ 7	**.80**	**.88**	**6.40**	**.23**	PPP	.43	.74	.87
					NPP	.98	.91	.81
RDS ≤ 6	**.63**	**1.00**	**N/A**	**.37**	PPP	1.00	1.00	1.00
					NPP	.96	.86	.73
RDS ≤ 7	**.70**	**.93**	**9.33**	**.32**	PPP	.53	.81	.91
					NPP	.97	.88	.76
Verbal Fluency								
FAS T-score ≤ 25	.23	.90	2.33	.85	PPP	.20	.50	.70
					NPP	.91	.73	.54
FAS T-score ≤ 26	.23	.88	1.87	.88	PPP	.18	.45	.66
					NPP	.91	.73	.53
Animals T-score ≤ 15	.33	.95	6.67	.70	PPP	.42	.74	.87
					NPP	.93	.77	.59
Animals T-score ≤ 19	.37	.90	3.67	.70	PPP	.29	.61	.79
					NPP	.93	.77	.59
LRE$_{Sugarman}$ ≥ .43	.40	.85	2.67	.71	PPP	.23	.53	.73
					NPP	.93	.77	.59
LRE$_{Sugarman}$ ≥ .48	.37	.93	4.89	.68	PPP	.37	.69	.84
					NPP	.93	.78	.60
LRE$_{Johnson}$ ≥ .70	.17	.85	1.11	.98	PPP	.11	.33	.53
					NPP	.90	.70	.51
LRE$_{Johnson}$ ≥ .72	.13	.90	1.33	.96	PPP	.13	.36	.57
					NPP	.90	.71	.51
CIM								
Raw ≤ 3	.27	.90	2.60	.82	PPP	.23	.54	.73
					NPP	.92	.74	.55
Raw ≤ 4	.30	.87	2.34	.80	PPP	.20	.50	.70
					NPP	.92	.74	.55
T-score ≤ 2	.17	.87	1.30	.96	PPP	.13	.36	.57
					NPP	.90	.71	.51
T-score ≤ 3	.30	.87	2.34	.80	PPP	.20	.50	.70
					NPP	.92	.74	.55
Stroop								
Color ≤ 2	.41	.95	8.28	.62	PPP	.48	.78	.89
					NPP	.94	.79	.62
Color ≤ 3	.48	.85	3.22	.61	PPP	.26	.58	.76
					NPP	.94	.79	.62
Word ≤ 6	**.59**	**.90**	**5.86**	**.46**	PPP	.40	.72	.86
					NPP	.95	.84	.69

	Signal detection properties					Hypothetical base rates		
Test cutoff	SENS	SPEC	+LR	-LR		.10	.30	.50
Word ≤ 7	**.66**	**.88**	**5.24**	**.39**	PPP	.38	.70	.85
					NPP	.96	.86	.72
INT ≤ 3	.30	.90	3.00	.78	PPP	.25	.56	.75
					NPP	.92	.75	.56
INT ≤ 4	.30	.88	2.40	.80	PPP	.22	.52	.71
					NPP	.92	.75	.56

Note: PVT: Performance validity test; LEP: Limited English proficiency; SENS: Sensitivity; SPEC: Specificity, +LR: Positive likelihood ratio; -LR: Negative likelihood ratio; PPP: Positive predictive power; NPP: Negative predictive power; Digit Span$_{\mathit{WAIS\text{-}III}}$: WAIS-III Digit Span subtest; RDS: Reliable Digit Span; FAS: Letter fluency test; Animals: Category animal fluency test; LRE$_{\mathit{Johnson}}$: Logistical regression equation combining overall letter fluency output and pattern of performance (Johnson et al., 2012); LRE$_{\mathit{Sugarman}}$: Logistical regression equation combining FAS and Animal Fluency T-scores (Sugarman & Axelrod, 2015); CIM: Complex Ideational Material; Stroop: D-KEFS Color-Word Interference Test; Color: Color Condition ACSS; Word: Word Condition ACSS; INT: Interference Condition ACSS, **Bolded** = Best cutoffs when SENS ≥.50 and SPEC ≥.85 (liberal) and ≥.90 (conservative).

Table 22

Classification Accuracy of Freestanding PVT$_{HVM}$ with Experimental Condition (Control vs. Malingering) as the Criterion Variable in LEP Sample (n = 70)

	Signal detection properties					Hypothetical base rates		
Test cutoff	SENS	SPEC	+LR	-LR		.10	.30	.50
WCT								
Accuracy ≤ 44	.70	.95	14.00	.32	PPP	.61	.86	.93
					NPP	.97	.88	.76
Accuracy ≤ 47	**.77**	**.93**	**10.22**	**.25**	PPP	.55	.83	.92
					NPP	.97	.90	.80
T2C ≤ 129	.48	.86	3.57	.60	PPP	.28	.60	.77
					NPP	.94	.79	.62
T2C ≤ 173	.24	.92	2.98	.83	PPP	.25	.56	.75
					NPP	.92	.74	.55
WRT								
Accuracy ≤ 6	**.50**	**.90**	**5.00**	**.56**	PPP	.36	.68	.83
					NPP	.94	.81	.64
Accuracy ≤ 7	**.63**	**.85**	**4.22**	**.43**	PPP	.32	.64	.81
					NPP	.95	.84	.70
Combination ≤ 6	.53	1.00	-	.47	PPP	1.00	1.00	1.00
					NPP	.95	.83	.68
Combination ≤ 7	**.57**	**.93**	**7.56**	**.47**	PPP	.48	.78	.89
					NPP	.95	.83	.68
Combination ≤ 8	**.63**	**.83**	**3.62**	**.44**	PPP	.29	.61	.79
					NPP	.95	.84	.69

Note: PVT: Performance validity test; LEP: Limited English proficiency; SENS: Sensitivity; SPEC: Specificity, +LR: Positive likelihood ratio; -LR: Negative likelihood ratio; PPP: Positive predictive power; NPP: Negative predictive power; WCT: Word Choice Test; T2C: Time to completion; WRT: Word Recognition Test, **Bolded** = Best cutoffs when SENS ≥.50 and SPEC ≥.85 (liberal) and ≥.90 (conservative).

Part 3: Does Malingering Manifest Differently as a Function of Language Proficiency?

Experimental-malingering profiles were examined *post-hoc* to determine whether malingering presents differently in individuals with LEP compared to NSE.

Results revealed that LEP-EM and NSE-EM participants performed similarly on PVT$_{LVM}$, with the exception of the RCFT (Copy and Immediate Recall trials), on which LEP participants produced significantly higher BR$_{Fail}$ (RR: 1.84-7.06; Table 23). In contrast, the malingering profile for LEP-EM and NSE-EM groups diverged on PVT$_{HVM}$, with the LEP-EM group having a higher BR$_{Fail}$ across most cutoffs (RDS, DS-ACSS, WRT Combination, Animals, LRE$_{Sugarman}$, and CIM) than the NSE-EM group (RR: 1.57-17.33; Table 24). No significant differences were observed on the WCT, WRT Recognition, FAS, LRE$_{Johnson}$, or Stroop. These findings were replicated with these PVTs as continuous variables (Tables 25-26).

In comparing the EM versus NC conditions within the LEP sample, it is evident that malingering is not captured on several PVT$_{HVM}$. Aside from the WCT, WRT, Digit Span, and Stroop indicators, BR$_{Fail}$ among LEP participants on PVT$_{HVM}$ at published cutoffs are very high (up to 82%) even in the NC condition, thus masking any effects of experimental malingering (Table 27). This is especially pronounced for FAS, Animals, and CIM, in which NC and EM groups are indistinguishable (RR: 1.14-1.48). This contrasts with the experimental malingering profile of NSE participants (Table 28), in which all PVT$_{HVM}$, except the LRE$_{Johnson}$, are able to capture malingering in the EM compared to the NC group. For PVT$_{LVM}$, LEP participants have an experimental-malingering profile (Table 29) largely comparable to NSE participants (Table 30), such that the EM group performed worse on all tests than the NC group. These findings were replicated using PVT$_{LVM}$ as continuous variables in both the LEP (Tables 31-32) and NSE (Tables 33-34) groups and are consistent with classification accuracy data (AUC, sensitivity and specificity) presented earlier.

Table 23

Comparing Instrument-Level BR$_{Fail}$ on Tests of Low Verbal Mediation as a Function of the English Proficiency Group in the Experimental Malingering Condition (n = 60)

			English Proficiency					
			LEP	NSE				
			(*n* = 30)	(*n* = 30)				
PVT	Score	Cutoff^a	BR$_{Fail}$	BR$_{Fail}$	RR	χ^{2b}	*p*	Φ^2
TOMM	T1	≤44	83.3	90.0	0.93		.71	.01
		≤39	80.0	80.0	1.00	-	-	-
DCT	E-score	≥15	70.0	63.3	1.11	0.30	.58	.01
		≥17	66.7	46.7	1.43	2.44	.12	.04
FIT	Recall	<10	24.1	13.3	1.81	1.14	.29	.02
		<9	20.7	10.0	2.07		.30	.02
	Recognition	<11	37.9	26.7	1.42	0.86	.36	.01
		<10	27.6	20.0	1.38	0.47	.49	.01
	Combined	<21	27.6	23.3	1.18	0.14	.71	<.01
		<20	27.6	20.0	1.38	0.47	.49	.01
TMT	A T-score	≤39	86.7	83.3	1.04		1.00	<.01
		≤34	70.0	70.0	1.00	-	-	-
	B T-score	≤37	63.3	56.7	1.12	0.28	.60	.00
		≤30	46.7	23.3	2.00	3.59	.06	.06
	A + B Raw	≥137	46.7	36.7	1.27	0.62	.43	.01
		≥170	40.0	20.0	2.00	2.86	.09	.05
CD$_{WAIS-III}$	ACSS	≤5	40.0	43.3	0.92	0.07	.79	<.01
		≤4	30.0	30.0	1.00	-	-	-
SS$_{WAIS-IV}$	ACSS	≤6	43.3	36.7	1.18	0.28	.60	<.01
		≤5	33.3	36.7	0.91	0.07	.79	<.01
RCFT	Copy	≤26	36.7	20.0	1.84	2.05	.15	.03
		≤23	23.3	3.3	7.06		.05	.09
	IR	≤10	23.3	6.7	3.48		.15	.05
		≤9.5	23.3	3.3	7.06		.05	.09
	Recog	≤16	36.7	16.7	2.20	3.07	.08	.05
		≤15	16.7	16.7	1.00	-	-	-
	Equation	≤47	36.7	30.0	1.22	0.30	.58	.01
		≤45	33.3	23.3	1.43	0.74	.39	.01

^{a}First row = Liberal cutoff; Second row = Conservative cutoff; ^{b}Fisher's Exact Test calculated when Chi-Square assumptions violated (e.g., expected frequencies > 5).

Note: PVT: Performance Validity Test; LEP: Limited English proficiency; NSE: Native speakers of English; BR$_{Fail}$: Base rate of failure; TOMM T1: Test of Memory Malingering Trial 1; DCT E-Score: Dot Counting Test Effort-Score; FIT: Rey 15-Item Test; TMT A T-score, B T-score: Trail Making Test Part A and Part B T-score; A + B: Trail Making Test Trial A & B Total Combined Time Score; CD$_{WAIS-III}$: WAIS-III Digit Symbol subtest; SS$_{WAIS-IV}$: WAIS-IV Symbol Search subtest; RCFT: Rey-Osterrieth Complex Figure Test; Copy: Copy Trial raw score; IR: Immediate Recall raw score; Recog: Recognition Trial raw; Equation: CT raw score + (true positive recognition – Atypical recognition errors) x 3 (Lu et al., 2003).

Table 24

Comparing Instrument-Level BR$_{Fail}$ on Tests of High Verbal Mediation as a Function of the English Proficiency Group in the Experimental Malingering Condition (n = 60)

			English Proficiency					
			LEP	NSE				
			(*n* = 30)	(*n* = 30)				
PVT	Score	Cutoff^a	BR$_{Fail}$	BR$_{Fail}$	RR	χ^{2b}	*p*	Φ^2
WCT	Accuracy	≤47	76.7	76.7	1.00	-	-	-
		≤43	70.0	63.3	1.11	0.30	.58	.01
	T2C	≥156	24.1	34.5	0.70	0.75	.39	.01
		≥171	24.1	27.6	0.87	0.09	.76	<.01
Digit Span$_{WAIS-III}$	RDS	≤7	70.0	50.0	1.40	2.50	.11	.04
		≤6	63.3	30.0	2.11	6.70	.01	.11
	ACSS	≤6	70.0	36.7	1.91	6.70	.01	.11
		≤5	50.0	26.7	1.87	3.46	.06	.06
WRT	Recognition	≤7	63.3	40.0	1.58	3.27	.07	.05
		≤5	33.3	23.3	1.43	0.74	.39	.01
	Combination	≤10	66.7	46.7	1.43	2.44	.12	.04
		≤8	63.3	36.7	1.72	4.27	.04	.07
FAS	T-score	≤33	46.7	36.7	1.27	0.62	.43	.01
		≤31	40.0	23.3	1.72	1.93	.17	.03
Animals	T-score	≤33	73.3	46.7	1.57	4.44	.04	.07
		≤31	73.3	40.0	1.83	6.79	.01	.11
	LRE$_{Johnson}$	≥.45	40.0	30.0	1.33	0.66	.42	.01
		≥.475	30.0	13.3	2.26	2.46	.12	.04
	LRE$_{Sugarman}$	≥.5	36.7	20.0	1.84	2.05	.15	.03
		≥.6	36.7	13.3	2.76	4.36	.04	.07
CIM	Raw	≤9	93.3	43.3	2.15	17.33	<.01	.29
		≤8	83.3	36.7	2.27	13.61	<.01	.23
	T-score	≤29	93.3	46.7	2.00	15.56	<.01	.26
		≤23	93.3	43.3	2.15	17.33	<.01	.29
Stroop	Color	≤7	82.8	70.0	1.18	1.33	.25	.02
		≤5	62.1	60.0	1.04	0.03	.87	<.01
	Word	≤7	65.5	66.7	0.98	0.01	.93	<.01
		≤5	55.2	63.3	0.87	0.41	.52	.01
	INT	≤7	53.3	56.7	0.94	0.07	.80	<.01
		≤5	53.3	33.3	1.60	2.44	.12	.04

^{a}First row = Liberal cutoff; Second row = Conservative cutoff; ^{b}Fisher's Exact Test calculated when Chi-Square assumptions violated (e.g., expected frequencies > 5).

Note: PVT: Performance Validity Test; LEP: Limited English proficiency; NSE: Native speakers of English; BR$_{Fail}$: Base rate of failure; WCT: Word Choice Test; T2C: Time to completion (seconds); Digit Span$_{WAIS-III}$: WAIS-III Digit Span subtest; RDS: Reliable Digit Span; ACSS: Age-corrected scaled-score; WRT: Word Recognition Test; FAS: Letter fluency test; Animals: Category animal fluency test; LRE$_{Johnson}$: Logistical regression equation combining overall letter fluency output and pattern of performance (Johnson et al., 2012); LRE$_{Sugarman}$: Logistical regression equation combining FAS and Animal Fluency T-scores (Sugarman & Axelrod, 2015) ; CIM: Complex Ideational Material; Stroop: D-KEFS Color-Word Interference Test; Color: Color Condition ACSS; Word: Word Condition ACSS; INT: Interference Condition ACSS.

Table 25

Descriptive Statistics & Independent t-Tests for Tests of High Verbal Mediation as a Function of the English Proficiency Group in the Experimental Malingering Condition (n = 60)

		English Proficiency						
		LEP (*n* = 30)		NSE (*n* = 30)				
Measure	Score	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
WCT	Accuracy	34.8	12.4	38.0	11.5	1.03	.31	0.27
	T2C	135.0	72.5	137.4	50.3	0.14	.89	0.04
Digit Span$_{WAIS\text{-}III}$	Total Raw	10.6	4.3	13.3	4.5	2.45	.02	0.61
	RDS	6.2	2.4	7.6	2.3	2.32	.02	0.60
	ACSS	5.5	2.6	7.3	2.9	2.52	.01	0.65
WRT	Recognition	7.0	2.6	8.1	3.3	1.40	.17	0.37
	Combination	7.3	5.6	10.5	5.5	2.26	.03	0.58
FAS	Raw	28.1	11.2	32.2	10.0	1.49	.14	0.39
	T-score	32.0	10.8	36.5	9.3	1.73	.09	0.45
Animals	Raw	13.2	5.6	17.1	5.3	2.74	.01	0.72
	T-score	23.2	13.3	33.2	13.3	2.92	<.01	0.75
Emotional Fluency	Raw	7.0	3.3	10.3	5.0	3.00	<.01	0.78
CIM	Raw	5.7	2.6	8.5	3.3	3.67	<.01	0.94
	T-score	9.4	9.5	26.5	17.3	4.75	<.01	1.23
Stroop	Color Raw	50.1	22.8	43.1	17.1	-1.35	.18	0.35
	Color ACSS	4.2	3.2	5.3	4.2	1.08	.28	0.29
	Word Raw	36.0	18.6	35.8	14.8	-0.05	.96	0.01
	Word ACSS	5.2	4.4	5.4	5.8	0.17	.87	0.04
	INT Raw	75.7	40.0	68.9	27.8	-0.77	.45	0.20
	INT ACSS	6.5	4.2	6.9	3.9	0.35	.73	0.10
BNT-15	Accuracy	6.3	3.4	12.8	3.1	7.85	<.01	2.00
	T2C	200.9	57.8	87.9	62.4	-7.27	<.01	1.88
Reading$_{WRAT\text{-}4}$	SS	84.7	11.4	100.1	15.4	4.39	<.01	1.14

Note: LEP: Limited English proficiency; NSE: Native speakers of English; WCT: Word Choice Test; T2C: Time to completion (seconds); Digit Span$_{WAIS\text{-}III}$: WAIS-III Digit Span subtest; RDS: Reliable Digit Span; ACSS: Age-corrected scaled score (*M* = 10, *SD* = 3); WRT: Word Recognition Test; FAS: Letter fluency test; Animals: Category animal fluency test; T-score (*M* = 50, *SD* = 10); Emotional Fluency: Category emotional fluency test; CIM: Complex Ideational Material; Stroop: D-KEFS Color-Word Interference Test; INT: Interference Condition; BNT-15: Boston Naming Test 15-Item Short Form; Reading$_{WRAT\text{-}4}$: Wide Range Achievement Test 4th Edition Reading subtest; SS: Standard score (*M* = 100; *SD* = 15).

Table 26

Descriptive Statistics & Independent t-Tests for Tests of Low Verbal Mediation as a Function of the English Proficiency Group in the Experimental Malingering Condition (n = 60)

		English Proficiency						
		LEP (*n* = 30)		NSE (*n* = 30)				
Measure	Score	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
TOMM	T1	29.3	11.0	32.1	10.2	1.04	.30	0.26
DCT	E-score	21.2	9.8	18.1	7.0	-1.40	.17	0.36
FIT	Recall	12.7	3.2	12.9	2.9	0.22	.83	0.07
	Recognition	9.1	5.0	11.1	3.8	1.78	.08	0.45
	Combined	21.4	8.6	23.9	6.9	1.23	.22	0.32
TMT	A Raw	63.2	49.7	50.9	34.5	-1.11	.27	0.29
	A T-Score	23.7	15.0	28.2	13.1	1.23	.22	0.32
	B Raw	118.1	71.5	83.3	38.8	-2.34	.02	0.60
	B T-Score	30.3	14.2	38.0	11.6	2.27	.03	0.59
CD$_{WAIS\text{-}III}$	Raw	55.0	19.9	57.8	20.4	0.55	.58	0.10
	ACSS	6.4	2.9	6.7	3.2	0.34	.74	0.39
	Recognition	5.3	1.6	6.1	2.4	1.56	.12	0.09
SS$_{WAIS\text{-}IV}$	Raw	23.8	11.9	24.9	12.2	0.34	.73	0.08
	ACSS	6.9	3.6	7.2	4.0	0.30	.76	0.39
RCFT	Copy	27.0	7.3	29.3	3.9	1.54	.13	0.39
	T2C	126.8	53.6	129.5	59.6	0.18	.85	0.05
	IR Raw	15.2	6.4	17.9	5.4	1.78	.08	0.46
	IR T-Score	30.9	11.7	33.6	12.4	0.88	.38	0.22
	DR Raw	13.9	7.0	16.5	6.0	1.58	.12	0.40
	DR T-Score	28.9	11.5	31.7	11.8	0.92	.36	0.24
	Recog Raw	17.3	4.2	18.2	3.0	0.92	.36	0.25
	Recog T-Score	31.2	12.1	31.5	10.8	0.11	.91	0.03
CDT	Raw	7.7	2.4	8.7	1.5	1.95	.06	0.50

Note: LEP: Limited English proficiency; NSE: Native speakers of English; TOMM T1: Test of Memory Malingering Trial 1; DCT E-Score: Dot Counting Test Effort-Score; FIT: Rey 15-Item Test; TMT-A, TMT-B: Trail Making Test Part A and Part B; T-score (*M* = 50, *SD* = 10); CD$_{WAIS\text{-}III}$: WAIS-III Digit Symbol subtest; ACSS: Age-corrected scaled score (*M* = 10, *SD* = 3); SS$_{WAIS\text{-}IV}$: WAIS-IV Symbol Search subtest; RCFT: Rey-Osterrieth Complex Figure Test; T2C: Time to completion; IR: Immediate Recall; DR: Delayed Recall; Recog: Recognition; CDT: Clock Drawing Test.

Table 27

Comparing Instrument-Level BR$_{Fail}$ on Tests of High Verbal Mediation as a Function of the Experimental Condition in the LEP Sample (n = 70)

			Exp. Condition					
			NC	EM				
			(*n* = 40)	(*n* = 30)				
PVT	Score	Cutoff[a]	BR$_{Fail}$	BR$_{Fail}$	RR	χ^{2b}	*p*	Φ^2
WCT	Accuracy	≤47	7.5	76.7	10.23	35.13	<.01	.50
		≤43	2.5	70.0	28.00	36.24	<.01	.52
	T2C	≥156	8.1	24.1	2.98	3.25	.07	.05
		≥171	8.1	24.1	2.98	3.25	.07	.05
Digit Span$_{WAIS-III}$	RDS	≤7	7.5	70.0	9.33	29.72	<.01	.43
		≤6	0.0	63.3	-	34.77	<.01	.50
	ACSS	≤6	2.5	70.0	28.00	36.24	<.01	.52
		≤5	0.0	50.0		25.46	<.01	.36
WRT	Recognition	≤7	15.0	63.3	4.22	17.44	<.01	.25
		≤5	0.0	33.3	-	15.56	<.01	.22
	Combination	≤10	25.0	66.7	2.67	12.15	<.01	.17
		≤8	17.5	63.3	3.62	15.43	<.01	.22
FAS	T-score	≤33	40.0	46.7	1.17	0.31	.58	.00
		≤31	30.0	40.0	1.33	0.76	.38	.01
Animals	T-score	≤33	62.5	73.3	1.17	0.91	.34	.01
		≤31	57.5	73.3	1.27	1.87	.17	.03
	LRE$_{Johnson}$	≥.5	35.0	40.0	1.14	0.18	.67	.00
		≥.6	20.0	30.0	1.50	0.93	.33	.01
	LRE$_{Sugarman}$	≥.45	10.0	36.7	3.67	7.24	<.01	.10
		≥.475	7.5	36.7	4.89	9.12	<.01	.13
CIM	Raw	≤9	82.1	93.3	1.14	1.90	.17	.03
		≤8	56.4	83.3	1.48	5.66	.02	.08
	T-score	≤29	82.1	93.3	1.14	1.90	.17	.03
		≤23	79.5	93.3	1.17	2.62	.11	.04
Stroop	Color	≤7	52.5	82.8	1.58	6.79	<.01	.10
		≤5	27.5	62.1	2.26	8.25	<.01	.12
	Word	≤7	12.5	65.5	5.24	20.83	<.01	.30
		≤5	7.5	55.2	7.36	19.15	<.01	.28
	INT	≤7	30.0	53.3	1.77	3.89	.05	.06
		≤5	20.0	53.3	2.66	8.45	<.01	.12

[a]First row = Liberal cutoff; Second row = Conservative cutoff; [b]Fisher's Exact Test calculated when Chi-Square assumptions violated (e.g., expected frequencies < 5).

Note: LEP: Limited English proficiency; Exp. Condition: Experimental condition; NC: Non-Malingering Control; EM: Experimental Malingering; BR$_{Fail}$: Base rate of failure; WCT: Word Choice Test; T2C: Time to completion (seconds); Digit Span$_{WAIS-III}$: WAIS-III Digit Span subtest; RDS: Reliable Digit Span; ACSS: Age-corrected scaled-score; WRT: Word Recognition Test; FAS: Letter fluency test; Animals: Category animal fluency test; LRE$_{Johnson}$: Logistical regression equation combining overall letter fluency output and pattern of performance (Johnson et al., 2012); LRE$_{Sugarman}$: Logistical regression equation combining FAS and Animal Fluency T-scores (Sugarman & Axelrod, 2015) ; CIM: Complex Ideational Material; Stroop: D-KEFS Color-Word Interference Test; Color: Color Condition ACSS; Word: Word Condition ACSS; Inhibition: Inhibition Condition ACSS.

Table 28

Comparing Instrument-Level BR$_{Fail}$ on Tests of High Verbal Mediation as a Function of the Experimental Condition in the NSE Sample (n = 70)

			Exp. Condition					
			NC	EM				
			(*n* = 40)	(*n* = 30)				
PVT	Score	Cutoff[a]	BR$_{Fail}$	BR$_{Fail}$	RR	χ^{2b}	*p*	Φ^2
WCT	Accuracy	≤47	2.5	76.7	30.68	41.85	<.01	.60
		≤43	0.0	63.3		34.77	<.01	.50
	T2C	≥156	0.0	34.5		15.40	<.01	.23
		≥171	0.0	27.6			<.01	.18
Digit Span$_{WAIS-III}$	RDS	≤7	10.0	50.0	5.00	13.87	<.01	.20
		≤6	2.5	30.0	12.00	10.59	<.01	.15
	ACSS	≤6	12.5	36.7	2.94	5.68	.02	.08
		≤5	5.0	26.7	5.34	6.57	.01	.09
WRT	Recognition	≤7	2.5	40.0	16.00	15.94	<.01	.23
		≤5	0.0	23.3	-		<.01	.15
	Combination	≤10	5.0	46.7	9.34	16.88	<.01	.24
		≤8	0.0	36.7	-	17.40	<.01	.25
FAS	T-score	≤33	15.0	36.7	2.45	4.38	.04	.06
		≤31	12.5	23.3	1.86	1.42	.23	.02
Animals	T-score	≤33	15.0	46.7	3.11	8.42	<.01	.12
		≤31	12.5	40.0	3.20	7.05	<.01	.10
	LRE$_{Johnson}$	≥.5	15.0	30.0	2.00	2.29	.13	.03
		≥.6	5.0	13.3	2.66		.39	.02
	LRE$_{Sugarman}$	≥.45	2.5	20.0	8.00		.04	.08
		≥.475	2.5	13.3	5.32		.16	.04
CIM	Raw	≤9	7.5	43.3	5.77	12.48	<.01	.18
		≤8	7.5	36.7	4.89	9.12	<.01	.13
	T-score	≤29	10.0	46.7	4.67	12.07	<.01	.17
		≤23	7.5	43.3	5.77	12.48	<.01	.18
Stroop	Color	≤7	15.0	70.0	4.67	21.89	<.01	.31
		≤5	0.0	60.0	-	32.31	<.01	.46
	Word	≤7	10.0	66.7	6.67	24.43	<.01	.35
		≤5	5.0	63.3	12.66	27.78	<.01	.40
	INT	≤7	2.5	56.7	30.68	26.33	<.01	.38
		≤5	2.5	33.3	-	12.31	<.01	.18

[a]First row = Liberal cutoff; Second row = Conservative cutoff; [b]Fisher's Exact Test calculated when Chi-Square assumptions violated (e.g., expected frequencies < 5).

Note: NSE: Native speakers of English; Exp. Condition: Experimental condition; NC: Non-Malingering Control; EM: Experimental Malingering; BR$_{Fail}$: Base rate of failure; WCT: Word Choice Test; T2C: Time to completion (seconds); Digit Span$_{WAIS-III}$: WAIS-III Digit Span subtest; RDS: Reliable Digit Span; ACSS: Age-corrected scaled-score; WRT: Word Recognition Test; FAS: Letter fluency test; Animals: Category animal fluency test; LRE$_{Johnson}$: Logistical regression equation combining overall letter fluency output and pattern of performance (Johnson et al., 2012); LRE$_{Sugarman}$: Logistical regression equation combining FAS and Animal Fluency T-scores (Sugarman & Axelrod, 2015) ; CIM: Complex Ideational Material; Stroop: D-KEFS Color-Word Interference Test; Color: Color Condition ACSS; Word: Word Condition ACSS; INT: Interference Condition ACSS.

Table 29

Comparing Instrument-Level BR_{Fail} on Tests of Low Verbal Mediation as a Function of the Experimental Condition in the LEP Sample (n = 70)

			Exp. Condition					
			NC	EM				
			(*n* = 40)	(*n* = 30)				
PVT	Score	Cutoff[a]	BR_{Fail}	BR_{Fail}	RR	χ^{2b}	*p*	Φ^2
TOMM	T1	≤44	5.0	83.3	16.66	44.40	<.01	.63
		≤39	0.0	80.0	-	48.70	<.01	.70
DCT	E-score	≥15	7.5	70.0	9.33	29.72	<.01	.43
		≥17	2.5	66.7	26.68	33.61	<.01	.48
FIT	Recall	<10	0.0	24.1	-		<.01	.16
		<9	0.0	20.7	-		<.01	.13
	Recognition	<11	5.0	37.9	7.58	11.92	<.01	.17
		<10	2.5	27.6	11.04	9.33	<.01	.14
	Combined	<21	2.5	27.6	11.04	9.34	<.01	.14
		<20	2.5	27.6	11.04	9.33	<.01	.14
TMT	A T-score	≤39	50.0	86.7	1.73	10.23	<.01	.15
		≤34	32.5	70.0	2.15	9.65	<.01	.14
	B T-score	≤37	40.0	63.3	1.58	3.73	.05	.05
		≤30	5.0	46.7	9.34	16.88	<.01	.24
	A + B Raw	≥137	5.0	46.7	9.34	16.88	<.01	.24
		≥170	2.5	40.0	16.00	15.94	<.01	.23
$CD_{WAIS\text{-}III}$	ACSS	≤5	0.0	40.0	-	19.31	<.01	.28
		≤4	0.0	30.0	-	13.77	<.01	.20
$SS_{WAIS\text{-}IV}$	ACSS	≤6	2.5	43.3	17.32	17.87	<.01	.26
		≤5	2.5	33.3	13.32	12.31	<.01	.18
RCFT	Copy	≤26	0.0	36.7	-	17.40	<.01	.25
		≤23	0.0	23.3	-	10.37	<.01	.15
	IR	≤10	0.0	23.3	-		<.01	.15
		≤9.5	0.0	23.3	-		<.01	.15
	Recog	≤16	5.3	36.7	6.92	10.69	<.01	.16
		≤15	5.3	16.7	3.15		.23	.03
	Equation	≤47	2.6	36.7	14.12	13.36	<.01	.20
		≤45	2.6	33.3	12.81	11.65	<.01	.17

[a] First row = Liberal cutoff; Second row = Conservative cutoff; [b] Fisher's Exact Test calculated when Chi-Square assumptions violated (e.g., expected frequencies < 5).

Note: LEP: Limited English proficiency; Exp. Condition: Experimental condition; NC: Non-Malingering Control; EM: Experimental Malingering; BR_{Fail}: Base rate of failure; TOMM T1: Test of Memory Malingering Trial 1; DCT E-Score: Dot Counting Test Effort-Score; FIT: Rey 15-Item Test; TMT A T-score, B T-score: Trail Making Test Part A and Part B T-score; A + B: Trail Making Test Trial A & B Total Combined Time Score; $CD_{WAIS\text{-}III}$: WAIS-III Digit Symbol subtest; $SS_{WAIS\text{-}IV}$: WAIS-IV Symbol Search subtest; RCFT: Rey-Osterrieth Complex Figure Test; Copy: Copy Trial raw score; IR: Immediate Recall raw score; Recog: Recognition Trial raw; Equation: CT raw score + (true positive recognition – Atypical recognition errors) x 3 (Lu et al., 2003).

Table 30

Comparing Instrument-Level BR_{Fail} on Tests of Low Verbal Mediation as a Function of the Experimental Condition in the NSE Sample (n = 70)

			English Proficiency					
			NC	EM				
			(*n* = 40)	(*n* = 30)				
PVT	Score	Cutoff[a]	BR_{Fail}	BR_{Fail}	RR	χ^{2b}	*p*	Φ^2
TOMM	T1	≤44	10.0	90.0	9.00	44.45	<.01	.64
		≤39	5.0	80.0	16.00	41.30	<.01	.59
DCT	E-score	≥15	5.0	63.3	12.66	23.78	<.01	.40
		≥17	2.5	46.7	18.68	19.82	<.01	.28
FIT	Recall	<10	0.0	13.3	-		.03	.08
		<9	0.0	10.0	-		.07	.06
	Recognition	<11	0.0	26.7	-		<.01	.17
		<10	0.0	20.0	-		<.01	.13
	Combined	<21	0.0	23.3	-	10.37	<.01	.15
		<20	0.0	20.0	-		<.01	.13
TMT	A T-score	≤39	32.5	83.3	2.56	17.85	<.01	.26
		≤34	25.0	70.0	2.80	14.07	<.01	.20
	B T-score	≤37	22.5	56.7	2.52	8.57	<.01	.12
		≤30	5.0	23.3	4.66	51.43	.02	.07
	A + B Raw	≥137	7.5	36.7	4.89	9.11	<.01	.13
		≥170	2.5	20.0	8.00		.04	.08
$CD_{WAIS-III}$	ACSS	≤5	2.5	43.3	17.32	17.87	<.01	.26
		≤4	2.5	30.0	12.00	10.59	<.01	.15
$SS_{WAIS-IV}$	ACSS	≤6	2.5	36.7	14.68	14.09	<.01	.20
		≤5	0.0	36.7	-	17.40	<.01	.25
RCFT	Copy	≤26	5.0	20.0	4.00	3.81	.05	.05
		≤23	0.0	3.3	-		.43	.02
	IR	≤10	7.5	6.7	.89		1.00	.00
		≤9.5	7.5	3.3	.44		.63	.01
	Recog	≤16	2.5	16.7	6.68		.04	.06
		≤15	0.0	16.7	-		.01	.10
	Equation	≤47	2.5	30.0	12.00	10.59	<.01	.15
		≤45	2.5	23.3	9.32		.02	.10

[a] First row = Liberal cutoff; Second row = Conservative cutoff; [b] Fisher's Exact Test calculated when Chi-Square assumptions violated (e.g., expected frequencies < 5).

Note: NSE: Native speakers of English; Exp. Condition: Experimental condition; NC: Non-Malingering Control; EM: Experimental Malingering; BR_{Fail}: Base rate of failure; TOMM T1: Test of Memory Malingering Trial 1; DCT E-Score: Dot Counting Test Effort-Score; FIT: Rey 15-Item Test; TMT A T-score, B T-score: Trail Making Test Part A and Part B T-score; A + B: Trail Making Test Trial A & B Total Combined Time Score; $CD_{WAIS-III}$: WAIS-III Digit Symbol subtest; $SS_{WAIS-IV}$: WAIS-IV Symbol Search subtest; RCFT: Rey-Osterrieth Complex Figure Test; Copy: Copy Trial raw score; IR: Immediate Recall raw score; Recog: Recognition Trial raw; Equation: CT raw score + (true positive recognition – Atypical recognition errors) x 3 (Lu et al., 2003).

Table 31

Descriptive Statistics & Independent t-Tests Comparing Scores Across Tests of High Verbal Mediation as a Function of the Experimental Condition in the LEP Sample (n = 70)

		Exp. Condition						
		NC (*n* = 40)		EM (*n* = 30)				
Measure	Score	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
WCT	Accuracy	49.1	1.5	34.8	12.4	6.29	<.01	1.62
	T2C	95.8	38.1	135.0	72.5	-2.64	.01	0.68
Digit Span$_{WAIS\text{-}III}$	Total Raw	17.0	3.7	10.6	4.3	6.70	<.01	1.60
	RDS	9.8	2.1	6.2	2.4	6.69	<.01	1.60
	ACSS	9.7	2.6	5.5	2.6	6.61	<.01	1.62
WRT	Recognition	10.1	2.3	7.0	2.6	5.23	<.01	1.26
	Combination	14.2	4.3	7.3	5.6	5.83	<.01	1.38
FAS	Raw	30.5	6.5	28.1	11.2	1.06	.30	0.26
	T-score	34.0	5.9	32.0	10.8	0.90	.37	0.23
Animals	Raw	16.3	3.7	13.2	5.6	2.60	.01	0.65
	T-score	30.0	9.9	23.2	13.3	2.35	.02	0.58
Emotional Fluency	Raw	8.1	2.7	7.0	3.3	1.48	.14	0.36
CIM	Raw	7.4	2.7	5.7	2.6	2.55	.01	0.64
	T-score	16.3	15.4	9.4	9.5	2.30	.02	0.54
Stroop	Color Raw	33.9	6.9	50.1	22.8	-3.72	<.01	0.96
	Color ACSS	7.2	3.0	4.2	3.2	3.98	<.01	0.97
	Word Raw	22.1	3.8	36.0	18.6	-3.96	<.01	1.04
	Word ACSS	10.0	2.3	5.2	4.4	5.39	<.01	1.37
	INT Raw	57.1	17.1	75.7	40.0	-2.39	.02	0.60
	INT ACSS	8.7	3.6	6.5	4.2	2.31	.02	0.56
BNT-15	Accuracy	6.5	3.1	6.3	3.4	0.27	.79	0.06
	T2C	185.8	57.2	200.9	57.8	-1.09	.28	0.26
Reading$_{\text{WRAT-4}}$	SS	85.4	10.2	84.7	11.4	0.26	.80	0.06

Note: LEP: Limited English proficiency; Exp. Condition: Experimental condition; NC: Non-Malingering Control; EM: Experimental Malingering; WCT: Word Choice Test; T2C: Time to completion (seconds); Digit Span$_{WAIS\text{-}III}$: WAIS-III Digit Span subtest; RDS: Reliable Digit Span; ACSS: Age-corrected scaled score (*M* = 10, *SD* = 3); WRT: Word Recognition Test; FAS: Letter fluency test; Animals: Category animal fluency test; T-score (*M* = 50, *SD* = 10); Emotional Fluency: Category emotional fluency test; CIM: Complex Ideational Material; Stroop: D-KEFS Color-Word Interference Test; INT: Interference Condition; BNT-15: Boston Naming Test 15-Item Short Form; Reading$_{\text{WRAT-4}}$: Wide Range Achievement Test 4th Edition Reading subtest; SS: Standard score (*M* = 100; *SD* = 15).

Table 32

Descriptive Statistics & Independent t-Tests Comparing Scores Across Tests of Low Verbal Mediation as a Function of the Experimental Condition in the LEP Sample (n = 70)

		Exp. Condition						
		NC (*n* = 40)		EM (*n* = 30)				
Measure	Score	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
TOMM	T1	48.6	2.0	29.3	11.0	9.53	<.01	2.44
DCT	E-score	11.1	2.3	21.2	9.8	-5.56	<.01	1.42
FIT	Recall	14.6	1.1	12.7	3.2	3.03	<.01	0.79
	Recognition	14.1	2.0	9.1	5.0	5.10	<.01	1.31
	Combined	28.7	2.7	21.4	8.6	4.42	<.01	1.15
TMT	A Raw	32.5	18.8	63.2	49.7	-3.21	<.01	0.82
	A T-Score	37.9	10.8	23.7	15.0	4.41	<.01	1.09
	B Raw	62.9	19.3	118.1	71.5	-4.11	<.01	1.05
	B T-Score	42.9	9.6	30.3	14.2	4.17	<.01	1.04
CD$_{WAIS\text{-}III}$	Raw	86.7	13.1	55.0	19.9	8.02	<.01	1.88
	ACSS	11.5	2.6	6.4	2.9	7.68	<.01	1.85
	Recognition	7.8	1.2	5.3	1.6	7.41	<.01	1.77
SS$_{WAIS\text{-}IV}$	Raw	36.3	6.3	23.8	11.9	5.22	<.01	1.31
	ACSS	11.0	2.3	6.9	3.6	5.39	<.01	1.36
RCFT	Copy	34.2	1.7	27.0	7.3	5.35	<.01	1.36
	T2C	171.4	113.8	126.8	53.6	1.99	.05	0.50
	IR Raw	23.0	4.9	15.2	6.4	5.79	<.01	1.37
	IR T-Score	45.2	12.4	30.9	11.7	4.89	<.01	1.19
	DR Raw	22.6	4.6	13.9	7.0	5.92	<.01	1.47
	DR T-Score	44.2	11.2	28.9	11.5	5.58	<.01	1.35
	Recog Raw	20.1	2.5	17.3	4.2	3.42	<.01	0.81
	Recog T-Score	41.5	14.3	31.2	12.1	3.16	<.01	0.78
CDT	Raw	8.4	1.5	7.7	2.4	1.59	.12	.35

Note: LEP: Limited English proficiency; Exp. Condition: Experimental condition; NC: Non-Malingering Control; EM: Experimental Malingering; TOMM T1: Test of Memory Malingering Trial 1; DCT E-Score: Dot Counting Test Effort-Score; FIT: Rey 15-Item Test; TMT-A, TMT-B: Trail Making Test Part A and Part B; T-score (*M* = 50, *SD* = 10);CD$_{WAIS\text{-}III}$: WAIS-III Digit Symbol subtest; ACSS: Age-corrected scaled score (*M* = 10, *SD* = 3); SS$_{WAIS\text{-}IV}$: WAIS-IV Symbol Search subtest; RCFT: Rey-Osterrieth Complex Figure Test; T2C: Time to completion; IR: Immediate Recall; DR: Delayed Recall; Recog: Recognition; CDT: Clock Drawing Test.

Table 33

Descriptive Statistics & Independent t-Tests Comparing Scores Across Tests of High Verbal Mediation as a Function of the Experimental Condition in the NSE Sample (n = 70)

		Exp. Condition						
		NC (*n* = 40)		EM (*n* = 30)				
Measure	Score	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
WCT	Accuracy	49.7	0.7	38.0	11.5	5.54	<.01	1.44
	T2C	72.4	20.6	137.4	50.3	-6.56	<.01	1.69
Digit Span$_{\textit{WAIS-III}}$	Total Raw	17.4	4.0	13.3	4.5	4.01	<.01	0.96
	RDS	10.0	2.1	7.6	2.3	4.55	<.01	1.09
	ACSS	10.0	2.7	7.3	2.9	3.96	<.01	0.96
WRT	Recognition	10.8	1.7	8.1	3.3	4.06	<.01	1.03
	Combination	15.8	3.3	10.5	5.5	4.72	<.01	1.17
FAS	Raw	39.7	10.8	32.2	10.0	2.96	<.01	0.72
	T-score	43.0	9.9	36.5	9.3	2.80	.01	0.68
Animals	Raw	23.6	5.3	17.1	5.3	5.07	<.01	1.23
	T-score	46.9	10.9	33.2	13.3	4.70	<.01	1.13
Emotional Fluency	Raw	13.3	6.1	10.3	5.0	2.25	.03	0.54
CIM	Raw	11.1	1.2	8.5	3.3	4.55	<.01	1.05
	T-score	44.2	13.8	26.5	17.3	4.76	<.01	1.13
Stroop	Color Raw	27.5	4.6	43.1	17.1	-5.53	<.01	1.25
	Colors ACSS	10.1	2.1	5.3	4.2	6.22	<.01	1.45
	Word Raw	20.9	4.1	35.8	14.8	-5.37	<.01	1.37
	Word ACSS	10.7	2.4	5.4	5.8	4.72	<.01	1.19
	INT Raw	43.8	8.6	68.9	27.8	-4.78	<.01	1.22
	INT ACSS	11.8	2.0	6.9	3.9	6.29	<.01	1.58
BNT-15	Accuracy	13.9	1.2	12.8	3.1	1.76	.09	0.47
	T2C	43.4	27.8	87.9	62.4	-3.64	<.01	0.92
Reading$_{\text{WRAT-4}}$	SS	102.7	11.7	100.1	15.4	0.82	.41	0.19

Note: NSE: Native speakers of English; Exp. Condition: Experimental condition; NC: Non-Malingering Control; EM: Experimental Malingering; WCT: Word Choice Test; T2C: Time to completion (seconds); Digit Span$_{\textit{WAIS-III}}$: WAIS-III Digit Span subtest; RDS: Reliable Digit Span; ACSS: Age-corrected scaled score (*M* = 10, *SD* = 3); WRT: Word Recognition Test; FAS: Letter fluency test; Animals: Category animal fluency test; T-score (*M* = 50, *SD* = 10); Emotional Fluency: Category emotional fluency test; CIM: Complex Ideational Material; Stroop: D-KEFS Color-Word Interference Test; INT: Interference Condition; BNT-15: Boston Naming Test 15-Item Short Form; Reading$_{\text{WRAT-4}}$: Wide Range Achievement Test 4th Edition Reading subtest; SS: Standard score (*M* = 100; *SD* = 15).

Table 34

Descriptive Statistics & Independent t-Tests Comparing Scores Across Tests of Low Verbal Mediation as a Function of the Experimental Condition in the NSE Sample (n = 70)

		Exp. Condition						
		NC (*n* = 40)		EM (*n* = 30)				
Measure	Score	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
TOMM	T1	47.4	3.1	32.1	10.2	7.94	<.01	2.03
DCT	E-score	10.3	2.7	18.1	7.0	-5.77	<.01	1.47
FIT	Recall	14.9	.47	12.9	2.9	3.75	<.01	0.96
	Recognition	14.5	1.0	11.1	3.8	4.75	<.01	1.22
	Combined	29.5	1.4	23.9	6.9	4.37	<.01	1.12
TMT	A Raw	32.4	18.8	50.9	34.5	-2.89	.01	0.67
	A T-Score	41.0	14.9	28.2	13.1	3.76	<.01	0.91
	B Raw	58.4	23.6	83.3	38.8	-3.34	<.01	0.78
	B T-Score	47.8	11.8	38.0	11.6	3.48	<.01	0.84
CD*WAIS-III*	Raw	87.0	13.7	57.8	20.4	6.77	<.01	1.68
	ACSS	11.6	2.4	6.7	3.2	7.41	<.01	1.73
	Recognition	8.3	1.1	6.1	2.4	4.38	<.01	1.18
SS*WAIS-IV*	Raw	37.0	7.1	24.9	12.2	4.87	<.01	1.21
	ACSS	11.3	2.7	7.2	4.0	4.83	<.01	1.20
RCFT	Copy	33.5	2.5	29.3	3.9	5.28	<.01	1.28
	T2C	146.3	54.2	129.5	59.6	1.23	.22	0.29
	IR Raw	23.6	6.8	17.9	5.4	3.85	<.01	0.93
	IR T-Score	48.6	14.8	33.6	12.4	4.47	<.01	1.10
	DR Raw	23.1	6.7	16.5	6.0	4.23	<.01	1.04
	DR T-Score	46.9	14.0	31.7	11.8	4.80	<.01	1.17
	Recog Raw	21.3	2.0	18.2	3.0	5.11	<.01	1.22
	Recog T-Score	48.1	12.9	31.5	10.8	5.67	<.01	1.40
CDT	Raw	9.7	0.6	8.7	1.5	3.50	<.01	0.88

Note: NSE: Native speakers of English; Exp. Condition: Experimental condition; NC: Non-Malingering Control; EM: Experimental Malingering; TOMM T1: Test of Memory Malingering Trial 1; DCT E-Score: Dot Counting Test Effort-Score; FIT: Rey 15-Item Test; TMT-A, TMT-B: Trail Making Test Part A and Part B; T-score (*M* = 50, *SD* = 10);CD*WAIS-III*: WAIS-III Digit Symbol subtest; ACSS: Age-corrected scaled score (*M* = 10, *SD* = 3); SS*WAIS-IV*: WAIS-IV Symbol Search subtest; RCFT: Rey-Osterrieth Complex Figure Test; T2C: Time to completion; IR: Immediate Recall; DR: Delayed Recall; Recog: Recognition; CDT: Clock Drawing Test.

CHAPTER VI: DISCUSSION

Summary of Results

The current study had three objectives: (1) to examine the effect of LEP on PVT performance; (2) to examine whether current PVT cutoffs are useful in detecting non-credible performance in individuals with LEP; and (3) to develop new PVT cutoffs for this population.

Main findings. Consistent with the *a priori* hypotheses, participants with LEP had a higher BR_{Fail} on and failed more PVT_{HVM} compared to NSE. The effect of language proficiency was large: the LEP group produced an overall BR_{Fail} that was 6.5 times greater than that of the NSE group. In contrast, LEP and NSE groups had a similar univariate and multivariate BR_{Fail} on PVT_{LVM}. A formal interaction analysis confirmed that the higher BR_{Fail} for LEP compared to NSE participants was specific to PVT_{HVM}. Finally, both self-reported and objective measures of English proficiency were highly correlated with BR_{Fail} on PVT_{HVM} but not on PVT_{LVM}.

Taken together, the present findings suggest that, under normal conditions (i.e., not instructed to malinger), individuals with LEP perform clinically and significantly worse on PVT_{HVM} than NSE, resulting in a disproportionally higher BR_{Fail} for the LEP group. As will be discussed below, the most plausible explanation of this pattern of findings is that the elevated BR_{Fail} on PVT_{HVM} among individuals with LEP reflects false-positive errors.

As predicted, the utility of PVTs for individuals with LEP depended on the type of measure: PVT_{LVM} demonstrated good classification accuracy, while the majority of PVT_{HVM} did not. The following tests resulted in poor classification accuracy (sensitivity <.50 at specificity ≥.85), and therefore, may have limited utility in this population: WCT

T2C, FAS, Animals, Verbal Fluency LREs, CIM, and Stroop Color and Interference conditions (sensitivity: .13-.48).

At the same time, many PVTs show promise in this population using adjusted cutoffs. The following PVTs demonstrated high sensitivity (.50-.90) while maintaining ≥.85 specificity: TOMM-1, DCT E-Score, Coding, RCFT Copy Trial, WCT Accuracy, Digit Span, FIT, Trail Making Test, Symbol Search, RCFT (IR, Equation), WRT, and Stroop Word condition. Notably, the majority of these are $\text{PVT}_{\mathit{LVM}}$.

Additional findings. Consistent with the main results, LEP and NSE participants had similar experimental-malingering profiles on $\text{PVT}_{\mathit{LVM}}$, while the profiles diverged on $\text{PVT}_{\mathit{HVM}}$. In the LEP group, the difference between NC and EM conditions was most pronounced on the WCT, Digit Span, WRT, $\text{LRE}_{\mathit{Sugarman}}$, Stroop Word condition (RR: 2.67-63.00), and to a smaller degree, Stroop Color and Interference conditions (RR: 1.58-2.66). In contrast, experimental malingering on the FAS, Animals, $\text{LRE}_{\mathit{Johnson}}$, and CIM was masked by LEP; Scores on these instruments were indistinguishable between NC and EM conditions (RR: 1.14-1.50). As will be discussed, it is recommended that these instruments not be used as PVTs for this population.

In general, freestanding PVTs that have been reported to be the most specific to malingering in the literature (e.g., TOMM) performed well across both LEP and NSE groups. In contrast, some tests produced poor classification accuracy regardless of English proficiency status. The TMT, for example, emerged as an instrument with poor specificity for both LEP and NSE participants due to a high $\text{BR}_{\mathit{Fail}}$ in the NC condition. Additionally, the RCFT produced poor sensitivity across participants, with low $\text{BR}_{\mathit{Fail}}$ in the EM condition. Finally, LEP participants showed *more* pronounced malingering than

NSE participants on the Digit Span. This may reflect differences in malingering strategies (e.g., prioritizing poor accuracy versus slower response time), or as discussed below, cultural differences between LEP and NSE groups.

Divergent findings. The current results are generally consistent with *a priori* hypotheses. However, three PVT$_{HVM}$ (Digit Span, WCT, Stroop Word) produced results in the opposite direction of the hypotheses. Specifically, LEP and NSE participants had similar BR$_{Fail}$ on the WCT and Stroop Word, while BR$_{Fail}$ on the Digit Span was *lower* for LEP than NSE participants (RR:1.33-5.00). Although the reason for the anomalies is unclear, a few potential explanations are noteworthy.

The WCT and Stroop Word were unique amongst the PVT$_{HVM}$ as these instruments contained visual stimuli of the text. As noted, LEP participants had higher self-rated proficiency in reading than speaking English, and thus may have greater automaticity with reading tasks. It is possible that LEP participants struggled on PVT$_{HVM}$ that required greater demand on oral skills (e.g., verbal comprehension and word generation), while written stimuli provided an alternative mechanism to process information to mitigate the deficits in overall language proficiency.

Previous investigations of the Digit Span in different languages have examined the word-length effect on performance, such that syllable length affects immediate verbal recall (Baddeley, Thomson, & Buchanan, 1975). Although a popular explanation of Digit Span score decrements in other languages (e.g., López, Steiner, Hardy, IsHak, & Anderson, 2016; Ostrosky-Solís & Lozano, 2006), the word-length effect is an unlikely mechanism in the current study, as LEP participants completed all measures in English.

Cultural differences between LEP and NSE participants offer a more plausible explanation. Specifically, certain cultures have a greater emphasis on numerical knowledge, which may result in greater fluency on numerical tasks. For example, studies have shown that Chinese children have a greater exposure to numbers and rote math training, and consistently outperform North American children in math tasks even prior to formal education (Geary, Bow-Thomas, Fan, & Siegler, 1993; Huntsinger, Jose, Liaw, & Ching, 1997; Siegler & Mu, 2008). Chinese speakers have also been found to score higher on the Digit Span compared to NSE, a finding that is well-replicated (Chen & Stevenson, 1988; Cheung & Kemper, 1993; Chincotta & Underwood, 1997; Hedden et al., 2002; Stigler, Lee, & Stevenson, 1986; Yang et al., 2012). The current study is consistent with this literature, as Chinese-speaking participants were overrepresented in the LEP group, and Digit Span was the only numerically-based PVT$_{HVM}$.

More recent research suggests that the working memory advantage for Chinese samples is not limited to digits; superior performance have also been found for immediate serial recall of words that is not attributed to the word-length effect (Mattys, Baddeley, & Trenkic, 2018). As such, Mattys and colleagues proposed that cultural differences extend beyond exposure to numerical concepts. Specifically, rote memorization is heavily emphasized in learning to read and write Chinese, which uses a logographic system with no sound-symbol correspondence. Greater educational demand and training with rote memory may result in increased verbal working memory capacity or more efficient rehearsal processes (Mattys et al.). Hence, the reversal in the expected BR$_{Fail}$ direction on the Digit Span may reflect cultural differences in the development of certain cognitive processes. Finally, it is possible that the anomaly results on the Digit Span reflect sample-specific findings, with further replication studies required.

Relevance to Previous Literature

The current study is the first to investigate PVT performance in individuals with LEP using an experimental-malingering design. Aside from one study (Salazar et al., 2007), previous research was conducted in countries outside of North America and tested participants in their native language. Overall, findings from the current study is consistent with past literature. Specifically, previous studies reported that many PVTs adequately distinguish between credible and non-credible performance with linguistic and cultural minorities, although adjustments in cutoffs are often necessary (e.g., Spanish-speaking samples: Burton et al, 2012; Vilar-López et al., 2008a, 2008b). These findings mostly apply to PVT*LVM*, as few studies included PVT*HVM.*

In the sole North American study comparing LEP and NSE on a battery including both PVT*LVM* and PVT*HVM*, Salazar and colleagues (2007) reported that the only difference between LEP and NSE groups was on the RDS and RCFT, in contrast to the current findings**.** However, Salazar and colleagues similarly found that using published cutoffs produced low specificity for individuals with LEP, and that adjustments in cutoffs were necessary. Cutoffs from this retrospective study, however, were much more conservative (Digit Span ACSS: ≤4, RDS: ≤5, DCT: ≤19, FIT: ≤12, RCFT equation: ≤45) and did not account for sensitivity.

The recurrent finding of NSE outperforming individuals with LEP on PVT*HVM* is consistent with previous literature on neuropsychological testing in cultural and linguistic minorities (Boone et al., 2007). Poorer performance on neuropsychological measures in ethnic and linguistic minorities compared to White-Anglo NSE have been well-replicated in both non-clinical (e.g., Jacobs et al., 1997) and patient samples (e.g., Boone et al., 2002) across various cognitive domains, including memory (e.g., Norman, Evans, Miller, & Heaton, 2000), executive function (e.g., Coffey, Marmol, Schock, & Adams, 2005), and visuomotor processing speed measures (e.g., Mehta et al., 2004).

While cultural variables were not analyzed in the present study, the findings are generally consistent with the long-standing assumption that nonverbal cognitive tests are less culturally biased. Indeed, this assumption is reflected in clinical practice: when neuropsychologists were asked to identify their approach to working with individuals with LEP, "administering tests designed to be culturally unbiased", such as nonverbal measures, was the third most common response (Elbulok-Charcape et al., 2014). At the same time, no instrument, regardless of the level of verbal mediation, can be assumed to be "culture-free", as non-linguistic cultural differences may influence performance (Fasfous et al., 2013; Rosselli & Ardila, 2003).

In the current study, for example, individuals with LEP performed significantly worse on the Clock Drawing Test (not included as a formal PVT) than NSE. A qualitative examination of the drawings suggested differences in conceptualization of the clock for some LEP participants (e.g., drawing a square clock, only including anchor numbers), which is penalized based on the scoring system. Such differences may not be fully accounted for by LEP, but likely reflect cultural differences that should not be overlooked when administering PVT$_{LVM}$. These findings highlight the importance of researching the validity of individual instruments and the importance of continuing to investigate the effects of cultural variables other than language on performance.

Implications of Findings & Practice Recommendations

No psychometric solution for PVT$_{HVM}$. PVT$_{LVM}$ tended to close the BR$_{Fail}$ gap between LEP and NSE participants, with similarly low BR$_{Fail}$ (0.0-2.5%) observed for failing ≥3 PVT$_{LVM}$ in both groups. In contrast, LEP participants had a higher BR$_{Fail}$ on PVT$_{HVM}$ regardless of the level of cutoff used. As such, English proficiency and performance validity appear to be psychometrically indistinguishable on PVT$_{HVM}$, suggesting that there is no solution for using PVT$_{HVM}$ with an LEP population. A failure on a PVT$_{HVM}$ for an individual with LEP may indicate performance invalidity, but also may indicate poor English proficiency, poor English proficiency coupled with brain injury, or a mix of these conditions. Simply examining Pass/Fail on PVT$_{HVM}$ cannot disentangle these various aspects. While there are exceptions to this apparent singularity, it is generally recommended that examiners do not use PVT$_{HVM}$ in evaluating performance validity for individuals with LEP.

Use of LEP-specific cutoffs. As discussed, many PVT$_{HVM}$ were not specific to malingering. A desirable classification accuracy could not be achieved by simply adjusting cutoffs. Indeed, inadequate specificity is a problem commonly found in neuropsychology research with ethnic minorities, resulting in increased rates of false positives and a risk to over-pathologize within this population (Rivera-Mindt et al., 2010). Although PVT$_{LVM}$ were found to have adequate classification accuracy, cutoffs on some tests had to be adjusted to maintain an optimal balance of sensitivity and specificity. While further research is necessary to validate these new cutoffs in other samples and settings, it is recommended that cutoffs normed with an LEP sample is used for this population, rather than relying on published cutoffs developed with NSE.

Instrument specificity. Despite an overall trend for better classification accuracy for PVT$_{\mathit{LVM}}$ than PVT$_{\mathit{HVM}}$ for individuals with LEP, not all PVTs conformed to this pattern. As mentioned, anomalies were found such that certain PVT$_{\mathit{LVM}}$ (e.g., RCFT) were not sensitive across any groups, while certain PVT$_{\mathit{HVM}}$ (e.g., Digit Span) performed unexpectedly well in detecting malingering for all groups, including LEP participants. The general rule of "BR$_{\mathit{Fail}}$ on PVT$_{\mathit{HVM}}$ > PVT$_{\mathit{LVM}}$ for LEP" largely applies, but ignores the idiosyncratic findings of the current research. To arrive at more accurate conclusions, classification accuracy of individual instruments should be examined in addition to interpreting higher-order findings. Hence, caution against the application of broad conclusions based solely on level of verbal mediation is recommended, and consideration of the properties of individual instruments is warranted when such research is available.

Relation to current practice guidelines. While current practice guidelines (e.g., AACN, 2007; APA, 2003) recognize and highly recommend the consideration of cultural, linguistic, and individual and social factors in neuropsychological assessments, few guidelines are provided on the actual implementation of such recommendations. For clients with LEP, best practice recommendations focus on referring to neuropsychologists proficient in the native language of the client, if one is not competent in working with the client's cultural or linguistic background (AACN, 2007). However, as discussed, this "gold standard" guideline may be difficult to attain in practice, due to the scarcity of neuropsychologists with native proficiency in other languages and highlights a broader issue of underrepresentation of ethnic and linguistic minorities in the field (Rivera-Mindt et al, 2010). While the current research does little to change the status quo of this systemic issue, use of the new PVT norms developed with individuals with LEP offers a practical solution to address the present needs of neuropsychologists.

Strengths & Limitations

Strengths. The current study utilized a strong experimental design with meticulous control of extraneous variables between conditions. RAs were trained over several weeks, with multiple assessments of their psychometric skills to ensure competence, standardization and uniformity in test administration. Scoring and audio-recordings were reviewed throughout, and feedback was provided to RAs on a regular basis. Furthermore, a single-blind procedure was used to diminish potential demand characteristics, and a comprehension check of malingering instructions was incorporated at two points of the protocol. This level of experimental rigour minimized administration errors and common misunderstandings among participants, resulting in a clean, highly controlled dataset with little to no missing data.

The prospective design of the study allowed for the inclusion of a broad array of instruments and for the most pertinent variables to be examined, in contrast to previous retrospective studies (e.g., Salazar et al., 2007). For example, an equal number of $\text{PVT}_{\textit{HVM}}$ and $\text{PVT}_{\textit{LVM}}$ were selected to assess a broad range of cognitive abilities (e.g., verbal memory, visual-constructional, executive attention, processing speed). Additionally, both freestanding and embedded indicators were included, with a balance of well-established (e.g., TOMM) and novel PVTs (e.g., Stroop). English proficiency was assessed with both an objective performance-based measure (BNT-15) and by subjective self-report (LEAP-Q), as recommended in the literature (Gollan et al., 2012).

One of the unique contributions of the current research is the use of an experimental-malingering design to establish a well-defined condition of prescribed invalid performance as a criterion. Many studies use known-groups (e.g., comparing groups based on litigation status; Meyers, & Volbrecht, 1999) or other established PVTs (e.g., comparing groups separated by scores on the TOMM; Curtis et al., 2009) as criteria to calculate classification accuracy. Although using known-groups and established PVTs both have their own advantages, these methods rest on the presumption that the non-credible and credible participants are adequately differentiated. Some litigating participants, for example, may not be actively malingering, and non-litigating participants may also have incentive to exaggerate poor performance or little incentive to perform well (An et al., 2017). An experimental-malingering design minimizes such overlap between the two groups, thus increasing power to detect differences.

The sample included in the current study also presents many strengths. Specifically, inclusion and exclusion criteria were strictly enforced to maximize internal validity. Balanced bilinguals were excluded, so as to maximize the differentiation between LEP and NSE groups on language proficiency.

Limitations. The findings of the current study must be interpreted in the context of its limitations. The high internal validity of the experimental design comes at a trade-off to external validity. Lack of real-world incentives to perform poorly while avoiding detection for the EM condition may not generalize to patients feigning cognitive impairment. Research has suggested that undergraduate participants with no contingencies on performance may not exert optimal effort on PVTs (An et al., 2012; An et al., 2017). It is possible that tangible, life-altering rewards (e.g., millions of dollars in compensation) may produce different patterns of malingering (e.g., more consistent, believable, or impaired) than the current study.

Additionally, while most participants in the EM condition appeared to have complied with malingering instructions, it was evident through the post-session questionnaire that there were a few exceptions. Qualitative data from speaking with these participants post-session revealed that one of the main reasons for non-compliance was forgetting to follow the malingering instructions due to the increased cognitive demands on some tests. This was especially apparent for LEP participants, who had the additional language processing demands. For LEP participants, it is possible that comprehending test instructions placed increased demands on cognitive resources to the extent that less resources were available to attend to the malingering task. Indeed, research has shown that deception and lying are cognitively demanding (Bigler, 2015; Vrij, Fisher, Mann, & Leal, 2006), with evidence of increased activation of prefrontal regions responsible for executive attention (Spence et al, 2004). Examinees must, for example, keep track of how many items they answered incorrectly and monitor their response times to present a believable impairment profile. Patients with cognitive impairment, similar to individuals with LEP, may also have difficulty maintaining a consistent impairment profile throughout a testing session. Hence, although a few participants in the current study struggled with sustaining the malingering task, this difficulty may be comparable to real-world malingering.

The current study included LEP participants who had at least a level of English proficiency that was adequate to apply for undergraduate studies at an English-speaking Canadian university. It is speculated that the difference between LEP and NSE groups would be even greater for individuals with lower levels of English proficiency in the community, following the trend observed in the current study. Caution is warranted when applying results to individuals with very low levels of English proficiency or LEP in combination with low levels of education, non-English speaking monolinguals, and balanced bilinguals, as these groups were not included in the study.

The use of a non-clinical, university-student population also restricts the generalizability of the findings to other populations. The sample was homogeneous such that participants were mostly young adults from educated families. Although exclusion of psychiatric and neurological conditions ensured tight control in the study, results from this sample may not be applicable to a clinical or forensic population, where the prevalence of neuropsychiatric disorders and hence, genuine cognitive impairment is significantly higher.

LEP and NSE groups differed on some demographic variables that could have affected the findings. Specifically, the LEP group contained numerous STEM graduate students, which traditionally is underrepresented by females (Vogt et al., 2007; Wang & Degol, 2017). This resulted in a greater number of higher-educated males in the LEP group – a potential issue as some neuropsychological tests are affected by gender and education (Saykin et al., 1995; Wiederholt et al., 1993). Women, for example, have been found to outperform men on verbal tasks (e.g., Verbal fluency; Loonstra, Tarlow, & Sellers, 2001; Verbal memory; Lewin, Wolgers, & Herlitz, 2001). Because the NSE group consisted of more females, this may have magnified the difference on PVT$_{HVM.}$

Similarly, higher education has been associated with better performance on certain measures (e.g., Digit Span; Walker, Batchelor, & Shores, 2009; Verbal fluency;

Tombaugh, Kozak, & Rees, 1999), which may have diminished the between-group difference on PVT$_{HVM}$. Furthermore, even with comparable years of education, educational attainment may not be equivalent between LEP and NSE participants due to differences in the education system across nations. Given that the majority of LEP participants immigrated to Canada in the past few years, education attainment in their home country may reflect different education experiences, making years of education challenging to equate. Thus, differences in demographic variables may have worked in favour of the hypotheses for some tests and against the hypotheses on others.

Finally, cultural variables were not investigated in the present research, despite differing between LEP and NSE groups. Variables such as acculturation level, ethnic identity, and years immigrated likely play a role on test performance, and parsing out LEP from culture ignores their interaction in affecting test performance. The LEP group was also overrepresented by one ethnicity and language group (Chinese), despite recruiting broadly, affecting generalizability of findings. Unfortunately, the small *n* of other cultures and languages in the sample precluded any *post-hoc* comparisons between these groups.

Future Directions & Conclusion

PVT research with individuals with LEP is sparse and disproportional to the growing interest in the field. Greater emphasis on including cultural and linguistic minorities in neuropsychological research is recommended in general, and particularly in the PVT field, so that appropriate normative data are available for more instruments.

The cutoffs reported in the current study need to be replicated using different populations, settings, and methodology. As the current dissertation is one of the first studies to investigate PVT performance in individuals with LEP, internal validity was emphasized to ensure that the effects observed were not influenced by extraneous variables. Studies with clinical, forensic, and community LEP samples are an important next step to assess whether these findings apply to the real-world and whether the proposed cutoffs can differentiate between performance invalidity and brain injury. Moreover, future research should strive to include a greater range of ages, ethnic backgrounds, education and SES levels to increase the generalizability of the findings.

Future studies may explore patterns of PVT performance for individuals with LEP and contrast different types of tests beyond level of verbal mediation (e.g., freestanding versus embedded). Isolated findings in the current study, such as the reversal of the $\mathrm{BR}_{\mathit{Fail}}$ pattern on Digit Span between LEP and NSE groups, would benefit from replication to further examine the relative merits of competing explanations. Differences in time-to-completion of the battery or changes in performance throughout the testing session between individuals with LEP and NSE also pose as interesting patterns to investigate.

Malingering strategies, although collected solely as a compliance check in the current study, deserve much greater attention in future studies, as they may provide an explanation for atypical patterns of performance (Cottingham et al., 2014) or the increased within-group validity among non-credible examinees (Erdodi et al., 2014). Studies have also suggested differences in deceptive and socially desirable responding between cultures (Fell & König, 2016; Lalwani, Shavitt, & Johnson, 2006). However, there is little literature on how different cultures approach the task of feigning impairment on cognitive tests. For example, cultural values may influence the extent to which one exaggerates or feigns impairment, or there may be cross-cultural differences in the type of malingering strategies believed to be most effective.

Another trend from the current study worth investigating is whether $\text{PVT}_{\mathit{HVM}}$ with a visual component protects against the deleterious effects of LEP. As described above, $\text{BR}_{\mathit{Fail}}$ was higher for LEP participants on $\text{PVT}_{\mathit{HVM}}$ that relied on oral comprehension or expression skills in English. The implications of these findings are clinically significant if such results are replicated. Although no psychometric solution for $\text{PVT}_{\mathit{HVM}}$ was found within this study, $\text{PVT}_{\mathit{HVM}}$ may nevertheless be utilized for this population if the task includes visual stimuli of the text.

The findings from the current dissertation point to a common theme: LEP increases false positives on $\text{PVT}_{\mathit{HVM}}$. Nevertheless, many standardized tests at modified cutoffs provide a valid assessment of non-credible performance in individuals with LEP. Although the sparsity of research on cultural and linguistic minorities in neuropsychology remains a pressing systemic issue, individuals with LEP, in the meantime, should not be precluded from accessing a valid neuropsychological evaluation, including an assessment of performance validity. The field of neuropsychology has increasingly recognized the need to include cultural and linguistic minorities in research to match the changing North American demographic, and it is hoped that a similar shift will occur in PVT research.

REFERENCES

Abeare, C. A., Freund, S., Kaploun, K., McAuley, T., & Dumitrescu, C. (2017). The Emotion Word Fluency Test (EWFT): Initial psychometric, validation, and physiological evidence in young adults. *Journal of Clinical and Experimental Neuropsychology*, *39*(8), 738-752.

American Academy of Clinical Neuropsychology (AACN) Board of Directors (2007). AACN Practice Guidelines for Neuropsychological Assessment and Consultation. *The Clinical Neuropsychologist, 21*, 209–231.

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

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, *57*(12), 1060-1073.

American Psychological Association. (2003). Guidelines on Multicultural Education, Training, Research, Practice, and Organizational Change for Psychologists. *The American Psychologist*, *58*(5), 377.

An, K. Y., Charles, J., Ali, S., Enache, A., Dhuga, J., & Erdodi, L. A. (2019). Reexamining performance validity cutoffs within the Complex Ideational Material and the Boston Naming Test–Short Form using an experimental malingering paradigm. *Journal of Clinical and Experimental Neuropsychology*, *41*(1), 15-25.

An, K. Y., Kaploun, K., Erdodi, L. A., & Abeare, C. A. (2017). Performance validity in undergraduate research participants: a comparison of failure rates across tests and cutoffs. *The Clinical Neuropsychologist*, *31*(1), 193-206.

An, K. Y., Zakzanis, K. K., & Joordens, S. (2012). Conducting research with non-clinical healthy undergraduates: Does effort play a role in neuropsychological test performance? *Archives of Clinical Neuropsychology, 27*(8), 849-857.

Ardila, A. (2005). Cultural values underlying psychometric cognitive testing. *Neuropsychology Review*, *15*(4), 185-195.

Armistead-Jehle, P., & Buican, B. (2013). Comparison of select Advanced Clinical Solutions embedded effort measures to the Word Memory Test in the detection of suboptimal effort. *Archives of Clinical Neuropsychology*, *28*(3), 297-301.

Armistead-Jehle, P., & Hansen, C. L. (2011). Comparison of the Repeatable Battery for the Assessment of Neuropsychological Status Effort Index and stand-alone symptom validity tests in a military sample. *Archives of Clinical Neuropsychology*, *26*(7), 592-601.

Artiola i Fortuny, L., & Mullaney, H. A. (1998). Assessing patients whose language you do not know: Can the absurd be ethical?. *The Clinical Neuropsychologist*, *12*(1), 113-126.

Axelrod, B. N., Fichtenberg, N. L., Millis, S. R., & Wertheimer, J. C. (2006). Detecting Incomplete Effort with Digit Span from the Wechsler Adult Intelligence Scale—Third Edition. *The Clinical Neuropsychologist*, *20*(3), 513-523.

Babikian, T., Boone, K. B., Lu, P., & Arnold, G. (2006). Sensitivity and specificity of various digit span scores in the detection of suspect effort. *The Clinical Neuropsychologist*, *20*(1), 145-159.

Backhaus, S. L., Fichtenberg, N. L., & Hanks, R. A. (2004). Detection of sub-optimal performance using a floor effect strategy in patients with traumatic brain injury. *The Clinical Neuropsychologist*, *18*(4), 591-603.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 575-589.

Baldessarini, F, R. J., Finklestein, S., & Arana, G. W. (1983). The predictive power of diagnostic tests and the effect of prevalence of illness. *Archives of General Psychiatry*, *40*(5), 569-573.

Barhon, L. I., Batchelor, J., Meares, S., Chekaluk, E., & Shores, E. A. (2015). A comparison of the degree of effort involved in the TOMM and the ACS Word Choice Test using a dual-task paradigm. *Applied Neuropsychology: Adult*, *22*(2), 114-123.

Bashem, J. R., Rapport, L. J., Miller, J. B., Hanks, R. A., Axelrod, B. N., & Millis, S. R. (2014). Comparisons of five performance validity indices in bona fide and simulated traumatic brain injury. *The Clinical Neuropsychologist*, *28*(5), 851-875.

Bell-Sprinkel, T. L., Boone, K. B., Miora, D., Cottingham, M., Victor, T., Ziegler, E., ... & Wright, M. (2013). Re-Examination of the Rey Word Recognition Test. *The Clinical Neuropsychologist*, *27*(3), 516-527.

Benton, A. L., Hamsher, K., de S., & Sivan, A. B. (1994). *Multilingual Aphasia Examination.* Iowa City: AJA Associates.

Bianchini, K. J., Mathias, C. W., & Greve, K. W. (2001). Symptom validity testing: A critical review. *The Clinical Neuropsychologist*, *15*(1), 19-45.

Bigler, E. D. (2012). Symptom validity testing, effort, and neuropsychological assessment. *Journal of the International Neuropsychological Society*, *18*(04), 632-640.

Bigler, E. D. (2014). Effort, symptom validity testing, performance validity testing and traumatic brain injury. *Brain injury*, *28*(13-14), 1623-1638.

Bigler, E. D. (2015). Neuroimaging as a biomarker in symptom validity and performance validity testing. *Brain Imaging and Behavior*, *9*(3), 421-444.

Blaskewitz, N., Merten, T., & Brockhaus, R. (2009). Detection of suboptimal effort with the Rey Complex Figure Test and recognition trial. *Applied Neuropsychology*, *16*(1), 54-61.

Boone, K. B. (Ed.). (2007). *Assessment of feigned cognitive impairment: A neuropsychological perspective*. NY: The Guilford Press.

Boone, K. B. (2009). The need for continuous and comprehensive sampling of effort/response bias during neuropsychological examinations. *The Clinical Neuropsychologist*, *23*(4), 729-741.

Boone, K. B., Lu, P., Back, C., King, C., Lee, A., Philpott, L., ... & Warner-Chacon, K. (2002a). Sensitivity and specificity of the Rey Dot Counting Test in patients with suspect effort and various clinical samples. *Archives of Clinical Neuropsychology*, *17*(7), 625-642.

Boone, K., Lu, P., & Herzberg, D. (2002b). *The Dot Counting Test.* Los Angeles: Western Psychological Services.

Boone, K. B., Salazar, X., Lu, P., Warner-Chacon, K., & Razani, J. (2002c). The Rey 15-item recognition trial: A technique to enhance sensitivity of the Rey 15-item memorization test. *Journal of Clinical and Experimental Neuropsychology*, *24*(5), 561-573.

Boone, K. B., Victor, T. L., Wen, J., Razani, J., & Pontón, M. (2007). The association between neuropsychological scores and ethnicity, language, and acculturation variables in a large patient population. *Archives of Clinical Neuropsychology*, *22*(3), 355-365.

Brickman, A. M., Cabo, R., & Manly, J. J. (2006). Ethical issues in cross-cultural neuropsychology. *Applied Neuropsychology*, *13*(2), 91-100.

Burton, V., Vilar-López, R., & Puente, A. E. (2012). Measuring effort in neuropsychological evaluations of forensic cases of Spanish speakers. *Archives of Clinical Neuropsychology*, *27*(3), 262-267.

Busse, M., & Whiteside, D. (2012). Detecting suboptimal cognitive effort: classification accuracy of the Conner's continuous performance test-II, brief test of attention, and trail making test. *The Clinical Neuropsychologist*, *26*(4), 675-687.

Cavé, J., & Grieve, K. (2009). Quality of education and neuropsychological test performance. *New Voices in Psychology*, *5*(1), 29-48.

Chang, S. Y. S. (2006). *Development of a test battery for assessing memory malingering in Hong Kong and its application on depressed patients.* (Unpublished doctoral dissertation). The Chinese University of Hong Kong.

Chen, C., & Stevenson, H. W. (1988). Cross-linguistic differences in digit span of preschool children. *Journal of Experimental Child Psychology*, *46*, 150–158.

Cheung, H., & Kemper, S. (1993). Recall and articulation of English and Chinese words by Chinese–English bilinguals. *Memory & Cognition, 21,* 666–670.

Chin, A. L., Negash, S., Xie, S., Arnold, S. E., & Hamilton, R. (2012). Quality, and not just quantity, of education accounts for differences in psychometric performance between african americans and white non-hispanics with Alzheimer's disease. *Journal of the International Neuropsychological Society*, *18*(2), 277-285.

Chincotta, D., & Underwood, G. (1997). Digit span and articulatory suppression: A cross-linguistic comparison. *European Journal of Cognitive Psychology, 9,* 89–96.

Coffey, D. M., Marmol, L., Schock, L., & Adams, W. (2005). The influence of acculturation on the Wisconsin Card Sorting Test by Mexican Americans. *Archives of Clinical Neuropsychology, 20*, 795–803.

Cole, M. (1999). Culture-free versus culture-based measures of cognition. In R. J. Sternberg (Ed.), *The nature of cognition* (pp. 645–664). MIT Press.

Cottingham, M. E., Victor, T. L., Boone, K. B., Ziegler, E. A., & Zeller, M. (2014). Apparent effect of type of compensation seeking (disability versus litigation) on performance validity test scores may be due to other factors. *The Clinical Neuropsychologist*, *28*(6), 1030-1047.

Curtis, K. L., Greve, K. W., & Bianchini, K. J. (2009). The Wechsler Adult Intelligence Scale—III and Malingering in Traumatic Brain Injury Classification Accuracy in Known Groups. *Assessment, 16*(4), 401-414.

Curtis, K. L., Thompson, L. K., Greve, K. W., & Bianchini, K. J. (2008). Verbal fluency indicators of malingering in traumatic brain injury: Classification accuracy in known groups. *The Clinical Neuropsychologist, 22*(5), 930-945.

Davis, J. J. (2014). Further Consideration of Advanced Clinical Solutions Word Choice: Comparison to the Recognition Memory Test-Words and Classification Accuracy in a Clinical Sample. *The Clinical Neuropsychologist, 28*(8), 1278-1294.

Davis, J. J., McHugh, T. S., Axelrod, B. N., & Hanks, R. A. (2012). Performance validity and neuropsychological outcomes in litigants and disability claimants. *The Clinical Neuropsychologist*, *26*(5), 850-865.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan Executive Function System (D-KEFS)*. Psychological Corporation.

DenBoer, J. W., & Hall, S. (2007). Neuropsychological test performance of successful brain injury simulators. *The Clinical Neuropsychologist*, *21*(6), 943-955.

Denning, J. H. (2012). The efficiency and accuracy of the Test of Memory Malingering Trial 1, errors on the first 10 items of the Test of Memory Malingering, and five embedded measures in predicting invalid test performance. *Archives of Clinical Neuropsychology*, *27*(4): 417-432.

DuAlba, L., & Scott, R. L. (1993). Somatization and malingering for workers' compensation applicants: A cross-cultural MMPI study. *Journal of Clinical Psychology*, *49*(6), 913-917.

Elbulok-Charcape, M. M., Rabin, L. A., Spadaccini, A. T., & Barr, W. B. (2014). Trends in the neuropsychological assessment of ethnic/racial minorities: A survey of clinical neuropsychologists in the United States and Canada. *Cultural Diversity and Ethnic Minority Psychology, 20*(3), 353.

Erdodi, L. A., Abeare, C. A., Lichtenstein, J. D., Tyson, B. T., Kucharski, B., Zuccato, B. G., & Roth, R. M. (2017). Wechsler Adult Intelligence Scale-Fourth Edition

(WAIS-IV) processing speed scores as measures of noncredible responding: The third generation of embedded performance validity indicators. *Psychological Assessment*, *29*(2), 148-157.

Erdodi, L. A., Jongsma, K. A., & Issa, M. (2017). The 15-item version of the Boston Naming Test as an index of English proficiency. *The Clinical Neuropsychologist*, *31*(1), 168-178.

Erdodi, L. A., Kirsch, N. L., Lajiness-O'Neill, R., Vingilis, E., & Medoff, B. (2014). Comparing the recognition memory test and the word choice test in a mixed clinical sample: Are they equivalent?. *Psychological Injury and Law*, *7*(3), 255-263.

Erdodi, L., & Roth, R. (2017). Low scores on BDAE Complex Ideational Material are associated with invalid performance in adults without aphasia. *Applied Neuropsychology: Adult*, *24*(3), 264-274.

Erdodi, L. A., Sagar, S., Seke, K., Zuccato, B. G., Schwartz, E. S., & Roth, R. M. (2018). The Stroop test as a measure of performance validity in adults clinically referred for neuropsychological assessment. *Psychological Assessment*, *30*(6), 755-766.

Erdodi, L. A., Tyson, B. T., Abeare, C. A., Zuccato, B. G., Rai, J. K., Seke, K. R., ... & Roth, R. M. (2018). Utility of critical items within the Recognition Memory Test and Word Choice Test. *Applied Neuropsychology: Adult*, *25*(4), 327-339.

Erdodi, L. A., Tyson, B. T., Shahein, A. G., Lichtenstein, J. D., Abeare, C. A., Pelletier, C. L., ... & Roth, R. M. (2017). The power of timing: Adding a time-to-completion cutoff to the Word Choice Test and Recognition Memory Test improves classification accuracy. *Journal of Clinical and Experimental Neuropsychology*, *39*(4), 369-383.

Erdodi, L. A., Tyson, B. T., Abeare, C. A., Lichtenstein, J. D., Pelletier, C. L., Rai, J. K., & Roth, R. M. (2016). The BDAE Complex Ideational Material—a Measure of Receptive Language or Performance Validity?. *Psychological Injury and Law*, *9*(2), 112-120.

Etherton, J. L., Bianchini, K. J., Greve, K. W., & Heinly, M. T. (2005). Sensitivity and specificity of reliable digit span in malingered pain-related disability. *Assessment*, *12*(2), 130-136.

Etherton, J. L., Bianchini, K. J., Heinly, M. T., & Greve, K. W. (2006). Pain, malingering, and performance on the WAIS-III Processing Speed Index. *Journal of Clinical and Experimental Neuropsychology*, *28*(7), 1218-1237.

Fasfous, A. F., Puente, A. E., Pérez-Marfil, M. N., Cruz-Quintana, F., Peralta-Ramirez, I., & Pérez-García, M. (2013). Is the color trails culture free?. *Archives of Clinical Neuropsychology*, *28*(7), 743-749.

Fell, C. B., & König, C. J. (2016). Cross-cultural differences in applicant faking on personality tests: a 43-nation study. *Applied Psychology*, *65*(4), 671-717.

Flynn, J. R. (2007). *What is intelligence?: Beyond the Flynn effect*. Cambridge University Press.

Fyffe, D. C., Mukherjee, S., Barnes, L. L., Manly, J. J., Bennett, D. A., & Crane, P. K. (2011). Explaining differences in episodic memory performance among older African Americans and whites: the roles of factors related to cognitive reserve and test bias. *Journal of the International Neuropsychological Society*, *17*(4), 625-638.

Gasquoine, P. G. (1999). Variables moderating cultural and ethnic differences in neuropsychological assessment: The case of Hispanic Americans. *The Clinical Neuropsychologist*, *13*(3), 376-383.

Geary, D. C., Bow-Thomas, C. C., Fan, L., & Siegler, R. S. (1993). Even before formal instruction, Chinese children outperform American children in mental addition. *Cognitive Development*, *8*(4), 517-529.

Gervais, R. O., Rohling, M. L., Green, P., & Ford, W. (2004). A comparison of WMT, CARB, and TOMM failure rates in non-head injury disability claimants. *Archives of Clinical Neuropsychology*, *19*(4), 475-487.

Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., & Cera, C. M. (2012). Self-ratings of spoken language dominance: A multi-lingual naming test (MINT) and preliminary norms for young and aging Spanish-English bilinguals. *Bilingualism: Language and Cognition, 15*(3), 594-615.

Goodglass, H., Kaplan, E., & Barresi, B. (2001). *Boston Diagnostic Aphasia Examination* (3rd ed.). Philadelphia: Lippincott Williams & Wilkins.

Green, P. (2003). *Manual for the Word Memory Test.* Edmonton, Alberta, Canada: Green’s Publishing.

Green, P., Rohling, M. L., Lees-Haley, P. R., & Allen III, L. M. (2001). Effort has a greater effect on test scores than severe brain injury in compensation claimants. *Brain Injury*, *15*(12), 1045-1060.

Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment*, *6*(3), 218-224.

Greiffenstein, M. F., Baker, W. J., & Gola, T. (1996). Comparison of multiple scoring methods for Rey's malingered amnesia measures. *Archives of Clinical Neuropsychology*, *11*(4), 283-293.

Greve, K. W., Bianchini, K. J., & Doane, B. M. (2006). Classification accuracy of the Test of Memory Malingering in traumatic brain injury: Results of a known-groups analysis. *Journal of Clinical and Experimental Neuropsychology*, *28*(7), 1176-1190.

Haines, M. E., & Norris, M. P. (2001). Comparing student and patient simulated malingerers performance on standard neuropsychological measures to detect feigned cognitive deficits. *The Clinical Neuropsychologist*, *15*(2), 171-182.

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

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

Hedden, T., Park, D. C., Nisbett, R., Ji, L. J., Jing, Q., & Jiao, S. (2002). Cultural variation in verbal versus spatial neuropsychological function across the life span. *Neuropsychology*, *16*(1), 65-73.

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., Millis, S. R., & Conference Participants 1. (2009). American Academy of Clinical Neuropsychology Consensus Conference Statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist*, *23*(7), 1093-1129.

Huntsinger, C. S., Jose, P. E., Liaw, F. R., & Ching, W. D. (1997). Cultural differences in early mathematics learning: A comparison of Euro-American, Chinese-American, and Taiwan-Chinese families. *International Journal of Behavioral Development*, *21*(2), 371-388.

Huppert, F. A., & Piercy, M. (1976). Recognition memory in amnesic patients: effect of temporal context and familiarity of material. *Cortex*, *12*(1), 3-20.

Iverson, G. L., Lange, R. T., Green, P., & Franzen, M. D. (2002). Detecting exaggeration and malingering with the Trail Making Test. *The Clinical Neuropsychologist*, *16*(3), 398-406.

Jacobs, D. M., Sano, M., Albert, S., Schofield, P., Dooneief, G., & Stern, Y. (1997). Cross-cultural neuropsychological assessment: A comparison of randomly selected, demographically matched cohorts of English-and Spanish-speaking older adults. *Journal of Clinical and Experimental Neuropsychology*, *19*(3), 331-339.

Jasinski, L. J., Berry, D. T., Shandera, A. L., & Clark, J. A. (2011). Use of the Wechsler Adult Intelligence Scale Digit Span subtest for malingering detection: A meta-analytic review. *Journal of Clinical and Experimental Neuropsychology*, *33*(3), 300-314.

Johnson, S. C., Silverberg, N. D., Millis, S. R., & Hanks, R. A. (2012). Symptom validity indicators embedded in the Controlled Oral Word Association Test. *The Clinical Neuropsychologist*, *26*(7), 1230-1241.

Jones, A. (2013). Test of memory malingering: cutoff scores for psychometrically defined malingering groups in a military sample. *The Clinical Neuropsychologist*, *27*(6), 1043-1059.

Kim, N., Boone, K. B., Victor, T., Lu, P., Keatinge, C., & Mitchell, C. (2010). Sensitivity and specificity of a digit symbol recognition trial in the identification of response bias. *Archives of Clinical Neuropsychology*, *25*(5), 420-428.

Kirkwood, M. W., & Kirk, J. W. (2010). The base rate of suboptimal effort in a pediatric mild TBI sample: Performance on the Medical Symptom Validity Test. *The Clinical Neuropsychologist*, *24*(5), 860-872.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606-613.

Kulas, J. F., Axelrod, B. N., & Rinaldi, A. R. (2014). Cross-validation of supplemental test of memory malingering scores as performance validity measures. *Psychological Injury and Law*, *7*(3), 236-244.

Lacy, M. A., Gore, P. A., Pliskin, N. H., Henry, G. K., Heilbronner, R. L., & Hamer, D. P. (1996). Verbal fluency task equivalence. *The Clinical Neuropsychologist*, *10*(3), 305-308.

Larrabee, G. J. (2003). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *The Clinical Neuropsychologist*, *17*(3), 410-425.

Larrabee, G. J. (2012). Performance Validity and Symptom Validity in Neuropsychological Assessment. *Journal of the International Neuropsychological Society*, *18*(04), 625–631.

Lalwani, A. K., Shavitt, S., & Johnson, T. (2006). What is the relation between cultural orientation and socially desirable responding?. *Journal of Personality and Social Psychology*, *90*(1), 165.

Lewin, C., Wolgers, G., & Herlitz, A. (2001). Sex differences favoring women in verbal but not in visuospatial episodic memory. *Neuropsychology*, *15*(2), 165.

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

Liskin-Gasparro, J. E. (2003). The ACTFL proficiency guidelines and the oral proficiency interview: A brief history and analysis of their survival. *Foreign Language Annals*, *36*(4), 483-490.

Liu, R., Gao, B., Li, Y., & Sheng, L. (2001). Simulated malingering: A preliminary trial on Hiscock's Forced-Choice Digit Memory Test. *Chinese Journal of Clinical Psychology*, *9*(3), 173-175.

Loonstra, A. S., Tarlow, A. R., & Sellers, A. H. (2001). COWAT metanorms across age, education, and gender. *Applied neuropsychology*, *8*(3), 161-166.

López, E., Steiner, A. J., Hardy, D. J., IsHak, W. W., & Anderson, W. B. (2016). Discrepancies between bilinguals' performance on the Spanish and English versions of the WAIS Digit Span task: Cross-cultural implications. *Applied Neuropsychology: Adult*, *23*(5), 343-352.

Lu, P. H., Boone, K. B., Cozolino, L., & Mitchell, C. (2003). Effectiveness of the Rey-Osterrieth Complex Figure Test and the Meyers and Meyers recognition trial in the detection of suspect effort. *The Clinical Neuropsychologist*, *17*(3), 426-440.

Mack, W. J., Freed, D. M., Williams, B.W., & Henderson, V.W. (1992). Boston Naming Test: Shortened version for use in Alzheimer's disease. *Journal of Gerontology, 47*(3), 164–168.

Manly, J. J., Byrd, D. A., Touradji, P., & Stern, Y. (2004). Acculturation, reading level, and neuropsychological test performance among African American elders. *Applied Neuropsychology, 11*(1), 37-46.

Manly, J. J., Jacobs, D. M., Touradji, P., Small, S. A., & Stern, Y. (2002). Reading level attenuates differences in neuropsychological test performance between African American and White elders. *Journal of the International Neuropsychological Society, 8*(3), 341-348.

Manly, J. J., Touradji, P., Tang, M. X., & Stern, Y. (2003). Literacy and memory decline among ethnically diverse elders. *Journal of Clinical and Experimental Neuropsychology, 25*(5), 680-690.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50*(4), 940-967.

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*(6), 741-776.

Martin, P. K., Schroeder, R. W., Wyman-Chick, K. A., Hunter, B. P., Heinrichs, R. J., & Baade, L. E. (2018). Rates of Abnormally Low TOPF Word Reading Scores in

Individuals Failing Versus Passing Performance Validity Testing. *Assessment*, *25*(5), 640–652.

Mathias, C. W., Greve, K. W., Bianchini, K. J., Houston, R. J., & Crouch, J. A. (2002). Detecting malingered neurocognitive dysfunction using the reliable digit span in traumatic brain injury. *Assessment*, *9*(3), 301-308.

Mattys, S. L., Baddeley, A., & Trenkic, D. (2018). Is the superior verbal memory span of Mandarin speakers due to faster rehearsal?. *Memory & Cognition*, *46*(3), 361–369.

Mehta, K., Simonsick, E. M., Rooks, R., Newman, A. B., Pope, S. K., Rubin, S. M., et al. (2004). Black and white differences in cognitive function scores: What explains the difference? *American Geriatrics Society, 52*(12), 2120–2127.

Meyers, J. E., & Meyers, K. R. (1995). *Rey Complex Figure Test and recognition trial professional manual*. Psychological Assessment Resources.

Meyers, J. E., & Volbrecht, M. (1999). Detection of malingerers using the Rey Complex Figure and recognition trial. *Applied Neuropsychology*, *6*(4), 201-207.

Miele, A. S., Gunner, J. H., Lynch, J. K., & McCaffrey, R. J. (2012). Are embedded validity indices equivalent to freestanding symptom validity tests?. *Archives of Clinical Neuropsychology, 27*(1), 10-22.

Millon, T., Krueger, R.F., & Simonsen, E. (Eds) (2010). *Contemporary Directions in Psychopathology: Scientific Foundations of the DSM-V and ICD-11*, The Guilford Press.

Moore, B. A., & Donders, J. (2004). Predictors of invalid neuropsychological test performance after traumatic brain injury. *Brain Injury*, *18*(10), 975-984.

Nies, K. J., & Sweet, J. J. (1994). Neuropsychological assessment and malingering: A critical review of past and present strategies. *Archives of Clinical Neuropsychology*, *9*(6), 501–552.

Nitch, S., Boone, K. B., Wen, J., Arnold, G., & Alfano, K. (2006). The utility of the Rey Word Recognition Test in the detection of suspect effort. *The Clinical Neuropsychologist*, *20*(4), 873-887.

Norman, M., Evans, J., Miller, W., & Heaton, R. (2000). Demographically corrected norms for the California Verbal Learning Test. *Journal of Clinical and Experimental Neuropsychology*, *22*(1), 80–94.

O'Bryant, S. E., Hilsabeck, R. C., Fisher, J. M., & McCaffrey, R. J. (2003). Utility of the Trail Making Test in the assessment of malingering in a sample of mild traumatic brain injury litigants. *The Clinical Neuropsychologist*, *17*(1), 69-74.

Ojeda, N., Aretouli, E., Peña, J., & Schretlen, D. J. (2016). Age differences in cognitive performance: A study of cultural differences in Historical Context. *Journal of Neuropsychology*, *10*(1), 104-115.

Ostrosky-Solís, F., & Lozano, A. (2006). Digit span: Effect of education and culture. *International Journal of Psychology*, *41*(5), 333-341.

Pankratz, L. (1983). A new technique for the assessment and modification of feigned memory deficit. *Perceptual and Motor Skills*, *57*(2), 367-372.

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

Razani, J., Murcia, G., Tabares, J., & Wong, J. (2007). The effects of culture on WASI test performance in ethnically diverse individuals. *The Clinical Neuropsychologist*, *21*(5), 776-788.

Reedy, S. D., Boone, K. B., Cottingham, M. E., Glaser, D. F., Lu, P. H., Victor, T. L., ... & Wright, M. J. (2013). Cross validation of the Lu and colleagues (2003) Rey-Osterrieth Complex Figure Test effort equation in a large known-group sample. *Archives of Clinical Neuropsychology*, *28*(1): 30-37.

Reese, C. S., Suhr, J. A., & Riddle, T. L. (2012). Exploration of malingering indices in the Wechsler Adult Intelligence Scale-digit span subtest. *Archives of Clinical Neuropsychology*, *27*(2), 176-181.

Reitan, R. M. (1992). *Trail Making Test: Manual for administration and scoring*. Reitan Neuropsychology Laboratory.

Rey, A. (1964). *L'examen clinique en psychologic* [The clinical examination in psychology]. Paris: Presses Universitaires de France.

Rivera Mindt, M., Byrd, D., Saez, P., & Manly, J. (2010). Increasing culturally competent neuropsychological services for ethnic minority populations: A call to action. *The Clinical Neuropsychologist*, *24*(3), 429-453.

Rogers, R. (Ed.). (2008). *Clinical assessment of malingering and deception*. Guilford Press.

Romero, H. R., Lageman, S. K., Kamath, V., Irani, F., Sim, A., Suarez, P., ... & the Summit participants. (2009). Challenges in the neuropsychological assessment of ethnic minorities: summit proceedings. *The Clinical Neuropsychologist*, *23*(5), 761-779.

Rosselli, M., & Ardila, A. (2003). The impact of culture and education on non-verbal neuropsychological measurements: A critical review. *Brain and Cognition, 52*(3), 326-333.

Rouleau, I., Salmon, D. P., Butters, N., Kennedy, C., & McGuire, K. (1992). Quantitative and qualitative analyses of clock drawings in Alzheimer's and Huntington's disease. *Brain and Cognition*, *18*(1), 70-87.

Ruffolo, L. F., Guilmette, T. J., & Willis, G. W. (2000). Comparison of Time and Error Rates on the Trail Making Test Among Patients with Head Injuries, Experimental Malingerers, Patients with Suspect Effort on Testing, and Normal Controls. *The Clinical Neuropsychologist*, *14*(2), 223-230.

Saez, P. A., Bender, H. A., Barr, W. B., Rivera Mindt, M., Morrison, C. E., Hassenstab, J., ... & Vazquez, B. (2014). The impact of education and acculturation on nonverbal neuropsychological test performance among Latino/a patients with epilepsy. *Applied Neuropsychology: Adult, 21*(2), 108-119.

Salazar, X. F., Lu, P. H., Wen, J., & Boone, K. B (2007). The use of effort tests in ethnic minorities and in non-English-speaking and English as a second language populations. In K. B. Boone (Eds.), *Assessment of Feigned Cognitive Impairment: A Neuropsychological Perspective.* 405-427. New York, NY: Guildford Press.

Sawyer, R. J., Young, J. C., Roper, B. L., & Rach, A. (2014). Are verbal intelligence subtests and reading measures immune to non-credible effort? *The Clinical Neuropsychologist, 28*(5), 756-770.

Saykin, A. J., Gur, R. C., Gur, R. E., Shtasel, D. L., Flannery, K. A., Mozley, L. H., ... & Mozley, P. D. (1995). Normative neuropsychological test performance: effects of age, education, gender and ethnicity. *Applied Neuropsychology*, *2*(2), 79-88.

Schroeder, R. W., Martin, P. K., & Odland, A. P. (2016). Expert beliefs and practices regarding neuropsychological validity testing. *The Clinical Neuropsychologist*, *30*(4), 515-535.

Schroeder, R. W., Twumasi-Ankrah, P., Baade, L. E., & Marshall, P. S. (2012). Reliable Digit Span: A Systematic Review and Cross-Validation Study. *Assessment*, *19*(1), 21-30.

Schwartz, B. S., Glass, T. A., Bolla, K. I., Stewart, W. F., Glass, G., Rasmussen, M., ... & Bandeen-Roche, K. (2004). Disparities in cognitive functioning by race/ethnicity in the Baltimore Memory Study. *Environmental Health Perspectives*, *112*(3), 314-320.

Sheng, L., Lu, Y., & Gollan, T. H. (2014). Assessing language dominance in Mandarin–English bilinguals: Convergence and divergence between subjective and objective measures. *Bilingualism: Language and Cognition*, *17*(02), 364-383.

Shura, R. D., Miskey, H. M., Rowland, J. A., Yoash-Gantz, R. E., & Denning, J. H. (2016). Embedded performance validity measures with postdeployment veterans: Cross-validation and efficiency with multiple measures. *Applied Neuropsychology: Adult*, *23*(2), 94-104.

Siegler, R. S., & Mu, Y. (2008). Chinese children excel on novel mathematics problems even before elementary school. *Psychological Science*, *19*(8), 759-763.

Silverberg, N. D., Hanks, R. A., Buchanan, L., Fichtenberg, N., & Millis, S. R. (2008). Detecting response bias with performance patterns on an expanded version of the Controlled Oral Word Association Test. *The Clinical Neuropsychologist, 22*(1), 140-157.

Sisco, S., Gross, A. L., Shih, R. A., Sachs, B. C., Glymour, M. M., Bangen, K. J., ... & Manly, J. J. (2014). The role of early-life educational quality and literacy in explaining racial disparities in cognition in late life. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *70*(4), 557-567.

Slick, D. J., Sherman, E. M., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist*, *13*(4), 545-561.

Spence, S. A., Hunter, M. D., Farrow, T. F., Green, R. D., Leung, D. H., Hughes, C. J., & Ganesan, V. (2004). A cognitive neurobiological account of deception: evidence from functional neuroimaging. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*(1451), 1755-1762.

Spencer, R. J., Axelrod, B. N., Drag, L. L., Waldron-Perrine, B., Pangilinan, P. H., & Bieliauskas, L. A. (2013). WAIS-IV reliable digit span is no more accurate than age corrected scaled score as an indicator of invalid performance in a veteran sample undergoing evaluation for mTBI. *The Clinical Neuropsychologist*, *27*(8), 1362-1372.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine*, *166*(10), 1092-1097.

Statistics Canada. (2011). Linguistic Characteristics of Canadians. Retrieved from http://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-314-x/98-314-x2011001-eng.pdf.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797-811.

Stenclik, J. H., Miele, A. S., Silk-Eglit, G., Lynch, J. K., & McCaffrey, R. J. (2013). Can the Sensitivity and Specificity of the TOMM Be Increased with Differential Cutoff Scores?. *Applied Neuropsychology: Adult*, *20*(4), 243-248.

Stigler, J. W., Lee, S., & Stevenson, H. W. (1986). Digit memory in Chinese and English: Evidence for a temporally limited store. *Cognition, 23*(1), 1–20.

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

Sue, D. W., Arredondo, P., & McDavis, R. J. (1992). Multicultural counseling competencies and standards: A call to the profession. *Journal of Counseling & Development*, *70*(4), 477-486.

Sugarman, M. A., & Axelrod, B. N. (2015). Embedded measures of performance validity using verbal fluency tests in a clinical sample. *Applied Neuropsychology: Adult*, *22*(2), 141-146.

Sugarman, M. A., Holcomb, E. M., Axelrod, B. N., Meyers, J. E., & Liethen, P. C. (2016). Embedded measures of performance validity in the Rey complex figure test in a clinical sample of veterans. *Applied Neuropsychology: Adult*, *23*(2), 105-114.

Suhr, J. A., & Gunstad, J. (2000). The effects of coaching on the sensitivity and specificity of malingering measures. *Archives of Clinical Neuropsychology*, *15*(5), 415-424.

Tombaugh, T. N. (1996). *The test of memory malingering (TOMM)*. Toronto, ON: Multi-Health Systems.

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

Van Dyke, S. A., Millis, S. R., Axelrod, B. N., & Hanks, R. A. (2013). Assessing effort: Differentiating performance and symptom validity. *The Clinical Neuropsychologist*, *27*(8), 1234-1246.

Van Gorp, W.G., Humphrey, L.A., Kalechstein, A., Brumm, V., McMullen, W.J., Stoddard, M. & Pachana, N.A. (1999). How well do standard clinical neuropsychological tests identify malingering? A preliminary analysis. *Journal of Clinical and Experimental Neuropsychology*, *21*(2), 245-250.

Victor, T. L., Boone, K. B., Serpa, J. G., Buehler, J., & Ziegler, E. A. (2009). Interpreting the meaning of multiple symptom validity test failure. *The Clinical Neuropsychologist*, *23*(2), 297-313.

Vilar-López, R., Santiago-Ramajo, S., Gómez-Río, M., Verdejo-García, A., Llamas, J. M., & Pérez-García, M. (2007). Detection of malingering in a Spanish population using three specific malingering tests. *Archives of Clinical Neuropsychology*, *22*(3), 379-388.

Vilar- López, R., Gomez-Rio, M., Caracuel-Romero, A., Llamas-Elvira, J., & Perez-Garcia, M. (2008a). Use of specific malingering measures in a Spanish sample. *Journal of Clinical and Experimental Neuropsychology*, *30*(6), 710-722.

Vilar-López, R., Gómez-Río, M., Santiago-Ramajo, S., Rodríguez-Fernández, A., Puente, A. E., & Pérez-García, M. (2008b). Malingering detection in a Spanish population with a known-groups design. *Archives of Clinical Neuropsychology*, *23*(4), 365-377.

Vogt, C. M., Hocevar, D., & Hagedorn, L. S. (2007). A social cognitive construct validation: Determining women's and men's success in engineering programs. *The Journal of Higher Education*, *78*(3), 337-364.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences*, *10*(4), 141-142.

Walker, A. J., Batchelor, J., & Shores, A. (2009). Effects of education and cultural background on performance on WAIS-III, WMS-III, WAIS-R and WMS-R measures: Systematic review. *Australian Psychologist*, *44*(4), 216-223.

Wang, M. T., & Degol, J. L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review*, *29*(1), 119-140.

Warrington, E. K. (1984). *Recognition Memory Test: manual*. Berkshire, UK; NFER-Nelson.

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

Wechsler, D. (2009). *Advanced clinical solutions for the WAIS-IV and WMS-IV*. San Antonio, TX: Pearson Education.

Weiss, R., & Rosenfeld, B. (2010). Cross-cultural validity in malingering assessment: The Dot Counting Test in a rural Indian sample. *International Journal of Forensic Mental Health, 9*(4), 300-307.

Whiteside, D. M., Kogan, J., Wardin, L., Phillips, D., Franzwa, M. G., Rice, L., ... & Roper, B. (2015). Language-based embedded performance validity measures in traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology*, *37*(2), 220-227.

Whiteside, D., Wald, D., & Busse, M. (2011). Classification accuracy of multiple visual spatial measures in the detection of suspect effort. *The Clinical Neuropsychologist*, *25*(2), 287-301.

Whitney, K. A., Shepard, P. H., Mariner, J., Mossbarger, B., & Herman, S. M. (2010). Validity of the Wechsler Test of Adult Reading (WTAR): Effort considered in a clinical sample of US military veterans. *Applied Neuropsychology*, *17*(3), 196-204.

Wiederholt, W. C., Cahn, D., Butters, N. M., Salmon, D. P., Kritz-Silverstein, D., & Barrett-Connor, E. (1993). Effects of age, gender and education on selected neuropsychological tests in an elderly community cohort. *Journal of the American Geriatrics Society*, *41*(6), 639-647.

Wilkinson, G. S., & Robertson, G. J. (2006). *WRAT 4: Wide Range Achievement Test; Professional manual*. Psychological Assessment Resources, Incorporated.

Wong, T. M., & Fujii, D. E. (2004). Neuropsychological assessment of Asian Americans: Demographic factors, cultural diversity, and practical guidelines. *Applied Neuropsychology*, *11*(1), 23-36.

Yamaguchi, T. (2005). *Detecting malingered memory impairment using the Rey 15-Item Memory Test and the Wechsler Digit Span subtest in a Japanese population* (Unpublished Master’s thesis). Central Missouri State University, Warrensburg, Missouri.

Yang, C. C., Kao, C. J., Cheng, T. W., Yang, C. C., Wang, W. H., Yu, R. L., ... & Hua, M. S. (2012). Cross-cultural effect on suboptimal effort detection: An example of the digit span subtest of the WAIS-III in Taiwan. *Archives of Clinical Neuropsychology*, *27*(8), 869-878.

Young, J. C., Sawyer, R. J., Roper, B. L., & Baughman, B. C. (2012). Expansion and re-examination of Digit Span effort indices on the WAIS-IV. *The Clinical Neuropsychologist*, *26*(1), 147-159.

Zuccato, B. G., Tyson, B. T., & Erdodi, L. A. (2018). Early bird fails the PVT? The effects of timing artifacts on performance validity tests. *Psychological Assessment*, *30*(11), 1491-1498.

APPENDICES

Appendix A: Language Experience and Proficiency Questionnaire (LEAP-Q) Abbreviated

(1) Please list all the languages you know **in order of dominance**:

1	2	3	4	5

(2) Please list all the languages you know **in order of acquisition** (your native language first):

1	2	3	4	5

(3) Please list what percentage of the time you are *currently* and *on average* exposed to each language. (*Your percentages should add up to 100%*):

List language here:					
List percentage here:					

(4) When choosing to **read a text** available in all your languages, in what percentage of cases would you choose to read it in each of your languages? *(Your percentages should add up to 100%*):

List language here:					
List percentage here:					

(5) When choosing a **language to speak** with a person who is equally fluent in all your languages, what percentage of time would you choose to speak each language? (*Your percentages should add up to 100%*):

List language here					
List percentage here:					

(6) Please name the **cultures** with which you identify. On a scale from zero to ten, please <u>circle the extent to which you identify</u> with each culture. (Examples: US-American, Chinese, Jewish Orthodox):

Culture: _______________

0 1 2 3 4 5 6 7 8 9

No identification — Moderate identification — Complete identification

10

Culture: _______________

0 1 2 3 4 5 6 7 8 9

No identification — Moderate identification — Complete identification

10

(7) Date of immigration to Canada, if applicable: ____________________

Language: ______________________________

This is my (**native** **second** **third** **fourth** **fifth**) language.

(1) Age when you learned this language: ____________

(2) Please circle your *level of* ***proficiency*** in speaking, understanding, and reading in this language:

Speaking

0	1	2	3	4	5	6	7	8	9	10
None	Very low	Low	Fair	Slightly less than adequate	Adequate	Slightly more than adequate	Good	Very good	Excellent	Perfect

Understanding spoken language

0	1	2	3	4	5	6	7	8	9	10
None	Very low	Low	Fair	Slightly less than adequate	Adequate	Slightly more than adequate	Good	Very good	Excellent	Perfect

Reading

0	1	2	3	4	5	6	7	8	9	10
None	Very low	Low	Fair	Slightly less than adequate	Adequate	Slightly more than adequate	Good	Very good	Excellent	Perfect

Language: __________________________

This is my (**native** **second** **third** **fourth** **fifth**) language.

(3) Age when you learned this language: ____________

(4) Please circle your *level of* ***proficiency*** in speaking, understanding, and reading in this language:

Speaking

0	1	2	3	4	5	6	7	8	9	10
None	Very low	Low	Fair	Slightly less than adequate	Adequate	Slightly more than adequate	Good	Very good	Excellent	Perfect

Understanding spoken language

0	1	2	3	4	5	6	7	8	9	10
None	Very low	Low	Fair	Slightly less than adequate	Adequate	Slightly more than adequate	Good	Very good	Excellent	Perfect

Reading

0	1	2	3	4	5	6	7	8	9	10
None	Very low	Low	Fair	Slightly less than adequate	Adequate	Slightly more than adequate	Good	Very good	Excellent	Perfect

Appendix B: Intake Questionnaire

Gender: Female ☐ Male ☐ Other ☐

Age: _________________

Handedness: Right ☐ Left ☐ Ambidextrous ☐ (i.e., able to use both hands with equal ease)

Years of Education: ___________

1. Have you ever been diagnosed with one of the following?
   a) Neurological disorder (e.g. dementia, stroke, multiple sclerosis)
      Yes ☐ No ☐
   b) Developmental disability, intellectual disability, or autism spectrum disorders
      Yes ☐ No ☐
   c) Cancer treated with spinal/brain radiation and chemotherapy
      Yes ☐ No ☐
   d) Head injury with loss of consciousness
      Yes ☐ No ☐
   e) Schizophrenia (or other psychotic disorder)
      Yes ☐ No ☐
2. Have you ever been involved in a serious car accident?
   Yes ☐ No ☐

3. What is the highest education of your mother?
   ☐ Less than High School
   ☐ High School (Grade 12 equivalent diploma)
   ☐ College certificate or diploma
   ☐ Bachelor’s degree
   ☐ Master’s degree
   ☐ Doctoral degree

4. What is the highest education of your father:
   ☐ Less than High School
   ☐ High School (Grade 12 equivalent diploma)
   ☐ College certificate or diploma
   ☐ Bachelor’s degree
   ☐ Master’s degree
   ☐ Doctoral degree

Appendix C: Five-Variable Psychiatric Screener (V-5)

Participant: ______________ Date ______________ Time ______________

Please mark the lines below with an “X” to best capture how you feel *right now, at this moment.*

Energy

No energy at all ____ Full of energy

Depression

Not depressed at all ____ Extremely depressed

Anxiety

Not anxious at all ____ Extremely anxious

Fatigue

Not tired at all ____ Extremely tired

Pain

No pain at all ____ Extreme pain

Appendix D.1: Instructions for Experimental-Malingering Group

Imagine that you were in a car accident in which another driver hit your car. You were knocked unconscious, and woke up in the hospital. The doctors told you that you had some bleeding in your brain after the accident.

Because the other driver is at fault, you have decided to take legal action against the driver. Your lawyer said that you may get more money if you look like you have sustained significant injuries because of the accident. You have decided to fake or exaggerate symptoms of a brain injury in order to increase the settlement you will receive. You have been told that common symptoms after a brain injury include difficulties with memory, concentrating, and being slower in responding.

The other driver's lawyer requires you to complete cognitive testing to determine if you sustained significant symptoms because the car accident. You know you can win a better settlement if you can convince the examiner that you have experienced significant brain damage. But if the examiner detects that you are faking, you are likely to lose the lawsuit.

You are about to take a series of cognitive tests that would be used in such a situation. I would like you to pretend you have brain damage, but in a believable way, such that your examiner cannot tell that you are attempting to fake a brain injury.

We recognize that participants may feel uncomfortable being asked to answer questions inaccurately or to deceive someone, and this can cause some anxiety. If you do not want to continue the study, please feel free to let the researcher know. If you feel anxious when the study is over, please let the researcher know before you leave the lab.

Appendix D.2: Instructions for Non-Malingering Control Group

You are about to take a series of cognitive tests. Some of the tests are easy and some are hard. I would like you to try your best on all of the tests.

Appendix E.1: Pre- and Post-Session Questionnaires for Experimental-Malingering Condition

Pre-Session Questionnaire (EM)

1. What are you asked to do for this study?

   A) Try my best on all of the tests

   B) Answer questions truthfully about my academics or career

   C) Pretend I have brain injury when I complete the tests

   D) Complete computerized tests in which I must respond very quickly

2. In this scenario, the character I'm playing can get more money by:

   A) Telling the examiner that I need money

   B) Performing poorly on the memory tests

   C) Pretending my leg is broken

   D) Appearing distressed and uncooperative

3. In this scenario, what will happen if I get caught faking?

   A) Lose the lawsuit

   B) Win more money

   C) Will be hospitalized at the inpatient psychiatric unit

   D) Nothing will happen

Post-session Questionnaire (EM)

What were you asked to do in the beginning of the study?

A) Try my best on all of the tests

B) Answer questions truthfully about my academics or career

C) Pretend I have brain injury when I complete the tests

D) Complete computerized tests in which I must respond very quickly

How much did you try to follow the instructions during testing?

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

Did not try at all

Tried my
absolute best

How much could you <u>imagine</u> the motor vehicle accident scenario described?

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

Not at all

I could imagine
it very vividly

What did you do during testing to pretend that you had cognitive difficulties? (circle as many as applies)

A. I responded to questions and completed tasks slower than usual

B. I answered questions incorrectly even though I knew the answer

C. I acted confused on how to complete the task

D. I asked the examiner to repeat questions

E. I didn’t follow the test instructions

F. I didn’t pretend

G. Other (Explain)

Appendix E.2: Pre- and Post-Session Questionnaires for Non-Malingering Control Condition

Pre-Session Questionnaire (NC)

1. What are you asked to do for this study?

   A) Try my best on all of the tests

   B) Answer questions truthfully about my academics or career

   C) Pretend I have brain injury when I complete the tests

   D) Complete computerized tests in which I must respond very quickly

Post-session Questionnaire (NC)

What were you asked to do in the beginning of the study?

A) Try my best on all of the tests

B) Answer questions truthfully about my academics or career

C) Pretend I have brain injury when I complete the tests

D) Complete computerized tests in which I must respond very quickly

How much did you try to follow the instructions during testing?

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

Did not try at all

Tried my
absolute best

VITA AUCTORIS

NAME:	Kelly An
PLACE OF BIRTH:	Beijing, China
YEAR OF BIRTH:	1989
EDUCATION:	Fredericton High School, Fredericton, NB, 2007
	University of Toronto, H.B.Sc., Toronto, ON, 2011
	University of Windsor, M.A., Windsor, ON, 2014
	University of Windsor, Ph.D., Windsor, ON, 2019