10-30-2020

The Effect of Motivation Status on Performance Validity in Concussion Baseline Testing

Isabelle Messa-Hamidi
University of Windsor

The Effect of Motivation Status on Performance Validity in Concussion Baseline Testing

By

Isabelle Messa

A Thesis
Submitted to the Faculty of Graduate Studies
through the Department of Psychology
in Partial Fulfillment of the Requirements for
the Degree of Master of Arts
at the University of Windsor

Windsor, Ontario, Canada

2020

The Effect of Motivation on Performance Validity in Concussion Baseline Testing

by

Isabelle Messa

APPROVED BY:

_____

A. Bain
Department of Kinesiology

_____

L. Erdodi
Department of Psychology

_____

C. Abeare, Advisor
Department of Psychology

August 19, 2020

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

Baseline neurocognitive testing is routinely conducted in athletes to obtain a point of comparison in the event of a concussion. Differential motivation exists, however, between baseline and post-injury testing, so clinicians must ensure the validity of baseline performance to make valid comparisons post-injury. There is increasing evidence that the validity indicators embedded within the ImPACT, the most widely used test in this context, are insensitive to invalid performance. The objective of the current study was to assess the convergent validity of ImPACT-based EVIs against a battery of well-established PVT/EVIs in an experimental malingering paradigm. Participants were undergraduate students at a Canadian university. Data was collected from 18 participants, 94.4% of whom were female, with a mean age of 21.61 years (SD=4.57). Malingerers had higher base rates of failure on free-standing PVTs, independent EVIs, and ImPACT-based EVIs. Malingerers also had lower neurocognitive performance on all measures, with effect sizes ranging from small-medium to large. All ImPACT Composite scores except for the Reaction Time Composite were significantly lower for experimental malingerers than controls. As expected, the Default EVI demonstrated substantially lower sensitivity than all other ImPACT-based EVIs, though specificity was consistently perfect. Overall, the ImPACT-5 had the best classification accuracy among the ImPACT-based EVIs. Results suggest that clinicians should stay abreast of the literature and use alternate ImPACT-based EVIs when assessing performance validity on ImPACT.

ACKNOWLEDGEMENTS

Thank to my advisor, Dr. Christopher Abeare, for his unwavering support throughout this process; his flexibility, encouragement, and understanding have allowed me to maintain both my sanity and productivity over the course of the past two years, and for that I am truly grateful. I would also like to thank Dr. Erdodi for his kindness and support, and for providing me with every possible opportunity to succeed. Thank you as well to Dr. Bane for his insightful and invaluable contributions to this project.

Thank you to my friends in the program, who have commiserated with me at times when no one else in the world could understand, and specifically to Kassandra Korcsog, without whom I would be lost and drifting through the process without an anchor or a compass.

And a special thank you to my family, without whom none of this would be possible. Thank you to my parents for their constant support and willingness to help in any and all ways (but especially with childcare). Thank you to my husband who has supported me in pursuing my dreams despite the fact that I am well past my expiry date. And thank you to my children, who give me a reason to keep going and trudge forward even when I feel as though I have nothing left. I am so grateful for all of you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS/SYMBOLS

ACSS – Age-corrected scaled score

AUC – Area under the curve

BNT-15 – Boston Naming Test, 15-item short form

BR_{Fail} – Base rate of failure

CD - Coding

CON – Conservative Cut-off

COWAT – Controlled Oral Word Association Test

Cumul. % - Cumulative percent

DM – Design memory total percent correct

DMCD – Design memory correct distractors

DS – Digit Span

EVI – Embedded validity indicator

EWFT – Emotion Word Fluency Test

expMAL – Experimental malingering

FAS – Letter Fluency

f - Frequency

GAD-7 – Generalized Anxiety Disorder, 7-Item Scale

ImPACT – Immediate Post-Concussion Assessment and Cognitive Testing

LIB – Liberal cutoff

LRE – Logistic regression equation

MMPI-2-RF – Minnesota Multiphasic Personality Inventory-2-Restructured Form

MSVT – Medical Symptom Validity Test

NP – Neuropsychological

PHQ-9 – Patient Health Questionnaire

PVT – Performance validity test

ROC – Receiver operating characteristic

RTP – Return to play

SENS – Sensitivity

SPEC – Specificity

SRC – Sport-related concussion

SS – Symbol Search

TMT – Trail Making Test

TOMM – Test of Memory Malingering

VI – Validity Index

WAIS – Weschler Adult Intelligence Scale

WCT – Word Choice Test

WMCD – Word memory correct distractors

WMDM – Word memory delayed memory percent correct

WMLP – Word memory learning percent correct

WMT – Word Memory Test

WRAT-4 – Wide Range Achievement Test, Fourth Edition

XO – Xs and Os total correct interference

CHAPTER 1

Introduction

Sport-related concussion (SRC) is becoming an increasing public health concern, with 1.1 – 1.9 million sport- and recreation-related concussions occurring annually in children 18 years of age or younger in the United States alone (Bryan, Rowhani-Rahbar, Comstock, & Rivara, 2016). Overall, it has been estimated that between 1.6 and 3.8 million SRCs occur in the United States annually, though this number may be an underestimate as many injuries go unrecognized (Langlois et al., 2006). The Concussion in Sport Group defines concussion as a traumatic brain injury induced by biomechanical forces that typically results in the rapid onset of short-lived impairment of neurological function, and resolves spontaneously (McCrory et al., 2018). In some cases, however, symptoms may evolve over time and/or recovery may be more protracted. The injury may or may not involve loss of consciousness and can be caused by either a direct blow to the head or by indirect forces to other areas of the body that are transmitted to the head (McCrory et al., 2018). Symptoms may include somatic, cognitive, emotional, physical, behavioral, and/or sleep disturbances, though the nature, severity, and duration of symptoms are highly variable among individuals (McCrory et al., 2018).

Though there are thought to be microstructural changes underlying concussion symptomatology, the injury is considered largely functional in nature, and thus cannot be seen on conventional neuroimaging modalities (Giza & Hovda, 2014). Specifically, the symptoms of concussion are thought to reflect injury-induced alterations in the functioning of brain tissue, and the resolution of these symptoms, then, to reflect a return to homeostasis (Giza & Hovda, 2014). Because of the heterogeneity of the injury and

lack of objective diagnostic tests, diagnosis and management of concussion has proven extremely difficult. According to a recent consensus statement, if a concussion is suspected, the player should be removed from play and not be permitted to return to play on the day of the injury. It is recommended that athletes rest until they are no longer experiencing symptoms, at which point they can gradually become more cognitively and physically active, as long as their level of activity does not exacerbate symptoms. However, the recommendation of resting until the athlete is symptom-free has been recently called into question (Valovich McLeod, Lewis, Whelihan, & Welch Bacon, 2017). The process of return to play (RTP) should proceed in a graduated manner (McCrory et al., 2018). Though there is no gold-standard way of knowing whether a concussion has occurred/resolved, a multi-faceted approach is recommended, and current practice typically includes assessment of symptomatology, balance, and neurocognitive functioning (Broglio, Guskiewicz, & Norwig, 2017). Clinicians employed by institutions with limited resources, however, may rely more heavily on computerized neuropsychological testing to make decisions regarding management and RTP given the high level of automaticity and relatively fewer resources required to complete this type of assessment (Resch et al., 2013).

Neuropsychological (NP) testing has previously been described as the "cornerstone" of concussion management and has significant clinical value in SRC evaluation (McCrory et al., 2018) as it allows for the assessment of areas of neurocognitive function that are thought to be affected by SRC (e.g., working memory, concentration, processing speed and reaction time) (Schatz, Elbin, Anderson, Savage, & Covassin, 2017). The resolution of concussion symptoms may not always overlap with

recovery of neurocognitive performance (Broglio, Macciocchi, & Ferrara, 2007) and thus the latter adds important information in the context of an assessment, particularly concerning decision making regarding RTP (McCrory et al., 2018). Athletes who are cleared for RTP while brain function is still impaired are at increased risk of reinjury and prolonged symptoms following subsequent injury (Carson et al., 2014). Unfortunately, there is evidence that many athletes exhibit impairments in cerebral function for up to 28 days, well past the period at which most are cleared for RTP (Mayers & Redick, 2012). Moreover, although one should exercise caution in interpreting the often-sensationalized portrayal of possible extreme outcomes of repeated head-injury (Broglio et al., 2017), there is increasing evidence that multiple concussions can increase the risk of cognitive impairment and mental health problems in some athletes (Manley et al., 2017).

Baseline Testing in Concussion

Though norm-group referencing is the standard in NP assessment, it is common practice in the context of SRC to administer pre-injury testing to obtain a baseline level of performance to which post-injury data can then be compared (Broglio et al., 2017). Contrary to most cases of NP assessment, where first patient contact occurs only after the identification of a potential problem, SRC is a unique context where baseline testing prior to injury can be done in a group that is known to incur concussions at a higher rate than the general population. Originally proposed by Barth et al. (1989), this baseline testing model theoretically allows for a more individualized approach, with athletes serving as their own controls, and has been argued to be more suitable for athletes whose neurocognitive performance is either above- or below-average (Schatz & Robertshaw, 2014). Indeed, there are a number of factors that may influence performance on NP tests

including concussion and education history, developmental disorders, cultural and linguistic differences, attention-deficit disorders, and learning disabilities (Echemendia et al., 2012). However, despite its potential to control for individual differences in NP testing, the utility of baseline testing over traditional norm-group referencing has been debated.

Criticisms against baseline testing are broad-ranging, and include arguments such as a lack of evidence that the practice improves diagnostic accuracy (Mayers & Redick, 2012; Randolph & Kirkwood, 2009; Randolph, McCrea, & Barr, 2005), reduces risk associated with the injury (Randolph, 2011), or predicts cognitive decline better than normative comparison (Arnett, Meyer, Merritt, & Guty, 2016; Echemendia et al., 2012). In fact, Echemendia et al., (2012) found that the method of calculating reliable change from baseline used most commonly in the context of concussion predicted cognitive decline at a rate similar to that expected due to chance alone. Moreover, there is concern that test-retest reliability for tests used in this context is unknown for the time intervals over which baseline and post-injury testing are conducted (Arnett et al., 2016), and the test-retest reliabilities that are known are less than optimal, particularly for longer time periods that are most relevant to the baseline-post-injury testing model (Broglio, Ferrara, Macciocchi, Baumgartner, & Elliott, 2007; Iverson, Lovell, & Collins, 2003; Schatz, 2010). Other criticisms, such as the extensive demand on time and resources required to conduct baseline testing as well as concern about practice effects have also been raised (Arnett et al., 2016). Moreover, a recent systematic review concluded that baseline testing using computerized neurocognitive tests in children is not recommended as there is significant variability in their performance over time as a result of age-related cognitive

development (Davis et al., 2017). Despite these criticisms, the practice of baseline testing remains popular.

Another issue with the practice of baseline testing that has received increasing attention in recent years is the assumption that baseline data is an accurate reflection of athletes' ability level (Abeare, Messa, Zuccato, Merker, & Erdodi, 2018). Specifically, there is concern that, though athletes are uniquely motivated to perform well on post-injury assessments in order to be cleared for RTP, the same motivational incentive is absent at baseline (Rabinowitz, Merritt, & Arnett, 2015). This difference in motivation is important, as it has been shown to influence test scores. Bailey, Echemendia, & Arnett (2006), for example, showed that those athletes who were identified as having suspect motivation at baseline testing were more likely to have significant improvements in their scores post-injury than those who had high motivation at baseline. Given the implausibility of the notion that concussion would improve cognitive function, this finding demonstrates that the difference in motivational incentive between pre- and post-injury may render comparisons between the two timepoints meaningless. Rabinowitz, Merritt, & Arnett (2016) also found that athletes who exhibited poor effort toward testing were more likely to trigger indicators suggesting invalid performance on testing.

In addition to concerns about athletes being less motivated to put forward their best effort at baseline vs post-injury testing, there is also evidence that athletes may intentionally suppress performance in order to obtain more favorable post-injury comparisons in the event of an injury (Schatz & Glatts, 2013). Indeed, many athletes wish to avoid removal from play at all costs, as evidenced by research demonstrating that over 50% of football players at the high school and professional levels do not report

concussions or concussion-related symptoms (Schatz, 2018). Of course, lack of motivation and intentional suppression of performance are only a few of the many reasons that baseline data may be invalid, with other reasons including distraction, boredom, and misunderstanding of test instructions, among others. Whatever the reasons for invalid performance, though, if athletes' baseline test scores are not an accurate reflection of their ability level, then athletes may be deemed "recovered" and cleared for RTP prematurely, putting them at increased risk of reinjury and prolonged recovery than if baseline data had not been available.

Performance Validity

Performance validity is the assumption that individuals' performance on NP testing is representative of their actual ability level. If this assumption is violated, interpretation of test results is, at best, a suspect endeavor. It was originally thought that clinical impression was sufficient to determine whether an individual's performance was a valid reflection of their ability level, however this idea has long since been refuted (Heaton, Smith, Lehman, & Vogt, 1978). Though base rates of malingering vary widely across samples and contexts, alarmingly high rates of invalid performance (18.3-36.7%) have been found even in neurologically intact young adults who participate in academic research, with no apparent incentive to underperform (An, Kaploun, Erdodi, & Abeare, 2016). The past two decades have seen a proliferation of research on performance validity testing, and the practice has come to be accepted as a standard component of clinical practice (Heilbronner, Sweet, Morgan, Larrabee, & Millis, 2010). Moreover, it is recommended that multiple measures of performance validity be used throughout testing that tap varying cognitive domains (Heilbronner et al., 2010).

There are two types of measures used to assess performance validity: stand-alone performance validity tests (PVTs) and embedded validity indicators (EVIs). PVTs are tests that were developed with the specific purpose of assessing performance validity, and as such are purposefully insensitive to true cognitive dysfunction. Because the purpose of these tests is to distinguish between non-credible performance and genuine impairment, it is rare for individuals with bona fide disorder to fail PVTs (Larrabee, 2014). EVIs, on the other hand, are, as their name implies, embedded within standardized neuropsychological tests, and as such these tests serve the double purpose of assessing both cognitive function and credibility of performance. A number of EVIs have been developed in recent years within tests spanning various neuropsychological domains, including attention (Abeare et al., 2019), processing speed (Erdodi et al., 2017), visual perception (Rai et al., 2019), executive function (Abeare et al., 2019), motor function (Axelrod et al., 2014; Erdodi et al., 2017), and sensory functioning (Miele, Gunner, Lynch, & Mccaffrey, 2012). Because EVIs are nested within data already being collected for clinical purposes, they are more efficient in terms of time and other resources and may also be less vulnerable to coaching (Miele et al., 2012). EVIs typically have lower signal detection profiles than stand-alone PVTs, though recent research suggests that combining multiple EVIs into a single composite improves signal detection to a rate comparable to standalone PVTs (Erdodi & Lichtenstein, 2017). Importantly, the American Academy of Clinical Neuropsychology recommends the use of both PVTs and EVIs as part of the assessment of performance validity (Heilbronner et al., 2010).

Performance Validity Testing in Concussion

Until recently, relatively little focus has been placed on performance validity testing in the context of concussion baseline testing. In fact, one study demonstrated that only roughly half of athletic trainers examine baseline tests for validity (Covassin, Robert, Iii, Stiller-Ostrowski, & Kontos, 2009). The Immediate Post-Concussion and Cognitive Testing (ImPACT) is a computerized neurocognitive test that is by far the most commonly used test of its kind in the context of SRC. One study found that over 75% of NCAA member institutions use the ImPACT as part of their baseline assessment protocol, while no other neurocognitive test was found to have usage rates over 3% in this context (Kerr et al., 2015). The ImPACT model is based on baseline and post-injury testing, and it is recommended that the presence of meaningful change from baseline scores be assessed via a Reliable Change Index (Iverson et al., 2003; Lovell, 2018), though age- and gender-stratified norms are available for individuals who do not have baseline scores. The test output provides a series of scores, including composite scores for verbal and visual memory, visual motor speed, reaction time, and impulse control. The ImPACT also contains an EVI (which will be referred to as "Default ImPACT EVI" throughout the document) to identify invalid baseline performance (See Table 1 for components of this index). If a profile meets any of the criteria listed in Table 1, the test automatically flags the results as being of "questionable validity", and the ImPACT manual encourages a repeat administration of the baseline exam after discussing the results with the athlete and attempting to identify the reasons for invalid performance. The test manual provides little information on how the Default EVI was developed, though it does cite studies done by Erdal (2012) and Schatz & Glatts (2013) to support

the notion that it does successfully identify a large majority (89-100%) of experimental malingerers. The results of these studies, however, are not accurately portrayed in the manual.

Table 1
ImPACT-Based Validity Indicators

***Default ImPACT EVIs**	**ImPACT "Red Flags"**	**Schatz & Glatts Criteria**
X's and O's Total Incorrect + Color Match Total Commissions > 30	Processing Speed Composite < 25	Word Memory Correct Distractors (WMCD; Immediate + Delayed) < 22
Impulse Control Composite > 30	Reaction Time Composite > 0.8 s	Design Memory Correct Distractors (DMCD; Immediate + Delayed) <16
Word Memory Learning % Correct < 69%	Verbal Memory Composite < 70%	Visual Motor Speed Composite < 25
Design Memory Learning % Correct < 50%	Visual Memory Composite < 60%	Reaction Time Composite > 0.80
Three Letters Total Letters Correct < 8		

*Note that the Default ImPACT EVIs have changed somewhat over time, and the most current version of the EVIs are reported here

Specifically, Erdal (2012) used both the Default ImPACT EVI and "Red Flags" (see Table 1) as validity indicators. The indicators that were found to identify the largest number of experimental malingerers were among the "Red Flag" criteria, which are not automatically flagged by the ImPACT and were not included in the most recent version of the ImPACT manual. Schatz & Glatts (2013) used a combination of the Default EVI and a set of independently developed additional criteria that use a yes/no recognition paradigm more closely resembling traditional stand-alone PVTs (the EVI published by Schatz & Glatts (2013) will be referred to as the "Schatz & Glatts criteria" for the remainder of the document; see Table 1 for a description of these criteria), and again it was found that two of their additional criteria identified substantially higher proportions of naïve and coached malingerers than the Default EVI. In fact, Schatz & Glatts (2013)

report that the Default EVI detected only 70% of naïve and 65% of coached malingerers in their sample. The Default EVI was outperformed not only by the researchers' additional ImPACT-based measures, but by a well-validated stand-alone PVT, the Medical Symptom Validity Test (MSVT), suggesting that the Default EVI is not as sensitive to intentional underperformance as is suggested by the testing manual.

A systematic review was recently conducted to assess both the prevalence of invalid responding on the ImPACT as well as the effectiveness of ImPACT-based EVIs in detecting invalid performance (Gaudet & Weyandt, 2016). The authors reviewed twelve studies that contained information about prevalence rates of invalid performance on baseline testing using the ImPACT, as well as an additional four studies that used experimental malingering paradigms to assess the effectiveness of ImPACT-based EVIs. They found that, of the 12 studies that reported prevalence of invalid baseline data, most relied solely on the validity indicators embedded within the version of the ImPACT being used, and that the reported rates of invalid performance ranged from 2.7% to 27.9%. The weighted prevalence of invalid performance across the 12 studies was 6.1%. Notably, the study reporting the highest rate of invalid performance used both the "Red Flags" and the Default EVI to detect invalid performance (Szabo, Alosco, Fedor, & Gunstad, 2013), and found that the Default EVI alone flagged only 10.4% of their sample as invalid, whereas inclusion of the "Red Flag" criteria identified an additional 17.5%. The next highest reported rate of invalid performance for the online version of the ImPACT was 9.2% (Maerlender & Molfese, 2015), and this study also used other validity indicators in addition to the Default EVI (the authors used two of the additional indicators from the Schatz & Glatts criteria as well as unpublished "local validity criteria"). As such, relying

solely on the Default EVI is likely to artificially suppress rates of invalid performance, and thus the overall weighted estimate provided by Gaudet & Weyandt (2016) is likely an underestimate. Moreover, consensus is clearly lacking regarding what the most appropriate and effective indicators are to determine invalid performance.

In the time since Gaudet & Weyandt (2016) published their review, Higgins, Denney, & Maerlender (2017) developed a logistic regression equation (henceforth referred to as the "Higgins LRE") on which a cut-score of ≥0.23 demonstrated 90.1% specificity and 100% sensitivity in identifying experimental malingerers, whereas the Default EVI identified only 65% of these individuals in their sample. Like Schatz & Glatts (2013), this group also found that Word Memory Learning Percent Correct and Word Memory Delay Memory Correct were most the useful scores for identifying experimental malingerers. The authors postulate that this may be because of the relative ease of the task, where individuals providing their "best effort" normally perform exceptionally well and therefore missing even a few words may be indicative of invalid responding. They suggest that the seeming inability to remember words may be a particularly sensitive indicator of malingering in the context of concussion.

Higgins, Caze, & Maerlender (2018) conducted a follow-up study in which they compared the rates of failure across different validity indicators including the Default EVI, two of the Schatz and Glatts criteria, and the LRE developed by their group and found that the rate of failure using the Default EVI alone (2.2-2.8%) was substantially lower than that determined by all other indicators (10.9-38.8%). Across indicators, they found that 31-39% of the athletes in their sample failed at least one indicator of invalid performance, and between 17-21% failed two or more indicators of invalidity.

Manderino, Zachman, & Gunstad (2018) also compared failure rates between the Default EVI and the Shatz & Glatts criteria in a large sample of NCAA division one athletes (N=1727). They found that, while the Default EVI flagged only 5.8% of their protocols as invalid, the Schatz and Glatts criteria flagged a substantially higher proportion of their protocols as such (25.7%-31.8%). Moreover, higher rates of invalid performance were identified by the Schatz & Glatts criteria even when more conservative cutoffs were used (6.7%-7.3%) (Manderino et al., 2018). These result support the notion that the Schatz & Glatts criteria are more sensitive than the existing Default ImPACT EVI to invalid performance, and the authors posit that this may be due to their reliance on a yes/no recognition paradigm as opposed to a threshold low score. It is also possible that these indicators produce an increased rate of false positive errors. However, the cost of false positives in this context (the need to re-administer the test) seem to outweigh the cost of false negatives (prematurely clearing an athlete for RTP), and so the argument could be made that validity indicators should seek to maximize sensitivity even if this comes at somewhat of a cost to specificity (Manderino et al., 2018). Manderino & Gunstad (2018) recently examined the classification accuracy and concurrent validity of the Default EVI and three proposed validity indices (word memory correct distractors (WDCD) and design memory correct distractors (DMCD) as proposed by (Schatz & Glatts, 2013), as well as total symptom score) using an experimental malingering paradigm. In addition to the ImPACT, they administered the Word Memory Test (WMT) and the Minnesota Multiphasic Personality Inventory – 2 – Restructured Form (MMPI-2-RF). The authors found that the Default EVI had the highest specificity, but that this came at the expense of significantly lower sensitivity than all other validity indices tested.

Even when specificity was held to the standard of .90, however, the sensitivity of the WMCD outperformed the Default EVI.

Raab, Peak, & Knoderer (2019) also conducted a study using an experimental malingering paradigm and found that 50% of experimental malingerers were not identified by the Default EVI. The authors propose the use of any composite score at or below the first percentile as another indicator of potentially invalid performance, as this had a superior signal detection profile to the Default EVI in their sample. Walton, Broshek, Freeman, Cullum, & Resch (2017), however, also conducted a study involving 769 athletes completing baseline assessments using the ImPACT. Though only 1% of their sample was flagged as invalid by the Default EVI, they required all individuals scoring below the 16th percentile relative to normative data (14.6% of their sample) on any neurocognitive index to retake the test. After readministration, 88% of those who previously scored below the 16th percentile subsequently scored above this threshold, suggesting that the original baseline data was not indicative of their true ability level. Some caution is warranted in the interpretation of these findings, however, as the reliable change index must be considered in order to determine the degree to which changes in scores between administrations exceed what would be expected based on random variability alone; it is not clear if this was accounted for by Walton et al. (2017).

Overall, the evidence suggests that the Default EVI is not sufficiently sensitive to invalid performance, and thus the prevalence of invalid performance on baseline testing using the ImPACT is likely substantially higher than reported by Gaudet & Weyandt (2016). Our group recently compared rates of failure across several validity indicators in a large sample of athletes undergoing baseline testing (N=7897; Abeare et al., 2018).

Interestingly, we found that the rate of invalid performance as determined by the Default EVI was 6.4%, which is remarkably similar to the rate reported by Gaudet & Weyandt (2016). Unsurprisingly, the rate of invalid performance identified by all other indicators was substantially higher (31.8% for the Reg Flags, 34.9% for the Higgins LRE, and 47.6% for the Schatz & Glatts criteria). The cumulative base rate of failure in our sample was 55.7%, though there was a remarkable difference between younger and older age groups (83.6% cumulative rate of failure in 10-year-olds vs 29.2% in 21-year-olds).

Relatively little work has been done to examine the convergent validity of the Default EVI with other well-validated PVTs and EVIs used in neuropsychological assessment. As previously mentioned, Schatz & Glatts (2013) administered the MSVT along with the ImPACT and found that, whereas the Default ImPACT EVIs identified only 60% of naïve and 75% of coached malingerers, the MSVT identified 80%, and 90% of these individuals, respectively. Manderino & Gunstad (2018) administered the ImPACT and the WMT within an experimental malingering paradigm and found that the WMT had significantly higher sensitivity and lower specificity than the Default EVI. Both the MSVT and WMT are free-standing PVTs that are based on forced-choice recognition paradigms, and both use a threshold low score as the measure of invalid performance. Given the limited scope of these PVTs, our group recently administered an extensive battery of well-validated PVTs/EVIs along with the ImPACT to a group of collegiate football players as part of their baseline testing protocol to examine the convergent validity of the Default EVI and other ImPACT-based validity indicators (Abeare et al., 2019). We found that the base rate of failure on both free-standing PVTs and EVIs was variable (between 1.2 and 12% for PVTs and between 1.2 and 19.3% for

EVIs), but, when these measures were combined, roughly half of the athletes (49.4%) had one or more indicators of invalid performance. Conversely, the Default EVI identified only 1.2% of our sample as invalid, though the alternative ImPACT-based EVIs flagged considerably higher proportions of the sample (Red Flags, 24.1%; Higgins LRE, 39.8%, Schatz & Glatts criteria, 41%). Together, the ImPACT-based EVIs identified 51.8% of the sample as having one or more indicators of invalid performance, which was strikingly similar to the cumulative percentage of the external PVT/EVIs (49.4%). Another important aspect of this study was the inclusion of an incentivized control group. Specifically, a group of 140 undergraduate students from the same university were administered a highly overlapping set of neuropsychological tests as part of a classroom exercise, allowing for a comparison of their performance with that of the student athletes. As an incentive to perform well, students were required to earn participation points as part of their final grade based on their performance on these tasks (i.e., failing validity cutoffs resulted in a loss of points). As a result, this group not only lacked any apparent incentive to underperform but was also expressly motivated to perform well in order to maximize their grade in the course. We found that, though their base rate of failure (BR_{Fail}) on free-standing PVTs was similar to that of their athlete peers (failure rates of 1.2-12.0% for athletes vs 1.4-7.7% for controls), the nonathlete controls had noticeably lower BR_{Fail} on EVIs (1.2-10.8% for athletes vs 0.0-2.7% for controls). Moreover, on measures of cognitive ability, the controls outperformed even athletes who passed all PVTs, which underscores the point that the absence of motivation to perform poorly is fundamentally different from the presence of motivation to perform well.

In a follow-up study using the same athlete sample, Erdodi et al. (2020) examined the classification accuracy of the existing ImPACT-based EVIs against a multivariate criterion PVT and found that the Default EVI had perfect specificity, but that this came at the expense of extremely low sensitivity (0.04). The Red Flags had acceptable specificity (0.85) with moderate sensitivity (0.43), while the Schatz and Glatts criteria and Higgins LRE had very similar classification accuracies (sensitivity of 0.68 and specificity of 0.73-0.75). However, though the two independently developed EVIs had the highest sensitivities, they also had high false-positive rates (15.5-19.2%). The authors proposed two new ImPACT-based EVIs (ImPACT 5A and B) which had more favorable signal detection properties relative to existing ImPACT-based EVIs. Moreover, the ImPACT 5A and B are based on composite scores rather than subtest scores, which improves reliability, and were calibrated against a multivariate criterion PVT which combined several different EVIs and different types of detection methods. In addition, they provide both liberal and conservative cutoffs, which allow flexibility when deciding whether to prioritize sensitivity or specificity in any given context (Erdodi et al., 2020).

The current study extends on our previous work by employing a battery of well-validated PVTs and EVIs alongside the ImPACT in an experimental malingering paradigm. This addresses a gap in the literature, as there has been very limited use of independent performance validity measures in previous studies that have used an experimental malingering paradigm to evaluate ImPACT-based EVIs. Moreover, no previous experimental malingering study has assessed the relative effectiveness of all existing ImPACT-based EVIs in one sample (See Table 2 for components of each ImPACT-based EVI). Given their recent publication, this is also the first study to assess

the classification accuracy of the ImPACT-5 in an experimental malingering paradigm, as well as the first to assess the classification accuracy of the Higgins LRE outside of the original sample of experimental malingerers from which it was conceived.

Table 2
Components of ImPACT-Based Validity Indicators

Indicator	**Scale**	**Cutoff**
Default ImPACT EVIs	X's and O's Total Incorrect + Color Match Total Commissions	>30
	Impulse Control Composite	>30
	Word Memory Learning % Correct	< 69%
	Design Memory Learning % Correct < 50%	< 50%
	Three Letters Total Letters Correct < 8	<8
"Red Flags"	Processing Speed Composite	< 25
	Reaction Time Composite	>0.8 s
	Verbal Memory Composite	< 70%
	Visual Memory Composite	< 60%
Schatz & Glatts Criteria	Word Memory Correct Distractors (WMCD; Immediate + Delayed)	< 22
	Design Memory Correct Distractors (DMCD; Immediate + Delayed)	< 16
Higgins LRE	$e^{(56.74-(0.15*WM\ LP)-(0.18*WM\ DM)-(0.13*DM)-(0.17*XO))}/1+e^{(56.74-(0.15*WM\ LP)-(0.18*WM\ DM)-(0.13*DM)-(0.17*XO))}$	≥ .23
ImPACT-5A(B)	Verbal Memory Composite	≤ 78 (≤ 76)
	Visual Memory Composite	≤ 65 (≤ 57)
	Visuomotor Composite	≤ 34 (≤ 33)
	Reaction Time Composite	≥ .67 (≥ .71)
	Impulse Control Composite	≥ 8 (≥ 11)

Note: WM LP = Word memory learning percent correct; WM DP = Word memory delayed memory percent correct; DM = Design memory total percent correct; XO = X's and O's total correct (interference).

Another important point that has not been addressed in previous experimental malingering studies assessing ImPACT-based EVIs is the base rate of failure in the control group. Unlike studies such as Erdodi et al (2020), which employ independent measures of performance validity in order to distinguish valid vs invalid performance, experimental malingering studies generally classify performance based only on the set of instructions given to participants. It follows that, because control participants are told

some variation of "do your best", any controls flagged by the EVIs under examination are categorized as false positives and diminish the measure's resulting classification accuracy. For this to be true, however, the actual rate of invalid performance in the control group must be zero, which is highly unlikely; as previously mentioned, rates of invalid performance in cognitively intact undergraduate research participants with no incentive to underperform have been shown to be relatively high, ranging from 18-36% (An et al., 2018). These individuals may underperform for many reasons, including boredom, inattention, or a failure to appreciate the importance of giving their best effort. Regardless of the etiology of invalid performance, however, it can be reasonably expected that invalid performance from controls contaminates criterion groups in experimental malingering paradigms. The current study attempts to address this limitation by calculating classification accuracy for ImPACT-based EVIs against criterion groups of both experimentally and psychometrically defined invalid performance.

Study Objectives and Hypotheses

1) To determine the classification accuracy (i.e., sensitivity and specificity) of various ImPACT-based EVIs in distinguishing between valid vs invalid performance. Based on previous findings, the Default EVI was expected to have the lowest sensitivity, but the highest specificity, to both experimentally and psychometrically defined invalid performance. Based on the findings of Erdodi et al. (2020), we anticipated that the Schatz and Glatts criteria and Higgins LRE would demonstrate similar classification accuracy, with the highest sensitivity, but lowest specificity among the ImPACT-based EVIs. The Red Flags and

ImPACT-5 were expected to demonstrate higher sensitivity, but lower specificity than the Default EVI, while demonstrating lower sensitivity, but higher specificity, than both the Higgins LRE and Schatz and Glatts criteria.

2) To determine the effect of performance validity on neurocognitive performance. We predicted that individuals demonstrating invalid performance profiles would perform significantly worse on both ImPACT-based and non-ImPACT-based measures of neurocognitive performance.

CHAPTER 2

Methods

Undergraduate students from the University of Windsor were recruited from the University's participant pool. Participants were randomly assigned to either the experimental malingering or control condition. Controls were explicitly told about the importance of exerting their best effort on testing and were asked to do so. Experimental Malingerers, on the other hand, were presented with a scenario in which they were asked to imagine they are a varsity athlete whose prospects for a career in professional sports depend on remaining in play for the duration of the season. They were told that the testing is intended to measure their baseline level of cognitive functioning and would be used as a comparison in the event of a head injury to determine whether they need to be removed from play. They were then told to intentionally underperform on testing to ensure that, if they did sustain a head injury, their post-injury scores would not be lower than their baseline, and they would not be removed from play. They were warned not to underperform so egregiously that it becomes obvious that they were trying to "trick the test". As an incentive to malinger without being detected, they were also told that those

who most successfully underperform without detection would receive higher monetary compensation in a future study, should they choose to participate. In reality, all individuals who choose to participate in the future study will be compensated equally (See Appendix A for scripts of instructions given to both controls and experimental malingerers).

Unfortunately, due to the restrictions associated with COVID-19, data was collected from only 18 participants (nine in each condition) before data collection was no longer possible. Participants were randomly assigned to conditions at a 1:1 ratio and the researcher conducting the testing was blind to experimental condition.

Measures

Each participant completed a battery of tests comprised of the following:

Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) – ImPACT is a computer-based neurocognitive test that is designed to measure aspects of attention, memory, visuo-spatial processing, impulse control, and processing speed in individuals from 12 to 59 years of age. The normative sample consisted of 16,566 athletes, though the older age groups were comprised teachers, coaches, school administrators and adult athletes. The test begins with a collection of demographic information, followed by a self-report concussion symptom scale. The neurocognitive test modules are then administered in the following order: Word Memory, Design Memory, X's and O's, Symbol Match, Color Match, Three Letters, Word Memory Delayed Recall, Design Memory Delayed Recall (see Appendix B for a description of each subtest). Test administration can generally be completed within 20 minutes, and all scoring is

automatically completed by the software. In addition to specific scores that are provided for each module, the following composite scores are also reported: Verbal Memory, Visual Memory, Visual Motor Speed, Reaction Time, and Impulse Control. A Total Symptom Composite Score is also provided, in addition to a Cognitive Efficiency Index. The test has been found to be sensitive to the effects of concussion with high sensitivity (0.82) and specificity (0.89)(Schatz, Pardini, Lovell, Collins, & Podell, 2006). Convergent validity of the ImPACT has been demonstrated against traditional neuropsychological tests (Maerlender et al., 2010), though a recent meta-analysis found unacceptably low test-retest reliabilities (intraclass correlation coefficients ranging from 0.52 for Verbal Memory to 0.77 for Visual-motor Speed) (Farnsworth, Dargo, Ragan, & Kang, 2017).

Letter Fluency – Letter Fluency from the Controlled Oral Word Association Test (COWAT; Benton & Hamsher, 1978; Gladsjo et al., 1999) is a task in which individuals are given a letter of the alphabet and asked to generate as many words as possible in one minute. Heaton norms were used, which correct for age, education, and ethnicity (Heaton et al., 2004). Curtis, Thompson, Greve, & Bianchini (2008) found that a Total Correct word T-score accurately differentiated malingered neurocognitive dysfunction from non-malingered neurocognitive dysfunction in mild TBI patients, with malingerers 4.3 times more likely to score at or below a cutoff of 33 than non-malingerers. In a sample of undergraduate students, Hurtubise et al. (2020) found that a T-score of ≤ 29 produced a good combination of sensitivity (0.40-0.42) and specificity (0.89-0.95).

Animal Fluency – The Animal Fluency task, also from the COWAT (Benton & Hamsher, 1978; Gladsjo et al., 1999), asks participants to name as many animals as

possible in one minute and is a measure of semantic fluency. Heaton norms were also used for Animal fluency, once again correcting for age, education, and ethnicity (Heaton et al., 2004). Hurtubise et al. (2020) found that a T-score cutoff of ≤31 demonstrated a good combination of sensitivity (0.53-0.71) and specificity (0.86-0.93) in their sample of undergraduate students.

Coding (CD) – CD is a subtest of the Weschler Adult Intelligence Scale – Fourth Edition (WAIS-IV) (Weschler, 2008) that requires participants to rapidly transcribe symbols associated with number-symbol pairs. Erdodi et al., (2017) found that a scaled score of ≤ 5 on the CD subtest of the WAIS-IV identified invalid performance with a specificity of .94-1.0 and sensitivity of .04-.28.

Symbol Search (SS) – SS is also a subtest of the WAIS-IV (Wechsler, 2008), which requires participants to search for and identify target symbols among distractors as quickly as possible. Erdodi et al. (2017) also found that a scaled score of ≤ 6 on the SS subtest of the WAIS-IV identified invalid performance with a sensitivity of .38-.64 and a specificity of .88-.93. The researchers also found that a CD minus SS (|CD-SS|) scaled score difference of ≥ 5 had a specificity of .89-.91 and sensitivity of .08-.12 in identifying invalid performance.

Digit Span (DS) – The DS subtest of the WAIS-III (Wechsler, 1997) requires participants to listen to and repeat back lists of digits of increasing length both forwards and backwards. Erdodi & Lichtenstein (2017) found that an Age-Corrected Scaled Score (ACSS) cutoff of ≤ 6 on this subtest identified invalid performance with acceptable specificity (.87-.90) but low sensitivity (.28-.32). Shura et al., (2020) found that, though

the ACSS produced the highest area under the curve (AUC) of all Digit Span-based EVIs, sensitivity at the best cutoff of <7 was quite low (0.17).

Word Choice Test (WCT) – The WCT (Pearson, 2009) is a recognition memory task involving the serial presentation 50 words followed by a forced-choice recognition task. The words are highly imageable and concrete, making the discrimination of targets from foils a simple task. Even in clinical settings, credible patients tend to perform near ceiling, with means above 49 (Davis, 2014). The technical manual provides a cutoff for invalid performance of ≤ 32-47, with 32 representing the upper limit of theoretical chance-level responding (Erdodi et al., 2018). Erdodi et al. (2017) found that a cutoff of ≤ 47 achieved the best classification accuracy in their study, with a sensitivity of 0.57 and a specificity of 0.87.

Rey-15 with Recognition– They Rey 15-item Memorization Test (Rey, 1964) was developed as a measure of performance validity and requires participants to memorize a page of 15 symbols. The symbols are related in various ways, making the task relatively simple, though it appears on its surface to be a somewhat challenging task because of the fairly large number of items to remember. Boone, Salazar, Lu, Warner-Chacon, & Razani (2002) found that a free recall cutoff of < 9 had good specificity (0.97-1.00), but modest sensitivity (0.47). However, using a combined recall and recognition score (free recall + [recognition-false positives] < 20) greatly increased sensitivity (0.71) and maintained high specificity (≥0.92). Poynter et al. (2019) found that a combined recall and recognition score of ≤ 22 produced adequate sensitivity (0.61) and high specificity (0.93).

Test of Memory Malingering Trial 1 (TOMM-1) – The TOMM (Tombaugh, 1996) is a recognition memory task, and trial one involves the serial presentation of 50 line-drawings of common items followed by a forced-choice recognition task. TOMM-1 has demonstrated good classification accuracy at cutoffs ranging from ≤ 35 to ≤ 45 against various criterion groups (Rai & Erdodi, 2019). Martin et al. (2019) recently conducted a meta-analysis and found that a cutoff of <42 for Trial 1 produced the highest sensitivity (0.59-0.70) while maintaining specificity at≥ 0.90.

B**oston Naming Test (BNT) -15** – The BNT-15 (Mack et al., 1992) is a 15-item short-form of the original Boston Naming Test, in which a series of 15 line drawings of objects are shown to an individual who is asked to name the object. The BNT-15 has been shown to function as an index of English language proficiency and predict the poorer performance of individuals with limited English proficiency on neuropsychological tests with high verbal mediation (Erdodi, Jongsma, & Issa, 2016).

Emotion Word Fluency Test (EWFT) - The EWFT (Abeare et al., 2017) asks participants to name as many emotion words as possible in one minute and is a measure of semantic fluency.

Trail Making Test (TMT) A and B – TMT (Reitan, 1955) is a neuropsychological test that is commonly used to assess executive functioning, attention, and visuomotor skills. Heaton norms were used for the TMT, once again correcting for age, education, and ethnicity (Heaton et al., 2004). TMT A presents individuals with a page of randomly dispersed numbers and asks them to connect them with a line, in order, as quickly as possible, and on TMT B numbers and letters are dispersed together, and

individuals are asked to alternate from numbers to letters, in order. Abeare et al. (2019) found that cutoffs of T ≤ 33 on both TMT A and B had superior classification accuracy to raw score cutoffs reported in the literature, eliminating age and education bias observed in raw score cutoffs.

Wide Range Achievement Test, Fourth Edition (WRAT-4), Reading Subtest – The Reading subtest of the WRAT-4 (Wilkinson & Robertson, 2006) is a test of word reading that is often used as a measure of pre-morbid functioning in the context of brain injury (Orme et al., 2004).

Generalized Anxiety Disorder 7-Item (GAD-7) Scale – The GAD-7 (Spitzer et al., 2006) is a seven-item self-report scale that is intended to identify probably cases of generalized anxiety disorder. The measure has been found to be a valid and efficient tool for screening and assessing the severity of generalized anxiety disorder in both clinical practice and research (Spitzer et al., 2006).

Patient Health Questionnaire (PHQ-9) – The PHQ-9 (Kroenke et al., 2001) is a nine-item, self-report scale that has been shown to be a valid measure of depression severity.

V8 – The V8 is an eight-variable psychiatric screener measuring energy, depression, anxiety, pain, fatigue, happiness, stress, and motivation on a visual analog scale (Erdodi et al., 2020). The individual is asked to mark an X along a 10 cm line, indicating the point that best captures how they are feeling in the moment.

Post-Assessment Survey – After completion of the test battery, participants were asked to complete a survey in order to assess the degree to which they understood and

complied with their particular set of instructions. This was to serve as a manipulation check.

Data Analysis

Descriptive Statistics

Descriptive statistics were calculated for demographic variables such as gender, age, education, and race. Base rates of failure (i.e., the proportion of participants whose scores fell below the respective cutoffs) were calculated using appropriate cutoffs for each of the five ImPACT-based EVIs, and at both liberal and conservative cutoffs for each of the non-ImPACT-based PVT/EVIs (See Tables 3 and 4 for cutoffs used for free-standing PVTs and EVIs, respectively). Though failure rates are traditionally compared using the Chi Square test of independence, our small sample size precluded us from performing this statistical comparison between groups as the expected frequency of many cells was lower than the minimum of five required to conduct the test. As such, failure rates are presented only as descriptive frequencies.

Cumulative Failure Rate

Cumulative failure rates were also calculated for non-ImPACT-based PVT/EVIs for each group at both liberal and conservative cutoffs. In order to determine whether there was a difference in the overall number of independent PVT/EVI failures between groups, a dummy variable was created for each test such that 0=Pass and 1=Fail. Four composite scores were then created: the "Validity Index 11" (VI-11, liberal and conservative) were created by summing the dummy variables for each of the 11 non-ImPACT-based PVT/EVIs at liberal and conservative cutoffs, and the VI-9 (liberal and

conservative) were created by doing the same, but excluding TMT A and B. The latter was done because data for 22% of the experimental malingering group were missing for TMT A and B due to errors in administration. As a result, for the purposes of calculating of the VI-11, both measures were coded as "Pass" for these participants in the absence of other information. The practice of coding missing data as "Pass" in this context, however, has the potential to inflate false negative rates, contaminate criterion groups, and compromise classification accuracy (Erdodi, 2017). As such, the VI-9 represents the cumulative failure rate on all tests for which there was complete data for the entire sample. Group scores were compared using t-tests, and effect sizes are reported as Hedge's g, as this measure of effect size is most appropriate with small samples.

Given the small sample size, a power analysis was conducted to estimate the power to detect a difference in cumulative failure rate if one was indeed present. A conservatively estimated effect size of d=1.0 was used. This estimate was informed by An et al (2019) and Hurtubise et al (2020), who demonstrated significant differences on multivariate validity indices between experimental malingerers and controls with effect sizes of d=1.34 and d=1.49, respectively. At an alpha level of 0.05, the current study was found to be adequately powered (0.83).

Neurocognitive performance

Neurocognitive performance was compared across experimental groups on both independent EVIs and ImPACT composite scores using t-tests. Once again, a power analysis was conducted with a conservatively estimated effect size of d=1.00. This estimate was based on a previous study by Hurtubise et al (2020), which demonstrated significant differences between experimental malingerers and controls on many of the

same neurocognitive tests used in the current study. Effect sizes in that study ranged from 0.62-1.69, with a mean of 1.14. As with cumulative failure rate, the current study was adequately powered (0.83) to detect a difference in neurocognitive performance between experimental malingerers and controls if one was indeed present, at an alpha level of 0.05.

In order to analyze the effect of PVT/EVI failure on neurocognitive performance independent of group assignment, experimental malingering and control groups were collapsed and the sample was split into groups based on the number of PVT/EVIs failed by each participant. This was done by creating a dichotomous criterion of ≤ 1 = Pass and ≥2 = Fail on both the VI-11 and VI-9 (at both liberal and conservative cutoffs), and then comparing neurocognitive performance between these groups using t-tests.

Classification Accuracy

Sensitivity and specificity for ImPACT-based EVIs were calculated using standard formulas against criteria of experimental group as well as both VI-11 and VI-9 at liberal and conservative cutoffs. Classification accuracy of ImPACT-based EVIs were calculated using AUC of receiver operator characteristic (ROC) curves. All statistical analyses were conducted using SPSS 26.0.

Chapter 3

Results

Demographics

Data for a total of 18 participants was collected (nine in each group). ImPACT data for one participant in the experimental malingering group was lost due to technical

difficulties with the online test, leaving ImPACT data for only eight participants in this group. In addition, as previously mentioned, two individuals in the experimental malingering group had missing data for TMT A and B due to errors in administration. The vast majority of participants were female (94.4%), the mean age was 21.61 years (SD=4.57), and the mean number of years of education was 13.61 (SD=1.38). The self-identified racial composition of the sample was 38.9% White, 27.8% Black, 5.6% Asian, and 27.8% Other. None of the participants endorsed ever having been diagnosed with a learning disability, ADD/ADHD, or autism. Fifteen participants (83%) indicated that English was their native language, while three (16.7%) participants indicated languages other than English as their native language (two participants were native Arabic speakers, and one participant was a native speaker of Kinyarwanda). The BNT-15 was used as a measure of English language proficiency, and there were no significant differences found between experimental malingerers (M=11.67, SD=1.66) and controls (M=12.11, SD=2.34; t(16)=0.46, p=0.65, g=0.22). There were also no differences between those who scored ≤1 (M=12.40, SD=2.30) vs. ≥2 (M=11.69, SD=1.93) on the VI-11 at liberal cutoffs (t(16)=0.66, p=0.52, g=0.35), or between those who scored ≤1 (M=12.63, SD=2.01) vs. ≥2 (M=11.30, SD=1.83) on the VI-11 at conservative cutoffs (t(16)=1.44, p=0.17, g=0.70). There were, however, significant differences between those who scored ≤1 (M=12.89, SD=1.90) vs. ≥2 (M=10.89, SD=1.62) on the VI-9 at liberal cutoffs (t(16)=2.41, p=0.03, g=1.13), as well as between those who scored ≤1 (M=12.67, SD=1.72) vs. ≥2 (M=10.33, SD=1.63) on the VI-9 at conservative cutoffs (t(16)=2.75, p=0.01, g=1.38).

Base Rates of Failure

Free-standing PVTs

As expected, base rates of failure on free-standing PVTs were considerably higher for experimental malingerers than controls at both liberal (ranging from 0-77.8% for malingerers and 0-33.3% for controls) and conservative cutoffs (ranging from 0-55.6% for malingerers and 0-11.1% for controls). See Table 3 for a summary of performance and base rates of failure on free-standing PVTs.

Table 3
Descriptive Statistics for Free-Standing Performance Validity Tests

		Mean (SD)	Range	Mean (SD)	Range		BRfail			
Test	Scale	Control		expMAL		Cutoff	Control (%)	expMAL (%)	Sens	Spec
Rey-15	FR+REC	28.44 (2.17)	24-30	28.78 (1.64)	25-30	≤ 23[a]	0 (0.0)	0 (0.0)	0.00	1.00
						≤ 20[b]	0 (0.0)	0 (0.0)	0.00	1.00
WCT	Raw score	49.78 (.44)	49-50	44.11 (6.66)	32-50	≤ 47[a]	0 (0.0)	5 (55.6)	0.56	1.00
						≤ 45[b]	0 (0.0)	4 (44.4)	0.44	1.00
TOMM-1	Raw score	46.67 (4.06)	39-50	39.11 (6.59)	27-49	≤ 43[a]	3 (33.3)	7 (77.8)	0.78	0.67
						≤ 40[b]	1 (11.1)	5 (55.6)	0.56	0.89

Note: expMAL = Experimental malingerers; BRfail = Base rate of failure; Sens = Sensitivity; Spec = Specificity; [a]Liberal cut-offs; [b]Conservative cut-off

Embedded Validity Indicators

Base rates of failure on EVIs were also higher for experimental malingerers than controls at both liberal and conservative cutoffs (ranging from 0-71.4% for malingerers and 0-55.6% for controls at both liberal and conservative cutoffs). See Table 4 for a summary of performance and base rates of failure on free-standing EVIs.

Table 4

Descriptive Statistics for Embedded Performance Validity Indicators

		Mean	Range	Mean	Range		BRfail			
Test	Scale	Control		expMAL		Cutoff	Control	expMAL	Sens	Spec
FAS	T-score	45.56 (7.33)	34-57	43 (7.37)	36-60	≤ 33[a]	0 (0.0)	0 (0.0)	0.00	1.00
						≤ 29[b]	0 (0.0)	0 (0.0)	0.00	1.00
Animals	T-score	44.06 (7.56)	32-55	40.5 (10.34)	22-57	≤ 31[a]	0 (0.0)	1 (11.1)	0.11	1.00
						≤ 29[b]	0 (0.0)	1 (11.1)	0.11	1.00
CD	ACSS	8.89 (3.30)	5-13	7.89 (2.03)	5-10	≤ 5[a]	2 (22.2)	2 (22.2)	0.22	0.78
						≤ 4[b]	0 (0.0)	0 (0.0)	0.00	1.00
SS	ACSS	7.89 (3.02)	4-13	6.44 (4.0)	1-14	≤ 6[a]	3 (33.3)	5 (55.6)	0.56	0.67
						≤ 5[b]	3 (33.3)	4 (44.4)	0.44	0.67
|CD-SS|	ACSS	3.00 (2.40)	0-6	3.22 (1.48)	1-5	≥ 3[a]	3 (33.3)	5 (55.6)	0.56	0.67
						≥ 5[b]	3 (33.3)	2 (22.2)	0.22	0.67
DS	ACSS	10.67 (2.35)	7-15	8.00 (2.83)	5-12	≤ 6[a]	0 (0.0)	4 (44.4)	0.44	1.00
						≤ 5[b]	0 (0.0)	2 (22.2)	0.22	1.00
TMT-A	T-score	41.33 (12.07)	27-64	34.00 (7.07)	24-44	≤ 37[a]	5 (55.6)	5 (71.4*)	0.71	0.44
						≤ 35[b]	5 (55.6)	5 (71.4*)	0.71	0.44
TMT-B	T-score	48.94 (8.32)	37-64	43.07 (11.94)	23-56	≤ 35[a]	0 (0.0)	2 (28.6*)	0.29	1.00
						≤ 33[b]	0 (0.0)	2 (28.6*)	0.29	1.00

Note: expMAL = Experimental malingerers; BRfail = Base rate of failure; Sens = Sensitivity; Spec = Specificity; FAS = Letter fluency; Animals = Animal fluency; [a]liberal cut-offs; [b]conservative cut-offs; *TMT A and B data were available for only 7 participants in the experimental malingering group

Cumulative Failures

When all 11 independent PVT/EVIs were considered, 77.8% of the control group failed at least one PVT/EVI at liberal cut-offs, and 66.7% failed at least one PVT/EVI at conservative cut-offs. Two thirds of the control group (66.7%) failed two or more PVT/EVIs at liberal cut-offs and 44.4% failed two or more at conservative cut-offs. In the experimental malingering group, 100% of the sample failed at least one PVT/EVI at both

cutoffs, 77.8% failed two or more at liberal cutoffs, and 66.7% failed two or more at conservative cutoffs (Table 5).

Table 5
Frequency Distribution of PVT+EVI Failures

	Liberal Cutoffs						Conservative Cutoffs					
	Control			expMAL			Control			expMAL		
# failed	f	%	Cumul. %	f	%	Cumul. %	f	%	Cumul. %	f	%	Cumul. %
0	2	22.2	22.2	0	0.0	0.0	3	33.3	33.3	0	0.0	0.0
1	1	11.1	33.3	2	22.2	22.2	2	22.2	55.6	3	33.3	33.3
2	3	33.3	66.7	0	0.0	22.2	3	33.3	88.9	1	11.1	44.4
3	0	0.0	66.7	2	22.2	44.4	0	0.0	88.9	2	22.2	66.7
4	2	22.2	88.9	1	11.1	55.6	1	11.1	100.0	2	22.2	88.9
5	1	11.1	100.0	2	22.2	77.8	0	0.0	100.0	0	0.0	88.9
6	0	0.0	100.0	1	11.1	88.9	0	0.0	100.0	1	11.1	100.0
7	0	0.0	100.0	0	0.0	88.9						
8	0	0.0	100.0	0	0.0	88.9						
9	0	0.0	100.0	0	0.0	88.9						
10	0	0.0	100.0	1	11.1	100.0						

Note: expMAL = experimental malingerers; f = Frequency; Cumul. % = Cumulative percent.

When Trails A and B were excluded, 77.8% of the control group failed at least one PVT/EVI at liberal cut-offs, and 55.6% failed at least one PVT/EVI at conservative cut-offs. One third of the control group (33.3%) failed two or more PVT/EVIs at liberal cut-offs and 11.1% failed two or more at conservative cut-offs. In the experimental malingering group, 100% of the sample failed at least one PVT/EVI at both cutoffs, 66.7% failed two or more at liberal cutoffs, and 44.4% failed two or more at conservative cutoffs (Table 6).

Table 6

Frequency Distribution of PVT+EVI Failures (Excluding TMT A and B)

	Liberal Cutoffs						Conservative Cutoffs					
	Control			expMAL			Control			expMAL		
# failed	f	%	Cumul. %	f	%	Cumul. %	f	%	Cumul. %	f	%	Cumul. %
0	2	22.2	22.2	0	0.0	0.0	4	44.4	44.4	1	11.1	11.1
1	4	44.4	66.7	3	33.3	33.3	4	44.4	88.9	3	33.3	44.4
2	0	0.0	66.7	0	0.0	33.3	0	0.0	88.9	1	11.1	55.6
3	3	33.3	100.0	2	22.2	55.6	1	11.1	100.0	2	33.3	88.9
4	0	0.0	100.0	2	22.2	77.8	0	0.0	100.0	1	11.1	100.0
5	0	0.0	100.0	1	11.1	88.9						
6	0	0.0	100.0	0	0.0	88.9						
7	0	0.0	100.0	1	11.1	100.0						

Note: expMAL = experimental malingerers; f = Frequency; Cumul. % = Cumulative percent.

Number of Independent PVT/EVI Failures

When the VI-11 was used, the overall number of PVT/EVI failures did not reach statistical significance between experimental malingerers and controls at either liberal ($Mean_{malinger}$=4.22 (SD=2.77), $Mean_{Control}$=2.22 (SD=1.79), t(16)=-1.82, p=0.09, g=0.90) or conservative cutoffs ($Mean_{malinger}$ =2.78 (SD=1.72), $Mean_{Control}$=1.33 (SD=1.32), t(16)=-2.00, P=0.06, g=0.95). Effect sizes, however, were found to be large for number of PVT/EVIs failed at both liberal and conservative cutoffs.

When the VI-9 was used (i.e., omitting TMT A and B from the analyses), the experimental malingering group had significantly more overall PVT/EVI failures than controls at both liberal ($Mean_{malinger}$=3.22 (SD=2.05), $Mean_{control}$= 1.44 (SD=1.24, t(16)=-2.23, p=0.04, g=1.05) and conservative cutoffs ($Mean_{malinger}$=2.00 (SD=1.32), $Mean_{Control}$=0.78 (SD=0.97), t(16)=-2.23, p=0.04, g=1.05), with large effect sizes for both comparisons.

ImPACT-Based EVIs

On ImPACT EVIs, the lowest base rate of failure was observed for the Default EVI (0.0% for controls and 25% for malingerers). This was followed by the Higgins LRE (22.2% for controls and 75% for malingerers), the Schatz & Glatts criteria (33.3% for controls and 75% for malingerers), and the Red Flags (44.4% for controls and 75% for malingerers). Finally, on the ImPACT 5A, 66.7% of controls and 87.5% of malingerers had one or more failures, whereas on the ImPACT 5B, 55.6% of controls and again 87.5% of malingerers had one or more failures. As the failure threshold increased, controls demonstrated proportionally fewer failures while the rate of failure of experimental malingerers remained relatively constant (Table 7).

Table 7
Base Rates of Failure for ImPACT-Based EVIs

EVI	Scale	Cutoff	BRfail		Sens	Spec
			Controls	expMAL		
Default EVI	X's and O's + Color Match	> 30	0 (0.0)	0 (0.0)	0.00	1.00
	Impulse Control	>30	0 (0.0)	0 (0.0)	0.00	1.00
	WMLPC	<69	0 (0.0)	1 (12.5)	0.13	1.00
	DMLPC	<50	0 (0.0)	1 (12.5)	0.13	1.00
	Three Letters	<8	0 (0.0)	2 (25.0)	0.25	1.00
	Overall		**0 (0.0)**	**2 (25.0)**	**0.25**	**1.00**
Red Flags	Processing Speed	<25	0 (0.0)	4 (50.0)	0.50	1.00
	Reaction Time	>0.8	2 (22.2)	4 (50.0)	0.50	0.78
	Verbal Memory	<70	0 (0.0)	5 (62.5)	0.63	1.00
	Visual Memory	<60	2 (22.2)	4 (50.0)	0.50	0.78
	Overall		**4 (44.4)**	**6 (75.0)**	**0.75**	**0.56**
Schatz & Glatts	WMCD	<22	1 (11.1)	5 (62.5)	0.63	0.89
	DMCD	<16	3 (33.3)	6 (75.0)	0.75	0.67
	Overall		**3 (33.3)**	**6 (75.0)**	**0.75**	**0.67**
Higgins LRE	**Overall**	≥0.23	**2 (22.2)**	**6 (75.0)**	**0.75**	**0.78**
ImPACT 5A + B	Verbal Memory	≤78[A]	1 (11.1)	6 (75.0)	0.75	0.89
		≤76[B]	1 (11.1)	6 (75.0)	0.75	0.89
	Visual Memory	≤65[A]	2 (22.2)	6 (75.0)	0.75	0.78
		≤57[B]	2 (22.2)	4 (50.0)	0.50	0.78
	Visuomotor Speed	≤34[A]	1 (11.1)	7 (87.5)	0.88	0.89
		≤33[B]	1 (11.1)	7 (87.5)	0.88	0.89
	Reaction Time	≥.67[A]	4 (44.4)	7 (87.5)	0.88	0.56
		≥.71[B]	3 (33.3)	5 (62.5)	0.63	0.67
	Impulse Control	≥8[A]	2 (22.2)	4 (50.0)	0.50	0.78
		≥11[B]	1 (11.1)	2 (25.0)	0.25	0.89
	Overall	≥1 Fail	**6 (66.7)[A]**	**7 (87.5)[A]**	**0.88**	**0.44**
			5 (55.6)[B]	**7 (87.5)[B]**	**0.88**	**0.56**
		≥2 Fail	**3 (33.3)[A]**	**7 (87.5)[A]**	**0.88**	**0.67**
			3 (33.3)[B]	**7 (87.5)[B]**	**0.88**	**0.67**
		≥3 Fail	**1 (11.1)[A]**	**7 (87.5)[A]**	**0.88**	**0.89**
			0 (0.0)[B]	**6 (75.0)[B]**	**0.75**	**1.00**

Note: expMAL = Experimental malingerers; BRfail = Base rate of failure; [A]ImPACT 5A; [B]ImPACT 5B

Neurocognitive Performance

Non-ImPACT-Based Measures

Neurocognitive performance on non-ImPACT-based measures was compared between experimental groups (Table 8). Mean performance for controls (M=10.67,

SD=2.4) was only significantly better than experimental malingerers (M=8.00, SD=2.8) on DS (t(16)=2.18, p=0.04), and the effect size was large (g=1.02). There were no significant differences between groups on any other non-ImPACT-based measures of neurocognitive performance. However, the mean performance for the experimental malingering group was consistently lower than controls on all measures of neurocognitive performance, with small-to-medium effect sizes for Letter Fluency, Animal Fluency, CD, and SS, medium effect sizes for WRAT-4 Reading and TMT-B, a medium-to-large effect size for TMT-A, and a large effect size for EWFT.

Table 8
Comparison Between Experimental Malingerers and Controls on Independent Measures of Neurocognitive Performance

		Mean (SD)			
Test	Scale	Control	expMAL	p	g
TMT A	T-Score	41.33 (7.1)	34.00 (12.1)	0.18	0.77
TMT B	T-Score	48.94 (8.3)	43.07 (11.9)	0.27	0.59
DS	ACSS	10.67 (2.4)	8.00 (2.8)	0.04*	1.02
SS	Scaled score	7.89 (3.0)	6.44 (4.0)	0.40	0.41
CD	Scaled score	8.89 (3.3)	7.89 (2.0)	0.45	0.37
FAS	T-Score	45.56 (7.3)	43.00 (7.4)	0.47	0.35
Animals	T-Score	44.06 (7.6)	40.50 (10.3)	0.42	0.39
EWFT	Raw score	14.00 (3.4)	10.98 (2.8)	0.05	0.97
WRAT-4 Reading	Scaled score	106.22 (12.1)	98.89 (11.1)	0.20	0.63

Note: expMAL = experimental malingering; FAS = Letter fluency; Animals = Animal fluency; *p < 0.05

ImPACT Composite Scores

ImPACT Composite scores were also compared, and significant differences were found between groups on all composite scores except the Reaction Time Composite (Table 9). In addition, no group differences were found on the Total Symptom Score.

Table 9
Comparison Between Experimental Malingerers and Controls on ImPACT Composite Scores

	Mean (SD)			
Test	Control	expMAL	p	g
Verbal Memory	94.22 (9.19)	65.63 (15.04)	<.01**	2.33
Visual Memory	73.78 (10.63)	56.63 (14.72)	0.01*	1.35
Visuomotor Speed	39.21 (7.22)	27.20 (10.44)	0.01*	1.35
Reaction Time	0.71 (0.15)	0.87 (0.27)	0.13	0.75
Impulse Control	4.56 (3.36)	10.75 (7.18)	0.04*	1.13
Total Symptom Score	30.67 (25.40)	35.88 (23.63)	0.67	0.21

Note: expMAL = experimental malingering; *p<0.05, **p<0.01

The Effect of PVT Failure on Neurocognitive Performance

Experimental groups were collapsed to examine the effects of PVT/EVI failure on neurocognitive performance. When VI-11 scores of ≤1 = Pass and ≥2 = Fail were used as the group criterion, comparisons based on liberal cutoffs yielded significant differences on TMT-A, TMT-B, DS, and SS (Table 10). At conservative cutoffs, there were significant differences on TMT-A, TMT-B, SS, and EWFT.

Table 10

Effects of Failing Two or More PVT/EVIs on neurocognitive performance (V-11)

		Liberal Cutoffs				Conservative Cutoffs			
		Mean (SD)				Mean (SD)			
Test	Scale	≤1 failure (N=5)	≥2 failures (N=13)	p	g	≤1 failure (N=10)	≥2 failures (N=8)	p	g
[a]TMT A	T-Score	52.25 (8.57)	33.42 (5.98)	<.01**	2.84	47.14 (9.60)	31.11 (3.76)	<.01**	2.09
[a]TMT B	T-Score	55.88 (5.72)	43.21 (9.37)	0.03*	1.45	52.79 (6.16)	41.39 (10.03)	0.02*	1.42
DS	ACSS	11.60 (2.30)	8.46 (2.63)	0.03*	1.23	9.75 (3.20)	9.00 (2.71)	0.60	0.25
SS	Scaled score	10.60 (2.19)	5.85 (3.02)	<.01**	1.67	9.25 (2.66)	5.50 (3.31)	0.02*	1.27
CD	Scaled score	9.40 (2.51)	8.00 (2.77)	0.34	0.51	9.25 (2.71)	7.70 (2.63)	0.24	0.58
FAS	T-Score	42.50 (4.37)	44.96 (8.14)	0.54	0.34	42.69 (4.64)	45.55 (8.86)	0.42	0.42
Animals	T-Score	45.50 (4.77)	41.04 (10.03)	0.36	0.50	45.56 (5.73)	39.65 (10.46)	0.17	0.73
EWFT	Raw score	14.40 (2.51)	11.69 (3.52)	0.14	0.82	14.63 (3.38)	10.70 (2.41)	0.11*	1.31
WRAT Reading	Scaled score	101.40 (13.32)	103.00 (11.81)	0.81	0.13	102.38 (14.11)	102.70 (10.53)	0.96	0.03

Note: FAS = Letter fluency; Animals = Animal fluency; [a]Sample size for TMT A and B was 4 and 12 at liberal cutoffs and 7 and 9 at conservative cutoffs due to missing data. *p<0.05, **p<0.01.

When VI-9 scores of ≤1 = Pass and ≥2 = Fail were used as the group criterion, TMT-A and B were no longer significantly different across groups at either cutoff. Significant differences remained on SS and DS at both cutoffs, and there were also significant differences on EWFT at both cutoffs (Table 11). There was no effect of PVT/EVI failure on performance for CD, Letter Fluency, Animal Fluency, or WRAT-4 Reading, regardless of criterion group or level of cutoff.

Table 11

Effects of Failing Two or More PVTs on Neurocognitive Performance (VI-9)

Test	Scale	Liberal Cutoffs				Conservative Cutoffs			
		Mean (SD)				Mean (SD)			
		≤1 failure (N=9)	≥2 failures (N=9)	p	g	≤1 failure (N=12)	≥2 failures (N=6)	p	g
[a]TMT A	T-Score	42.13 (12.26)	34.13 (7.26)	0.14	0.79	41.09 (11.33)	31.60 (4.56)	0.10	0.96
[a]TMT B	T-Score	48.69 (10.42)	44.06 (9.99)	0.38	0.45	47.68 (9.03)	43.50 (12.92)	0.46	0.41
DS	ACSS	11.22 (2.05)	7.44 (2.30)	<.01**	1.74	10.33 (2.71)	7.33 (2.16)	0.03*	1.18
SS	Scaled score	9.89 (2.42)	4.44 (1.94)	<.01**	2.48	8.92 (2.81)	3.67 (1.63)	<.01**	2.10
CD	Scaled score	9.44 (2.79)	7.33 (2.29)	0.10	0.83	8.75 (2.96)	7.67 (2.16)	0.44	0.39
FAS	T-Score	45.28 (5.79)	43.28 (8.72)	0.57	0.27	43.63 (6.39)	45.58 (9.27)	0.60	0.26
Animals	T-Score	43.17 (5.86)	41.39 (11.62)	0.69	0.19	43.29 (6.54)	40.25 (13.13)	0.52	0.33
EWFT	Raw score	14.33 (2.40)	10.56 (3.36)	0.01*	1.29	13.83 (3.07)	9.67 (2.34)	0.01*	1.45
WRAT-4 Reading	Scaled score	104.33 (10.48)	100.78 (13.49)	0.54	0.29	104.50 (12.00)	98.67 (11.61)	0.34	0.49

Note: FAS = Letter fluency; Animals = Animal fluency; [a]Sample size for TMT A and B was 8 and 8 at liberal cutoffs and 11 and 5 at conservative cutoffs due to missing data. *p<0.05, **p<0.01.

Classification Accuracy

Sensitivities and specificities of the ImPACT-based EVIs were calculated first against a criterion of experimental group, and then against dichotomized VI-11 and VI-9 scores. For the latter comparisons, participants were once again separated into groups based on scores of ≤1 = Pass vs ≥2 = Fail on the VI-11 and VI-9, at both liberal and conservative cutoffs.

As expected, the Default EVI demonstrated substantially lower sensitivity than all other ImPACT-based EVIs, regardless of the criterion against which classification accuracy

was calculated (0.17-0.33; Table 12). Specificity, however, was consistently perfect. Against the criterion of experimental group, the highest sensitivities were demonstrated by thresholds of ≥1 and ≥2 failures on the ImPACT 5-A and B, as well as ≥3 failures on the ImPACT 5-A. Each of these indicators demonstrated sensitivities of 0.88, though specificities ranged from 0.44 for ≥1 failure on ImPACT 5-A to 0.89 for ≥3 failures on ImPACT 5-A. Overall, ≥3 failures on the ImPACT-5A and B had the best classification accuracy among the ImPACT-based EVIs, with the ImPACT-5A maximizing sensitivity and the ImPACT-5B maximizing specificity, as expected.

Against a criterion of ≥2 failures on the VI-11 at liberal cutoffs, the highest sensitivities were demonstrated by a threshold of ≥1 failure on the ImPACT 5-A and B (0.83-0.92), though specificity was unacceptably low (0.60). Increasing the threshold to ≥2 failures produced a slight decrease in sensitivity (0.75) but brought specificity to a more acceptable level (0.80). Increasing the threshold to ≥3 failures further reduced sensitivity (0.58), though specificity remained constant (0.80). Overall, ≥2 failures on the ImPACT-5A and B once again had the best classification accuracy among the ImPACT-based EVIs, with equal sensitivities (0.75) and specificities (0.80). A similar pattern was seen when ≥2 failures at conservative cutoffs on the VI-11 was used as the criterion measure, though almost every indicator showed a decrease in specificity with little or no increase in sensitivity.

Given the inherent error within the VI-11, classification accuracy was also calculated against a criterion of ≥2 failures on the VI-9 at liberal and conservative cutoffs. At liberal cutoffs, the highest sensitivities were once again demonstrated by thresholds of ≥1 and ≥2 failures on the ImPACT 5A and B (1.00). Specificity was

unacceptably low at ≥1 failure for both ImPACT 5A and B (0.50-0.63) but increased for the threshold of ≥2 failures (0.88). When the threshold was increased to ≥3 failures on the ImPACT 5 A and B, specificity remained the same (0.88), though sensitivity was reduced (0.56-0.78). Overall, the best classification accuracy was once again demonstrated by a threshold of ≥2 failures on the ImPACT 5A and B, with perfect sensitivity and high specificity (0.88). A similar pattern was seen for ≥2 failures at conservative thresholds on the VI-9, though, once again, most indicators showed a decrease in specificity, with little or no increase in sensitivity.

Table 12

Classification Accuracy of ImPACT-Based EVIs

			expMAL vs Control		VI-11 (LIB)		VI-11 (CON)		VI-9 (LIB)		VI-9 (CON)	
EVI	Scale	Cut	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
Default EVIs	X's and O's + Color Match	>30	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
	Impulse Control	>30	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
	WMLPC	<69	0.13	1.00	0.08	1.00	0.11	1.00	0.11	1.00	0.17	1.00
	DMLPC	<50	0.13	1.00	0.08	1.00	0.11	1.00	0.11	1.00	0.17	1.00
	Three Letters	<8	0.25	1.00	0.17	1.00	0.22	1.00	0.22	1.00	0.33	1.00
	Overall		**0.25**	**1.00**	**0.17**	**1.00**	**0.22**	**1.00**	**0.22**	**1.00**	**0.33**	**1.00**
Red Flags	Processing Speed	<25	0.50	1.00	0.33	1.00	0.33	0.88	0.44	1.00	0.50	0.91
	Reaction Time	>0.8	0.50	0.78	0.42	0.80	0.44	0.75	0.56	0.88	0.67	0.82
	Verbal Memory	<70	0.63	1.00	0.50	0.80	0.33	0.75	0.44	0.88	0.50	0.82
	Visual Memory	<60	0.50	0.78	0.42	0.80	0.33	0.63	0.55	0.88	0.33	0.64
	Overall		**0.75**	**0.56**	**0.67**	**0.60**	**0.67**	**0.50**	**0.89**	**0.75**	**0.83**	**0.55**
Schatz & Glatts	WMCD	<22	0.63	0.89	0.42	0.80	0.44	0.75	0.63	0.88	0.50	0.73
	DMCD	<16	0.75	0.67	0.67	0.80	0.67	0.63	0.89	0.88	0.83	0.64
	Overall		**0.75**	**0.67**	**0.67**	**0.80**	**0.67**	**0.63**	**0.89**	**0.88**	**0.83**	**0.64**
Higgins LRE	**LRE**	**≥0.23**	**0.75**	**0.78**	**0.58**	**0.80**	**0.67**	**0.75**	**0.78**	**0.88**	**0.83**	**0.73**
ImPACT 5A(B)	Verbal Memory (A)	≤78	0.75	0.89	0.50	0.80	0.56	0.75	0.67	0.88	0.67	0.73
	(B)	≤76	0.75	0.89	0.50	0.80	0.56	0.75	0.67	0.88	0.67	0.73
	Visual Memory (A)	≤65	0.75	0.78	0.58	0.80	0.56	0.63	0.78	0.88	0.67	0.64
	(B)	≤57	0.50	0.78	0.42	0.80	0.33	0.63	0.56	0.88	0.33	0.64
	Visuomotor Speed (A)	≤34	0.88	0.89	0.58	0.80	0.67	0.75	0.78	0.88	1.00	0.82
	(B)	≤33	0.88	0.89	0.58	0.80	0.67	0.75	0.78	0.88	1.00	0.82
	Reaction Time (A)	≥.67	0.88	0.56	0.75	0.60	0.89	0.63	0.77	0.50	1.00	0.56
	(B)	≥.71	0.63	0.67	0.58	0.80	0.67	0.75	0.67	0.75	0.83	0.73
	Impulse Control (A)	≥8	0.50	0.78	0.42	0.80	0.33	0.63	0.56	0.88	0.50	0.73
	(B)	≥11	0.33	0.89	0.25	1.00	0.11	0.75	0.33	1.00	0.17	0.82
	Overall	**≥1 fail (A)**	**0.88**	**0.44**	**0.92**	**0.60**	**1.00**	**0.50**	**1.00**	**0.50**	**1.00**	**0.34**
		(B)	**0.88**	**0.56**	**0.83**	**0.60**	**0.89**	**0.50**	**1.00**	**0.63**	**1.00**	**0.45**
		≥2 fail (A)	**0.88**	**0.67**	**0.75**	**0.80**	**0.78**	**0.63**	**1.00**	**0.88**	**1.00**	**0.64**
		(B)	**0.88**	**0.67**	**0.75**	**0.80**	**0.78**	**0.63**	**1.00**	**0.88**	**1.00**	**0.64**
		≥3 fail (A)	**0.88**	**0.89**	**0.58**	**0.80**	**0.67**	**0.75**	**0.78**	**0.88**	**1.00**	**0.82**
		(B)	**0.75**	**1.00**	**0.58**	**0.80**	**0.44**	**0.75**	**0.56**	**0.88**	**0.67**	**0.82**

Note: expMAL = Experimental malingerers; LIB = Liberal cutoffs; CON = Conservative cutoffs; Sens = Sensitivity; Spec = specificity

Area Under the Curve

Area under the ROC curves were calculated for each component of the ImPACT-based EVIs against each of the five criterion groups previously discussed (Table 13). Against a criterion of experimental group, AUCs for the components of the Default EVI ranged from 0.778-0.917, with all but the Delayed Memory Learning Percent Correct criterion reaching statistical significance. AUCs for the components of the Schatz & Glatts criteria ranged from 0.785-0.819, with both criteria reaching statistical significance. The Higgins LRE had an AUC of 0.847 and was statistically significant, and the ImPACT's composite scores, which comprise the ImPACT-5 and the Red Flags, had AUCs ranging from 0.708-0.958, with only Reaction Time not reaching statistical significance. When compared against criterion groups of ≤1 or ≥2 failures on the VI-11 at either liberal or conservative cutoffs, almost all AUCs decreased considerably, and none reached statistical significance. When compared against a criterion of ≤1 or ≥2 failures on the VI-9, however, AUCs were generally only slightly lower than when experimental group was used as the criterion.

Table 13

Areas Under the ROC Curve for ImPACT-based EVIs

		expMAL vs Control		VI-11 (LIB)		VI-11 (CON)		VI-9 (LIB)		VI-9 (CON)	
EVI	Scale	AUC	p	AUC	p	AUC	p	AUC	p	AUC	p
Default EVIs[a]	X's and O's + Color Match	.819	0.03*	.617	0.46	.528	0.85	.812	0.03*	.742	0.11
	WMLPC	.792	0.04*	.492	0.96	.486	0.92	.618	0.41	.621	0.42
	DMLPC	.778	0.05	.683	0.25	.611	0.44	.868	0.01*	.811	0.04*
	Three Letters	.917	<.01**	.600	0.53	.625	0.39	.750	0.08	.818	0.04*
Schatz & Glatts	WMCD	.819	0.03*	.600	0.53	.583	0.56	.771	0.06	.682	0.23
	DMCD	.785	0.04*	.767	0.09	.597	0.50	.938	<.01**	.780	0.06
Higgins LRE	LRE	.847	0.02*	.683	0.25	.681	0.21	.889	<.01**	.848	0.02*
ImPACT-5/ Red Flags[b]	Verbal Memory[b]	.958	<.01**	.683	0.25	.646	0.31	.875	<.01**	.773	0.07
	Visual Memory[b]	.854	0.01*	.700	0.21	.583	0.56	.903	<.01**	.750	0.10
	Visuomotor Speed[b]	.861	0.01*	.708	0.19	.701	0.16	.910	<.01**	.924	<.01**
	Reaction Time[b]	.708	0.15	.800	0.06	.778	0.05	.806	0.03*	.841	0.02
	Impulse Control[a]	.819	0.03*	.617	0.46	.528	0.85	.812	0.03*	.742	0.11

Note: expMAL = Experimental malingerers; LIB = Liberal cutoffs; CON = Conservative cutoffs; Sens = Sensitivity; Spec = Specificity; AUC = Area under the curve; [a]The Impulse Control Composite score is also a component of the Default EVI. [b]The Verbal Memory, Visual Memory, Visuomotor Speed, and Reaction Time Composites also comprise the Red Flags; *p<0.05, **p<0.01.

CHAPTER 4

Discussion

Despite our small sample size, the current study did demonstrate higher levels of PVT/EVI failures in experimental malingerers than controls, with large effect sizes at both liberal and conservative cutoffs. Interestingly, though experimental malingerers had significantly poorer performance on four out of the five ImPACT composite scores with large effect sizes, a difference in neurocognitive performance on non-ImPACT measures was only found for Digit Span. Of course, given the small sample size, effect sizes may be more informative than statistical significance when assessing group differences; despite the lack of statistical significance, effect sizes for other traditional neuropsychological measures ranged from small-medium (i.e., g=0.35 for Letter Fluency) to large (i.e., g=0.97 for Emotion Word Fluency). Effect sizes for ImPACT composite scores, however, were substantially larger, ranging from g=0.75 for the Reaction Time Composite, to g=2.33 for the Verbal Memory Composite. As such, the effect of experimental malingering on neurocognitive performance seemed to be more pronounced for ImPACT than for independent measures of neurocognitive performance employed in this study.

Interestingly, a recent study demonstrated that ImPACT composite scores did not differ between a positively incentivized group relative to controls, and the authors inferred from this that ImPACT composite scores are unaffected by incentives, supporting their validity as measures of cognitive function as opposed to measures of effort (Merritt et al., 2019). The authors also suggest that only a small proportion of athletes are likely to clearly “sandbag” their baseline performance and then substantially

improve after a concussion, and that the effect of a positive incentive (akin to the return-to-play incentive) is therefore more relevant to the context of concussion testing than attempts to perform poorly at baseline. Because it is not clear how the authors operationalize a "clear sandbagging pattern", rebutting this point with empirical evidence is difficult. However, results of the current study are in line with previous studies demonstrating lower scores on ImPACT composites in experimental malingerers than controls (Raab et al., 2019; Higgins et al., 2017; Schatz & Glattz, 2013). Despite this clear suppression of neurocognitive performance, only 25% of experimental malingerers in this study were flagged by the Default EVI as representing invalid profiles. As such, contrary to the conclusions of Merritt et al. (2019), our results suggest that it is possible to suppress neurocognitive performance on ImPACT without being flagged as "clearly sandbagging", and that ImPACT composite scores are indeed sensitive to effort. In line with this, Walton et al. (2017) recently introduced the concept of "valid but invalid" ImPACT profiles, demonstrating that, of the 16% of athletes in their sample who were either flagged by the Default EVI or obtained one or more composite scores below the 16th percentile, 88% scored above the 16th percentile upon retest, suggesting that their original performance was not reflective of their true abilities.

Our results are also consistent with previous literature demonstrating that the ImPACT's Default EVI is not sufficiently sensitive to invalid performance. In the literature, the reported sensitivities for the Default EVI when used in experimental malingering paradigms ranged from 0.42 to 0.70 (Erdal, 2012; Schatz & Glatts, 2013; Siedlik, 2016; Higgins et al, 2017; Manderino & Gunstad, 2018; Raab et al., 2019), with a weighted average of 0.61. In our sample, the Default EVI demonstrated a sensitivity of

only 0.25 to experimental group, which is unacceptably low, particularly given that experimental malingering paradigms often yield more exaggerated patterns of underperformance than naturalistic samples (Vickery et al, 2001). On the other hand, previous research has shown that the Default EVI typically produces high specificity, and this was also the case here. One reason that our study may have yielded considerably lower sensitivity than previously reported for the Default EVI was that, of studies employing an experimental malingering paradigm to evaluate the ImPACT EVIs, one could argue that the current study gave participants the strongest external incentive to malinger in a credible and sophisticated fashion. Of the six previous studies using an experimental malingering paradigm with the ImPACT, only two reported providing additional incentive for successful malingering over and above compensation for participation. Specifically, Erdal (2012) told participants that the top undetected sandbagger would be given a $20 gift certificate in addition to the $5 gift card they were being given for participating, and Manderino & Gunstad (2018)'s participants were told that the test contained indicators of effort and feigning, and only those successfully putting forth full effort or feigning without detection (depending on condition) would be entered into a $50 Visa gift card raffle. In our study, on the other hand, each individual was eligible for increased financial compensation at a later date, depending on how well they malingered; they were told that, depending on how well they simulated a concussion, they would each be paid between $10 and $40 dollars for an additional 40 minutes of their time. As such, unlike previous studies, participants could guarantee themselves substantially higher compensation at a later date by malingering in a credible fashion. Given that the Default EVI was the least sensitive of the ImPACT-based

indicators, it would be the least likely to detect a subtle or sophisticated malingering strategy.

It is also important to note that experimentally induced malingering is not a true independent variable, as the researcher controls only the instructions given to participants and not the degree to which they are carried out (Abeare et al, 2020). Moreover, the rate of invalid performance in the control group is likely much higher than the 0% assumed by the experimental malingering model and has been shown in previous research to fall somewhere between 18% and 37% (An et al., 2017). In the current study, 33% of controls had ≥2 VI-9 failures at liberal cutoffs, and 11.1% had ≥2 VI-9 failures at conservative cutoffs. As a result, we also calculated classification accuracy as a function of psychometrically defined invalid performance, operationalized by ≥2 independent PVT/EVI failures. This is the first study to use an extensive battery of well-validated, independent measures of performance validity alongside the ImPACT in an experimental malingering paradigm, and as such is the first to report classification accuracy of the ImPACT EVIs against both experimentally defined and psychometrically defined invalid performance in the same sample. Only one previous study has reported classification accuracy of ImPACT EVIs to psychometrically defined invalid performance, though this was in a naturalistic sample of collegiate athletes undergoing baseline testing (Erdodi et al., 2020). Notably, the set of independent performance validity measures used in that study was highly overlapping with the measures used here.

When invalid performance was defined by ≥2 failures on independent PVT/EVIs, Erdodi et al. (2020) found that the Default EVI produced a sensitivity of only 0.04, which is considerably lower than the 0.17-0.33 found in the current study. This discrepancy may

be a reflection of sample characteristics; namely, the majority of those with ≥2 failures on independent PVT/EVIs in the current study were in the experimental malingering group, and thus likely had a more exaggerated form of invalid performance (i.e., more easily detected by the Default EVI) than the athletes undergoing true baseline testing in the study by Erdodi et al. (2020). This is supported by the fact that the mean ImPACT composite scores of participants who failed ≥2 PVT/EVIs (VI-9, liberal cutoffs) in the current study were between 0.33-1.79 standard deviations lower than the scores of those who failed ≥2 PVT/EVIs in the study by Erdodi et al (2020). The consistently lower ImPACT Composite Scores in those with psychometrically defined invalid performance here vs in Erdodi et al. (2020)'s study serves as empirical confirmation that the effect size for underperformance from experimental malingerers is larger than for real-world athletes during baseline testing.

The Red Flags, Schatz & Glatts criteria, and Higgins LRE each demonstrated sensitivities of 0.75 against experimental group, though none met the minimum acceptable specificity of 0.84 (specificities ranged from 0.56-0.78). The overall classification accuracy of the Red Flags in differentiating between experimental malingerers and controls has not previously been reported in the literature. Erdal (2012) reported that a Verbal Memory Composite cutoff of <70% was the most sensitive Red Flag indicator in her study, with a sensitivity of 0.73 to experimental malingering. A control group was not employed, however, and as such, specificity was not reported. Schatz and Glatts (2013) did report both sensitivity and specificity for two of the four Red Flags (Processing Speed Composite <25 and Reaction Time Composite >0.8), with sensitivities ranging from 0.60-0.70, and perfect specificity. The overall sensitivity of

0.75 for the Red Flags in our study is therefore in line with what has been reported in the literature, however we found the Red Flags to have considerably lower specificity to experimental group (0.56) than the perfect specificity reported by Schatz and Glatts (2013). One reason for this may be that Schatz & Glatts (2013) used only two of the four Red Flag indicators, whereas all four were used here. Moreover, Schatz & Glatts (2013) did not employ the Verbal Memory Composite score cutoff, which was the indicator with the highest sensitivity in both the current study as well as in Erdal (2012). Generally, there is a trade-off between sensitivity and specificity, and thus by using less sensitive components of the Red Flags, Schatz & Glatts (2013) likely maximized specificity. When ≥2 VI-9 failures was instead used as the criterion for classification, the Red Flags demonstrated better classification accuracy at both liberal and conservative cutoffs (sensitivities of 0.89 and 0.83, respectively, and specificities of 0.75 and 0.55, respectively), though the minimum threshold for specificity was still not met. Conversely, Erdodi et al. (2020) reported lower sensitivity (0.43) and higher specificity (0.85) for the Red Flags against psychometrically defined invalid performance using a naturalistic sample. This difference may once again be at least partially attributable to differences in the magnitude of the effect size of underperformance in experimental malingerers vs. real-world athletes.

The Schatz & Glatts (2013) criteria have been used in three studies employing experimental malingering paradigms. Sensitivities and specificities reported for the WMCD criterion range from 0.74-1.00 and 0.66-1.00, respectively, and for the DMCD criterion range from 0.69-0.95 and 0.65-0.80, respectively. When the data is combined across studies, the WMCD and DMCD produce sensitivities of 0.82 and 0.75, and

specificities of 0.69 and 0.67, respectively. Together, we found that the Schatz and Glatts criteria produced an overall sensitivity of 0.75 and a specificity of 0.67 against experimental group, which is in line with previous findings. Against a criterion of ≥2 VI-9 failures, classification accuracy improved, particularly at liberal cutoffs (sensitivity = 0.89 and specificity = 0.88). Comparatively, Erdodi et al. (2020) reported a sensitivity of 0.68 and specificity of 0.73 for the Schatz & Glatts (2013) criteria against psychometrically defined invalid performance.

The Higgins LRE was developed from a study employing an experimental malingering paradigm, and as such the equation is necessarily calibrated to the response patterns specific to that sample. Like any psychometric measure, it is important for EVIs to be calibrated across different settings and samples in order to determine their generalizability outside of the original sample in which they were conceived. This is the first study to our knowledge to attempt to cross-validate the classification accuracy of the Higgins LRE with an independent sample of experimental malingerers and controls. As expected, both sensitivity (0.75) and specificity (0.78) were found to be lower than reported in the original study (1.00 and .91, respectively). Classification accuracy improved, however, for ≥2 VI-9 failures, particularly at liberal cutoffs (sensitivity = 0.78, specificity = 0.88). Erdodi et al. (2020) reported a sensitivity of 0.68 and a specificity of 0.75 for the Higgins LRE in their athlete sample.

The ImPACT-5 A and B differ from the other ImPACT-based EVIs in that they were not derived from, and have not previously been tested in, an experimental malingering paradigm. Instead, they were developed in a naturalistic sample of university athletes undergoing baseline testing, with independent measures of performance validity

used as criteria for distinguishing valid vs invalid profiles. This approach to EVI development is likely to be more ecologically valid, as invalidity is determined based on actual performance rather than a set of contrived group instructions. Moreover, Erdodi et al (2020) used a multivariate criterion comprised of well-validated performance validity measures to determine invalid performance, further increasing the psychometric rigor of their proposed EVIs relative to the other existing ImPACT-based EVIs. Given this, it is not surprising that the ImPACT-5A and B demonstrated superior classification accuracy to both experimental group and psychometrically defined invalid performance than all other ImPACT-based EVIs in the current study. Against a criterion of experimental group, ≥3 failures on the ImPACT-5A produced high sensitivity (0.88) with good specificity (0.89), and ≥3 failures on the ImPACT 5B had somewhat lower sensitivity (0.75), with perfect specificity (1.00). Classification accuracy against a criterion of ≥2 VI-9 failures at liberal cutoffs was even higher, with ≥2 ImPACT-5 failures producing high sensitivity (0.88) and perfect specificity with both A and B versions. Overall, our results suggest that a sufficiently conservative threshold on the ImPACT-5 may provide "the best of both worlds", with both the A and B versions offering the highest combinations of sensitivity and specificity among the ImPACT-based EVIs.

In addition to the ImPACT-5 demonstrating the best classification accuracy of the ImPACT-based EVIs, two notable trends emerged with regards to classification accuracy. First, classification accuracy of ImPACT-based EVIs was generally superior to the criterion of psychometrically defined invalid performance (as measured by ≥2 VI-9 failures) than to experimental group. This is consistent with the limitations of experimental malingering paradigms previously discussed, in that experimental groups do

not reflect true independent variables that guarantee valid vs. invalid performance profiles. As such, when a performance-based measure of invalid performance was used as the classification criterion as opposed to a criterion based on the instructions given to participants, classification accuracy improved. The second notable trend is that classification accuracy of all ImPACT-based EVIs to psychometrically defined invalid performance was generally superior in our experimental sample than what was demonstrated by Erdodi et al. (2020) in real-world athletes completing baseline testing. The latter trend is consistent with experimental malingerers yielding larger effect sizes on effort measures, as previously discussed.

There are many limitations to the current study that must be considered, the most significant of which is our small sample size. Because our data collection was interrupted due to COVID-19, all of our results should be considered preliminary until replicated with a larger sample. Another limitation is that the vast majority of participants were female, and although most studies have found no difference in rates of invalid performance between males and females during baseline testing (Lichtenstein et al., 2013; Nelson et al., 2015; Tsushima et al., 2019; French et al., 2019), no previous study has investigated whether gender influences the way in which one approaches experimental malingering on ImPACT. Limitations associated with experimental malingering paradigms in general also apply to the current study. As mentioned previously, one such limitation is that experimentally-induced malingering only allows the researcher to control the instructions given to participants, and not the degree to which they adhere to these instructions. Moreover, once cannot ensure that the control group is comprised only of valid profiles, and this is in fact very unlikely to be the case.

As such the criterion groups are almost certainly contaminated, leading to diminished classification accuracy (Abeare et al., 2020). The current study attempted to address this limitation by administering multiple independent validity measures, and performing analyses using both experimentally defined and psychometrically defined invalid performance for comparison. Future research should consider a similar approach in order to more thoroughly assess the convergent validity of ImPACT-based EVIs. Finally, the motivational incentive to malinger successfully is presumably much stronger in real-world athletes undergoing baseline testing than in undergraduates participating in research for course credit. Though we attempted to provide an enticing external incentive, it is not clear to what degree this incentive motivated participants to malinger convincingly. Only one study to date has used a naturalistic sample of athletes to examine the performance of ImPACT-based EVIs (Erdodi et al., 2020), and as such more research with real-world athletes undergoing baseline testing is needed to evaluate the classification accuracy of EVIs in a more ecologically valid manner.

In summary, the current study supported previous research demonstrating that the ImPACT's Default EVI is insufficiently sensitive to invalid performance, demonstrating the lowest sensitivity of all ImPACT-based EVIs. Seventy-five percent of experimental malingerers were not detected by the Default EVI in our sample, despite having significantly lower composite scores on ImPACT. As such, it is crucial that clinicians administering ImPACT as part of a concussion management protocol use alternative measures to assess performance validity. Of the ImPACT-based EVIs, the ImPACT 5-A and B demonstrated the most superior classification accuracy and offer clinicians the option to prioritize either sensitivity or specificity, depending on the circumstance. Of

course, our findings must be interpreted cautiously as a result of our small sample size, and future research should aim to replicate this finding with a larger sample. This study also provided empirical support for the notion that the effect size of underperformance is larger in experimental malingerers than in real-world athletes undergoing baseline testing, and as such future research investigating performance validity on ImPACT should do so in naturalistic athlete populations, using psychometrically defined invalid performance as a criterion.

REFERENCES/BIBLIOGRAPHY

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. *The Journal of Clinical and Experimental Psychology*, *39*(8), 738-752.

Abeare et al., C. A., Hurtubise, J. L., Cutler, L., Sirianni, C., Brantuo, M., Makhzoum, N., & Erdodi, L. A. (2020). Introducing a forced choice recognition trial to the Hopkins Verbal Learning Test – Revised. *The Clinical Neuropsychologist*, 1-29.

Abeare, C. A., Messa, I., Zuccato, B. G., Merker, B., & Erdodi, L. (2018). Prevalence of invalid performance on baseline testing for sport-related concussion by age and validity indicator. *JAMA Neurology*, *75*(6), 697–703.

Abeare, C., Messa, I., Whitfield, C., Zuccato, B., Casey, J., Rykulski, N., & Erdodi, L. (2018). Performance Validity in Collegiate Football Athletes at Baseline Neurocognitive Testing. *Journal of Head Trauma Rehabilitation*, *34*(4), E20-E31.

Abeare, C., Sabelli, A., Taylor, B., Holcomb, M., Dumitrescu, C., & Kirsch, N. (2019). The importance of demographically adjusted cutoffs : Age and education bias in raw score cutoffs within the Trail Making Test. *Psychological Injury and Law*, *12*(2), 170–182. https://doi.org/10.1007/s12207-019-09353-x

An, K. Y., Charles, J., Ali, S., Enache, A., Dhuga, J., & Laszlo, 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.

https://doi.org/10.1080/13803395.2018.1483488.

An, K. Y., Kaploun, K., Erdodi, L. A., & Abeare, C. A. (2016). *Performance validity in undergraduate research participants: a comparison of failure rates across tests and cutoffs. The Clinical Neuropsychologist, 31*(1), 193-206.

Arnett, P., Meyer, J., Merritt, V., & Guty, E. (2016). Neuropsychological Testing in Mild Traumatic Brain Injury What to Do When Baseline Testing Is Not Available. *Sports Medicine Arthroscopy Review, 24*(3), 116-122.

Axelrod, B. N., Meyers, J. E., & Davis, J. J. (2014). Finger Tapping Test performance as a measure of performance validity. *The Clinical Neuropsychologist, 28*(5), 876–888. https://doi.org/10.1080/13854046.2014.907583

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

Bailey, C. M., Echemendia, R. J., & Arnett, P. A. (2006). The impact of motivation on neuropsychological performance in sports-related mild traumatic brain injury. J*ournal of the International Neuropsychological Society, 12*, 475-484.

Barth, J. T., Alves, W. M., Ryan, T. V., Macciocchi, S. N., Rimel, R. W., Jane, J. A., & Nelson, W. E. (1989). Head injury in sports: Neuropsychological sequelae and recovery of function. In H. S. Levin, H. M. Eisenberg, & A. L. Benton (Eds.), *Mild Head Injury* (pp. 257–275). New York: Oxford Press.

Benton, A. L., & Hamsher, K. deS. (1978). *Multilingual aphasia examination: Manual.*

Iowa City: University of Iowa.

Boone, K., Salazar, X., Lu, P., Warner-Chacon, K., & Razani, J. (2002). 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.

Broglio, S. P., Ferrara, M. S., Macciocchi, S. N., Baumgartner, T. A., & Elliott, R. (2007). Test-Retest Reliability of Computerized Concussion Assessmemt Programs. *Journal of Athletic Training*, *42*(2), 509-514.

Broglio, S. P., Guskiewicz, K. M., & Norwig, J. (2017). If you're not measuring, you're guessing: The advent of objective concussion assessments. *Journal of Athletic Training*, *52*(3), 160–166.

Broglio, S. P., Macciocchi, S. N., & Ferrara, M. S. (2007). Neurocognitive performance of concussed athletes when symptom free. *Journal of Athletic Training*, *42*(4), 504–508.

Bryan, M. A., Rowhani-Rahbar, A., Comstock, R. D., & Rivara, F. (2016). Sports- and Recreation-Related Concussions in US Youth. *Pediatrics*, *138*(1), 1-8.

Carson, J. D., Lawrence, D. W., Kraft, S. A., Garel, A., Snow, C. L., … Frémont, P. (2014). Premature return to play and return to learn after a sport-related concussion. Canadian Family Physician, 60, e310-e315.

Covassin, T., Robert, ;, Iii, J. E., Stiller-Ostrowski, J. L., & Kontos, A. P. (2009). Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) practices

of sports medicine professionals. *Journal of Athletic Training*, *44*(6), 639–644.

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, G., Anderson, V., Babl, F., Gioia, G. A., Giza, C. C., Meehan, W. P., … Zemek, R. (2017). What is the difference in concussion management in children as compared with adults? A systematic review. *British Journal of Sports Medicine*, *0*, 1–12.

Echemendia, R. J., Bruce, J. M., Bailey, C. M., Sanders, J. F., Arnett, P., & Vargas, G. (2012). The utility of post-concussion neuropsychological data in identifying cognitive change following sports-related MTBI in the absence of baseline data. *The Clinical Neuropsychologist*, *26*(7), 1077–1091.

Erdal, K. (2012). Neuropsychological testing for sports-related concussion: How athletes can sandbag their baseline testing without detection. *Archives of Clinical Neuropsychology*, *27*, 473–479.

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., Korcsog, K., Considine, C., Casey, J., Scoboria, A., & Abeare, C. (2020).

Introducing the ImPACT-5: An Empirically Derived Multivariate Validity Composite. *The Journal of Head Trauma Rehabilitation*.

Erdodi, L. A., Sabelli, A. G., An, K. Y., Hastings, M., McCoy, C., & Abeare, C. A. (2020). Introducing a Five-Variable Psychiatric Screener based on the Visual Analog Scale (V-5). *Psychology & Neuroscience*. Advance online publication.

Erdodi, L. A, Seke, K. R., Tyson, B. T., & Roth, R. M. (2017). Low scores on the Grooved Pegboard Test are associated with invalid responding and psychiatric symptoms. *Psychology & Neuroscience, 10*(3), 325–344. https://doi.org/10.1037/pne0000103

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), 327–329.

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.

Farnsworth, J. L., Dargo, L., Ragan, B. G., & Kang, M. (2017). Reliability of computerized neurocognitive tests for concussion assessment: A meta-analysis. *Journal of Athletic Training*, *52*(9), 826–833.

French, J., Huber, P., Mcshane, J., Holland, C. L., Elbin, R. J., & Kontos, A. P. (2019).

Influence of Test Environment , Age , Sex , and Sport on Baseline Computerized Neurocognitive Test Performance. 3263–3269. https://doi.org/10.1177/0363546519875137

Gaudet, C. E., & Weyandt, L. L. (2016). Immediate Post-Concussion and Cognitive Testing (ImPACT): A systematic review of the prevalence and assessment of invalid performance. *The Clinical Neuropsychologist, 31*(1), 43-58.

Giza, C. C., & Hovda, D. A. (2014). The new neurometabolic cascade of concussion. *Neurosurgery*, *75*(4), S24–S33.

Gladsjo, J. A., Schuman, C. C., Evans, J. D., Peavy, G. M., Miller, S. W., & Heaton, R. K. (1999). Norms for letter and category fluency: Demographic corrections for age, education, and ethnicity. *Assessment, 6*(2), 147-178.

Green, P., Montijo, J., & Brockhaus, R. (2011). High specificity of the Word Memory Test and Medical Symptom Validity Test in groups with severe verbal memory impairment. *Applied Neuropsychology*, *18*(2), 86–94.

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.

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

Heaton, R. K., Miller, S. W., Taylor, M. J., & Grant, I. (2004). *Revised comprehensive norms for an expanded Halstead-Reitan bettery: Demographically adjusted*

neuropsychological norms for African American and Caucasian adults. Lutz, FL: Psychological Assessment Resources.

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., & Millis, S. R. (2010). 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.

Heinly, M. T., Greve, K. W., Bianchini, K. J., Love, J. M., & Brennan, A. (2004). WAIS Digit Span-based indicators of malingered neurocognitive dysfunction: Classification accuracy in traumatic brain injury. *Assessment*, *12*(4), 429–444.

Higgins, K. L., Caze, T., & Maerlender, A. (2018). Validity and Reliability of Baseline Testing in a Standardized Environment. *Archives of Clinical Neuropsychology*, *33*, 437–443. https://doi.org/10.1093/arclin/acx071

Higgins, K. L., Denney, R. L., & Maerlender, A. (2017). Sandbagging on the Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) in a high school athlete population. *Archives of Clinical Neuropsychology*, *32*, 259–266.

Hurtubise, J., Baher, T., Messa, I., Cutler, L., Shahein, A., Hastings, M., Carignan-Querqui, M., & Erdodi, L. A. (2020). Verbal fluency and digit span variables as performance validity indicators in experimentally induced malingering and real world patients with. *Applied Neuropsychology: Child*, 1–18. https://doi.org/10.1080/21622965.2020.1719409

Iverson, G. L., Lovell, M. R., & Collins, M. W. (2003). Interpreting change on ImPACT

following sport concussion. *The Clinical Neuropsychologist*, *17*(4), 460–467.

Kerr, Z. Y., Snook, E. M., Lynall, R. C., Dompier, T. P., Sales, L., Parsons, J. T., & Hainline, B. (2015). Concussion-related protocols and preparticipation assessments used for incoming student-athletes in national collegiate athletic association member institutions. *Journal of Athletic Training*, *50*(11), 1174–1181.

Kim, M. S., Boone, K. B., Victor, T., Marion, S. D., Amano, S., Cottingham, M. E., … Zeller, M. A. (2010). The Warrington Recognition Memory Test for Words as a measure of response bias: Total score and response time cutoffs developed on "real world" credible and noncredible subjects. *Archives of Clinical Neuropsychology*, *25*, 60–70.

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

Langlois, J. A., Rutland-Brown, W., & Wald, M. M. (2006). The epidemiology and impact of traumatic brain injury: a brief overview. *The Journal of head trauma rehabilitation*, *21*(5), 375-378.

Larrabee, G. J. (2014). False-positive rates associated with the use of multiple performance and symptom validity tests. *Archives of Clinical Neuropsychology*, *29*, 364–373.

Lichtenstein, J. D., Psyd, Y., Moser, R. S., & Schatz, P. (2013). Age and Test Setting Affect the Prevalence of Invalid Baseline Scores on Neurocognitive Tests. *The American Journal of Sports Medicine*, *42*(2), 479–484.

https://doi.org/10.1177/0363546513509225

Lovell, M. R. (2018). *ImPACT Administration and Interpretation Manual*. San Diego, CA: ImPACT Applications, Inc.

Mack, W. J., Freed, D. M., Williams, B. W., & Henderson, V. W. (1992). Boston Naming Test: Shortened versoins for use in Alzheimer's disease. Journal of Gerontology, 47, 154-158.

Maerlender, A, Flashman, L., Kessler, A., Kumbhani, S., Greenwald, R., Tosteson, T., & Mcallister, T. (2010). Examination of the Construct Validity of Impact™ Computerized Test, Traditional, and Experimental Neuropsychological Measures. *The Clinical Neuropsychologist*, *24*(8), 1309–1325.

Maerlender, Arthur, & Molfese, D. L. (2015). Repeat baseline assessment in college-age athletes. *Developmental Neuropsychology*, *40*(2), 69–73.

Manderino, L. M., & Gunstad, J. (2018). Performance of the Immediate Post-Concussion Assessment and Cognitive Testing Protocol Validity Indices. *Archives of Clinical Neuropsychology*, *33*, 596–605.

Manderino, L., Zachman, A., & Gunstad, J. (2018). Novel ImPACT validity indices in collegiate student-athletes with and without histories of ADHD or academic difficulties. *The Clinical Neuropsychologist*. DOI:10.1080/13854046.2018.1539191

Manley, G., Gardner, A. J., Schneider, K. J., Guskiewicz, K. M., Bailes, J., Cantu, R. C., … Iverson, G. L. (2017). A systematic review of potential long-term effects of sport-related concussion. *Br J Sports Med*, *51*, 969–977.

Martin, P. K., Schroeder, R. W., Olsen, D. H., Maloy, H., Boettcher, A., Ernst, N., & Okut, H. (2020). A systematic review and meta-analysis of the Test of Memory Malingering in adults: Two decades of deception detection. *The Clinical Neuropsychologist*, *34*(1), 88-119.

Mayers, L., & Redick, T. (2012). Clinical utility of ImPACT assessment for postconcussion return-to-play counseling: Psychometric issues. *Journal of Clinical and Experimental Neuropsychology*, *34*(3), 235–242.

McCrory, P., Meeuwisse, W., Dvorak, J., Aubry, M., Bailes, J., Broglio, S., … Vos, P. E. (2018). Consensus statement on concussion in sport-the 5th international conference on concussion in sport held in Berlin, October 2016. *British Journal of Sports Medicine*, *51*, 838–847.

Merritt, V. C., Rabinowitz, A. R., Guty, E., Meyer, J. E., Greenberg, L. S., & Arnett, P. A. (2019). Financial incentives influence ImPACT validity indices but not cognitive composite scores. *Journal of Clinical and Experimental Neuropsychology*, *41*(3), 312–319. https://doi.org/10.1080/13803395.2018.1551519

Meyers, J., & Volbrecht, M. (1998). Validation of Reliable Digits for detection of malingering. *Assessment*, *5*(3), 303–307.

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

Nelson, L. D., Pfaller, A. Y., Rein, L. E., & McCrea, M. A. (2015). Rates and predictors

of invalid baseline test performance in high school and collegiate athletes for 3 computerized neurocognitive tests: ANAM, Axon Sports, and ImPACT. *The American Journal of Sports Medicine, 43*(8), 2018–2026. https://doi.org/10.1177/0363546515587714

Orme, D. R., Johnstone, B., Hanks, R., & Novack, T. (2004). The WRAT-3 Reading Subtest as a Measure of Premorbid Intelligence Among Persons With Brain Injury. Rehabilitation Psychology, 49(3), 250–253. https://doi.org/10.1037/0090-5550.49.3.250.

Pearson. (2009). *Advanced Clinical Solutions for the WAIS-IV and WMS-IV - Technical Manual*. San Anotnio, TX: Author.

Poynter, K., Boone, K. B., Ermshar, A., Miora, D., Cottingham, M., Victor, T. L., ... & Wright, M. (2019). Wait, there'sa baby in this bath water! Update on quantitative and qualitative cut-offs for Rey 15-Item Recall and Recognition. *Archives of Clinical Neuropsychology*, *34*(8), 1367-1380.

Raab, C. A., Peak, A. S., & Knoderer, C. (2019). Half of purposeful baseline sandbaggers undetected by ImPACT's embedded invalidity indicators. *Archives of Clinical Neuropsychology*, DOI:10.1093/arclin/acz001

Rabinowitz, A. R., Merritt, V., & Arnett, P. A. (2016). A pilot investigation of the Motivation Behaviors Checklist (MBC): An observational rating scale of effort towards testing for baseline sports-concussion assessment. *Journal of Clinical and Experimental Neuropsychology*, *38*(6), 599–610.

Rabinowitz, A. R., Merritt, V. C., & Arnett, P. A. (2015). The return-to-play incentive and the effect of motivation on neuropsychological test performance: Implications for baseline concussion testing. *Developmental Neuropsychology*, *40*(1), 29–33.

Rai, J., & Erdodi, L. (2019). Impact of criterion measures on the classification accuracy of TOMM-1. *Applied Neuropsychology: Adult*, 1–12. https://doi.org/10.1080/23279095.2019.1613994

Rai, J., An, K., & Erdodi, L. A. (2019). Introducing a forced choice recognition trial to the Rey. *Psychology & Neuroscience, 12*(4), 1–22. https://doi.org/10.1037/pne0000175

Randolph, C. (2011). Baseline neuropsychological testing in managing sport-related concussion: Does it modify risk? *Current Sports Medicine Reports, 10*(1), 21–26.

Randolph, C., & Kirkwood, M. W. (2009). What are the real risks of sport-related concussion, and are they modifiable? *Journal of the International Neuropsychological Society*, *15*(4), 512–520.

Randolph, C., McCrea, M., & Barr, W. B. (2005). Is neuropsychological testing useful in the management of sport-related concussion? *Journal of Athletic Training*, *40*(3), 139–152.

Reitan, R. M. (1955). The relation of the trail making test to organic brain damage. *Journal of Consulting Psychology, 19*, 393-394.

Resch, J., Driscoll, A., McCaffrey, N., Brown, C., Ferrara, M. S., Macciocchi, S., … Walpert, K. (2013). ImPact test-retest reliability: Reliably unreliable?. *Journal of*

Athletic Training, *48*(4), 506–511.

Rey, A. (1964). *The Clinical Examination in Psychology*. Paris: University Press of France.

Schatz, P. (2010). Long-term test-retest reliability of baseline cognitive assessments using ImPACT. *The American Journal of Sports Medicine, 38*(1), 47-53.

Schatz, P. (2018). Role of baseline testing. In V. Musahl, J. Karlsson, W. Krutsch, B. Mandelbaum, J. Espregueira-Mendes, & P. D'Hooghe (Eds.), *Return to Play in Football* (pp. 671–682). Springer, Berlin, Heildelberg.

Schatz, P., Elbin, R. J., Anderson, M. N., Savage, J., & Covassin, T. (2017). Exploring sandbagging behaviors, effort, and perceived utility of the ImPACT baseline assessment in college athletes. *Sport, Exercise and Performance Psychology, 6*(3), 243–251.

Schatz, P., & Glatts, C. (2013). "Sandbagging" baseline test performance on ImPACT, without detection, is more difficult than it appears. *Archives of Clinical Neuropsychology, 28,* 236–244.

Schatz, P., Pardini, J., Lovell, M., Collins, M., & Podell, K. (2006). Sensitivity and specificity of the ImPACT Test Battery for concussion in athletes. *Archives of Clinical Neuropsychology*, *21,* 91–99.

Schatz, P., & Robertshaw, S. (2014). Comparing post-concussive neurocognitive test data to normative data presents risks for under-classifying "above average" athletes. *Archives of Clinical Neuropsychology*, *29,* 625–632.

Shura, R. D., Martindale, S. L., Taber, K. H., Higgins, A. M., & Rowland, J. A. (2020). Digit Span embedded validity indicators in neurologically-intact veterans. *The Clinical Neuropsychologist*, *34*(5), 1025-1037.

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

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.

Szabo, A. J., Alosco, M. L., Fedor, A., & Gunstad, J. (2013). Invalid performance and the ImPACT in National Collegiate Athletic Association Division I football players. *Journal of Athletic Training*, *48*(6), 851–855.

Tombaugh, T. N. (1996) *Test of memory malingering*. New York, NY: Multi-Health Systems.

Tsushima, W. T., Yamamoto, M. H., Ahn, H. J., Siu, A. M., Choi, S. Y., Murata, N. M. (2019). Does sandbagging occur with high school athletes ? Invalid baseline testing with ImPACT. *Applied Neuropsychology: Child*, 1–10. https://doi.org/10.1080/21622965.2019.1642202.

Valovich McLeod, T. C., Lewis, J. H., Whelihan, K., & Welch Bacon, C. E. (2017). Rest and return to activity after sport-related concussion: A systematic review of the literature. *Journal of Athletic Training*, *52*(3), 262–287.

Vickery, C. D., Berry, D. T. R., Inman, T. H., Harris, M. J., & Orey, S. A. (2001). Detection of inadequate effort on neuropsychological testing : A meta-analytic review of selected procedures. *Archives of Clinical Neuropsychology, 16*, 45–73.

Walton, S., Broshek, D., Freeman, J., Cullum, C. M., & Resch, J. E. (2017). Valid but invalid: Suboptimal ImPACT baseline performance in university athletes. *British Journal of Clinical Psychology*, *51*(11), A12–A13.

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Slotkin, J., … Gershon, R. (2014). The cognition battery of the NIH toolbox for assessment of neurological and behavioral function: Validation in an adult sample. *Journal of the International Neuropsychological Society*, *20*(6), 567–578.

Wechsler, D. (1997). *WAIS-III/WMS-III technical manual.* San Antonio, TX: Psychological Corporation.

Wechsler, D., (2008). Technical and interpretive manual for the Weschler adult intelligence scale (Fourth ed.). San Antonio, TX: Pearson.

Wilkinson, G. S., & Robertson, G. J. (2006). Wide range achievement test 4. Lutz, FL: Psychological Assessment Resources, Inc.

APPENDICES

Appendix A – Scripts for Malingerers and Controls

Control Script

There has been a lot of research to show that a person's level of motivation and effort is a big contributor to their scores on neuropsychological tests like the ones that you're going to be completing today. Because the purpose of our study is to look at the influence of peoples' level of motivation and effort on the tests that you're about to complete, we ask that you really try to perform to the best of your abilities. Of course, some of the tests are going to be more difficult than others, and no one is expected to get everything right. We just really ask that you put in your best effort so that we get a good measure of peoples' performance when they are trying their best.

Does that make sense? Do you have any questions?

Malingering Script

I would like you to imagine that you are an athlete whose prospects for a career in professional sports depend on your ability to play your sport for the duration of the upcoming athletic season. Recently, there has been increased awareness about sport-related concussions and, as a result, your team is required to undergo baseline cognitive testing to measure everyone's performance at the beginning of the season. Anyone who sustains a concussion will have to retake the tests and will not be able to return to play until their performance has returned to baseline levels. This means that if you do well on the tests now but are not able to perform as well after a head injury, you will not be allowed to return to play until your performance on these tests is back to its original level. It is therefore NOT in your best interest to perform to the best of your ability on the tests that you are about to take. This way, you will be more likely to remain in play if you do sustain a concussion at some point during the season.

You have been a competitive athlete for a number of years and have sustained a concussion in the past; you remember that after your concussion you experienced persistent headaches, occasional dizziness, as well as memory lapses for about a month or so. Being removed from play for any number of games would be very damaging to your athletic career, so I would like you, in the best way you know how, to respond to the tests in a manner that is similar to how you would perform

after a concussion. However, you do not want to perform so poorly as to make it obvious that you are “tricking” the test.

Do you have any questions?

I also want to remind you that this is part 1 of a 2-part study. So, at the end of the testing session, you will be asked whether you consent to be contacted for Part 2 after data collection for Part 1 has been completed. Part 2 will take about 40 minutes of your time, and, if you choose to participate, you will be paid between $10 and $40 based on how well you manage to successfully fake a mild brain injury today. So, the more closely your performance today resembles what we would expect from an athlete with a concussion, the more money you will be paid later if you choose to participate in Part 2.

Does that make sense? Do you have any questions?

Appendix B – A Description of ImPACT Subtests

Word Memory – the Word Memory subtest is designed to measure attention and verbal recognition memory. The individual is presented with a list 12 words, twice, for 750 ms per word. They are then presented with a list of 24 words and asked to identify which words they had seen as part of the original list by clicking "yes" or "no" on the screen. Distractor words are chosen from the same semantic category as target words. Five versions of the word list are available to minimize practice effects. After a 20-minute delay (during which the participant completes other subtests), the individual is again asked to identify the words that were part of the original list.

Design Memory – the Design Memory subtest is designed to measure attention and visual recognition memory. The individual is presented with a series of 12 designs, twice, for 750 ms per design. They are then presented with a series of 24 designs and asked to identify which designs they had seen before by clicking "yes" or "no" on the screen. Distractor designs are target designs that have been rotated in space. The designs were selected in order to make verbal encoding difficult, and different subsets of designs are available to reduce practice effects. After a 20-minute delay (during which the participant completes other subtests), the individual is again asked to identify the designs they had seen as part of the group of designs.

X's and O's – The X's and O's subtest is designed to measure visual working memory and visual processing/visual motor speed. The individual is presented with a distractor task, in which they are asked to press a specific key based on the image they see on the screen (e.g., "if you see a blue circle, press the "p" key on the keyboard"). After

completing the distractor task, they are presented with a screen of randomly assorted X's and O's which is displayed for 1.5 seconds. Each time the X's and O's are presented, three X's or O's are highlighted in yellow, and the subject is asked to remember the location of the highlighted letters on the screen. Following the presentation of the letters, the distractor task is presented again to interfere with rehearsal. After completing the distractor task, the individual is once again presented with a screen of X's and O's and asked to indicate which letters were previously highlighted. This process is repeated for four trials.

Symbol Match – The Symbol Match subtest is designed to measure visual processing speed, learning, and memory. The individual is presented with a grid of the digits 1-9 paired with a common symbol. Symbols are readily identifiable (e.g., triangle, square, arrow). With the grid available to them, the individual is presented with a symbol and asked to click, as quickly as possible, on the number that corresponds with that symbol. After 27 trials, the symbols from the grid are removed. The individual is then again shown a series of symbols and asked to indicate, from memory, the number that was matched with each symbol.

Color Match – The Color Match subtest is designed to measure impulse control/response inhibition. The individual is first asked to click a red, blue, or green button on the screen to ensure adequate color vision. After this, the individual is presented with color words presented in a box in either the same color as the word, or in a different color (e.g., the word RED would be presented in red on color-congruent trials, and in another color on incongruent trials). The subject is asked to click in the box as quickly as possible, but only if the word appears in the matching color.

Three Letters – The Three Letters subtest is designed to measure working memory and visual-motor response speed. The individual is first presented with a distractor task, where they are presented with a randomly scattered grid of the numbers 1-25 and asked to count backwards from 25 by clicking on each successive number. Three consonants are then presented on the screen. The distractor task is then presented again for 18 seconds, after which the individual is asked to recall the three letters by typing them on the keyboard. This process is repeated five times.

VITA AUCTORIS

NAME:	Isabelle Messa
PLACE OF BIRTH:	Lasalle, Quebec
YEAR OF BIRTH:	1988
EDUCATION:	Bramalea Secondary School, Brampton, ON, 2006
	University of Waterloo, B.Sc., Waterloo, ON, 2010
	University of Waterloo, M.Sc., Waterloo, ON, 2014