2019

DETECTION OF COGNITIVE CHANGE: EXAMINATION OF APPROACHES FOR IMPROVING THE ACCURACY OF IMPACT

Charles Edward Gaudet III
University of Rhode Island, chad_gaudet@my.uri.edu

DETECTION OF COGNITIVE CHANGE: EXAMINATION OF APPROACHES FOR IMPROVING THE ACCURACY OF IMPACT

BY

CHARLES EDWARD GAUDET III

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR IN PHILOSOPHY

IN

CLINICAL PSYCHOLOGY

UNIVERSITY OF RHODE ISLAND

2019

DOCTOR OF PHILOSOPHY DISSERTATION

OF

CHARLES EDWARD GAUDET III

APPROVED:

Dissertation Committee:

Major Professor: David Faust

Kathleen Webster

Jeff Konin

Nasser H. Zawia
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2019

ABSTRACT

Concussion is an increasingly recognized public health concern. Proper assessment and management of concussion are critical factors in mitigating adverse effects associated with this injury. Neuropsychological assessment has demonstrated utility in identifying cognitive symptoms related to concussion and monitoring their resolution. Early methods involved administering paper and pencil tests to appraise cognitive domains thought to be most affected by concussion. As interest in concussion and methods of assessment evolved, baseline testing became an integral component of assessment in sport-related concussion (SRC). Baseline testing consists of administering a healthy, or non-injured, individual a battery of cognitive tests that subsequently serve as a reference point to evaluate the individual's performance on the same tests following a suspected injury or change in cognitive status. Recent advances, spurred by an interest in increasing access to baseline testing, contributed to the adoption of computerized neurocognitive tests (CNTs). CNTs allow for baseline testing of groups of individuals, in one setting, and in a short amount of time. Immediate Post Concussion and Cognitive Testing (ImPACT) has emerged as the most commonly used CNT in the assessment and management of SRC. This body of research aimed to explore ImPACT's reliability and validity to appraise its efficacy in accurately detecting cognitive change associated with concussion and explore potential improvements. The first chapter is devoted to examining ImPACT's test-retest reliability, which refers to the expected consistency in results over time in healthy individuals. This study examines ImPACT score reports for 107 healthy individuals

that included testing at two time points. Results reveal less than adequate test-retest reliability attributable to, at least in part, a restricted range of possible scores, or the presence of a ceiling effect, on numerous subscales. Additional discussion includes corrective measures, such as proactively identifying individuals producing maximum scores on baseline testing, extending the length of subscales, and incorporating adaptive testing. The second chapter evaluates ImPACT's validity, and specifically, its classification accuracy in differentiating between individuals with and without concussion. This study incorporates a novel approach through its use of standardized regression based (SRB) reliable change index (RCI) scores to measure post-injury testing deviations from baseline scores. The SRB methodology, coupled with discriminant function analyses (DFAs), is compared to current interpretive procedures. The study includes 129 individuals without concussion whose SRB RCI scores are compared to 81 individuals with concussion. Results of analyses suggest that the current interpretive procedure performs at a chance level in accurately identifying individuals with concussion; conversely, the SRB method and DFA approach yield positive predictive values exceeding 80%, however sensitivities below 50%. Additionally, the Post Concussion Symptom Scale (PCSS), a self-report measure of symptoms, is largely equivalent to ImPACT's cognitive measures in classification accuracy. Collectively, these results raise considerable concern regarding ImPACT's efficacy as a measure to aid in the assessment and management of concussion.

ACKNOWLEDGMENTS

I owe a tremendous debt of gratitude to my major professor, Dr. David Faust. Although I have undoubtedly benefited from your seemingly infinite knowledge, your enduring impact on my development resides with your diligence, work ethic, and compassion for others; characteristics that I aspire to emulate.

I want to thank my committee: Dr. Kathleen Webster, Dr. Jeff Konin, and Dr. William Renehan, for your support and guidance throughout my graduate studies. I have been incredibly fortunate to engage in this process with a group of open-minded and talented scientists and researchers.

I would also like to thank Dr. Lisa Weyandt and Dr. Lisa Harlow, who have afforded me ample opportunities to pursue and advance my knowledge in neuroscience and methodology, respectively. Additionally, I would like to thank the University of Rhode Island's Psychology Department. It is an extraordinary collection of individuals, and I have been fortunate to learn from so many of you.

Lastly, I would like to especially thank Deena Mandes and her colleagues Helen Pagliaro and Jill Wainwright. They consistently provide answers to my endless stream of questions. It has been a pleasure working with you.

DEDICATION

I dedicate this dissertation to my family.

To my wife Hilary, your love and support have been inspirational. Thank you for allowing me the time to take on this pursuit.

To my parents, Jody and Chuck, thank you for your unconditional belief.

To my brothers, Ross and Cam, thank you for relenting on puns at my expense for not having “a real job” for the past 5 years.

To Laura Darby-McNally, nearly 15 years later and still reliably reliable.

And most importantly, to my son Charlie, you are relentless in your pursuits, and I am incredibly grateful to witness them.

PREFACE

This dissertation is presented in Manuscript Format. These studies stem from an ongoing body of research in collaboration with Dr. David Faust at the University of Rhode Island, endeavoring to explore methodological improvements to advance the practice of clinical neuropsychology. Each of the two manuscripts presented will be submitted for publication to the specified journals highlighted on each manuscript title page upon final dissertation submission. The methodological considerations examined within, I believe, possess applications well beyond concussion. I am optimistic regarding neuropsychology's future, and I am privileged to offer even a small contribution to the field's advancement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

Immediate Post Concussion and Cognitive Testing (ImPACT): An assessment of factors affecting test-retest reliability

by

Charles E. Gaudet[1]

To be submitted to the *Archives of Clinical Neuropsychology* (or a comparable outlet)

[1]PhD Candidate, Department of Psychology, The University of Rhode Island, Kingston, RI 02881. Email: chad_gaudet@my.uri.edu

Abstract

Objective: Computerized neurocognitive tests (CNTs) have been widely adopted and occupy a prominent position in concussion assessment and management. The most commonly used CNT in this endeavor, Immediate Post Concussion and Cognitive Testing (ImPACT), has demonstrated a range of test-retest reliabilities. This study aimed to further examine the reliability of this measure, along with factors potentially affecting its reliability.

Method: A retrospective file review was conducted for 300 consecutively selected, ImPACT score reports generated between 2010-2015 by individuals attending a secondary school. Test-retest reliability for composite and subscale scores was analyzed using two statistics: Pearson product moment correlations (*r*) and Intraclass Correlation Coefficients (ICCs). Additionally, to appraise potential ceiling effects, we calculated the number of individuals obtaining maximum possible scores and scores within 10% of the maximum possible score. Lastly, subscales producing test-retest reliabilities greater than 0.60 were combined to determine whether a supplemental index might demonstrate improved reliability over the current composite scores.

Results: Of the score reports, 107 included multiple baseline assessments without an intervening concussion. Test-retest reliabilities ranged from 0.42 to 0.69 for composite scores and 0.18 to 0.68 for subscales. Multiple subscales evidenced ceiling effects, with the most prominent appearing on the Word Memory, Design Memory, and X's and O's subscales. The supplemental index produced test-retest reliabilities ranging from 0.57 to 0.74.

Conclusions: These results are consistent with a large segment of the literature and raises considerable concern regarding ImPACT's reliability. Further, this study identified factors potentially adversely affecting the reliability of composite scores through its examination of subscale reliabilities and ceiling effects. Lastly, this study generated a supplemental index that produced reliability values exceeding the current composite scores.

Keywords: ImPACT; test-retest reliability; concussion; serial assessment; neurocognitive testing; baseline testing

Introduction

Background

Neuropsychological assessment refers to "the systematic assessment of cognitive abilities that often also evaluates patterns of behavior, affect, personality, and major psychiatric disorders" (Loring, 2015, p. 260). Such systematic evaluation often includes structured tests that permit comparisons between an individual's performance with populations or reference groups of interest, or with that same individual's prior performance to assess for such matters as possible change over time. (Heilbronner et al., 2010; Lareau & Ahern, 2012; Strauss, Sherman, & Spreen, 2006). In either case, neuropsychological assessment can be no better than the quality of the data upon which it is based, thus rendering the psychometric quality of tests a critical matter. For example, although perhaps overstated at times (for technical reasons to be described later), a test result derived from a measure with poor reliability will be masked by error, thereby obfuscating a clinician's ability to estimate an examinee's level of functioning.

Reliability generally refers to the consistency of a measure, and there are multiple forms of reliability, such as test-retest, internal, and inter-rater (American Education Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Test-retest reliability, or the consistency of scores at two time points, occupies particular clinical relevance as it provides a proxy for the generalizability of test results over a specified time period among other considerations such as most likely true score, score comparisons between tests, and

pattern analysis (American Education Research Association, 2014; Calamia, Markon, & Tranel, 2013; Larrabee, 2018). For example, if an individual obtained a score of 100 on a test at Time 1, a clinician may use the test-retest reliability coefficient to predict the individual's score at Time 2. As such, substantial deviations from the expected score at Time 2 may be indicative of meaningful improvement or decline in abilities. As test-rest reliability declines, the accuracy of Time 2 score predictions lessens, and the proportion of variance in observed scores attributable to error is higher. In the context of neuropsychological assessment, the example described above illustrates the significance of test-retest reliability in determining whether change is attributable to such factors or conditions as acquired brain injury, neurodegenerative conditions, other psychiatric or neurological changes, or simply measurement error.

Reliability is often a precursor for validity. Validity generally refers to the accuracy of measurement, or whether a test measures what it purports to measure (American Education Research Association, 2014). As such, for a test to be valid, it requires a reasonable degree of reliability to ensure consistency in test results. For example, consider a test that is purported to measure a relatively stable ability, such as word reading, for which the same individual produces two vastly different scores (not attributable to a practice effect or an intervening change in neurological status). It would be difficult to conclude that this measure accurately measures word reading abilities. Moreover, to the extent a measure is valid, improving its reliability will correspondingly enhance its validity.

Test-retest reliability is generally reported in administrative or technical manuals accompanying neuropsychological tests, or in peer-reviewed studies. When

appraising test-retest reliability for a given measure, several considerations warrant careful review. The magnitude of the test-retest reliability coefficient is a critical starting point, and potential endpoint, when evaluating whether a measure is appropriate for use. At present, there is not an agreed upon system for classifying test-retest reliabilities, and qualifying factors vary by test purpose and use (Calamia et al., 2013). For purposes of this discussion, reliabilities greater than 0.70 will be considered "adequate" (Strauss et al., 2006). Overall, neuropsychological measures generally demonstrate adequate reliability and several measures high or very high reliability. A meta-analysis of commonly used neuropsychological measures revealed test-retest reliabilities ranging from 0.28 (Auditory Verbal Learning Test – Recognition score) to 0.92 (Wechsler Adult Intelligence Scale-IV – Information score), with most reported test-retest reliabilities greater than .70, or at least in the adequate range; however, not all reliabilities for each score produced by a given measure were reported (Calamia et al., 2013).

Additional factors necessitating further consideration in appraising test-retest reliability include the length of the test-retest interval, size of the sample included in the reliability study, heterogeneity of the sample, and clinical status of the sample included in the reliability study (i.e., healthy vs. clinical) among others (Calamia et al., 2013; Duff, 2012). Longer test-retest intervals are associated with reduced test-retest reliabilities (Calamia et al., 2013; Duff, 2012). However, in serial neuropsychological assessment, there is ongoing debate regarding whether the selection of test-retest reliability coefficients for use in clinical practice should align with an individual's time since injury, or the value reported in the manual, which is often based on a much

shorter interval (Chin, Nelson, Barr, McCrory, & McCrea, 2016). Test-retest reliability coefficients derived from larger samples are associated with higher levels of reliability (Bridges & Holler, 2007; Crawford & Garthwaite, 2008). Additionally, the status of the sample used to generate test-retest reliabilities is an important yet under-researched factor (Duff, 2012). For example, groups with schizophrenia have produced lower test-retest reliabilities relative to nonpsychiatric controls (Granholm, Link, Fish, Kraemer, & Jeste, 2010). As such, test-retest reliability should not be considered a static psychometric of a test, but rather a dynamic quality that may vary in relation to multiple factors and specifics.

The role of neuropsychological assessment in concussion

Concussion refers to a rotational, acceleration-deceleration force, or pressure wave sufficient to disrupt brain function resulting in an acute and subacute pathophysiological metabolic response (Elder, Mitsis, Ahlers, & Cristian, 2010; Romeu-Mejia, Giza, & Goldman, 2019; Signoretti, Lazzarino, Tavazzi, & Vagnozzi, 2011). This transient form of mild head injury is generally not detectable using basic computed tomography ([CT] Riggio & Jagoda, 2016) and magnetic resonance imaging ([MRI] McCrory et al., 2013). More advanced physiological measures, including functional MRI (fMRI), diffusion tensor imaging (DTI), magnetic resonance spectroscopy (MRS), and transcranial magnetic stimulation (TMS) have revealed incongruent findings regarding the duration of cerebral dysfunction following imaging (Kamins et al., 2017). Moreover, research suggests the time course of cognitive recovery and symptom recovery may not overlap with the course of physiological recovery (McCrory et al., 2017). As such, neuropsychological assessment, given its

application in appraising psychological and cognitive status, has occupied an important role in concussion diagnosis and management (Barr & McCrea, 2001; Barth et al., 1989; McCrea et al., 2005; Van Kampen, Lovell, Pardini, Collins, & Fu, 2006).

Neuropsychological assessment can make various contributions to concussion diagnosis and management. At a basic level, assessment includes a self-report post-injury symptom checklist, which is not unique to neuropsychology; however, given the overlap between concussive symptoms and those of everyday stress (e.g. headache) and depression (e.g. reduced motivation) it may be difficult to differentiate among etiologies (Iverson et al., 2015; Riegler, Guty, & Arnett, 2018). A more comprehensive approach is to combine symptom checklists with cognitive or formal neuropsychological tests. When such assessment is limited to the post-injury period, the interpretation of results may partly or mainly rely on comparison to normative groups matched along critical sociodemographic features, such as age and education. In contrast, the repeated-measures approach, or comparison between baseline and post-injury functioning, has long been considered the optimal method for appraising neuropsychological functioning in the context of concussion (Barth et al., 1989; McCrea et al., 2005). However, given concerning psychometric characteristics associated with measures commonly employed in the repeated-measures approach to concussion assessment, this assertion has recently been questioned (Alsalaheen, Stockdale, Pechumer, & Broglio, 2016a, 2016b; Echemendia et al., 2013; McCrory et al., 2017).

Computerized neurocognitive tests (CNTs) have assumed a prominent role in the neuropsychological assessment of concussion, particularly sport-related

concussion (SRC). Athletic trainers have identified CNTs as an important component of assessment, evidenced by rapidly increasing use in the past 15 years (Lynall, Laudner, Mihalik, & Stanek, 2013). Commonly used CNTs include Immediate Post Concussion Assessment and Cognitive Testing (ImPACT), Cogsport or Axon Sport, Automated Neuropsychological Assessment Metrics (ANAM), and Headminder. Among such measures, ImPACT is the most frequently used by a wide margin; surveys indicate that approximately 90% of athletic trainers use the measure (Covassin, Elbin, & Stiller-Ostrowski, 2009; Lynall et al., 2013; Meehan, Collins, Taylor, & Dawn Comstock, 2012). Given its widespread use in concussion assessment, a thorough understanding of its psychometric characteristics is essential to appraise clinical utility.

ImPACT is designed to function as a serial, or longitudinal, measure in which examinees undergo assessment when presumed healthy and functioning normally (i.e., baseline) and again following suspected concussion (i.e., post-injury). In the absence of a baseline test, individuals may undergo post-injury assessment in isolation, and results are interpreted using percentiles based on normative data. The measure consists of six test modules that assess aspects of attention, processing speed, reaction time, and memory. Subscale scores within these modules are combined to form four primary cognitive composite scores.

The four cognitive composite scores were, "derived logically rather than through factor analysis and were designed to provide summary level information to the healthcare provider using the test" (Lovell, 2016, pp. 31-32). Composite scores are not presented in a standardized metric (e.g., T-score, z-score), but rather an averaged raw

score consisting of components of varying subscales. The Verbal Memory composite includes the following module scores: Word Memory – Total Percent Correct, Symbol Match – Total Correct (hidden), and Three Letters – Percentage of Total Letters Correct module scores. The Visual Memory composite includes the following module scores: Design Memory – Total Percent Correct and X's and O's – Total Correct (memory) module scores. The Visual Motor Speed composite includes the following module scores: X's and O's Total Correct (interference) and Three Letters – Average Counted Correctly module scores. The Reaction Time (RT) composite includes the following module scores: X's and O's – Average Correct RT (interference), Symbol Match – Average Correct RT (visible), and Color Match – Average Correct RT.

Additionally, the measure includes the Post Concussion Symptom Scale (PCSS), a self-report form with 22 items that are each rated on a 7-point scale (0-6), and address symptoms commonly associated with concussion. Decrements between baseline and post-injury scores, in the absence of mitigating factors, are generally ascribed to cognitive deficits or symptoms associated with the intervening injury (Lovell, 2016). The ImPACT Administration Manual does not offer clear interpretive guidelines or decision rules (Lovell, 2016); however, recent research suggests the presence of a score exceeding the reliable change interval on at least one of the cognitive composite scores and, or, a PCSS score exceeding the reliable change interval, constitutes the likely presence of ongoing concussion symptoms (Van Kampen et al., 2006).

Considering the critical importance of test-retest reliability, especially using a pre-post comparison strategy, it is essential to examine ImPACT's standing on this

measurement property. The earliest direct comparison between test-retest reliability of ImPACT relative to traditional paper and pencil tests appeared in 2005 (Randolph, McCrea, & Barr, 2005). The authors concluded that ImPACT's test-retest reliabilities, as reported in its manual (0.54 – 0.76), were inadequate for concussion assessment and screening. Additionally, the review also questioned the utility of traditional paper and pencil tests for concussion assessment, given their lack of comportment with specified criteria. Among such criteria, the authors cited, "establishing test-retest reliability over time intervals that are practical for this clinical purpose. Because baseline testing is likely to precede postinjury testing by a period of weeks to months (or even years), test-retest reliability should be established for all applicable time periods" (Randolph et al., 2005, p. 150). This criterion provides a framework for evaluating subsequent research examining ImPACT's test-retest reliability.

The most comprehensive review of ImPACT's test-retest reliability appeared in 2016 (Alsalaheen et al., 2016a). This systematic review reported test-retest reliabilities, using both Pearson product correlation coefficients (*r*) and intraclass correlation coefficients (ICCs), for the four cognitive composite scores from 10 studies that met inclusion criteria. Inclusion criteria included study samples with participants who completed ImPACT at least twice without sustaining a concussion between assessments and reliability statistics reported as either a Pearson *r* or ICC; test-retest intervals ranged from 24 hours to 2 years. Overall, test-retest reliabilities fell below the adequate range. For example, of the 36 Pearson product correlation coefficients reported, only three (~8%) were greater than 0.80; in contrast, nearly half of the coefficients (17 of 36, or ~47%), were less than 0.60. Given the extent of measurement

error associated with ImPACT, coupled with the high percentage of false positive errors reported across studies, the review urged clinicians to exercise caution when incorporating ImPACT data into clinical decisions.

Concerns surrounding ImPACT's test-retest reliability persist. A recent study compared reliabilities for three CNTs across multiple time points (Resch, Schneider, & Munro, 2018). In a sample of 41 non-athlete university students, the authors reported the following reliabilities over a 47-day test-retest interval: Verbal Memory = 0.19, Visual Memory = 0.45, Visual Motor Speed = 0.81, and Reaction Time = 0.57; at a 54-day test-retest interval, they reported the following reliabilities: Verbal Memory = 0.32, Visual Memory = 0.54, Visual Motor Speed = 0.74, and Reaction Time = 0.53. The authors also presented reliabilities over a 7-day interval, composed of results from Times 2 and 3: Verbal Memory = 0.63, Visual Memory = 0.47, Visual Motor Speed = 0.89, and Reaction Time = 0.59. Across time intervals, the Verbal Memory composite's reliability was among the lowest, and the Visual Motor Speed among the highest.

A meta-analysis attempted to synthesize test-retest reliabilities for three CNTs (Farnsworth, Dargo, Ragan, & Kang, 2017). Test-retest intervals ranged from 1 day to 2 years. Results of this study revealed the following test-retest reliabilities for ImPACT's four cognitive composite scores: Verbal Memory = 0.52, Visual Memory = 0.56, Visual Motor Speed = 0.77, and Reaction Time = 0.65. Of note, the meta-analysis only included studies that reported the ICC and excluded studies that only included Pearson product correlation coefficients. Relative to the other CNTs analyzed (Axon/Cogsport & ANAM), ImPACT was deemed a less desirable measure as it

displayed poorer test-retest reliabilities than Axon and required nearly twice the time to administer.

The National Collegiate Athletic Association (NCAA) and Department of Defense (DoD) initiative to study concussion, the CARE Consortium, recently published results from a longitudinal study evaluating the test-retest reliability of concussion assessment tools (Broglio et al., 2018). The sample included over 4,000 university student-athletes and cadets at participating military service academies. At a 1-year interval, the authors reported the following reliabilities for composite scores: Verbal Memory = 0.50, Visual Memory = 0.58, Visual Motor Speed = 0.72, and Reaction Time = 0.47. At a 2-year interval, the authors reported the following reliabilities: Verbal Memory = 0.47, Visual Memory = 0.47, Visual Motor Speed = 0.66, and Reaction Time = 0.34. Given the relatively similar results observed among the other concussion assessment measures, such as Computerized Neurocognitive Software Vital Signs (CNS Vital Signs), Cogstate Computerized Cognitive Assessment Tool (CCAT), and the Standardized Assessment of Concussion (SAC), the authors stated, "The reliance on consensus and clinical experience to implement these measures is at odds with reliability metrics presented herein" (p. 1265). Moreover, the authors questioned the utility of repeated baseline assessments given variability of scores over time, as observed in the study.

Although considerable research documents ImPACT's apparent weaknesses in test-retest reliability, there is ongoing debate regarding contributing factors. Alsalaheen and colleagues (2016a) suggested that methodological factors stemming from statistical differences inherent to the use of Pearson product correlation

coefficients and varying ICC models have contributed somewhat to discrepancies between studies on test-retest reliabilities. Specifically, studies that have used the two-way mixed model with *average measures* to calculate an ICC value have produced higher reliabilities on average. For example, Schatz and Ferris (2013) used this procedure and reported the following reliabilities over a one-month test-retest interval: Verbal Memory = 0.79, Visual Memory = 0.60, Visual Motor Speed = 0.88, and Reaction Time = 0.77 (Schatz & Ferris, 2013). Elbin, Schatz, and Covassin (2011) produced similarly high results using this procedure over a 1-year interval: Verbal Memory = 0.62, Visual Memory = 0.70, Visual Motor Speed = 0.85, and Reaction Time = 0.76 (Elbin, Schatz, & Covassin, 2011). However, use of the *average measures* model does not appear appropriate as it is intended when the test is administered to the same individual *multiple times* at each time point, which was not consistent with the methodology used in the studies described above (Alsalaheen et al., 2016a).

Additionally, the time interval between baseline testing and the re-baseline testing, in healthy individuals without an intervening concussion, has been cited as a factor, with longer test-retest intervals associated with poorer test-retest reliabilities (Farnsworth et al., 2017; Resch et al., 2018); yet this finding has not been uniform across studies (Alsalaheen et al., 2016a). Additional factors potentially influencing reduced test-retest reliability include effort, group versus individual administration setting, and demographic characteristics of individuals undergoing assessment (see Gaudet & Weyandt, 2017; Resch et al., 2018).

The factors above include essential considerations in determining the applicability of test-retest reliabilities for specific individuals, in certain settings, and at varying time points. However, more fundamental analysis of ImPACT's intrinsic components and their contribution to its apparent deficiencies in test-retest reliability is lacking as evidenced by the dearth of research analyzing individual subscale scores (Henry & Sandel, 2015). Moreover, there has not been sufficient examination of how features of ImPACT's test construction contribute to the documented problems with test-retest reliability. A better understanding of the factors diminishing test-retest reliability may also aid in designing corrective steps.

Mayers and Redick (2012) have raised concerns regarding one of ImPACT's cognitive composite scores. The authors noted a negatively skewed distribution of scores on the Verbal Memory composite, which in turn may contribute to inflated ICC values. Allen and Gfeller's (2011) factor analytic study raised concerns about another potential design problem. When discussing their outcomes, they indicated that a ceiling effect obtained on the Color Match Commissions score distribution likely resulted from problems in test design rather than the expected score distribution of the sampled population.

Skewed distributions are likely to distort measures of relative consistency, such as a Pearson product correlation, as the percentage of the sample obtaining maximum scores at Time 1, will at best, only be able to match their score at Time 2, or decline. Assuming some degree of practice effects between administrations, the positive increase in scores from Time 1 to Time 2 for individuals obtaining less than a maximum score at Time 1 counteracts the consistency between scores observed for

individuals obtaining maximum scores at Time 1 and again at Time 2. As such, this further reduces the likelihood of obtaining desirable test-retest reliability.

Objectives

This study aimed to assess ImPACT's test-retest reliability in a population of high-achieving secondary school students. It was designed to extend the existing literature in two critical regards. First, analyses examined population- (e.g., age & gender) and administration-based (time interval) factors contributing to test-retest reliability. Second, analyses also evaluated aspects associated with test design, such as ceiling effects, in the context of test-retest reliability across not only composite but also subscale scores. The inclusion of ImPACT subscale component scores represents a novel and understudied aspect of the measure. Lastly, as an exploratory analysis, a combination of subscales demonstrating high test-retest reliabilities was examined to determine whether it might produce improved reliability relative to the existing composites.

Methods

Study Design

This study is an archival, or retrospective, cross-sectional chart review.

Setting

A secondary school located in the U.S. provided ImPACT data drawn from its student population. The school serves students in grades 9-12. Reported mean scores on the Scholastic Aptitude Test (SAT) for the 2017 graduating class were as follows: Critical Reading = 690/800 (>90th percentile), Math = 700/800 (>90th percentile), Writing = 690/800 (>90th percentile), Overall Score = 2080/2400 (>90th percentile).

Participants

Individuals completed ImPACT as part of their school's standard concussion management protocol. File review was conducted for 300 consecutively selected, ImPACT score reports generated between 2010-2015. ImPACT reports included scores from a single baseline test, multiple baseline tests, a single baseline test and post-injury test, or multiple baseline tests and a post-injury test. Baseline tests were administered in a group setting; post-injury tests were administered in a one-on-one setting. All tests were administered by a certified athletic trainer.

Data were de-identified before analyses. As such, the Institutional Review Board determined the study was exempt from full review for human subjects research.

Measure

ImPACT protocols included either version 2.0 or 2.1. Version 2.1 provides data integration with version 1.0; subscale and composite scores are equivalent between versions. Twenty-four percent of participants completed version 2.0. ImPACT includes six cognitive performance modules: Word Memory, Design Memory, X's and O's, Symbol Match, Color Match, and Three Letters. Individual scores from these modules are combined to form five composite scores: Verbal Memory, Visual Memory, Visual Motor Speed, Reaction Time, and Impulse Control. Of note, the Impulse Control composite examines protocol validity rather than cognitive change (Lovell, 2016). The test also includes the PCSS. This self-report measure solicits examinee ratings using a 7-point scale (0-6) for symptoms commonly associated with concussion (e.g., headache, nausea, irritability). The inventory queries 22 total symptoms.

Procedures and Analyses

ImPACT reports yielded the following information for collection: basic demographic characteristics (gender, country of origin, first language, attention-deficit/hyperactivity disorder status, learning disability status), all composite and subscale scores for the first two baseline tests completed, and first post-injury test completed as available. Administration dates were also collected to allow for the calculation of intervals between administrations.

Analyses included only participants who had undergone multiple baseline assessments. Analyses did not include: individuals that produced invalid baseline results according to embedded ImPACT validity criteria as defined in the test manual (Lovell, 2016); Of note, research examining the accuracy of the embedded validity indicators is underdeveloped (Gaudet & Weyandt, 2017). Additional exclusionary criteria consisted of: participants with a reported history of concussion, ADHD, or learning disability (Cook et al., 2017; Elbin et al., 2013); participants born outside of the U.S. and for whom English was not the first language; and participants who sustained a concussion between baseline administrations. These criteria were applied to reduce factors that have been found to contribute to variance on ImPACT (Alsalaheen et al., 2016b).

We examined the normality of score distributions. Specifically, skewness values exceeding |1.0| were set as the criterion for negative or positive skewness (Harlow, 2014).

Test-retest reliability was analyzed using two statistics: Pearson *r* correlation coefficients and ICCs. The inclusion of both reliability statistics sustains continuity

with common ImPACT reporting practices (Alsalaheen et al., 2016a). Pearson *r* correlation coefficients are commonly used for measuring the test-retest reliability of neuropsychological measures (Strauss et al., 2006). Moreover, the Pearson *r* is input into the equation that generates reliable change intervals for ImPACT (Iverson, Lovell, & Collins, 2004; Lovell, 2016). Additionally, one-way random and two-way mixed ICCs for single, as opposed to average, measures were calculated (Landers, 2015; Shrout & Fleiss, 1979). The ICC model most appropriate for appraising ImPACT's test-retest reliability is uncertain; although single measures, as opposed to average, are suggested given an individual completes only a single test at each time point (Alsalaheen et al., 2016a).

For test-retest reliability analyses, we examined the four primary cognitive composite scores used to inform clinical decisions: Verbal Memory, Visual Memory, Visual Motor Speed, and Reaction Time. These analyses also included module subscale scores that comprise each composite. Additionally, we analyzed several factors potentially affecting composite and module subscale test-retest reliability, specifically age, gender, and test interval.

The number of individuals obtaining maximum scores and scores within 10% of the maximum scores served as a proxy to appraise ceiling effects. The following subscales included scores that allowed for the production of a maximum score: Word Memory, Design Memory, X's and O's, Symbol Match, and Three Letters.

Lastly, an exploratory analysis was conducted to evaluate the creation of an alternative composite score comprised of subscales with higher test-retest reliabilities. For this analysis, module subscale scores displaying test-retest reliabilities greater than

0.60 were combined to create a "High Reliability Index." A potential benefit of combining more reliable test components is to create a composite score that reduces obfuscation attributable to measurement error.

Results

Demographics

ImPACT score reports from 300 individuals were collected and analyzed. Of the reports, 217 included data from multiple baseline assessments. A check on exclusionary criteria yielded the following results: 74 (34.1%) participants reported a prior history of concussion, 21 (9.7%) reported an ADHD diagnosis, 9 (4.1%) reported a diagnosed learning disability, 17 (7.8%) reported non-U.S. place of origin and a first language other than English, and 73 (33.6%) participants sustained a concussion between baseline administrations; in addition, three baseline administrations (1.4%) were invalid based on embedded validity indicators set forth in the manual (Lovell, 2016). After all exclusions, 107 score reports remained that included a total of 214 baseline assessments.

The sample consisted of 50 females and 57 males; the mean age at first and second baseline administrations was 14.5 years and 16.3 years, respectively; the mean time interval between baseline administrations was 21.1 months. Table 1 displays frequencies, means, standard deviations, and ranges for demographic variables across the 107 individuals.

Table 2 displays score distribution characteristics for composites and module subscales. As can be seen by examining this table, multiple score distributions exceeded the |1.0| threshold for negative or positive skewness, or non-normality. For

composite scores, the Visual Memory composite was negatively skewed (-1.00) on the second baseline assessment; the Verbal Memory composite approached a negative skew (-0.70) on the second baseline assessment. For subscale scores included in the Verbal Memory Composite, Word Memory – Total Percentage Correct was negatively skewed (-1.32) on the second baseline assessment; Three Letters –Percentage of Total Letters Correct was negatively skewed (-1.22, -1.46) on the first and second baselines, respectively. For module subscales included in the Visual Memory composite, both Design Memory – Total Percentage Correct (-0.95) and X's and O's – Total Correct (memory) (-0.97) approached a negative skew on the second baseline assessment. For module subscales included in the Reaction Time composite, Symbol Match – Average Correct RT (visual) scores were positively skewed on both baseline assessments (1.48, 2.81); Color match – Average Correct RT was positively skewed (1.18) on the second baseline assessment.

Test-Retest Reliability

Table 3 lists reliabilities for the total sample ($N = 107$). Pearson r test-retest correlations ranged from 0.43 to 0.69 across composites and 0.19 to 0.68 across module subscales that form the composites. One-way random ICCs ranged from 0.43 to 0.56 across composites and 0.19 to 0.62 across module subscales. Two-way mixed ICCs ranged from 0.42 to 0.68 across composites and 0.18 to 0.67 across module subscales.

Gender

The total sample included 57 males and 50 females. Table 4 displays reliabilities stratified by gender. The mean age at initial baseline assessment was 14.4

(SD = 0.77) years for females and 14.5 (SD = 0.66) years for males; this difference was not statistically significant, $t(105) = 0.55$, $p = 0.59$. The mean interval time between baseline administrations was 20.7 (SD = 5.16) months for females and 21.5 (SD = 4.64) months for males; also a non-significant difference, $t(105) = 0.85$, $p = 0.40$. Test-retest reliability coefficients (r) ranged from 0.43 to 0.68 for females and 0.35 to 0.75 for males across composites, and 0.06 to 0.68 for females and 0.21 to 0.70 for males across module subscales that form the composites.

Age

Table 5 lists reliabilities divided by age group. The sample was divided by the average age at the midway point of secondary school (i.e., second, or sophomore, year). The sample included 67 participants under age 15 ("younger group") and 40 participants age 15 or older ("older group"). There were 30 females and 37 males in the younger group and 20 females and 20 males in the older group; this difference in proportion of males and females was not statistically significant, $\chi^2(2, N = 107) = 0.27$, $p = 0.60$. The mean interval time between baseline administrations was 23.4 (SD = 1.6) months for the younger group and 17.3 (SD = 6.02) months for the older group a difference that reached statistical significance, $t(105) = 7.86$, $p < 0.01$. Test-retest reliability coefficients (r) ranged from 0.41 to 0.65 for the younger group and 0.37 to 0.78 for the older group across composites, and from 0.10 to 0.67 for the younger group and 0.23 to 0.79 for the older group across module subscales that form the composites.

Test Interval

Sample size limitations precluded an analysis of the effect of time interval. Specifically, only 22 participants had a 1-year interval between baseline administrations; research suggests this is insufficient size to allow for a rigorous analysis of this factor (Bridges & Holler, 2007; Crawford & Garthwaite, 2008).

Ceiling Effect

Table 6 displays the percentage of individuals obtaining scores approaching the ceiling — a substantial number of cases obtained maximum possible scores across numerous subscales. On the Word Memory subscale, approximately 25% of cases identified 12/12 stimulus words accurately both during the learning and delayed trials, with more than 75% obtaining scores of 90% or higher. On the Design Memory subscale, approximately 5% of cases accurately identified 12/12 stimulus designs both during the learning and delayed trials; more than 30% of cases obtained scores of 90% or higher. On the X's and O's subscale, approximately 8% of cases obtained maximum scores; more than 15% of the cases obtained scores of 90% or higher. On the Symbol Match subscale, approximately 25% of cases accurately matched symbols and digits during the hidden trial. On the Three Letters subscale, approximately 50% of cases accurately the stimulus letters; more than 60% of cases obtained scores of 90% or higher.

High Reliability Index

Two module subscale scores produced test-retest reliabilities greater than 0.60: Three Letters – Average Counted Correctly (3L:ACC; r=0.68) and X's and O's Average Correct Reaction Time (interference) [XO:RT; r=0.68]). To compute a composite comprised of these subscale scores, we performed a crude score

transformation. The transformations allowed the scores to obtain relatively equal weights before summation, as the 3L:ACC raw score was a whole number, and the XO:RT raw score was fractional. 3L:ACC scores were multiplied by four to expand the range from 0-100. XO:RT scores were multiplied by 100 and then subtracted by 100 to expand its range and better equate its scores to 3L:ACC scores. Following the transformation, 3L:ACC and XO:RT scores were summed and divided by two.

Means and standard deviations for the High Reliability Index were 55.7 (SD = 9.4) at Time 1 and 61.5 (SD = 8.3) at Time 2. Skewness values were 0.26 at Time 1 and -0.31 at Time 2; Kurtosis values were -0.40 at Time 1 and -0.11 at Time 2. The Pearson *r* value was 0.74; the two-way mixed ICC was 0.73, and the one-way random ICC was 0.57.

Discussion

This study's primary aims were to advance knowledge of ImPACT's test-retest reliability and potential factors influencing standing on this critical psychometric parameter. Results showed that test-retest reliabilities for most composite and subscale scores fell below adequate levels, with many scores falling at low levels. Additionally, several composite and subscale score distributions displayed evidence of non-normal distributions, indicating a potential ceiling effect, at least with high functioning groups. Multiple adverse effects of ceiling effects will be addressed in the following discussion. Taken together, results of the present study raise significant concern regarding the ongoing use of ImPACT, particularly given its intended design as a serial measure and a growing body of research outcomes that raise serious questions about the stability of the scores it yields.

Test-retest reliabilities for the overall sample, as calculated by both Pearson *r* and ICC, fell below adequate levels for the composite scores. Pearson *r* values for the composite scores were as follows: Verbal Memory = 0.43, Visual Memory = 0.52, Visual Motor Speed = 0.69, and Reaction Time = 0.59. The accompanying ICC values were largely similar. These results comport with the majority of findings in the research literature (Alsalaheen et al., 2016a; Broglio et al., 2018; Farnsworth et al., 2017; Resch et al., 2018). Correspondingly, the 10 subscale scores also produced test-retest reliabilities falling below adequate levels (< 0.70). When the error component of most scores exceeds 40%, then various problems relating to diagnostic and predictive accuracy are likely to follow.

Interestingly, the Visual Motor Speed composite score has consistently produced the highest relative test-retest reliability among the composite scores; it yielded a Pearson *r* value of 0.69 in the present study. The high value appears at least partly attributable to the higher relative test-retest reliability of the two underlying subscale scores: X's and O's – Total Correct (interference) [*r*=0.59] and Three Letters – Average Counted Correctly (*r*=0.68). Of note, the score distributions were also among the least skewed of all the subscales, with values near zero. For the remaining three composite scores, only Reaction Time included a subscale component with a test-retest reliability greater than 0.60 (X's and O's – Average Counted Correct [Reaction Time – interference], *r*=0.68).

Conversely, the Verbal and Visual Memory test-retest reliabilities appear to have been affected adversely by the skewed distributions of their underlying subscale scores. For example, the Verbal Memory composite score distributions displayed

skewness values ranging from -0.37 to -0.70 at Times 1 and 2, perhaps consequently generating low test-retest reliability (r=0.43). Further investigation reveals even greater levels of negative skews among the component subscales, ranging from -0.72 to -1.46. Similar to the overarching composite score, the test-retest reliabilities for these subscales were low as well, ranging from 0.18 to 0.54.

The results of the study also align with prior research suggesting that demographic factors, such as age and gender, may affect ImPACT's test-retest reliability (Covassin, Schatz, & Swanik, 2007; Lichtenstein, Moser, & Schatz, 2014; Moser, Davis, & Schatz, 2018; Schatz & Robertshaw, 2014). In the present study, ImPACT composite and subscale scores appeared slightly less reliable for males as compared to females. Specifically, the mean composite score test-retest reliability (r) for the males was 0.53 versus 0.58 for females. Males and females evidenced larger discrepancies (0.06-0.15) on several subscales. Research involving gender differences in performance on neuropsychological testing has yielded inconsistent findings (Ardila, Rosselli, Matute, & Inozemtseva, 2011; Cromer, Schembri, Harel, & Maruff, 2015; Rosselli, Ardila, Matute, & Vélez-Uribe, 2014). As such, it is unclear whether these observed gender discrepancies are attributable to gender-based differences or random error.

The effect of age on ImPACT's test-retest reliability was variable. The mean composite score test-retest reliability (r) for the younger age group (< 15 years) was 0.54 versus 0.59 for the older age group (≥ 15 years). Three of the composite scores produced discrepancies from 0.10 to 0.16 between the age groups. For example, for the Visual Motor Speed composite score, the test-retest reliability for the younger

group was 0.65, whereas, for the older group, it was 0.78; Verbal Memory was the only composite score for which the test-retest reliability for the younger group was higher (0.47 to 0.37). Research suggests a general reduction in reaction time throughout adolescence (Cromer et al., 2015). Interestingly, subscale scores measuring reaction time tended to have higher test-retest reliabilities as four of the six values across age groups were higher than 0.50. Similar to gender effects, it is unclear the extent to which random error is accounting for differences between age groups.

Ceiling effects appeared prominent on numerous subscales, at least with the present, high functioning sample, and thereby creates interpretive challenges and problems. For seven of the 10 subscales examined, more than 10% of individuals obtained the maximum possible score on these subscales. For these individuals, assuming the maximum score does not represent their full capacity, clinicians are unable to accurately estimate their abilities on these subscales as they may have scored higher than the subscale allowed. Further, nearly 50% of individuals obtained the maximum possible score for the Three Letters subscale. These results are consistent with prior research (Allen & Gfeller, 2011; Mayers & Redick, 2012). Limitations attributable to ceiling effects preclude the possibility of accurate baseline measurement, which is a prerequisite for reliably identifying change in serial assessment.

For example, if an individual is capable of learning and recalling 20 total words, the inclusion of only 12 words in ImPACT's Word Memory subscale will under-represent that person's word learning and recall abilities substantially. Consequently, except for those for whom perfect performance represents full capacity,

for all others, the maximum score on this subscale underrepresents baseline functioning to varying degrees, and very likely some to a substantial degree. For example, if 40% perform at the ceiling and one assumes a normal distribution in capacities, about 15% of these individuals should have capacities that fall at about the 85th percentile given a normal distribution, and about 5% at the 95th percentile or higher. Therefore, a result that falls just a little above the middle of a bell curve can underrepresent true capacities considerably. Although not intended as a statement on the worth of individuals, the end result will be to miss cognitive dysfunction most often in the most capable individuals.

Given the high adoption rate and use of ImPACT in collegiate settings, its ceiling effects place examinees with above average to superior cognitive abilities at risk. For an individual with a word learning and memory capacity of 20 words, if this individual sustained a concussion and was only able to learn and recall 12 words, or experienced a large decline in abilities, the individual's score still equates to the initial baseline score. Hence, a clinician relying on this measure to inform return to play/activity decisions may erroneously conclude that there is no evidence of cognitive impairment.

Moreover, there appears to be a systematic relationship between subscales demonstrating appreciable ceiling effects and the test-retest reliability of the composite scores which they comprise. Both the Verbal and Visual Memory composite scores are calculated using combinations of the subscale scores. It has been well-established in the research literature that these two measures possess minimal clinical utility in detecting concussion (Alsalaheen et al., 2016b). It seemingly is not coincidental that

the Three Letters subscale score demonstrated the most prominent ceiling effect and also displayed the lowest test-retest reliability ($r = 0.19$). This subscale score is one of three composing the Verbal Memory Composite, which also had the lowest test-retest reliability of the four composite scores ($r = 0.43$). Conversely, the Visual Motor Speed Composite did not include any of these subscales that demonstrating a restricted range and displayed the highest test-retest reliability ($r = 0.69$).

Unfortunately, ceiling effects are not uncommon amongst neuropsychological tests. Such effects are strongly associated with test misinterpretation and erroneous conclusions of brain damage (Russell & Russell, 2003). For example, the Neuropsychological Assessment Battery (NAB) Naming Test consists of 31 items; however, because a substantial portion of individuals are able to name all 31 items correctly, this perfect score is only considered "average" (Brooks, Sherman, Strauss, Iverson, & Slick, 2009; Sachs, Rush, & Pedraza, 2016). There is also a risk that a neuropsychologist who does not look closely enough may misinterpret this result as reflecting a relative weakness in a high functioning individual who obtains higher scaled scores on other test components without such restriction in range. This potential underestimation, or misrepresentation, of cognitive abilities, is present among ImPACT composite and subscale scores and is especially problematic given its intended use as a serialized measure.

How then is a ceiling effect fixed? If a test is under construction or can be revised, the most basic approach consists of the basal-ceiling method, in which an examinee must answer a specified number of items either correctly or incorrectly to establish a basal or ceiling; commonly-used neuropsychological measures such as the

Boston Naming Test employ this method (Russell, 2003). Moreover, this approach is readily translatable to CNTs, in which more complex algorithms may be employed to determine when examinees obtain a score accurately reflecting their abilities (Russell, 2003).

However, when revisions to test-design are not possible, ceiling effects should be accounted for in alternative manners. One such option includes exercising caution when interpreting a score that is at the ceiling on retesting and not accepting it as evidence of stability. In the context of ImPACT, the Word Memory test, in which many individuals were able to accurately learn and recognize all 12 stimulus words resulting in a ceiling effect, represents one such test warranting caution. If such a performance is identified preemptively at baseline testing, an alternative list learning measure that includes more words, such as the California Verbal Learning Test-III, may be administered to more accurately appraise learning and memory abilities.

The calculation of the High Reliability Index yielded the highest test-retest reliability among both composite and subscale scores as computed using the Pearson *r*, 0.72. Additionally, this is the only test-retest reliability value among ImPACT indicators that would qualify as "adequate" according to Strauss et al.'s (2006) criteria. Given the concerns surrounding the reliability of ImPACT's existing composite scores, this result is encouraging as it provides preliminary evidence for an alternative approach to reducing measurement error in this widely adopted measure.

Lastly, the generally poor test-retest reliabilities observed in this study adversely affect ImPACT's current interpretive procedure. This procedure determines the presence of cognitive impairment or ongoing symptomatology based on an

individual producing one or more reliable change(s) on a composite score. As such, given the high rate of measurement error present in these composite scores, interpretive error is compounded when these scores are considered in conjunction. For example, consider the reliabilities for Verbal Memory ($r = 0.43$) and Visual Memory ($r = 0.52$). When these scores are considered conjunctively, reliability drops to 0.22, or results in nearly 0.80 measurement error. When considered in the context of multiple composite scores, ImPACT's utility in enhancing clinicians' decision-making accuracy is severely restricted.

Limitations

This study was subject to numerous limitations, and several caveats apply to the interpretation of the results. First, only 107 cases met the inclusion criteria for the study. This figure allowed for the calculation of test-retest reliabilities; however, relative to other more commonly used measures, this is a small number. Additionally, sample size limitations precluded an analysis of the effect of time interval between administrations. Specifically, only 22 cases had a 1-year interval between baseline administrations, thereby precluding a reliable analysis of this factor. As such, in accordance with Randolph, McCrea, and Barr's (2005) recommendation, the test-retest reliabilities reported in this study are only applicable to individuals sustaining a concussion more than 1-year following their baseline assessment. As a result, the interval may be longer than is clinically relevant in some cases.

Demographic characteristics of the sample warrant further discussion. The sample averaged scores above the 90th percentile on the SAT. As such, relative to scholastic aptitude, these results apply to a small segment of the general adolescent

population. This consideration is especially pertinent to the finding of ceiling effects on multiple subscales. Given the high aptitude of this sample, it may be fair to infer that these individuals may have been more likely to obtain maximum possible scores than individuals drawn from a more general population (e.g., a group averaging scores at the 50th percentile on the SAT).

Moreover, the high-achievement status of this sample reduced its variability. This reduction in variability is likely to suppress reliability relative to samples with more normally distributed characteristics, such as scores on achievement testing. As such, the very nature of this sample is a likely contributor to the lower reliabilities observed. Further, the high achievement status of the sample likely increased the frequency with which ceiling effects were present.

Cross-cultural applications of this study are limited as well. Data collection efforts were restricted to archival chart review, thereby preventing the collection of important cultural variables to assess their effect on ImPACT scores and interpretation. For example, research has documented differences in both pre- and post-injury performance between African-American and white athletes (Kontos, Elbin, Covassin, & Larson, 2010). Given, the diversity among athletic populations, where this measure is commonly employed, additional research appraising cultural effects is warranted.

Conclusion

In conclusion, this study replicated and advanced the accumulating body of research documenting concerns regarding ImPACT's test-retest reliability. Overall, none of the composite or subscale scores yielded adequate test-retest reliabilities.

Consequently, there is a wide range of error associated with many ImPACT scores, thereby reducing the clinical utility of this measure. This finding is particularly concerning given ImPACT's intended use as a serial measure. Given the range of error associated with baseline scores attributable to its poor test-retest reliability, the sensitivity of the measure to subtle changes associated with concussion is severely lessened. Additional results aligned with prior research indicating a trend toward lower test-retest reliability among male and younger age populations.

In addition, this study potentially increases understanding of factors that account for weak test-retest reliability, namely ceiling effects. Ceiling effects were present on numerous subscales, indicating skewed distributions. Such effects are attributable to a restricted range of possible scores and reduced test difficulty. Consequently, these factors result in both a potential underestimation of an individual's true abilities in a given domain and reduce variability, thereby adversely affecting test-retest reliability. Alternative approaches to design, such as adaptive testing, might alleviate this issue to an extent.

The High Reliability Index provided a reason for optimism as it introduced an alternative indicator composed of ImPACT's most reliable subscale scores. This example of an alternative approach to data combination represents a possible pathway toward improving ImPACT's psychometric deficiencies and enhancing its clinical utility. Additional research is warranted to evaluate whether this newly generated index provides discriminative utility in differentiating between individuals with and without concussion symptomatology.

Taken together, caution is warranted when interpreting ImPACT results. The extent to which ImPACT's cognitive measures provide incremental utility above and beyond existing protocols, such as self-report measures in concussion assessment and management, remains uncertain. ImPACT is a widely adopted measure, and following its approval by the Food and Drug Administration (FDA), the public perception of its efficacy in concussion assessment and management may well be high. However, given the ongoing concerns and growing evidence raising serious questions regarding ImPACT's psychometric shortcomings, future research targeting methodologies to reduce error will likely yield the greatest clinical utility.

Funding

The author did not receive funding for this project.

Acknowledgments

The author would like to thank his major professor, Dr. David Faust and his dissertation defense committee, Dr. Jeff Konin, Dr. Kate Webster, and Dr. William Renehan, for their guidance and support.

References

Allen, B. J., & Gfeller, J. D. (2011). The Immediate Post-Concussion Assessment and Cognitive Testing battery and traditional neuropsychological measures: A construct and concurrent validity study. *Brain Injury*, *25*(2), 179–191. https://doi.org/10.3109/02699052.2010.541897

Alsalaheen, B., Stockdale, K., Pechumer, D., & Broglio, S. P. (2016a). Measurement error in the Immediate Postconcussion Assessment and Cognitive Testing (ImPACT): Systematic review. *Journal of Head Trauma Rehabilitation*, *31*(4), 242–251. https://doi.org/10.1097/HTR.0000000000000175

Alsalaheen, B., Stockdale, K., Pechumer, D., & Broglio, S. P. (2016b). Validity of the Immediate Post Concussion Assessment and Cognitive Testing (ImPACT). *Sports Medicine*, *46*(10), 1487–1501. https://doi.org/10.1007/s40279-016-0532-y

American Education Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Ardila, A., Rosselli, M., Matute, E., & Inozemtseva, O. (2011). Gender differences in cognitive development. *Developmental Psychology*, *47*(4), 984–990. https://doi.org/10.1037/a0023819

Barr, W. B., & McCrea, M. (2001). Sensitivity and specificity of standardized neurocognitive testing immediately following sports concussion. *Journal of the International Neuropsychological Society*, *7*(6), 693–702. https://doi.org/10.1017/S1355617701766052

Barth, J. T., Alves, W. M., Ryan, T. V., Macciocchi, S. N., Rimel, R. W., Jane, J. A.,

& Nelson, W. E. (1989). Mild 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, NY: Oxford University Press.

Bridges, A. J., & Holler, K. A. (2007). How many is enough? Determining optimal sample sizes for normative studies in pediatric neuropsychology. *Child Neuropsychology*. https://doi.org/10.1080/09297040701233875

Broglio, S. P., Katz, B. P., Zhao, S., McCrea, M., McAllister, T., Reed Hoy, A., … Lintner, L. (2018). Test-retest reliability and interpretation of common concussion assessment tools: Findings from the NCAA-DoD CARE Consortium. *Sports Medicine*, *48*(5), 1255–1268. https://doi.org/10.1007/s40279-017-0813-0

Brooks, B. L., Sherman, E. M. S., Strauss, E., Iverson, G. L., & Slick, D. J. (2009). Developments in neuropsychological assessment: Refining psychometric and clinical interpretive methods. *Canadian Psychology*, *50*(3), 196–209. https://doi.org/10.1037/a0016066

Calamia, M., Markon, K., & Tranel, D. (2013). The robust reliability of neuropsychological measures: Meta-analyses of test-retest correlations. *Clinical Neuropsychologist*, *27*(7), 1077–1105. https://doi.org/10.1080/13854046.2013.809795

Chin, E. Y., Nelson, L. D., Barr, W. B., McCrory, P., & McCrea, M. A. (2016). Reliability and validity of the Sport Concussion Assessment Tool-3 (SCAT3) in high school and collegiate athletes. *American Journal of Sports Medicine*, *44*(9), 2276–2285. https://doi.org/10.1177/0363546516648141

Cook, N. E., Huang, D. S., Silverberg, N. D., Brooks, B. L., Maxwell, B., Zafonte, R.,

… Iverson, G. L. (2017). Baseline cognitive test performance and concussion-like symptoms among adolescent athletes with ADHD: Examining differences based on medication use. *Clinical Neuropsychologist*, *31*(8), 1341–1352. https://doi.org/10.1080/13854046.2017.1317031

Covassin, T., Elbin, R., & Stiller-Ostrowski, J. L. (2009). Current sport-related concussion teaching and clinical practices of sports medicine professionals. *Journal of Athletic Training*, *44*(4), 400–404.

Covassin, T., Schatz, P., & Swanik, C. B. (2007). Sex differences in neuropsychological function and post-concussion symptoms of concussed collegiate athletes. *Neurosurgery*, *61*(2), 345–351. https://doi.org/10.1227/01.NEU.0000279972.95060.CB

Crawford, J. R., & Garthwaite, P. H. (2008). On the "optimal" size for normative samples in neuropsychology: Capturing the uncertainty when normative data are used to quantify the standing of a neuropsychological test score. *Child Neuropsychology*, *14*(2), 99–117. https://doi.org/10.1080/09297040801894709

Cromer, J. A., Schembri, A. J., Harel, B. T., & Maruff, P. (2015). The nature and rate of cognitive maturation from late childhood to adulthood. *Frontiers in Psychology*, *6*, 704. https://doi.org/10.3389/fpsyg.2015.00704

Duff, K. (2012). Current topics in science and practice evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology*, *27*(3), 248–261. https://doi.org/10.1093/arclin/acr120

Echemendia, R. J., Iverson, G. L., McCrea, M., Macciocchi, S. N., Gioia, G. A.,

Putukian, M., & Comper, P. (2013). Advances in neuropsychological assessment of sport-related concussion. *British Journal of Sports Medicine*, *47*(5), 294–298. https://doi.org/10.1136/bjsports-2013-092186

Elbin, R. J., Kontos, A. P., Kegel, N., Johnson, E., Burkhart, S., & Schatz, P. (2013). Individual and combined effects of LD and ADHD on computerized neurocognitive concussion test performance: Evidence for separate norms. *Archives of Clinical Neuropsychology*. https://doi.org/10.1093/arclin/act024

Elbin, R. J., Schatz, P., & Covassin, T. (2011). One-year test-retest reliability of the online version of ImPACT in high school athletes. *The American Journal of Sports Medicine*, *39*(11), 2319–2324. https://doi.org/10.1177/0363546511417173

Elder, G. A., Mitsis, E. M., Ahlers, S. T., & Cristian, A. (2010). Blast-induced mild traumatic brain injury. *Psychiatric Clinics of North America*, *33*(4), 757–781. https://doi.org/10.1016/j.psc.2010.08.001

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. https://doi.org/10.4085/1062-6050-52.6.03

Gaudet, C. E., & Weyandt, L. L. (2017). 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. https://doi.org/10.1080/13854046.2016.1220622

Granholm, E., Link, P., Fish, S., Kraemer, H., & Jeste, D. (2010). Age-related practice effects across longitudinal neuropsychological assessments in older people with

schizophrenia. *Neuropsychology*, *24*(5), 616–624. https://doi.org/10.1037/a0019560

Harlow, L. (2014). *The Essence of Multivariate Thinking: Basic Themes and Methods* (2nd ed.). New York, NY: Routledge.

Heilbronner, R. L., Sweet, J. J., Attix, D. K., Krull, K. R., Henry, G. K., & Hart, R. P. (2010). Official position of the American Academy of Clinical Neuropsychology on serial neuropsychological assessments: The utility and challenges of repeat test administrations in clinical and forensic contexts. *Clinical Neuropsychologist*, *24*(8), 1267–1278. https://doi.org/10.1080/13854046.2010.526785

Henry, L. C., & Sandel, N. (2015). Adolescent subtest norms for the ImPACT neurocognitive battery. *Applied Neuropsychology: Child*, *4*(4), 266–276. https://doi.org/10.1080/21622965.2014.911094

Iverson, G. L., Lovell, M. R., & Collins, M. W. (2004). Interpreting change on ImPACT following sport concussion. *The Clinical Neuropsychologist*, *17*(4), 460–467. https://doi.org/10.1076/clin.17.4.460.27934

Iverson, G. L., Silverberg, N. D., Mannix, R., Maxwell, B. A., Atkins, J. E., Zafonte, R., & Berkner, P. D. (2015). Factors associated with concussion-like symptom reporting in high school athletes. *JAMA Pediatrics*, *169*(12), 1132–1140. https://doi.org/10.1001/jamapediatrics.2015.2374

Kamins, J., Bigler, E., Covassin, T., Henry, L., Kemp, S., Leddy, J. J., … Giza, C. C. (2017). What is the physiological time to recovery after concussion? A systematic review. *British Journal of Sports Medicine*, *51*(12), 935–940. https://doi.org/10.1136/bjsports-2016-097464

Kontos, A. P., Elbin, R. J., Covassin, T., & Larson, E. (2010). Exploring differences in computerized neurocognitive concussion testing between African American and white athletes. *Archives of Clinical Neuropsychology*, *25*(8), 734–744. https://doi.org/10.1093/arclin/acq068

Landers, R. (2015). Computing intraclass correlations (ICC) as estimates of interrater reliability in SPSS. *The Winnower*. https://doi.org/10.15200/winn.143518.81744

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

Larrabee, G. J. (2018). Psychometric foundations of neuropsychological assessment. In J. E. Morgan & J. H. Ricker (Eds.), *Textbook of Clinical Neuropsychology* (2nd ed., pp. 22–38). New York, NY: Taylor & Francis.

Lichtenstein, J. D., Moser, R. S., & Schatz, P. (2014). Age and test setting affect the prevalence of invalid baseline scores on neurocognitive tests. *American Journal of Sport*, *42*, 479–484. https://doi.org/10.1177/0363546513509225

Loring, D. W. (2015). *INS Dictionary of Neuropsychology and Clinical Neurosciences*. New York, NY: Oxford University Press.

Lovell, M. R. (2016). *ImPACT Administration and Interpretation Manual*. ImPACT Applications, Inc. Retrieved from www.impacttest.com

Lynall, R. C., Laudner, K. G., Mihalik, J. P., & Stanek, J. M. (2013). Concussion-assessment and -management techniques used by athletic trainers. *Journal of Athletic Training*, *48*(6), 844–850. https://doi.org/10.4085/1062-6050-48.6.04

Mayers, L. B., & Redick, T. S. (2012). Authors' reply to "Response to Mayers and Redick: 'Clinical utility of ImPACT assessment for postconcussion return-to-play counseling: Psychometric issues.'" *Journal of Clinical and Experimental Neuropsychology*, *34*(4), 435–442. https://doi.org/10.1080/13803395.2012.667790

McCrea, M., Barr, W. B., Guskiewicz, K., Randolph, C., Marshall, S. W., Cantu, R., … Kelly, J. P. (2005). Standard regression-based methods for measuring recovery after sport-related concussion. *Journal of the International Neuropsychological Society*, *11*(01), 58–69. https://doi.org/10.1017/s1355617705050083

McCrory, P., Meeuwisse, W., Aubry, M., Cantu, R. C., Dvořák, J., Echemendia, R. J., … Turner, M. (2013). Consensus statement on concussion in sport: The 4th international conference on concussion in sport, held in Zurich, November 2012. *Journal of Athletic Training*, *48*(4), 554–575. https://doi.org/10.4085/1062-6050-48.4.05

McCrory, P., Meeuwisse, W., Dvořák, J., Aubry, M., Bailes, J., Broglio, S., … Vos, P. E. (2017). 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*(11), 838–847. https://doi.org/10.1136/bjsports-2017-097699

Meehan, W. P., Collins, C. L., Taylor, A. M., & Dawn Comstock, R. (2012). Computerized neurocognitive testing for the management of sport-related concussions. *Pediatrics*, *129*(1), 38–44. https://doi.org/10.1542/peds.2011-1972

Moser, R. S., Davis, G. A., & Schatz, P. (2018). The age variable in childhood

concussion management: A systematic review. *Archives of Clinical Neuropsychology*, *33*(4), 417–426. https://doi.org/10.1093/arclin/acx070

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), 139–154.

Resch, J. E., Schneider, M. W., & Munro, C. (2018). The test-retest reliability of three computerized neurocognitive tests used in the assessment of sport concussion. *International Journal of Psychophysiology*, *132*(December 2016), 31–38. https://doi.org/10.1016/j.ijpsycho.2017.09.011

Riegler, K. E., Guty, E. T., & Arnett, P. A. (2018). Validity of the ImPACT Post-Concussion Symptom Scale (PCSS) affective symptom cluster as a screener for depression in collegiate athletes. *Archives of Clinical Neuropsychology*, *34*(4), 563–574. https://doi.org/10.1093/arclin/acy081

Riggio, S., & Jagoda, A. (2016). Concussion and its neurobehavioural sequelae. *International Review of Psychiatry*, *28*(6), 579–586. https://doi.org/10.1080/09540261.2016.1220927

Romeu-Mejia, R., Giza, C. C., & Goldman, J. T. (2019). Concussion pathophysiology and injury biomechanics. *Current Reviews in Musculoskeletal Medicine*, *12*(2), 105–116. https://doi.org/10.1007/s12178-019-09536-8

Rosselli, M., Ardila, A., Matute, E., & Vélez-Uribe, I. (2014). Language development across the life span: A neuropsychological/neuroimaging perspective. *Neuroscience Journal*. https://doi.org/10.1155/2014/585237

Russell, E. (2003). Toward an explanation of Dodrill's observation: High

neuropsychological test performance does not accompany high IQs. *The Clinical Neuropsychologist*, *15*(3), 423–428. https://doi.org/10.1076/clin.15.3.423.10267

Russell, E. W., & Russell, S. L. K. (2003). Twenty ways and more of diagnosing brain damage when there is none. *Journal of Controversial Medical Claims*, *10*(1), 1–14.

Sachs, B. C., Rush, B. K., & Pedraza, O. (2016). Validity and reliability of the NAB Naming Test. *Clinical Neuropsychologist*, *30*(4), 629–638. https://doi.org/10.1080/13854046.2016.1149618

Schatz, P., & Ferris, C. S. (2013). One-month test-retest reliability of the ImPACT test battery. *Archives of Clinical Neuropsychology*, *28*(5), 499–504. https://doi.org/10.1093/arclin/act034

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*(7), 625–632. https://doi.org/10.1093/arclin/acu041

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.86.2.420

Signoretti, S., Lazzarino, G., Tavazzi, B., & Vagnozzi, R. (2011). The pathophysiology of concussion. *PM & R*, *3*(10), 359–368. https://doi.org/10.1016/j.pmrj.2011.07.018

Strauss, E. H., Sherman, E. M. S., & Spreen, O. (2006). *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*. New York, NY: Oxford University Press.

Van Kampen, D. A., Lovell, M. R., Pardini, J. E., Collins, M. W., & Fu, F. H. (2006). The "value added" of neurocognitive testing after sports-related concussion. *American Journal of Sports Medicine*, *34*(10), 1630–1635. https://doi.org/10.1177/0363546506288677

Tables

Table 1. Demographic Characteristics ($N = 107$)

	N (%) / Mean (SD)	Range
Female	50 (46.7)	
Male	57 (53.3)	
Age at Baseline 1 (years)	14.5 (0.7)	13.1 – 17.0
Age at Baseline 2 (years)	16.3 (0.5)	15.1 – 18.0
Test-Retest Interval (months)	21.1 (4.9)	11.0 – 24.0

Table 2. Composite and Module Subscale Score Distribution Characteristics ($N = 107$)

	Baseline 1				Baseline 2			
Composite / Subscale	Mean (SD)	Median	Skewness	Kurtosis	Mean (SD)	Median	Skewness	Kurtosis
Verbal Memory	87.1 (9.1)	88.0	-0.37	-0.81	86.7 (10.5)	88.0	-0.70	-0.12
WM: Tot. % Cor.	95.4 (4.5)	96.0	-0.89	-0.34	94.5 (5.6)	96.0	-1.32	1.77
SM: Tot Cor. (hid.)	6.7 (2.1)	7.0	-0.72	-0.28	6.7 (2.1)	7.0	-0.86	0.22
3L: % Tot. Let. Cor.	91.2 (10.7)	93.33	-1.22	0.66	91.5 (11.9)	100.0	-1.46	1.58
Visual Memory	76.9 (12.1)	78.0	-0.52	-0.00	78.8 (13.9)	83.0	-1.00	1.3
DM: Tot. % Cor.	83.1 (11.1)	83.5	-0.32	-0.98	84.5 (13.2)	87.5	-0.95	0.28
XO: Tot Cor. (mem.)	8.5 (2.1)	9.0	-0.38	-0.19	8.8 (2.3)	9.0	-0.97	1.43
Visual Motor Speed	37.2 (6.4)	37.2	0.18	-0.14	41.0 (5.7)	41.0	-0.23	-0.43
XO: Tot. Cor. (int.)	111.6 (6.8)	112.0	-0.08	0.18	113.9 (6.8)	115.0	-0.58	1.00
3L: Avg. Ct. Cor.	15.5 (4.0)	15.6	0.24	-0.35	17.8 (3.5)	17.8	-0.23	-0.50
Reaction Time	0.6 (0.1)	0.6	0.35	0.12	0.8 (0.6)	0.6	0.58	0.56
XO: Avg. Cor. RT (int.)	0.5 (0.1)	0.5	0.34	0.21	0.5 (0.1)	0.5	0.60	0.32
SM: Avg. Cor. RT (vis.)	1.5 (0.4)	1.5	1.48	8.7	1.5 (0.4)	1.5	2.81	11.3
CM: Avg. Cor. RT	0.8 (0.1)	0.7	0.84	0.86	0.7 (0.1)	0.7	1.18	2.58

Note: WM: Tot. % Cor. = Word Memory – Total Percentage Correct score; SM: Tot. Cor. (hid.) = Symbol Match – Total Correct (hidden) score; 3L: % Tot. Let. Cor. = Three Letters – Percentage of Total Letters Correct score; DM: Tot. % Cor. = Design Memory – Total Percentage Correct score; XO: Tot. Cor. (mem.) = X's and O's – Total Correct (memory) score; XO: Tot. Cor. (int.) = X's and O's – Total Correct (interference) score; 3L: Avg. Ct. Cor. = Three Letters – Average Counted Correctly score; XO: Avg. Cor. RT (int.) = X's and O's – Average Correct RT (interference) score; SM: Avg. Cor. RT (vis.) = Symbol Match – Average Correct RT (visible) score; CM: Avg. Cor. RT = Color Match – Average Correct RT score.

Table 3. Test-Retest Reliabilities for the Total Sample ($N = 107$)

		Intraclass Correlation Coefficient (95% confidence interval)	
Composite / Subscale	Pearson *r*	One-Way Random	Two-Way Mixed
Verbal Memory	0.43	0.43 (0.26 – 0.57)	0.42 (0.25 – 0.57)
WM: Tot. % Cor.	0.54	0.52 (0.36 – 0.64)	0.53 (0.37 – 0.65)
SM: Tot Cor. Hid.	0.47	0.48 (0.32 – 0.61)	0.47 (0.31 – 0.61)
3L: % Tot. Let. Cor.	0.19	0.19 (0.00 – 0.36)	0.18 (-0.01 – 0.36)
Visual Memory	0.52	0.51 (0.35 – 0.64)	0.51 (0.36 – 0.64)
DM: Tot. % Cor.	0.56	0.55 (0.40 – 0.67)	0.55 (0.41 – 0.67)
XO: Tot Cor. (Mem.)	0.32	0.32 (0.14 – 0.48)	0.32 (0.14 – 0.48)
Visual Motor Speed	0.69	0.53 (0.38 – 0.66)	0.68 (0.57 – 0.77)
XO: Tot. Cor. (Int.)	0.59	0.54 (0.40 – 0.67)	0.59 (0.45 – 0.70)
3L: Avg. Ct. Cor.	0.68	0.53 (0.38 – 0.65)	0.67 (0.55 – 0.76)
Reaction Time	0.59	0.56 (0.41 – 0.67)	0.59 (0.45 – 0.70)
XO: Avg. Cor. RT (Int.)	0.68	0.62 (0.48 – 0.72)	0.68 (0.56 – 0.77)
SM: Avg. Cor. RT (Vis)	0.41	0.42 (0.25 – 0.56)	0.41 (0.24 – 0.56)
CM: Avg. Cor. RT	0.59	0.55 (0.41 – 0.67)	0.59 (0.45 – 0.70)

Note: WM: Tot. % Cor. = Word Memory – Total Percentage Correct score; SM: Tot. Cor. (hid.) = Symbol Match – Total Correct (hidden) score; 3L: % Tot. Let. Cor. = Three Letters – Percentage of Total Letters Correct score; DM: Tot. % Cor. = Design Memory – Total Percentage Correct score; XO: Tot. Cor. (mem.) = X's and O's – Total Correct (memory) score; XO: Tot. Cor. (int.) = X's and O's – Total Correct (interference) score; 3L: Avg. Ct. Cor. = Three Letters – Average Counted Correctly score; XO: Avg. Cor. RT (int.) = X's and O's – Average Correct RT (interference) score; SM: Avg. Cor. RT (vis.) = Symbol Match – Average Correct RT (visible) score; CM: Avg. Cor. RT = Color Match – Average Correct RT score.

Table 4. Test-Retest Reliabilities Stratified by Gender ($N = 107$)

	Female ($N = 50$)			Male ($N = 57$)		
		ICC (95% confidence interval)			ICC (95% confidence interval)	
Composite / Subscale	Pearson *r*	One-Way Random	Two-Way Mixed	Pearson *r*	One-Way Random	Two-Way Mixed
Verbal Memory	0.48	0.49 (0.24 – 0.67)	0.48 (0.24 - 0.67)	0.36	0.36 (0.11 – 0.56)	0.35 (0.10 – 0.56)
WM: Tot. % Cor.	0.61	0.60 (0.40 – 0.75)	0.61 (0.40 – 0.76)	0.49	0.47 (0.24 – 0.65)	0.47 (0.25 – 0.65)
SM: Tot Cor. Hid.	0.47	0.47 (0.23 – 0.66)	0.47 (0.22 – 0.66)	0.44	0.45 (0.22 – 0.63)	0.44 (0.21 – 0.63)
3L: % Tot. Let. Cor.	0.06	0.07 (-0.21 – 0.34)	0.06 (-0.22 – 0.33)	0.28	0.29 (0.03 – 0.51)	0.28 (0.02 – 0.50)
Visual Memory	0.54	0.52 (0.29 – 0.70)	0.53 (0.30 – 0.70)	0.48	0.48 (0.25 – 0.65)	0.47 (0.24 – 0.65)
DM: Tot. % Cor.	0.53	0.53 (0.30 – 0.70)	0.53 (0.29 – 0.70)	0.59	0.57 (0.36 – 0.72)	0.57 (0.37 – 0.73)
XO: Tot Cor. (Mem.)	0.42	0.40 (0.14 – 0.61)	0.41 (0.16 – 0.62)	0.21	0.22 (-0.04 – 0.45)	0.21 (-0.05 – 0.44)
Visual Motor Speed	0.62	0.43 (0.18 – 0.63)	0.61 (0.40 – 0.76)	0.75	0.62 (0.44 – 0.76)	0.74 (0.60 – 0.84)
XO: Tot. Cor. (Int.)	0.57	0.43 (0.18 – 0.63)	0.57 (0.34 – 0.73)	0.61	0.61 (0.42 – 0.75)	0.61 (0.42 – 0.75)
3L: Avg. Ct. Cor.	0.69	0.54 (0.31 – 0.71)	0.68 (0.50 – 0.81)	0.67	0.53 (0.32 – 0.69)	0.66 (0.49 – 0.79)
Reaction Time	0.68	0.64 (0.45 – 0.78)	0.67 (0.49 – 0.80)	0.53	0.50 (0.28 – 0.67)	0.53 (0.32 – 0.69)
XO: Avg. Cor. RT (Int.)	0.65	0.55 (0.33 – 0.72)	0.65 (0.45 – 0.78)	0.70	0.66 (0.49 – 0.79)	0.70 (0.54 – 0.81)
SM: Avg. Cor. RT (Vis)	0.68	0.67 (0.49 – 0.80)	0.67 (0.49 – 0.80)	0.19	0.20 (-0.07 – 0.43)	0.19 (-0.07 – 0.43)

CM: Avg. Cor. RT	0.61	0.60 (0.39 – 0.75)	0.61 (0.40 – 0.76)	0.59	0.52 (0.31 – 0.69)	0.58 (0.38 – 0.73)

Note: ICC = Intraclass Correlation Coefficient; WM: Tot. % Cor. = Word Memory – Total Percentage Correct score; SM: Tot. Cor. (hid.) = Symbol Match – Total Correct (hidden) score; 3L: % Tot. Let. Cor. = Three Letters – Percentage of Total Letters Correct score; DM: Tot. % Cor. = Design Memory – Total Percentage Correct score; XO: Tot. Cor. (mem.) = X's and O's – Total Correct (memory) score; XO: Tot. Cor. (int.) = X's and O's – Total Correct (interference) score; 3L: Avg. Ct. Cor. = Three Letters – Average Counted Correctly score; XO: Avg. Cor. RT (int.) = X's and O's – Average Correct RT (interference) score; SM: Avg. Cor. RT (vis.) = Symbol Match – Average Correct RT (visible) score; CM: Avg. Cor. RT = Color Match – Average Correct RT score.

Table 5. Test-Retest Reliabilities Stratified by Age ($N = 107$)

	Age < 15 ($N = 67$)			Age ≥ 15 ($N = 40$)		
		ICC (95% confidence interval)			ICC (95% confidence interval)	
Composite / Subscale	Pearson *r*	One-Way Random	Two-Way Mixed	Pearson *r*	One-Way Random	Two-Way Mixed
Verbal Memory	0.47	0.47 (0.26 – 0.64)	0.46 (0.25 – 0.63)	0.37	0.37 (0.07 – 0.61)	0.37 (0.07 – 0.61)
WM: Tot. % Cor.	0.51	0.51 (0.31 – 0.67)	0.51 (0.30 – 0.66)	0.59	0.53 (0.26 – 0.72)	0.55 (0.30 – 0.74)
SM: Tot Cor. Hid.	0.61	0.62 (0.44 – 0.74)	0.61 (0.43 – 0.74)	0.23	0.24 (-0.08 – 0.51)	0.23 (-0.09 – 0.50)
3L: % Tot. Let. Cor.	0.10	0.11 (-0.13 – 0.34)	0.10 (-0.14 – 0.33)	0.33	0.34 (0.04 – 0.58)	0.33 (0.02 – 0.58)
Visual Memory	0.45	0.44 (0.23 – 0.61)	0.45 (0.24 – 0.62)	0.61	0.60 (0.37 – 0.77)	0.60 (0.35 – 0.76)
DM: Tot. % Cor.	0.46	0.45 (0.24 – 0.62)	0.45 (0.24 – 0.62)	0.72	0.71 (0.52 – 0.83)	0.71 (0.52 – 0.84)
XO: Tot Cor. (Mem.)	0.34	0.33 (0.10 – 0.53)	0.34 (0.11 – 0.54)	0.31	0.32 (0.11 – 0.57)	0.31 (0.00 – 0.56)
Visual Motor Speed	0.65	0.41 (0.19 – 0.59)	0.65 (0.48 – 0.77)	0.78	0.73 (0.55 – 0.85)	0.77 (0.61 – 0.87)
XO: Tot. Cor. (Int.)	0.54	0.48 (0.28 – 0.65)	0.53 (0.34 – 0.68)	0.68	0.65 (0.43 – 0.80)	0.67 (0.46 – 0.81)
3L: Avg. Ct. Cor.	0.64	0.41 (0.19 – 0.59)	0.64 (0.47 – 0.76)	0.76	0.72 (0.53 – 0.84)	0.76 (0.58 – 0.86)
Reaction Time	0.59	0.56 (0.37 – 0.70)	0.59 (0.41 – 0.73)	0.58	0.56 (0.30 – 0.74)	0.58 (0.33 – 0.75)
XO: Avg. Cor. RT (Int.)	0.61	0.57 (0.38 – 0.71)	0.61 (0.44 - 0.74)	0.79	0.70 (0.51 – 0.83)	0.79 (0.64 – 0.88)

SM: Avg. Cor. RT (Vis)	0.47	0.47 (0.26 – 0.64)	0.46 (0.25 – 0.63)	0.28	0.25 (-0.06 – 0.52)	0.25 (-0.07 – 0.51)
CM: Avg. Cor. RT	0.67	0.54 (0.35 – 0.69)	0.62 (0.45 – 0.75)	0.55	0.55 (0.30 – 0.73)	0.55 (0.29 – 0.73)

Note: ICC = Intraclass Correlation Coefficient; WM: Tot. % Cor. = Word Memory – Total Percentage Correct score; SM: Tot. Cor. (hid.) = Symbol Match – Total Correct (hidden) score; 3L: % Tot. Let. Cor. = Three Letters – Percentage of Total Letters Correct score; DM: Tot. % Cor. = Design Memory – Total Percentage Correct score; XO: Tot. Cor. (mem.) = X's and O's – Total Correct (memory) score; XO: Tot. Cor. (int.) = X's and O's – Total Correct (interference) score; 3L: Avg. Ct. Cor. = Three Letters – Average Counted Correctly score; XO: Avg. Cor. RT (int.) = X's and O's – Average Correct RT (interference) score; SM: Avg. Cor. RT (vis.) = Symbol Match – Average Correct RT (visible) score; CM: Avg. Cor. RT = Color Match – Average Correct RT score.

Table 6. Number of Individuals Obtaining Scores at or Near the Maximum Possible Score (*N*=107)

	Time 1		Time 2	
Subscale	100%	>90%	100%	>90%
Word Memory				
Learning percent correct (%)	71 (66.4)	105 (98.1)	61 (57)	99 (92.5)
Delayed memory percent correct (%)	29 (27.1)	76 (71.0)	30 (28)	73 (68.2)
Total percent correct (%)	27 (25.2)	87 (81.3)	25 (23.4)	82 (76.6)
Design Memory				
Learning percent correct (%)	11 (10.3)	49 (45.8)	18 (16.8)	56 (52.3)
Delayed memory percent correct (%)	9 (8.4)	27 (25.2)	16 (15.0)	44 (41.1)
Total percent correct (%)	5 (4.7)	33 (30.8)	8 (7.5)	42 (39.3)
X's and O's				
Total correct (memory)	9 (8.4)	18 (16.8)	8 (7.5)	28 (26.2)
Symbol Match				
Total correct (visible)	96 (89.7)	104 (97.2)	99 (92.5)	106 (99)
Total correct (hidden)	29 (27.1)	29 (27.1)	28 (26.2)	28 (26.2)
Three Letters				
Percent of total letters correct	48 (44.9)	68 (63.6)	58 (54.2)	68 (63.6)

CHAPTER TWO

An examination of data integration approaches using ImPACT to aid in the detection of concussion symptomatology

by

Charles E. Gaudet[1]

To be submitted to the *Journal of the International Neuropsychological Society* (or a comparable outlet)

[1]PhD Candidate, Department of Psychology, The University of Rhode Island, Kingston, RI 02881. Email: chad_gaudet@my.uri.edu

Abstract

Background: Assessment and management of concussion are increasingly attracting public attention. Immediate Post Concussion Assessment and Cognitive Testing (ImPACT) is a widely adopted serial measure designed to aid in this pursuit. As such, the measure's ability to accurately differentiate between individuals with and without concussion is a fundamental component of its clinical utility. This study aimed to evaluate ImPACT's classification accuracy and potential approaches to improve it.

Methods: A retrospective file review was conducted for 300 consecutively selected, ImPACT score reports generated between 2010-2015 by individuals attending a secondary school with testing conducted at multiple time points. To appraise ImPACT's utility as a serialized measure, standard regression-based equations were formulated to compute reliable change index (RCI) scores. We conducted discriminant function analyses (DFAs) consisting of varying combinations of ImPACT composite scores and compared their accuracy to that produced by the standard interpretive procedure.

Results: The sample included 81 individuals who sustained a concussion following initial baseline testing and 129 individuals that did not. The DFAs yielded sensitivities ranging from 31% to 49%, specificities from 88% to 95%, positive predictive values (PPVs) from 61% to 83%, and negative predictive values (NPVs) from 67% to 75%. Conversely, the standard interpretive procedure yielded a sensitivity of 73%, specificity of 43%, PPV of 45%, and NPV of 72%.

Conclusion: The standard interpretive procedure produced a higher sensitivity rate than the DFAs; however, its PPV did not exceed chance levels. Conversely, the DFA

equations produced superior PPVs; however, their sensitivity hovered around 50%, leaving a substantial proportion of individuals with concussion undetected. A range of base rate conditions and psychometric factors appear to contribute to ImPACT's limited classification accuracy and resultant reduced clinical utility.

Keywords: ImPACT; concussion; validity; serial assessment; regression; neurocognitive testing; baseline testing, computer baseline neurocognitive testing (additional thoughts for key words)

Introduction

As the epidemic of mild head injury has attracted increasing national attention, the challenges healthcare providers face in curbing the adverse effects of such injuries are becoming increasingly apparent. Approximately 100 to 300 per 100,000 people annually are estimated to seek medical attention for concussion or mild traumatic brain injury (mTBI) worldwide; however, given the frequency with which individuals do not seek attention for mTBI or care is inaccessible, that figure may rise to as many as 600 per 100,000 (Cassidy et al., 2004), or a total of about 50 million. Almost by definition, individuals who sustain such injury experience initial symptoms. These symptoms can take a considerable short-term toll on a person's welfare. Furthermore, even if only a small percentage of these individuals experience persistent sequelae (Schretlen & Shapiro, 2003), such as 5%, it still leaves tens of thousands of individuals adversely affected and many more potentially at increased susceptibility to subsequent head injuries. Consequently, improved assessment and management of mild head injury can prove highly beneficial, not merely financially, but particularly in terms of human welfare. To this end, neuropsychology has demonstrated value in concussion and mTBI assessment and management.

As a foundation for the materials that follow, it is necessary to address the varying nomenclature commonly applied to mild head injuries, specifically, variations in the use of the terms *mild traumatic brain injury* and *concussion*. As Laker (2011) stated: "Typical methodological limitations in most studies of mTBI and concussion have included variability in definition and ascertainment methods (e.g., self-report, discharge diagnosis on retrospective review, and survey)" (p. S354). Some

practitioners and researchers use the terms concussion and mTBI interchangeably. Others place these descriptors along a spectrum of severity, in which concussion would be considered less severe than mTBI, with distinguishing features related to such factors as the presence or absence of loss of consciousness (LOC) or presumed underlying biological features (Laker, 2011). If the scientific literature does not provide detail on the selection and application of terminology, one can at least be explicit to reduce possible ambiguity. The current work, and the majority of literature cited in this study, refers to concussion. Hence, we will use the term *concussion*, unless the literature specifically refers to *mTBI,* or injury characteristics more consistent with mTBI.

Neuropsychology occupies an important role in the assessment and management of concussion, with sport-related concussion (SRC) garnering substantial interest (Lemonda, Tam, Barr, & Rabin, 2017). In the past 30 years, numerous consensus statements, position papers, and guidelines have been put forth. These statements generally endorse the inclusion of neuropsychological evaluation in some form in the assessment and management of concussion (Echemendia et al., 2013; Lemonda et al., 2017; McCrory et al., 2013, 2017). However, the extent to which neuropsychological evaluation is needed to ensure accurate and efficient assessment of concussion remains a subject of debate (Echemendia et al., 2013; Lemonda et al., 2017; McCrory et al., 2017).

In addition to testing, neuropsychological evaluation can include other sources of data gathering, such as interviews, record review, and neuroimaging. The clinician evaluates data from multiple sources to formulate judgments about such matters as

diagnosis, treatment, or forensic issues. In so doing, a critical but sometimes insufficiently appreciated step involves appraising the quality of the data collected. Ultimately, conclusions based on data derived from neuropsychological testing are only as sound as these measures' psychometric qualities, in particular, their reliability and validity. For a test to function as intended, for example, to detect the presence or absence of injury, it is paramount that it evidences adequate reliability and validity in its intended application and population of use.

As the neuropsychological assessment of concussion has advanced, shorter, computerized neurocognitive tests (CNTs) have been developed and widely adopted (Lynall, Laudner, Mihalik, & Stanek, 2013); Immediate Post Concussion and Cognitive Testing (ImPACT) is currently the most prominent CNT used in concussion assessment by a wide margin (Covassin, Elbin, & Stiller-Ostrowski, 2009; Lynall et al., 2013; Meehan, d'Hemecourt, Collins, Taylor, & Dawn Comstock, 2012). Of note, a recent survey of neuropsychologists that routinely assess patients with concussion revealed that only approximately 30% include CNTs in their battery, however, of those, 52% reportedly use ImPACT, with Automated Neuropsychological Assessment Metrics (ANAM) and Central Nervous System (CNS) Vital Signs occupying the next highest percentages, at 10% each (Lemonda et al., 2017).

ImPACT is designed to function as a serial, or longitudinal, measure in which examinees undergo assessment when healthy (i.e., baseline) and again following suspected concussion (i.e., post-injury). The term *baseline assessment* refers to an ImPACT administration conducted when individuals are presumed healthy, although this does not account for the possibility of a history of head injuries or other premorbid

factors that might affect performance on testing. Stated differently, *baseline assessment* is intended to capture individuals' cognitive functioning at a time point when they are functioning as they typically do in day-to-day activities. The measure consists of six test modules that assess aspects of attention, processing speed, reaction time, and memory. Subscale scores within these modules are combined to form four cognitive composite scores. Additionally, the measure includes the Post Concussion Symptom Scale (PCSS), which contains self-report items for 22 symptoms commonly associated with concussion rated along a 7-point scale (0-6).

Decrements between baseline and post-injury performance, in the absence of mitigating factors, are generally ascribed to ongoing cognitive deficits or symptoms associated with the injury (Lovell, 2016). To qualify as a reliable change, a score on a given composite must exceed a predetermined 80% confidence interval in a negative direction. For example, if a given baseline score is accompanied by a 10-point change interval, an individual obtaining a score on post-injury testing that is more than 10 points lower than the baseline score would demonstrate a "reliable" change. If an individual obtained scores that are within the confidence interval (i.e., within 10 points of the baseline score following the example cited above), this performance would be interpreted to reflect a return to the pre-injury level of cognitive functioning and likely indicative of the resolution of objective cognitive deficits.

Reliability generally refers to the consistency of a measure, and there are multiple forms of reliability, such as test-retest, internal, and inter-rater (American Educational Research Association, 2014). Test-retest reliability, which refers to the consistency of measurement over time, is most relevant to appraising a measure's

ability to detect changes when used in a serialized manner. To this end, the test-retest reliability coefficient is a core component of reliable change formulas that are intended to characterize the extent of differences between multiple test performances quantitively.

Multiple approaches have been developed to assess reliable change (i.e., is the difference in observed scores attributable to an actual change in function, as opposed to a chance occurrence or the psychometric properties of a measure?). Formal methods of appraising reliable change subscribe to one of two methods (with an overlap in some cases) – reliable change intervals and standardized regression-based (SRB) formulas that result in reliable change index (RCI) scores (for review see Duff, 2012). The reliable change interval method was originally developed to measure the effects of intervention in psychotherapy (Jacobson & Truax, 1991). In its most basic form, it is equal to the score at Time 2 minus the score at Time 1 divided by the standard error of the difference (Duff, 2012); it has been adapted to account for factors unique to neurocognitive testing such as practice effects (Chelune, Naugle, Lüders, Sedlak, & Awad, 1993) and standard error of the difference at Time 2 (Iverson, 2001). The interval approach relies on a predetermined symmetrical range surrounding a score at Time 1 as a criterion for determining whether performance at Time 2 represents a change (i.e., falls outside of the interval) or is unchanged (i.e., falls within the interval).

The SRB method performs a similar function. However, it is modeled using a regression equation to predict the Time 2 score using the Time 1 score (Crawford & Garthwaite, 2007; Crawford, Garthwaite, Denham, & Chelune, 2012; Hinton-Bayre,

2016; McSweeny, Naugle, Chelune, & Luders, 1993). The predicted score is then compared to the observed score to produce an RCI score indicative of the degree of change. The extent to which an RCI score deviates from expectation guides clinicians' interpretations regarding meaningful change. A unique advantage of the SRB method is that it allows for the incorporation of additional variables into the regression equation that might possess predictive utility such as age, gender, or race (Crawford et al., 2012; Duff, 2012). Hinton-Bayre's (2016) recent study, which examined the responsiveness of reliable change detection methods using the Wechsler Memory Scale-Fourth Edition (WMS-IV), found that the SRB method, as compared to reliable change interval approaches, was the most responsive to negative change. Superior sensitivity seemed, primarily attributable to its deliberate incorporation of the test-retest reliability coefficient into the regression equation, resulting in substantial corrections made for regression to the mean effects.

Given the extensive research documenting ImPACT's less than desirable test-retest reliability (Alsalaheen, Stockdale, Pechumer, & Broglio, 2016a; Broglio et al., 2018; Farnsworth, Dargo, Ragan, & Kang, 2017; Resch, Schneider, & Munro, 2018), failure to adjust for its effects will likely adversely affect classification accuracy. Evidence of poor classification accuracy stems from the high false-positive rate observed in a recent systematic review (Alsalaheen et al., 2016a). As such, an SRB approach appears appropriate to further our understanding of the interplay between test-retest reliability and classification accuracy.

Although ImPACT's Administration Manual cautions against its use as a stand-alone measure for diagnostic purposes, it does purport to aid clinicians in

determining an individual's concussion status and guiding future treatment strategy (Lovell, 2016). There is also an extensive literature showing that even single pieces of information or cues, especially salient ones, can have a surprisingly robust influence on conclusions despite what might be subjective impressions to the contrary (Arkes, González-Vallejo, Bonham, Kung, & Bailey, 2009; Faust, 1989; Ruscio, 2003). Given this context, a thorough understanding of ImPACT's efficacy in differentiating between individuals with and without concussion based on deviations from baseline on post-injury assessment is fundamental to implementing this measure in clinical practice.

Classification accuracy includes the use of several terms to characterize specific aspects of a test's accuracy that the public frequently misunderstands, as clinicians may as well (Gigerenzer, 2007; Labarge, Mccaffrey, & Brown, 2003). Sensitivity refers to the probability that a sign will be positive given that the disorder is present; specificity refers to the probability that a sign will be negative given that the disorder is not present; positive predictive value (PPV) refers to the likelihood that an individual has a disorder given a positive sign; and negative predictive value (NPV) refers to the likelihood that an individual does not have a disorder give a negative sign. PPV and NPV are contingent on the base rate of the disorder within the population under examination (and their calculation depend on knowing or estimating that base rate), whereas sensitivity and specificity are not.

Labarge et al. (2003) found that less than 65% of practitioners correctly answered a question pertaining to PPV, underscoring the persistent misunderstanding and poor application of base rate information. Despite, arguably, the seminal work on

this topic having been published about 65 years ago (Meehl & Rosen, 1955). The underutilization of base rate data can severely compromise diagnostic and predictive accuracy and continues to plague clinical neuropsychology.

For example, in the context of concussion, consider the case of a construction manager responsible for the safety of others. He hopes to return to work eight days after the injury and undergoes an evaluation on the eighth day. Assuming a base rate of approximately 35% of individuals remain impaired at 8 days post-injury (Nelson et al., 2016), clinicians would be correct 65% of the time if they concluded the manager was no longer impaired. Given this base rate, if clinicians administered a test with a sensitivity of 90% and specificity of 80%, the test would yield an approximate PPV of 70% and NPV of 95%. Hence, the predictive accuracy of this measure would exceed the accuracy of reliance on the base rate, thereby improving clinicians' decision-making in this instance.

Conversely, consider an alternative test with a sensitivity of 60% and specificity of 50%. Given an identical base rate of impairment, 35%, the approximate PPV declines to 40% and NPV to 70%. In this instance, clinicians' decision-making becomes less clear, as a positive result on the test is essentially no better than chance, although a negative result improves accuracy relative to the base rate by nearly 5%. As such, the incremental utility of incorporating this alternative test declines along with clinicians' confidence in clearing the construction worker to return to work in the presence of a positive test result.

In appraising a measure's clinical utility, base rate considerations are imperative. Consider a rare neurological condition such as Multiple System Atrophy

(MSA) with an estimated base rate of less than 1% (National Institute of Neurological Disorders and Stroke, 2014). For a test to improve clinicians' decision-making accuracy, it would need to have nearly perfect predictive accuracy. Moreover, using the sensitivity and specificity figures provided in the former example provided above (90% & 80%), one may observe how base rates substantially influence the accuracy of test results. If a base rate is as low as 5%, the PPV is 19%, and NPV is 99%; if the base rate is 50%, the PPV is 82%, and NPV is 89%; and if the base rate is 95%, the PPV is 99% and the NPV is 8%. In the context of concussion, base rate consideration becomes critically important as one appraises the presence or absence of symptoms relative to the time since injury as the base rates decline as the temporal proximity from the injury lengthens (McCrea et al., 2003; Nelson et al., 2016).

A recent review identified four studies that examined ImPACT's diagnostic accuracy (Alsalaheen, Stockdale, Pechumer, & Broglio, 2016b). When evaluating the measure, ImPACT's sensitivity in detecting concussed individuals ranged from 62.5% to 83.0%. Unfortunately, not all studies reported additional accuracy indicators such as specificity, PPV, and NPV. Varying a priori criteria for concussion groups further obscured comparisons among studies. As such, these four studies warrant a more thorough individual examination.

Broglio, Macciocchi, and Ferrara (2007) evaluated 24 Division 1 athletes who had recently sustained a physician-diagnosed concussion (specifics of diagnostic method not reported) within 24 hours of injury using ImPACT; the study did not include a healthy control group. The criterion for determining the presence of concussion symptomatology on ImPACT was at least one composite score exceeding

the 80% reliable change interval. Cognitive composite scores (Verbal Memory, Visual Memory, Visual Motor Speed, and Reaction Time) subjected to this criterion produced an overall sensitivity of 62.5%. When including the PCSS, along with the aforementioned composite scores, this figure increased to 79.2%. Interestingly, the PCSS's sensitivity, independent of the cognitive composite scores, was also 62.5%, suggesting that the inclusion of cognitive measures enhanced sensitivity by nearly 20% (Broglio, Macciocchi, & Ferrara, 2007). The absence of a control group and consideration of base rates severely limit the generalizability of this study and conclusions regarding ImPACT's classification accuracy in real-world settings.

Gardner, Shores, Batchelor, and Honan (2012) compared ImPACT performances between a sample of individuals with and without concussion. Those with concussion completed post-injury assessment within 72 hours of sustaining the injury. Given the study's cross-sectional design, or absence of serial assessment data, it was limited to providing information about ImPACT's utility in comparing those with and without concussion at one-time point, as opposed to the measure's intended use, which is to compare individuals' performance on post-injury testing to their performance on baseline testing. Additionally, this study did not examine ImPACT's standard interpretive procedure – that is, the production of one or more deviant composite scores relative to an individual's baseline performance. With this caveat in mind, ImPACT composite scores were entered in step two of a hierarchical logistic regression model and did not enhance classification accuracy beyond the following grouping of variables: age, Wechsler Test of Adult Reading (WTAR) score, Wechsler Adult Intelligence Scale – 3rd Edition (WAIS-III) Processing Speed Index (PSI), and

number of previous concussions. However, the study did find that the PCSS provided incremental utility when added to the demographic variables and in the absence of cognitive measures (Gardner, Shores, Batchelor, & Honan, 2012).

Van Kampen, Lovell, Pardini, Collins, and Fu (2006), unlike the prior two studies, came closer to applying ImPACT in lines with its designers' intent as a serialized measure. This study included a sample of 122 individuals with concussion (tested within two days of injury) and 70 without; both groups had completed baseline assessments. The criterion for the concussion group was a physician or athletic trainer's on-field diagnosis at the time of injury. Using a deviation exceeding the 80% RCI on the cognitive composite scores or PCSS as indicative of the presence of concussive symptomatology, the following classification accuracy statistics resulted: sensitivity = 93%, specificity = 70%, PPV = 84%, and NPV = 85%, assuming prior probability was equal to the distribution of the sample (64% concussion & 36% healthy). Additionally, the PPV of the PCSS in isolation was 93%; however, the NPV was 59%. The authors concluded that cognitive testing added value above and beyond the interpretation of the PCSS (Van Kampen, Lovell, Pardini, Collins, & Fu, 2006).

Schatz, Pardini, Lovell, Collins, and Podell (2006) used discriminant function analysis (DFA) to evaluate ImPACT's diagnostic accuracy. This study included a sample of 72 individuals with concussion (tested within 72 hours of injury) and 66 without extracted from a larger sample of approximately 1,500 individuals; both groups had completed baseline assessments. The criterion for the concussion group was an on-field diagnosis by a physician or athletic trainer at the time of injury. The authors conducted a stepwise DFA consisting of cognitive composites, impulse

control, and PCSS scores. The DFA yielded an equation that included the PCSS, Processing Speed, Visual Memory, and Impulse Control scores and correctly classified 85.5% of individuals overall. Accuracy for specific parameters were as follows: sensitivity = 81.9%, specificity = 89.4%, PPV = 89.4%, and NPV = 81.9%, assuming prior probability was equal to the distribution of the sample (52% concussion, 48% healthy). Given that the stepwise DFA selected only two of the cognitive composite scores, the authors suggested there may be a high degree of shared variance among them, thereby reducing the incremental utility of these indicators (Schatz, Pardini, Lovell, Collins, & Podell, 2006).

An additional study that was not included in the Alsalaheen et al. (2016b) review examined the diagnostic accuracy of ImPACT at multiple time points post-injury (Nelson et al., 2016). This study consisted of 166 individuals with concussion, as diagnosed by U.S. Department of Defense criteria, and 166 healthy individuals. The sample was examined at 24-hours, eight days, and 15 days post-injury using a criterion of one or more composite scores exceeding the 80% reliable change interval. The following base rates were assumed at each time point respectively, 89.4%, 35.4%, and 14.8%; clinical interview and symptom reporting on the Sport Concussion Assessment Tool-3 (SCAT3) determined base rates on concussion at the varying time points. Computations yielded the following predictive values (PPV/NPV) at the three respective time points: 91.7%/15.6%, 50.7%/80.9%, and 20.1%/90.3%. These results indicated that ImPACT's utility in detecting cognitive changes associated with concussion substantially declines once an individual reaches 8 days post-injury.

In addition to research explicitly examining diagnostic accuracy, test-retest reliability research provides additional insight into ImPACT's discriminative utility as a measure's reliability is often a prerequisite for determining its validity. Specifically, in a recent systematic review, the proportion of healthy participants producing one or more reliable changes across cognitive composite scores was aggregated; rates ranged from 22% to 46%. (Alsalaheen et al., 2016a). These data suggest a high percentage of false-positive errors when examining healthy groups in isolation.

Taken together, several critical questions regarding ImPACT's clinical utility remain unanswered or in need of further investigation. Of the literature reviewed, only two studies evaluated ImPACT's discriminative utility in a manner consistent with its intended purpose – serial assessment, or by comparing an individual's baseline and post-injury scores to determine whether there had been a change in neuropsychological status (Nelson et al., 2016; Van Kampen et al., 2006). Furthermore, optimal methods for clinical interpretation warrant further consideration as well. Of the studies that most closely resemble ImPACT's implementation in real-world settings, Van Kampen et al. (2006) and Nelson et al. (2016) examined the standard interpretive procedure, that is, identifying the presence of concussion based on a reliable change in one or more composite scores. Alternatively, Schatz et al. (2006) conducted a DFA to determine an efficient combination of variables to attain optimal accuracy. The results of these studies were mostly equivalent, as evidenced by overall accuracy rates in the 80% range.

The present study, which included exploratory elements, aimed to advance the current state of research involving ImPACT's diagnostic accuracy. First, the study

evaluated a novel procedure for detecting reliable change based on a regression-based method and compared it to the predetermined 80% reliable change interval standard. Second, the study used DFAs to compare data combinations, using SRB RCI scores, to determine whether this approach will improve accuracy relative to the current standard of one or more scores exceeding a predetermined 80% reliable change interval.

Methods

Participants

For a full description of the study sample's characteristics, please refer to Chapter 1, pp. 16-17.

Analyses included only individuals who completed multiple baseline assessments or sustained a concussion following the initial baseline. Analyses also included individuals with a reported history of concussion, ADHD, learning disability, individuals born outside of the U.S., and individuals for whom English was not the first language. Analyses did not include individuals that produced invalid baseline results according to embedded ImPACT validity criteria defined in the test manual (Lovell, 2016); although evidence for the sensitivity of these criteria is underdeveloped (Gaudet & Weyandt, 2017). Additionally, for the concussion group, analyses were restricted to individuals who underwent post-injury testing within six days of the injury.

Assessment

ImPACT protocols included either version 2.0 or 2.1. Version 2.1 provides data integration with version 1.0; subscale and composite scores are equivalent between versions. ImPACT includes six cognitive performance modules: Word

Memory, Design Memory, X's and O's, Symbol Match, Color Match, and Three Letters. Individual scores from these modules are combined to form five composite scores: Verbal Memory, Visual Memory, Visual Motor Speed, Reaction Time, and Impulse Control. Of note, the Impulse Control composite examines protocol validity rather than cognitive change (Lovell, 2016). The test also includes the PCSS. This self-report measure solicits examinee ratings using a seven-point scale (0-6) for symptoms commonly associated with concussion (e.g., headache, nausea, irritability). The inventory queries 22 total symptoms.

Statistical Analyses

Demographics: Chi-square and t-tests were used to examine differences between the concussion and healthy groups. Individuals diagnosed with concussion at some point following baseline assessment comprised the "concussion group." Individuals who did not sustain a reported concussion between baseline assessments comprised the "healthy group." Performances on ImPACT composite scores between the two groups were evaluated using a multivariate analysis of variance (MANOVA) model.

Calculation of SRB RCI scores: We used the procedures outlined by Duff (2012) to form SRB equations and calculate RCI scores. The calculation of RCI scores required the following data derived from a healthy population: Mean and standard deviation values at Times 1 and 2 and a test-retest correlation coefficient. This procedure consisted of generating a regression equation to produce a predicted score on testing at Time 2 based on test-retest data from a sample of healthy individuals. The predicted score is then subtracted from the observed, or actual score at Time 2 and

divided by the standard error of the estimate to produce an RCI score. The RCI score serves as a standardized indicator of the extent to which the observed score at Time 2 deviates from the predicted score; the smaller the score, the less deviant it is from the predicted score at Time 2.

We used Broglio, Katz, et al.'s (2018) test-retest reliability statistics for the four cognitive composite scores. These investigators tested a sample of over 3,100 participants across a 1-year test-retest interval; this study did not report test-retest reliability for the PCSS. Resultantly, for the PCSS, data computed from a sample of 56 participants using a 6-day test-retest interval were used (Iverson et al., 2004); these data were then used to calculate the reliable change intervals in the original ImPACT program. Supplemental Table 1 displays the data inputs for the regression and RCI equations. RCI scores between groups were evaluated using a MANOVA model.

Discriminant function analyses (DFAs): A series of DFAs appraised ImPACT's utility in differentiating between those with and without concussion. These DFAs evaluated alternative approaches to combining composite scores to improve classification accuracy. Additionally, the DFAs appraised the incremental value of including all composite scores in interpreting performance and discriminating between those with and without concussion. Classification accuracy metrics, including sensitivity, specificity, PPV, NPV, and overall classification accuracy, were computed and used as evaluative criteria.

We conducted and analyzed four DFAs. The first was a stepwise DFA that included the four cognitive composite scores and PCSS. Scores were combined to minimize the Wilks lambda statistic. The second DFA evaluated the PCSS in isolation,

as this is the only composite score based on a self-report of neuropsychological symptoms. The third DFA included only the cognitive-based composite scores (Verbal Memory, Visual Memory, Visual Motor Speed, and Reaction Time). The fourth DFA included all five composite scores – including cognitive scores and symptom report (PCSS).

Standard interpretive procedure: The current procedure for determining whether an individual's performance has significantly declined from baseline involves a reliable change interval. In accordance with the ImPACT Administration Manual, data derived from standardization and test-retest reliability allow for the calculation of a symmetrical interval surrounding the baseline score (Iverson et al., 2004; Lovell, 2016). When a composite score on post-injury testing falls outside of the interval, the composite score is presented in bold red font, "indicating that this score exceeds the RCI [reliable change interval] when compared to the baseline score" (Lovell, 2016, p. 45). These reliable change intervals were applied to baseline data for the present sample to determine the classification accuracy in differentiating between those with and without concussion. An a priori criterion of one or more composite score exceeding the 80% reliable change interval was used to identify the presence of possible concussive symptomatology.

Alpha levels of statistical significance were set at 0.05. Statistical analyses were conducted using SPSS version 23 and Microsoft Excel 365.

Results

Demographics

ImPACT score reports from 300 individuals were collected and analyzed. A check on exclusionary criteria yielded the following results: 78 (26%) included only an initial baseline assessment and no follow-up in the form of either a second baseline or post-injury assessment; one (0.3%) case met criteria for an invalid baseline performance as outlined in the Administration Manual (Lovell, 2016); and 11 (3.6%) individuals in the concussion group did not undergo assessment within 6 days of injury. Consequently, analyses included 210 individuals. The healthy group included 129 individuals, and the concussion group included 81 individuals. A physician or certified athletic trainer determined concussion diagnoses before the administration of post-injury ImPACT assessment. For the concussion group, measures of central tendency of time from injury to post-injury ImPACT assessment were as follows (in days): mean = 2.1, median = 2.0, standard deviation (SD) = 1.1, range = 1-6.

Table 1 displays differences between the healthy and concussion groups along demographic variables. As can be seen by examining this table, the concussion group was older than the healthy group at baseline, $t(208) = 2.1$, $p = 0.03$; however, the healthy group was older than the concussion group on follow-up testing, $t(208) = 4.1$, $p < 0.01$ (second baseline for the healthy group; post-injury assessment for the concussion group). Additionally, there was a higher proportion of females in the concussion group than the healthy group (59.3% to 43.4%). The remaining demographic variables did not produce statistically significant differences: U.S. born, English as the first language, Attention Deficit/Hyperactivity Disorder, and learning disability status ($p > 0.05$). Additionally, the test-retest interval for the time between

either obtaining two baselines for the healthy group or a baseline and post-injury test for the concussion group was not statistically significant between groups.

Table 2 displays ImPACT composite scores at baseline. A MANOVA revealed no statistically significant differences between concussion and healthy groups, Pillai's trace was 0.02, with $F(5, 204) = 0.72$ and $p = 0.61$. The largest Cohen's d effect size was 0.18 for the PCSS, suggesting that the concussion group reported higher levels of concussion symptoms at baseline.

Table 3 displays differences on follow-up testing. A MANOVA revealed statistically significant differences between groups across all ImPACT composite scores at Time 2, which was either a second baseline for the healthy group or post-injury assessment for the concussion group; Pillai's trace was 0.24, with $F(5,204) = 12.81$, $p < 0.01$. There was a large effect size, partial $\eta^2 = 0.24$ (Cohen, 1992). Follow-up ANOVAs revealed small to large effect sizes, ranging from Verbal Memory ($d = 0.19$) to PCSS ($d = 0.98$); the remaining composite scores evidenced medium effect sizes (Cohen, 1988).

SRB RCI scores: Table 4 displays differences in SRB RCI scores. A MANOVA revealed statistically significant differences between groups; Pillai's trace was 0.30, with $F(5,204) = 11.9$, $p < 0.01$. There was a large effect size, partial $\eta^2 = 0.23$ (Cohen, 1992). Follow-up ANOVAs revealed statistically significant differences between groups on the Visual Memory ($F(1,208) = 17.9$, $p < 0.01$, $d = 0.59$), Reaction Time ($F(1,208) = 11.2$, $p < 0.01$, $d = 0.44$), and PCSS ($F(1,208) = 43.5$, $p < 0.01$, $d = 0.85$) composites. The Verbal Memory and Visual Motor Speed composites did not produce statistically significant differences.

Discriminant function analyses: Table 5 displays sensitivities, specificities, PPVs, and NPVs for each DFA. Several analyses were undertaken to appraise the accuracy of ImPACT's composite indices in differentiating between individuals with and without concussion. Prior probabilities, or base rates, were set to align with the frequency of concussion present in the sample (i.e., 38.6% concussion; 61.4% healthy).

Results of the stepwise DFA selected PCSS and Visual Memory scores. These two variables significantly discriminated between groups, $F(2, 207) = 27.41$, $p < 0.01$, $\eta^2 = 0.21$. The pooled within structure loadings were 0.89 and 0.57 for PCSS and Visual Memory scores, respectively. The combination of these two predictor variables correctly classified 74.3% of individuals.

Results of the PCSS-only DFA revealed that this variable significantly discriminated between groups, $\chi^2 = 49.42$, $p < 0.01$, $\eta^2 = 0.17$. The PCSS, in isolation, correctly classified 73.8% of individuals.

Results of the cognitive composite scores only DFA revealed that this combination of variables significantly discriminated between groups, $\chi^2(4) = 23.21$, $p < 0.01$, $\eta^2 = 0.11$. The pooled within structure loadings were as follows: Visual Memory = 0.85, Reaction Time = 0.67, Verbal Memory = 0.11, and Visual Motor Speed = 0.09. The combination of these four predictor variables correctly classified 65.7% of individuals.

Results of the total combined composite scores DFA revealed that this combination of variables significantly discriminated between groups, $\chi^2 = 52.44$, $p < 0.01$, $\eta^2 = 0.23$. The pooled within structure loadings were as follows: PCSS = 0.85,

Visual Memory = 0.54, Reaction Time = 0.43, Verbal Memory = 0.07, and Visual Motor Speed = 0.06. The combination of these five predictor variables correctly classified 76.2% of individuals.

Standard interpretive procedure: The current interpretive procedure failed to classify 22 (27.2%) individuals in the concussion group as producing a score on post-injury assessment exceeding the reliable change interval. Fifty-nine (72.8%) individuals produced a score on at least one composite exceeding the reliable change interval, 11 (13.6%) individuals on two composites, 10 (12.3%) individuals on three composites, 10 (12.3%) individuals on four composites, and 2 (2.5%) individuals on all five composites. For the PCSS, 42 (51.8%) individuals exceeded the reliable change interval.

Conversely, for the healthy group, 56 (43.4%) individuals did not evidence a change on post-injury assessment using the reliable change interval as the criterion. Seventy-three individuals (56.6%) produced a score on at least one composite exceeding the reliable change interval, 16 (12.4%) individuals on two composites, and 4 (3.1%) individuals on three composites. For the PCSS, 10 (7.8%) individuals exceeded the reliable change interval.

Results of a chi-square test indicated that the proportion of individuals obtaining one or more scores exceeding a reliable change interval was significantly different between groups, $\chi^2 = 5.63$, $p = 0.02$. When results from both groups are combined, and assuming the same prior probabilities used in the DFAs, the classification accuracy rates for observing one or more score exceeding the reliable change interval as the criterion are as follows: sensitivity = 72.8%, specificity =

43.4%, positive predictive value = 44.7%, and negative predictive value = 71.8%. Figure 1 displays differences in accuracy rates among data integration approaches.

Discussion

This study aimed to evaluate ImPACT's diagnostic accuracy using a regression-based procedure for detecting change. Additionally, it contrasted the standard interpretive procedure versus data integration approaches derived from DFAs. The DFAs revealed that the cognitive composite scores provided minimal incremental value beyond relying on the PCSS score in isolation. Regarding differences between interpretive procedures, results suggested that the DFA equations were more accurate than the standard procedure. However, there was a substantial discrepancy in sensitivity, as none of the DFA equations produced a sensitivity higher than 50%; the sensitivity using the standard procedure was approximately 73%. Relative to prior research, ImPACT's classification accuracy rates were largely consistent with expectations.

Properties of assessment methods or relative efficacy of different methods can vary in relationship to base rates in the setting of application, and thus the base rates in the current study (i.e., about 61% healthy, 39% concussed) should be kept in mind when interpreting results. Under these conditions, the stepwise DFA, which included only the PCSS and Visual Memory scores, marginally exceeded the accuracy of the PCSS in isolation. A positive result on the DFA yielded respective accuracy rates in identifying concussion of 80% and 83% using the stepwise DFA versus the PCSS DFA alone. There was about a 1% difference in NPVs. As such, when relying on these DFA derived approaches in determining, or helping to appraise, the presence or

absence of concussion, a 22 symptom self-report rating scale (PCSS) appears to classify individuals as accurately in isolation as it does in conjunction with a battery of cognitive tests. Although not intended for use as a stand-alone diagnostic tool, this finding raises questions about the utility of ImPACT's cognitive tests and their incremental value in clinical decision-making.

It might be argued that administering cognitive tests that do not add to, but do not diminish from, specificity, and that could provide useful information about the level of cognitive functioning across a series of domains, has clinical value. However, shortcomings in the sensitivity of the DFA equations compromise their overall value and create serious concerns. As noted, at best, the ImPACT DFA equations were only able to detect concussion, when present, about 50% of the time. Failure to identify concussion when present, or a false-negative error, creates significant risk that the adverse consequences one is using such procedures to try to prevent in the first place will nevertheless occur. Clearing individuals to return to school, work, or other activities prematurely creates an elevated risk for them and possibly for others. The standard procedure may partly offset the DFA approach's deficient sensitivity, raising it to about 73%; however, the associated elevation in the false-positive rate creates its own set of problems and potential harms. Although the standard procedure has displayed some utility in detecting concussion, its PPV was less than chance, thereby rendering the interpretation of a positive sign as no more accurate than a coin flip. Falsely identifying concussion and removing an individual from a variety of mainstream or usual activities can devolve, for example, into chronic iatrogenic disorder (Mittenberg et al., 1992; Silverberg & Iverson, 2011).

These results would seem to add to the literature addressing ImPACT's classification accuracy. The methodological differences in prior research and the current study largely preclude comparison of accuracy rates. Schatz et al.'s (2006) methodology appears comparable to the present study, with one crucial distinction. Schatz et al. (2006) entered scores on post-injury testing for the concussion group and baseline testing for the healthy group into the DFA. As such, results from this study only relate to interindividual differences on ImPACT drawn from a single time point. The present study represents a novel approach in examining RCI scores in the DFA, rather than cross-sectional composite scores, thereby examining both intraindividual and interindividual differences in a manner more closely resembling the application of ImPACT in clinical practice.

The results of the present study most closely paralleled those of Nelson et al. (2016) and underscore the importance of base rate considerations. For example, at 24 hours post-injury, Nelson et al. (2016) reported a PPV of 91.7% and NPV of 15.6% relying on the standard interpretive procedure. Conversely, at approximately 3 days post-injury, following the standard interpretive procedure, the results of this study yielded a PPV of 44.7% and NPV of 71.8%. The differentiating factor between studies was the base rate of individuals with concussion – 89.4% in Nelson et al. (2016) and 38.6% in the present study. In Nelson et al. (2016), at eight days post-injury, when the base rate more closely matched the present study (35.4%), the PPV declined to 50.7%, and the NPV increased to 80.9%, which more closely align with the results of the present study.

In addition to the factors described above, test-retest reliability is a critical and often overlooked factor in appraising the accuracy of observed test scores. As a number of studies converge in showing, ImPACT's test-retest reliabilities for composite scores fall below desirable levels, or are frankly problematic (Alsalaheen et al., 2016a; Broglio et al., 2018; Farnsworth, Dargo, Ragan, & Kang, 2017; Resch, Schneider, & Munro Cullum, 2018). In a recent study, test-retest reliabilities for ImPACT's cognitive composite scores ranged from 0.47 to 0.72 over a 1-year interval. One can conceptualize test-retest reliability coefficients operating as levers in calculating RCI scores. The higher the reliability, the less the predicted score on testing at Time 2 will deviate from the observed score at Time 1 due to weaker regression to the mean effects. As this discrepancy narrows, it increases the probability of detecting subtle declines on post-injury testing. Even modest improvements in test-retest reliability, such as incorporating adaptive testing and extending the range of possible scores, might substantially enhance classification accuracy (see Chapter 1).

Although these results provide further insight into ImPACT's classification accuracy by using what is arguably more advanced methodology, it is still subject to substantial limitations. Namely, the underlying methodology is ultimately anchored to a subjective criterion (e.g., judgment of a physician or athletic trainer) to determine ImPACT's diagnostic accuracy. This approach is problematic as these professionals may not always agree on the diagnosis. For example, in a study consisting of 40 rugby medicine doctors, the physicians only agreed on 67.8% of concussion diagnoses (Fuller, Kemp, & Raftery, 2017). As such, it is apparent how reliance on a fallible indicator further undermines attempts to validate test measures; although our

understanding of more objective indicators such as fluid biomarkers and genetic testing is rapidly developing (McCrea et al., 2017).

Given the superior accuracy of methods reliant on more objective, rather than subjective criteria, and bootstrapping as opposed to human judgment, the shortcomings of current approaches to the study of concussion assessment are evident (Kaufmann & Wittmann, 2016). Dawes and Meehl (1966) proposed a potential solution to this problem worth considering in the design of future research. Specifically, Dawes and Meehl (1966) suggest a "mixed group" validation procedure, which in the context of ImPACT, results for two groups with different, but known, base rates of concussive symptomatology might be used to determine classification accuracy (Dawes & Meehl, 1966); Jewsbury (2018) has further advanced this approach. A mixed group validation procedure confers several advantages. First, differential rates of ongoing concussion symptomatology at varying time points is understood for certain discriminating variables (e.g., age; Kamins et al., 2017). Through mixed group validation, the accuracy of a positive result on testing is inferred based on an individual's match with a discriminating variable. Additionally, given the nonspecific clinical symptoms associated with concussion, it lends itself to an "open concept" approach in which various indicators (e.g., ImPACT cognitive composite scores, symptom endorsements) may provide an alternative approach to conceptualizing concussion assessment and management (Dawes & Meehl, 1966; Jewsbury, 2018; Meehl, 1965).

The matter of varying base rates of concussion symptomatology at differing time points post-injury represents another limitation of the current literature and a

future direction. Given ImPACT's use as a longitudinal measure designed to be administered at multiple time points to monitor recovery, most research has only examined its accuracy within a 72-hour window following injury (Alsalaheen et al., 2016b). As base rates of symptoms generally decline, and those that persist tend to reduce in severity in the days following injury, it likely becomes even more difficult to identify individuals with and without concussion accurately (Nelson et al., 2016). As such, additional research is warranted to determine the clinical utility, if any, of administering ImPACT beyond 72 hours post-injury.

This study was subject to additional limitations. It consisted of a relatively small sample, with fewer than 100 individuals with concussion available for comparison. The comparison groups (i.e., concussion versus healthy) were also of unequal size. Additionally, the healthy group was significantly older than the concussion group at the time of follow-up testing (post-injury or second baseline). Moreover, the sample was high achieving, as evidenced by average SAT scores exceeding the 90th percentile, thereby reducing the generalizability of these results to the broader adolescent population. Coupled with the expanding body of research suggestive of ceiling effects on several subscales, individuals in this sample were possibly less likely to produce detectable declines on retesting (Allen & Gfeller, 2011; Mayers & Redick, 2012; Chapter 1). Specifically, the restricted range of possible scores may have contributed to underestimates of abilities on baseline testing. For example, individuals with the ability to accurately recognize 20 out of 20 words following a delay were only able to demonstrate their ability to recognize 12 out of 12.

As such, subtle declines for these individuals, would not be detected due to the constraints in ImPACT's design on several subscales.

From a cultural perspective, this study was relatively unique in that approximately 10% of the sample was born outside of the U.S. and did not speak English as a first language; however, the sample was not diverse enough to examine the effects of such variables on ImPACT performance and diagnostic accuracy. As indicated in the prior studies examining differences in cultural variables, further research is necessary (Kontos, Elbin, Covassin, & Larson, 2010).

In conclusion, this study examined ImPACT's classification accuracy. It extended and advanced prior research through its use of regression-based RCI scores to appraise discriminative utility in a manner that closely resembles its use in clinical practice (i.e., comparing post-injury scores to baseline scores). Additionally, it compared standard interpretive procedures for determining the presence of concussive symptomatology to a data-driven approach using DFAs.

Results revealed that the standard interpretive procedure yielded a higher sensitivity rate than the DFA; however, the PPV did not exceed chance levels. Conversely, the DFA equations yielded superior PPVs; however, their sensitivity hovered around 50%, leaving a substantial proportion of concussion cases undetected. Regarding clinical utility, given the increase in PPV attributable to the DFA approach, clinicians may have higher confidence in concluding the presence of concussion upon obtaining a positive sign, but with full awareness that the absence of a positive sign does not rule out concussion.

When considering the composition of the DFA equations, it is worth noting that the cognitive composite scores evidenced marginal incremental utility in enhancing classification accuracy. The DFA consisting of only the TSS score yielded classification rates equivalent to the stepwise DFA comprised of the PCSS and Visual Memory scores. This result calls into question the utility of including ImPACT's cognitive testing component in clinical decision making as it did not appear to differentiate between those with and without concussion more effectively than a symptom scale, when using DFAs; although this result did not hold when using standard interpretive procedures. In sum, there appears to be a basis for the inclusion of cognitive testing within ImPACT; however, pending further refinement of its psychometric characteristics, these indicators provide minimal value in accurately detecting the presence or absence of concussion as applied in a serial design. Moreover, base rate conditions, such as the decline in ongoing concussion symptomatology relative to the proximity of time since injury, further reduce ImPACT's clinical utility.

Funding

The author did not receive funding for this project.

Acknowledgments

The author would like to thank his major professor, Dr. David Faust and his dissertation defense committee, Dr. JeffrKonin, Dr. Kate Webster, and Dr. William Renehan, for their guidance and support.

References

Allen, B. J., & Gfeller, J. D. (2011). The Immediate Post-Concussion Assessment and Cognitive Testing Battery and traditional neuropsychological measures: A construct and concurrent validity study. *Brain Injury*, *25*(2), 179–191. https://doi.org/10.3109/02699052.2010.541897

Alsalaheen, B., Stockdale, K., Pechumer, D., & Broglio, S. P. (2016a). Measurement error in the Immediate Postconcussion Assessment and Cognitive Testing (ImPACT): Systematic review. *Journal of Head Trauma Rehabilitation*, *31*(4), 242–251. https://doi.org/10.1097/HTR.0000000000000175

Alsalaheen, B., Stockdale, K., Pechumer, D., & Broglio, S. P. (2016b). Validity of the Immediate Post Concussion Assessment and Cognitive Testing (ImPACT). *Sports Medicine*, *46*(10), 1487–1501. https://doi.org/10.1007/s40279-016-0532-y

American Education Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Arkes, H. R., González-Vallejo, C., Bonham, A. J., Kung, Y.-H., & Bailey, N. (2009). Assessing the merits and faults of holistic and disaggregated judgments. *Journal of Behavioral Decision Making*, *23*(3), 250–270. https://doi.org/10.1002/bdm.655

Broglio, S. P., Katz, B. P., Zhao, S., McCrea, M., McAllister, T., Reed Hoy, A., … Lintner, L. (2018). Test-retest reliability and interpretation of common concussion assessment tools: Findings from the NCAA-DoD CARE Consortium. *Sports Medicine*, *48*(5), 1255–1268. https://doi.org/10.1007/s40279-017-0813-0

Broglio, S. P., Macciocchi, S. N., & Ferrara, M. S. (2007). Sensitivity of the

concussion assessment battery. *Neurosurgery*, *60*(6), 1050–1058. https://doi.org/10.1227/01.NEU.0000255479.90999.C0

Cassidy, J. D., Carroll, L. J., Peloso, P. M., Borg, J., Von Holst, H., Holm, L., … Coronado, V. G. (2004). Incidence, risk factors and prevention of mild traumatic brain injury: Results of the WHO Collaborating Centre Task Force on mild traumatic brain injury. *Journal of Rehabilitation Medicine*, *43*, 28–60. https://doi.org/10.1080/16501960410023732

Chelune, G. J., Naugle, R. I., Lüders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology*, *7*(1), 41–52. https://doi.org/10.1037/0894-4105.7.1.41

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. San Diego, CA: Academic Press.

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

Covassin, T., Elbin, R., & Stiller-Ostrowski, J. L. (2009). Current sport-related concussion teaching and clinical practices of sports medicine professionals. *Journal of Athletic Training*, *44*(4), 400–404.

Crawford, J. R., & Garthwaite, P. H. (2007). Using regression equations built from summary data in the neuropsychological assessment of the individual case. *Neuropsychology*, *21*(5), 611–620. https://doi.org/10.1037/0894-4105.21.5.611

Crawford, J. R., Garthwaite, P. H., Denham, A. K., & Chelune, G. J. (2012). Using regression equations built from summary data in the psychological assessment of the individual case: Extension to multiple regression. *Psychological Assessment*, *24*(4), 801–814. https://doi.org/10.1037/a0027699

Dawes, R. M., & Meehl, P. E. (1966). Mixed group validation: A method for determining the validity of diagnostic signs without using criterion groups. *Psychological Bulletin*, *66*(2), 63–67. https://doi.org/10.1037/h0023584

Duff, K. (2012). Current topics in science and practice evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology*, *27*(3), 248–261. https://doi.org/10.1093/arclin/acr120

Echemendia, R. J., Iverson, G. L., McCrea, M., Macciocchi, S. N., Gioia, G. A., Putukian, M., & Comper, P. (2013). Advances in neuropsychological assessment of sport-related concussion. *British Journal of Sports Medicine*, *47*(5), 294–298. https://doi.org/10.1136/bjsports-2013-092186

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. https://doi.org/10.4085/1062-6050-52.6.03

Faust, D. (1989). Data integration in legal evaluations: Can clinicians deliver on their promises? *Behavioral Sciences & the Law*, *7*(4), 469–483. https://doi.org/10.1002/bsl.2370070405

Fuller, G. W., Kemp, S. P. T., & Raftery, M. (2017). The accuracy and reproducibility of video assessment in the pitch-side management of concussion in elite rugby. *Journal of Science and Medicine in Sport*, *20*(3), 246–249. https://doi.org/10.1016/J.JSAMS.2016.07.008

Gardner, A., Shores, E. A., Batchelor, J., & Honan, C. A. (2012). Diagnostic

efficiency of ImPACT and CogSport in concussed rugby union players who have not undergone baseline neurocognitive testing. *Applied Neuropsychology: Adult*, *19*(2), 90–97. https://doi.org/10.1080/09084282.2011.643945

Gaudet, C. E., & Weyandt, L. L. (2017). 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. https://doi.org/10.1080/13854046.2016.1220622

Gigerenzer, G. (2007). Helping doctors and patients make sense of health statistics: Toward an evidence-based society. *Psychological Science in the Public Interest*, *8*(2), 53–96.

Hinton-Bayre, A. D. (2016). Clarifying discrepancies in responsiveness between reliable change indices. *Archives of Clinical Neuropsychology*, *31*(7), 754–768. https://doi.org/10.1093/arclin/acw064

Iverson, G. L. (2001). Interpreting change on the WAIS-III/WMS-III in clinical samples. *Archives of Clinical Neuropsychology*, *16*(2), 183–191.

Iverson, G. L., Lovell, M. R., & Collins, M. W. (2004). Interpreting change on ImPACT following sport concussion. *The Clinical Neuropsychologist*, *17*(4), 460–467. https://doi.org/10.1076/clin.17.4.460.27934

Jacobson, N. S. Y., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and. *Clinical Psychology*, *59*(1), 12–19.

Jewsbury, P. A. (2018). Diagnostic test score validation with a fallible criterion. *Applied Psychological Measurement*. https://doi.org/10.1177/0146621618817785

Kamins, J., Bigler, E., Covassin, T., Henry, L., Kemp, S., Leddy, J. J., … Giza, C. C. (2017). What is the physiological time to recovery after concussion? A systematic review. *British Journal of Sports Medicine*, *51*(12), 935–940. https://doi.org/10.1136/bjsports-2016-097464

Kaufmann, E., & Wittmann, W. W. (2016). The success of linear bootstrapping models: Decision domain-, expertise-, and criterion-specific meta-analysis. *PLoS ONE*, *11*(6), e0157914. https://doi.org/10.1371/journal.pone.0157914

Kontos, A. P., Elbin, R. J., Covassin, T., & Larson, E. (2010). Exploring differences in computerized neurocognitive concussion testing between African American and white athletes. *Archives of Clinical Neuropsychology*, *25*(8), 734–744. https://doi.org/10.1093/arclin/acq068

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

Laker, S. R. (2011). Epidemiology of concussion and mild traumatic brain injury. *PM and R*, *3*(10 SUPPL. 2), S354–S358. https://doi.org/10.1016/j.pmrj.2011.07.017

Lemonda, B. C., Tam, D., Barr, W. B., & Rabin, L. A. (2017). Assessment trends among neuropsychologists conducting sport-related concussion evaluations. *Developmental Neuropsychology*, *42*(2), 113–126. https://doi.org/10.1080/87565641.2016.1274315

Lovell, M. R. (2016). *ImPACT Administration and Interpretation Manual*. ImPACT Applications, Inc. Retrieved from www.impacttest.com

Lynall, R. C., Laudner, K. G., Mihalik, J. P., & Stanek, J. M. (2013). Concussion-

assessment and -management techniques used by athletic trainers. *Journal of Athletic Training*, *48*(6), 844–850. https://doi.org/10.4085/1062-6050-48.6.04

Mayers, L. B., & Redick, T. S. (2012). Authors' reply to "Response to Mayers and Redick: 'Clinical utility of ImPACT assessment for postconcussion return-to-play counseling: Psychometric issues.'" *Journal of Clinical and Experimental Neuropsychology*, *34*(4), 435–442. https://doi.org/10.1080/13803395.2012.667790

McCrea, M., Guskiewicz, K. M., Marshall, S. W., Barr, W., Randolph, C., Cantu, R. C., … Kelly, J. P. (2003). Acute effects and recovery time following concussion in collegiate football players. *JAMA*, *290*(19), 2556. https://doi.org/10.1001/jama.290.19.2556

McCrea, M., Meier, T., Huber, D., Ptito, A., Bigler, E., Debert, C. T., … Mcallister, T. (2017). Role of advanced neuroimaging, fluid biomarkers and genetic testing in the assessment of sport-related concussion: a systematic review. *British Journal of Sports Medicine*, *51*(12), 919–929. https://doi.org/10.1136/bjsports-2016-097447

McCrory, P., Meeuwisse, W., Aubry, M., Cantu, R. C., Dvořák, J., Echemendia, R. J., … Turner, M. (2013). Consensus statement on concussion in sport: The 4th international conference on concussion in sport, held in Zurich, November 2012. *Journal of Athletic Training*, *48*(4), 554–575. https://doi.org/10.4085/1062-6050-48.4.05

McCrory, P., Meeuwisse, W., Dvořák, J., Aubry, M., Bailes, J., Broglio, S., … Vos, P. E. (2017). 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*(11), 838–847. https://doi.org/10.1136/bjsports-2017-097699

McSweeny, A. J., Naugle, R. I., Chelune, G. J., & Luders, H. (1993). "T scores for change": An illustration of a regression approach to depicting change in clinical neuropsychology. *Clinical Neuropsychologist*, *7*(3), 300–312. https://doi.org/10.1080/13854049308401901

Meehan, W. P., d'Hemecourt, P. Collins, C. L., Taylor, A. M., & Comstock, D. (2012). Computerized neurocognitive testing for the management of sport-related concussions. *Pediatrics*, *129*(1), 38–44. https://doi.org/10.1542/peds.2011-1972

Meehl, P. E. (1965). *Detecting latent clinical taxa by fallible quantitative indicators lacking an accepted criterion*. Minneapolis, MN: University of Minnesota. Retrieved from http://meehl.umn.edu/sites/g/files/pua1696/f/065techrep1.pdf

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

Mittenberg, W., DiGiulio, D. V, Perrin, S., Bass, A. E., Diane DiGiulio, F. V, Sean Perrin, F., … Anthony Bass, U. E. (1992). Symptoms following mild head injury: expectation as aetiology. *J Neurol Neurosurg Psychiatry*, *55*, 200–204. https://doi.org/10.1136/jnnp.55.3.200

National Institute of Neurological Disorders and Stroke. (2014). Multiple System Atrophy Fact Sheet | National Institute of Neurological Disorders and Stroke. Retrieved June 25, 2019, from https://www.ninds.nih.gov/disorders/patient-

caregiver-education/fact-sheets/multiple-system-atrophy

Nelson, L. D., Laroche, A. A., Pfaller, A. Y., Lerner, E. B., Hammeke, T. A., Randolph, C., … McCrea, M. A. (2016). Prospective, head-to-head study of three computerized neurocognitive assessment tools (CNTs): Reliability and validity for the assessment of sport-related concussion. *Journal of the International Neuropsychological Society*, *22*(1), 24–37. https://doi.org/10.1017/S1355617715001101

Resch, J. E., Schneider, M. W., & Munro, C. (2018). The test-retest reliability of three computerized neurocognitive tests used in the assessment of sport concussion. *International Journal of Psychophysiology*, *132*(December 2016), 31–38. https://doi.org/10.1016/j.ijpsycho.2017.09.011

Ruscio, J. (2003). Holistic judgment in clinical practice: Utility or futility? *The Scientific Review of Mental Health Practice: Objective Investigations of Controversial and Unorthodox Claims in Clinical Psychology, Psychiatry, and Social Work*, *2*(1), 38–48.

Schatz, P., Pardini, J. E., Lovell, M. R., Collins, M. W., & Podell, K. (2006). Sensitivity and specificity of the ImPACT test battery for concussion in athletes. *Archives of Clinical Neuropsychology*, *21*(1), 91–99. https://doi.org/10.1016/j.acn.2005.08.001

Schretlen, D. J., & Shapiro, A. M. (2003). A quantitative review of the effects of traumatic brain injury on cognitive functioning. *International Review of Psychiatry*, *15*(4), 341–349. https://doi.org/10.1080/09540260310001606728

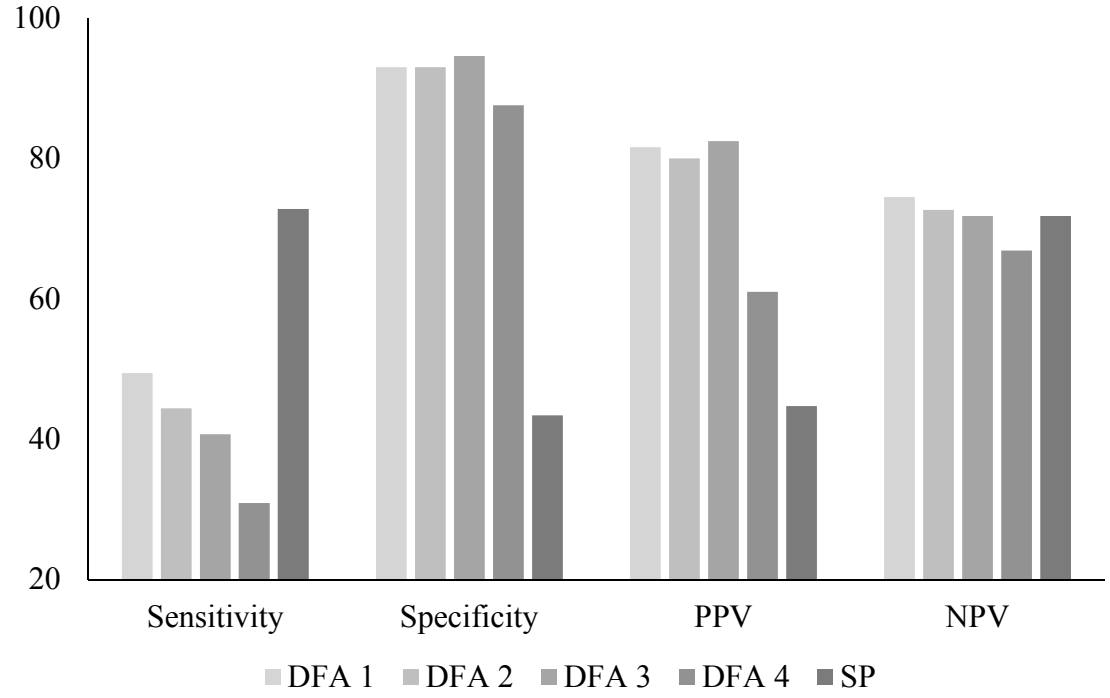
Silverberg, N. D., & Iverson, G. L. (2011). Etiology of the post-concussion syndrome:

Physiogenesis and psychogenesis revisited. *NeuroRehabilitation*, *29*, 317–329. https://doi.org/10.3233/NRE-2011-0708

Van Kampen, D. A., Lovell, M. R., Pardini, J. E., Collins, M. W., & Fu, F. H. (2006). The "value added" of neurocognitive testing after sports-related concussion. *American Journal of Sports Medicine*, *34*(10), 1630–1635. https://doi.org/10.1177/0363546506288677

Figures

Figure 1. Comparison of classification accuracy rates among approaches



Note: PPV = positive predictive value; NPV = negative predictive value; DFA 1 = all ImPACT composite scores; DFA 2 = stepwise procedure (Post Concussion Symptom Scale [PCSS]) and Visual Memory; DFA 3 = PCSS-only; DFA 4 = cognitive composite scores only; SP = standard interpretive procedure

Tables

Table 1. Comparison of demographic variables between groups with and without concussion

	No Concussion (*N*=129)	Concussion (*N*=81)		
	Mean (SD) / N (%)	Mean (SD) / N (%)	t / χ^2	p
Age at Time 1	14.5 (1.2)	14.8 (1.2)	2.1	0.03
Age at Time 2	16.3 (0.6)	15.8 (1.3)	4.1	<0.01
Female	56 (43.4)	48 (59.3)	5.0	0.03
U.S. Born	116 (89.9)	73 (90.1)	0.0	0.96
English First Language	118 (91.5)	72 (88.9)	0.4	0.53
ADHD	14 (10.9)	6 (7.4)	0.7	0.41
Learning Disability	6 (4.7)	3 (3.7)	0.1	0.74
Test-Retest Interval (years)	1.5 (1.2)	1.5 (0.8)	0.3	0.78

Note: ADHD = Attention Deficit/Hyperactivity Disorder self-reported diagnosis

Table 2. Comparison of ImPACT composite scores at baseline between groups with and without concussion

	No Concussion (*N*=129)	Concussion (*N*=81)			
	Mean (SD)	Mean (SD)	*d*	*F*	*p*
Verbal Memory	85.5 (9.0)	84.2 (10.7)	0.13	0.90	0.34
Visual Memory	76.1 (12.5)	76.7 (13.1)	0.04	0.09	0.77
Visual Motor Speed	37.9 (6.3)	37.2 (6.6)	0.11	0.51	0.48
Reaction Time	0.59 (0.1)	0.59 (0.1)	0.02	0.04	0.84
Post Concussion Symptom Scale	3.8 (5.9)	5.0 (7.3)	0.18	1.85	0.18

Note: *d* = Cohen's *d*; F and *p* values refer to one-way analysis of variance results.

Table 3. Comparison of ImPACT composite scores at Time 2 (post-injury or second baseline) between groups with and without concussion

	No Concussion (*N*=129)	Concussion (*N*=81)			
	Mean (SD)	Mean (SD)	*d*	*F*	*p*
Verbal Memory	85.6 (10.8)	83.4 (12.2)	0.19	1.91	0.17
Visual Memory	78.1 (13.1)	70.3 (16.2)	0.53	14.55	<0.01
Visual Motor Speed	40.9 (5.9)	37.2 (7.5)	0.55	15.77	<0.01
Reaction Time	0.58 (0.1)	0.64 (0.2)	0.38	13.19	<0.01
Post Concussion Symptom Scale	5.7 (9.7)	19.2 (16.8)	0.98	54.18	<0.01

Note: *d* = Cohen's *d*; F and *p* values refer to one-way analysis of variance results.

Table 4. Comparison of ImPACT reliable change index composite scores between groups with and without concussion

	No Concussion (*N*=129)	Concussion (*N*=81)			
	Mean (SD)	Mean (SD)	*d*	*F*	*p*
Verbal Memory	-0.03 (1.02)	-0.12 (1.20)	0.08	0.31	0.58
Visual Memory	0.04 (1.09)	-0.64 (1.21)	0.59	17.88	<0.01
Visual Motor Speed	-0.09 (1.09)	-0.17 (1.36)	0.06	0.22	0.64
Reaction Time	0.16 (0.87)	-0.48 (1.87)	0.44	11.21	<0.01
Post Concussion Symptom Scale	-0.29 (1.13)	-1.89 (2.4)	0.85	43.52	<0.01

Note: *d* = Cohen's *d*; F and *p* values refer to one-way analysis of variance results.

Table 5. ImPACT classification accuracy using discriminant function analyses

Procedure	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value
All Composite Scores (%)	49.4	93.0	81.6	74.5
Stepwise[1] (%)	44.4	93.0	80.0	72.7
Symptom Scores Only (%)	40.7	94.6	82.5	71.8
Cognitive Only (%)	30.9	87.6	61.0	66.9

Note: Prior probabilities were set to align with the frequency of concussion cases in the sample – 38.6% concussion & 61.4% healthy.[1]The stepwise procedure selected the Total Symptom Score and Visual Memory Score for the discriminant function.

Supplemental Tables

Table 1. Standardized regression-based and reliable change index score inputs by composite

	β	Constant	Standard Error of the Estimate
Verbal Memory	1.02	-0.86	10.44
Visual Memory	1.00	1.76	12.17
Visual Motor Speed	1.01	0.53	4.82
Reaction Time	1.12	-0.07	0.09
Post Concussion Symptom Scale	1.49	-2.01	7.17