DEVELOPMENT AND VALIDATION OF AN OBJECTIVE BALANCE ERROR SCORING SYSTEM

by

HARRISON JAMES BROWN

B.Sc, University of Guelph, 2010

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

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Kinesiology)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

September 2013

Abstract

Standing balance is an important unbiased indicator of concussion severity. However, limited accessibility to high-end technology and unreliability of simple balance assessment tools make it difficult to assess standing balance accurately outside of research laboratory settings. The objective of this thesis was to develop and validate a simple objective balance assessment tool that can provide an accurate, reliable, and affordable alternative to the currently available sideline methods. In Experiment 1, thirty healthy subjects were filmed performing the Balance Error Scoring System (BESS) while wearing inertial measurement units (IMUs) that measured linear accelerations and angular velocities from seven landmarks: forehead, chest, waist, right & left wrist, right & left shin. Each video was scored by four experienced BESS raters. Mean experienced rater scores were used to develop an algorithm to compute objective BESS (oBESS) scores solely from IMU data. oBESS was able to accurately fit and predict mean experienced rater BESS scores using acceleration data from only one IMU located at the forehead. In Experiment 2, twenty healthy subjects wore the same network of IMUs and serially performed 12 BESS tests in a hypoxic altitude chamber, aimed at increasing the number of balance errors. Each video was scored by three experienced raters and two athletic trainers. Similarly to Experiment 1, experienced rater scores were used along with IMU data to develop the oBESS algorithm. However, because experienced raters displayed low inter-rater and intra-rater reliability, algorithm training and analyses were performed only using trials where the raters had marginal scoring differences. The oBESS was able to fit mean experienced rater scores with greater accuracy than the two athletic trainers, but not at a level commonly associated with high clinical reliability. In summary, this thesis shows that the oBESS can reliably predict total BESS

scores in normal subjects, but only if trained using an accurate gold standard that allows the algorithm to overcome measurement error associated with the human-scored BESS. Pending further validation, the oBESS may represent a useful and valid tool to assess balance in athletes on the sideline by offering an objective alternative to the current scoring methods of the BESS.

Preface

The research presented in Chapter 4 (Experiment 1) was conducted in the Sensorimotor Physiology Lab at the University of British Columbia (UBC) using methods approved by UBC's Clinical Research Ethics Board (CREB; H11-02306). A manuscript describing this work has been prepared and will be submitted for publication to a scientific journal in early September 2013. For this study, I was the lead investigator and was responsible for all major areas of concept formation, data collection and analysis, algorithm development, and manuscript composition. Jean-Sébastien Blouin and Gunter Siegmund were supervisory authors involved in concept formation, analysis advisory, and manuscript edits. Kees Van den Doel was involved in the development of a pilot algorithm and advisory on the final algorithm. Kevin Guskiewicz provided advise on methods and access to experienced BESS raters. Edmond Cretu provided advise regarding the body sensors used for data collection.

The research presented in Chapter 5 (Experiment 2) was conducted in the Environmental Physiology Lab at UBC using methods approved by UBC's CREB (H12-00787). A manuscript describing this work has yet to be prepared. For this study, I was the lead investigator and was responsible for all major areas of concept formation, data collection, analysis, and algorithm development. Jean-Sébastien Blouin and Gunter Siegmund were supervisory authors involved in concept formation and analysis advisory. Michael Koehle was involved in concept formation, data collection, and analysis advisory. Kevin Guskiewicz provided advise on methods and access to experienced BESS raters.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my co-advisors, Dr. Jean-Sébastien Blouin and Dr. Gunter Siegmund, for inspiring me throughout my time at UBC. In addition to giving me the opportunity to research an area I am passionate about, they have taught me a number of valuable skills that are required to perform quality medical and natural sciences research. I am very grateful for time and effort they both have put into not only the work presented in this thesis, but also into my development as a researcher.

I would also like to thank my third committee member, Dr. Michael Koehle, for approaching my ideas and questions with a positive and enthusiastic attitude. His insights from the clinical perspective were integral to this thesis as they provided perspectives that I, and many other students in this field, rarely have access to. I would also like to thank Dr. Jim Rupert for providing access to the hypoxic altitude chamber, which allowed us to perform Experiment 2.

Lastly, I would like to thank my fellow students from the Sensorimotor Physiology Lab for their help, advise, encouragement, and in many cases involvement during the development of this thesis.

Dedication

To my parents.

1 Introduction

Human standing balance is an important unbiased indicator of concussion severity (Davis et al. 2009, Guskiewicz 2011, Riemann and Guskiewicz 2000). As a result, assessments of balance have been incorporated into leading sports-related concussion identification and management protocols that are used to make important clinical decisions on the sideline, such as whether it is safe for a concussed athlete to return-to-play (Guskiewicz et al. 2001, Johnson et al. 2011, Cavanaugh et al. 2005). However, limited accessibility to high-end technology and unreliability of simple balance assessment tools has rendered many medical professionals unable to accurately and reliably assess human standing balance outside of a research laboratory setting (Clark et al. 2010, Bell et al. 2011). Consequently, users have likely been unable to take full advantage of the utility balance assessment presents during sideline evaluation of sports-related concussions.

Standing balance can be assessed a number of different ways: from complex 3D motion tracking systems to very simple techniques, such as visually observing individuals as they stand. Though each method presents advantages and disadvantages, the medical device industry currently lacks an inexpensive tool to perform accurate balance assessment on the sideline (Clark et al. 2010). The current standard for assessing balance in concussed athletes on the sideline, the Balance Error Scoring System (BESS) is a simple test involving three balance-testing stances on both a firm and foam surface (Guskiewicz 2011, Valovich McLeod et al. 2012, Hunt et al. 2009). The BESS offers many benefits over high-end balance assessment tools: it takes less than five minutes to complete, is cost-effective, scientifically validated, and can be conducted in any environment. However, the BESS suffers from one critical limitation: it is scored by the human

judgment of a pre-defined set of "balance errors". The subjective judgement of these errors can lead to scoring differences between raters due to different interpretations and strictness of scoring criteria. While there is literature reporting high reliability of the BESS (ICC = 0.98; Valovich-McLoed et al., 2004), recent studies have reported poor reliability both between (ICC = 0.74) and within scorers judging the same test twice (ICC = 0.57; Finoff et al., 2009). Results have lead researchers to suggest that BESS scores must change by almost 50% for the change to be attributed to a change in balance behaviour and not scorer judgment error (Finoff et al., 2009). Unreliable information regarding balance from the BESS may lead medical professionals to make inappropriate clinical decisions, such as allowing concussed athletes to pre-emptively return to play. The potentially permanent or fatal consequences of sustaining subsequent concussive injuries before the first has resolved stresses the need to develop more accurate balance quantification methods for use on the sideline so that users can make more informed decisions regarding the status of potentially concussed athletes (Guskiewicz et al., 2003).

The objective of this thesis was to develop and validate a simple objective balance assessment tool that can provide an accurate, reliable, and affordable alternative to the currently available sideline methods. The proposed system, the Objective Balance Error Scoring System (oBESS), uses data collected from a network of kinematic sensors placed on the body and a custom-built algorithm to predict BESS scores. By quantifying the BESS using objective kinematic data characteristic of actual body movements during the test, outcome measures will be entirely dependent upon movement of the body, and not on the subjective interpretations of scoring criteria that limit the current human-scoring procedures. The oBESS will not intend to replace the BESS, rather address the limitations of this clinically useful test by objectively automating

it’s the scoring procedures, which if deemed accurate and reliable, could significantly improve balance assessments in the field.

2 Review of Literature

2.1 Sports-related concussions

Since the 1800's researchers and medical professionals have attempted to define concussion, yet there is currently no general consensus upon what exactly the injury is, how it can be diagnosed, or what effects it has on the body (Roozenbeek et al. 2013). It is thought that the term "concussion" originated from the Latin verb "*concutere*", meaning to shake violently or the action of striking together. One of the simplest definitions of a cerebral concussion is a reversible traumatic paralysis of nervous function which lasts for a variable period of time (Ropper and Brown, 2005). In this thesis, I will use a descriptive working definition when referring to the terms "concussion" and "sports-related concussion", originally published by the Concussion in Sport Group (McCrory et al., 2013):

"Concussion is a brain injury and is defined as a complex pathophysiological process affecting the brain, induced by biomechanical forces. Several common features that incorporate clinical, pathologic and biomechanical injury constructs that may be utilised in defining the nature of a concussive head injury include:

1. Concussion may be caused either by direct blow to the head, face, neck or elsewhere on the body with an "impulsive" force transmitted to the head.
2. Concussion typically results in the rapid onset of short-lived impairment of neurological function that resolves spontaneously. However, in some cases, symptoms and sign may evolve over a number of minutes to hours.

3. Concussion may result in neuropathological changes, but the acute clinical symptoms largely reflect a functional disturbance rather than a structural injury and, as such, no abnormality is seen on standard neuroimaging studies.
4. Concussion results in a graded set of clinical symptoms that may or may not involve loss of consciousness. Resolution of the clinical and cognitive symptoms typically follows a sequential course. However, it is important to note that in some cases symptoms may be prolonged."

Cerebral concussions are considered to be a subset, or type of mild traumatic brain injury (mTBI), however "mTBI" and "concussion" are among a number of terms used interchangeably to describe temporary sports-related brain injuries in literature. This ambiguity of terminology has likely impeded comparisons between relevant studies while confusing researchers and medical professionals (Mills and Leathem, 2000). Concussions are often described as a heterogeneous injury with variable presentation, meaning that each injury is unique and therefore must be diagnosed, managed, and treated differently (Eckner et al., 2011). Although presentation is variable, there are a number of hallmark signs of a concussion which can be categorized into three main groups: physical symptoms (e.g. loss of consciousness, balance deficits), cognitive impairments (e.g. poor memory or processing), and behavioural changes (e.g. irritability).

Originally demonstrated by Denny-Brown and Russel (1940), the optimal condition for the production of a concussion is a sudden change in momentum of the head. They performed a number of experiments on monkeys and cats showing that a concussion would result when a freely moving head was struck by a heavy mass, but if the head was prevented from moving at

the moment of impact, the same degree of force failed to produce a concussive injury (Denny-Brown and Russel, 1940). These results were later verified by Gennarelli et al. (1981), who induced concussions in primates by rapid accelerations of the head without an actual impact. There are a number of different occurrences in sport that may present the aforementioned scenarios and potentially lead to the generation of a concussive injury: collision with other athletes or impact with equipment, playing apparatus, or playing surface (see Appendix A). Sports organizations have attempted to minimize these occurrences by implementing rule changes or preventative safety measures including protective equipment such as helmets and mouth guards. However, these preventative measures have not necessarily resulted in a reduced risk of concussions in sport (Donaldson et al. 2013, Benson et al. 2009).

The use of gelatin models to investigate closed head injuries led researchers to discover that when the head is struck with force the brain must always lag behind as a result of its inertia (Gennarelli et al., 1981). Due to its attachment to the high midbrain and neck, the brain is exposed to rotational forces and shearing stresses immediately following the impact (See Figure 1.1). While early efforts to understand the biomechanical basis of concussion focused on linear accelerations, it is now suggested that shear deformations of neurons caused by rotational accelerations of the brain are the predominant biomechanical mechanism of injury (Gennarelli et al. 1982, Meaney and Smith 2011). Immediately following the stretching and shearing of axons in the brain, there is a disruption of neuronal membranes causing an indiscriminate release of neurotransmitters and unchecked ionic flux, leading to a characteristic cascade of neurometabolic events (Giza and Hovda 2001, see Appendix B). Because hallmark signs of a concussive injury occur with minimal detectable anatomic pathology and often completely resolve over time, it

suggests that concussions are the result of temporary neuronal dysfunction due to this cascade of neurometabolic events rather than cell death (Giza and Hovda, 2001).

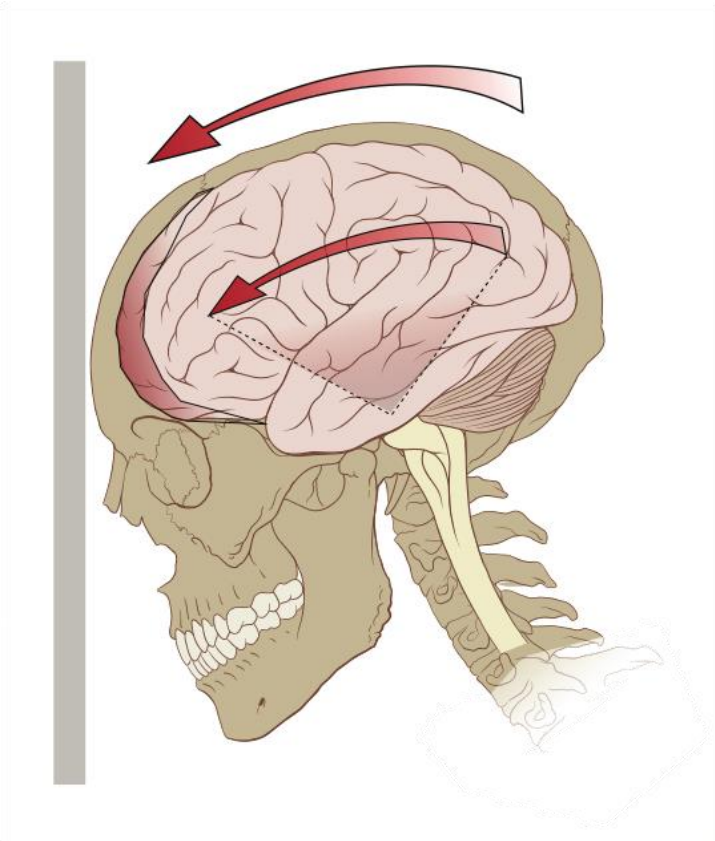


Figure 2.1. Biomechanical depiction of the forces generated on the brain during a concussive injury (© 2006 Patrick J. Lynch, Reproduced with permission from Author).

Some researchers have suggested that the reticular formation is the anatomic site of concussive injury, as activity in the medial reticular formation in concussed monkeys (found in the upper brainstem) was shown to be depressed longer and more severely than the cerebral cortex (Foltz

and Schmidt, 1956). Torque produced at the upper reticular formation, an area of the brain involved in consciousness, would explain the immediate loss of consciousness commonly observed in individuals who are severely concussed. While concussed individuals normally regain consciousness after only a short period of time, they often suffer from anterograde amnesia, a delay in the restoration of normal brain functioning preventing them from creating new memories or recalling the recent past. The duration of anterograde amnesia has been proposed as one of the most reliable indices of concussion severity (Yarnell and Lynch, 1970), however it's time-varying nature and extreme difficulty to estimate have led researchers to downplay its usefulness in concussion evaluations (McCrory et al. 2013, Ropper and Brown 2005). Numerous other methods have been developed to aid the identification and evaluation of concussions: from simple sideline tools to help team physicians and athletic trainers determine if an athlete can return-to-play, to complex algorithm-based neuroimaging techniques that quantify blood flow changes in different areas of the brain that may have occurred as a result of the injury.

2.1.1 Epidemiology

Literature suggests that between 1.6 and 3.8-million Americans experience a sports or recreation-related TBI annually, 300,000 of which are thought to be concussions, costing the American healthcare system an estimated US$60 billion (Gilchrist et al. 2007, Faul et al. 2010, Finkelstein et al. 2006, Gessel et al. 2007). The lack of applicable census data has resulted in broad estimations on the annual number of sports-related concussion occurrences in Canada, however a frequently reported conservative estimate is 31,900 (Gordon et al., 2006). In both cases the actual number is likely much larger due to issues with identification, underreporting by athletes, or

improper evaluations by medical professionals (Meehan and Bachur 2009, Williamson and Goodman 2005, McCrea et al. 2004). In fact, underreporting of concussions by athletes is thought to be as high as 53% (McCrea et al., 2004).

To date, majority of research on the incidence of sports-related concussions has been focused on American football, a sport reported to have among the highest concussion rates in collegiate athletes (0.61/1000 athlete-exposures; A-E: practice or game; Gessel et al. 2007). Other sports that display high concussion rates in collegiate athletes include: female soccer (0.63/1000 A-E), male soccer (0.49/1000 A-E), and female basketball (0.43/1000 A-E; Gessel et al. 2007, see Appendix C). Literature shows that male collegiate athletes generally experience higher rates of concussion (M=0.45/1000 A-E, F=0.38/1000 A-E; Gessel et al. 2007), while females tend to take longer to recover and return-to-play (see Appendix D & E; Gessel et al. 2007). Alarmingly, younger athletes are generally shown to be at higher risk, with athletes between the ages of 10- to 14-years reporting the highest incidence of sports-related concussions of all age groups (Gilchrist et al., 2007; see Figure 2.2).

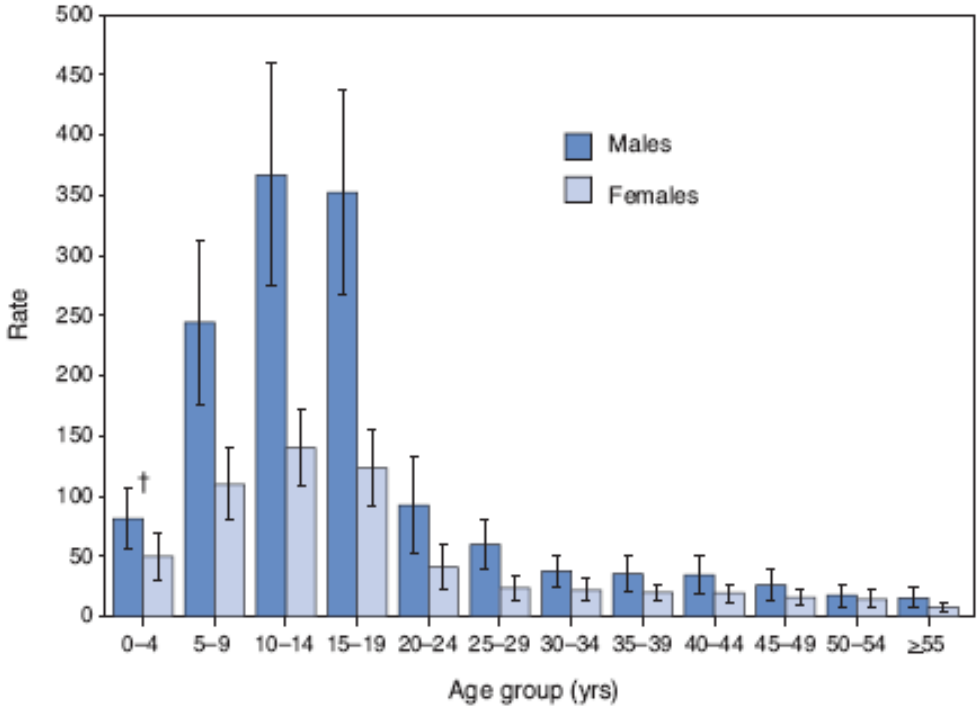


Figure 2.2. Estimated annual rate per 100,000 population of nonfatal, sports- and recreation-related traumatic brain injuries treated in emergency departments, by age group and sex – United States, 2001-2005. Figure 1 from © Gilchrist J, Thomas KE, Wald M, Langlois J. Nonfatal Traumatic Brain Injuries from Sports and Recreation Activities – United States, 2001-2005. Morbidity & Mortality Weekly Report. 2007;56(29):733-737. Page 735. Reproduced with permission from publisher.

2.2 Evaluation methods of sports-related concussions

Because sports-related concussions are a type of mTBI, tests used to evaluate them are generally directed towards quantifying and identifying abnormalities in brain function that may have occurred as a result of the injury. Since the human brain is involved in so many different

processes in the body, there is a large variety of functions clinicians may choose to target. Standardized concussion evaluation protocols used by major sporting organizations such as the National Collegiate Athletics Association (NCAA) generally employ a multi-faceted comprehensive approach to sports-related concussion evaluation that incorporates a number of different tests, each aimed at detecting irregularities in different aspects brain function. This comprehensive method is currently recommended as the appropriate way to evaluate sports-related concussive injuries, as no single test can accurately and reliably diagnose concussion in isolation (McCrory et al., 2013).

One key theme concussion evaluation protocols stresses is athletes should always complete a full recovery before returning to play (McCrory et al., 2013). Numerous studies have suggested that if a concussed athlete returns to play before they have completely recovered, they will be at a much higher risk for repeat injury, and the second concussive injury will generally be more severe and slower to resolve than the first (Laurer et al. 2001, Guskiewicz et al. 2003, Eckner et al. 2011). Some researchers have suggested there is an increased neuronal vulnerability in the brain for approximately 7-10 days following concussive injury, where a second injury could lead to a potentially permanent or fatal "second impact syndrome" (Wetjen et al., 2010). It is also thought that athletes with previous concussion history are more likely to sustain subsequent concussive injuries (Laurer et al., 2001). Guskiewicz and colleagues (2003) have reported that athletes with 3 or more previous concussions were 3 times more likely to sustain another concussion than an athlete with no concussion history whatsoever. To avoid repeat injury, sports-related concussion evaluation is critical in making appropriate return-to-play decisions, and therefore only the most reliable and accurate tests should be included in standardized protocols.

The three most highly recommended areas of brain functioning to target in sports-related concussion evaluations are: symptom evaluation, neurocognitive testing, and balance assessment (McCrory et al. 2013, Peterson et al. 2003). Athletes normally perform symptom evaluations by self-report using a checklist or rating scale, which can be quickly administered to detect hallmark symptoms such as headache, dizziness, confusion, or irritability (McCrory et al., 2013). Numerous symptom checklists are available to choose from, each with varying combinations of symptoms, however no single checklist is considered the "gold standard" by which to perform concussion evaluations (Dziemianowicz et al. 2012, Alla et al. 2009). While literature suggests the validity of using self-report symptom checklists to evaluate sports-related concussions (McCrory et al. 2013, Peterson et al. 2003), the biggest limitation of this type of testing is that it is inherently subject to underreporting of symptoms by athletes eager to return to play (Meehan and Bachur 2009, Williamson and Goodman 2005, McCrea et al. 2005, McCrea et al. 2004).

Neurocognitive testing involves getting athletes to perform tasks that assess specific functions linked to a particular brain structure or pathway. This type of testing is shown to have large clinical value in sports-related concussion evaluations as the deficits seen immediately following injury, and the time course of their resolution, generally overlap with symptom recovery (McCrory et al. 2013, Peterson et al. 2003, Macciocchi et al. 1996). However, literature questioning the accuracy and reliability of popular test batteries used in concussion evaluation protocols furthers the notion that neurocognitive testing alone cannot be used to determine the status of a concussive injury (Randolph 2011, Van Kampen et al. 2006).

Balance testing has become a key component of comprehensive concussion evaluation protocols due to its ability to objectively assess neurological functioning of the brain, which along with other components of the central nervous system (CNS), is tasked with regulating balance. Immediately follow a concussive injury, athletes display deficits in common postural stability measures (e.g. COP; centre of pressure) that last approximately 3-5 days (McCrea et al. 2003, Guskiewicz 2011). Some researchers have investigated more intensive analyses of postural stability measures, such as approximate entropy (ApEn), which has been able to detect balance deficits up to 10 days post-injury (Peterson et al. 2003, Cavanaugh et al. 2006). These results are significant as at 10 days post-injury other reliable metrics regarding deficits due to concussion (e.g. symptoms, neurocognitive measures) have likely resolved, suggesting that balance assessment may offer a more sensitive indicator of recovery. Unfortunately, many balance assessment methods require large and expensive equipment to operate, limiting their utility on the sideline. Simpler tools such as the current standard for assessing static postural stability in concussed athletes on the sideline, the Balance Error Scoring System (BESS), offer quicker and user-friendly methods to quantify balance. However, these simple balance assessment tools generally suffer from a multitude of issues such as poor test-retest reliability and practice effects, suggesting that more reliable sideline tools must be developed to take advantage of the value balance assessment presents to sports-related concussion evaluations.

Numerous tests and tools have been developed to aid sports-related concussion evaluations, from simple paper-based tests such as the Sport Concussion Assessment Tool (SCAT), to more complex instrumented tests such the computer-based Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) tool. While each method has its own strengths and weaknesses

in the evaluation process, one key trait that differentiates them is their potentiality to be utilized outside of a research laboratory setting and on the sideline.

2.2.1 Sport concussion assessment tool 3 (SCAT3)

The SCAT3 is a standardized comprehensive tool for evaluating and managing potentially concussed athletes on the sideline, however it is also used in other settings to help determine when previously concussed athletes may safely return-to-play. SCAT3 is employed and endorsed by a number of notable sporting organizations such as the International Olympic Committee (IOC), National College Athletics Association (NCAA), Federation Internationale de Football Association (FIFA), International Ice Hockey Federation (IIHF), and the National Football League (NFL). It was developed by a panel of experts during the 4th International Conference on Concussion in Sport (Zurich, November 2012) using modifications of existing concussion assessment tools (McCrory et al., 2013). This simple paper-based test is the third version of the SCAT, and is widely considered to be the current standard for comprehensive sideline concussion evaluations (Patricios et al. 2013). The test takes approximately 15-20 minutes to complete, exercising a comprehensive set of test components, including symptom, physical (e.g. balance deficits), neurocognitive, and behavioural evaluations. Although individual components of the SCAT have been validated in literature, the reliability and validity of the SCAT as a complete tool has yet to be investigated. The majority of literature on the SCAT is currently focused on collecting normative data to determine baseline values within specific populations (Valovich McLeod et al. 2012, Shehata et al. 2009). However, one distinct advantage the SCAT3 presents over other concussion testing tools is that it is a non-instrumented paper-based test, and therefore presents a simple yet standardized method to perform sideline evaluation.

2.2.2 Immediate post-concussion assessment and cognitive testing (ImPACT)

The ImPACT is a computer-administered neurocognitive test battery that takes about 20 minutes to assess specific aspects of cognitive function: attention, memory, processing speed, reaction time, and problem solving (Iverson et al., 2003). ImPACT scores not only correlate well with cognitive deficits and the resolution times associated with sports-related concussions, but also with other traditional neurocognitive testing methods (Schatz et al. 2006, Maerlender et al. 2010). While literature suggest that the ImPACT may provide a useful tool to assess the neurocognitive component of concussive injury, it is concerning that a majority of the investigations supporting its usefulness have been performed by researchers with conflicts of interest (Schatz et al. 2006, Iverson et al. 2002, Partridge and Hall 2013). There is also literature suggesting that the ImPACT possesses poor test-retest reliability leading to a 40% false positive rate (Randolph 2011, Broglio et al. 2007), or that it may actually not be able to detect cognitive abnormalities in concussed individuals (Van Kampen et al., 2006). One advantage ImPACT offers over other concussion evaluation methods is that it can be administered without the presence of medical personnel, however, the requirement of a computer with internet access limits usability on the sideline. Additionally, being a strictly neurocognitive test battery, it lacks the other components required for a proper comprehensive approach suggested by experts, and therefore alone cannot provide sufficient evaluation. In summary, while the ImPACT appears to be a valuable tool to quantify neurocognitive function, it presents a laborious method to assess a small portion of a proper sports-related concussion evaluation.

2.2.3 Neuroimaging

Neuroimaging techniques such as computed tomography (CT) and magnetic resonance imaging (MRI) have been shown to contribute little to the diagnosis and monitoring of concussions (McCrory et al. 2013, Pulsipher et al. 2011). The lack of clinical relevance offered by conventional imaging techniques can mainly be attributed to their results providing estimates of brain structure, whereas concussions are considered to be injuries of impaired neurologic function, and thus do not necessitate structural change. Traditional neuroimaging techniques are however useful in screening for more serious brain injuries caused by the concussive stimulus, such as a fracture to the skull, intracerebral lesion, or brain hemorrhage.

Functional neuroimaging techniques have recently attracted the attention of researchers for their ability to provide estimates of brain function. Functional MRI (fMRI), for example, is an imaging technique that measures brain activity by detecting changes in blood oxygenation caused by neuronal activation. This technique is similar to conventional MRI, but uses the change in magnetization between oxygen-rich and oxygen-poor blood to localize brain activity (Johnston et al., 2001). A number of fMRI studies have been published identifying changes in neuronal function following concussion, however, as with other functional neuroimaging techniques, it is still considered to be in its early stage of development and thus is not recommended for use outside of a research setting (Chen et al. 2004, Jantzen et al. 2004, Slobounov et al. 2011, Talavage et al. 2013, McCrory et al. 2013).

There are numerous other methods to evaluate concussions that I have not mentioned above. While few are clinically validated, others have mixed reviews in literature or no scientific

validation altogether. Although I have mentioned the testing methods of importance, and of relevance to this thesis, researchers and clinicians are continuously establishing new methods and refining old ones as the popularity of the sports-related concussion field grows.

2.3 Human standing balance

In simple terms, static standing balance is described as the process of maintaining the body's center of gravity (COG) within its base of support (BOS), the region bounded by the feet when in contact with a support surface. Postural control is achieved by complex interactions between the sensory systems of the body allowing the central nervous system (CNS) to perceive and predict incoming stimuli, generate necessary motor adjustments, and maintain the body's COG within the BOS. The definition of balance can be extended to include neuromuscular responses to destabilizing events such as externally triggered perturbations (e.g. a shove) or self-initiated changes in posture due to breathing or shifting of body weight. Standing balance is largely controlled automatically without conscious attention: sensory information from the vestibular, somatosensory, and visual systems help the body relay important information regarding the relative position of limbs in space, stimuli that may perturb the body, and the orientation of gravity.

The vestibular system detects linear and angular accelerations of the head from its position in the skull using two sets of high-sensitivity organs: the otoliths and the semicircular canals. Three orthogonal semicircular canals detect angular accelerations via the inertial lag of an enclosed fluid, which in turn activates hair cells allowing information regarding mechanical movement to

be converted into electrical signals and sent to other areas of the body. The otoliths consist of a utricle and a saccule, which are excited when hair cells surrounded by a gelatinous membrane weighted by crystals is displaced by linear accelerations. The tonic activation of hair cells due to the downward pull of gravity allow the otoliths to provide the CNS information regarding the orientation of gravity, even when vision is concealed. Sub-cortical connections between vestibular nuclei and muscles involved in balance allow the vestibular system to initiate quick reflex reactions and modulations of muscle tone via the spinal cord to maintain postural stability. An example is the excitatory pathways from the otolith organs to the extensor muscles of the trunk and limbs via the lateral vestibulospinal tract. This tonic activation of the extensor muscles suggests that this pathway is normally suppressed by descending projections from higher-levels of the CNS, where vestibular system also forms a number of important connections. Higher-level connections with structures such as the cerebellum or thalamus allow vestibular signals can be centrally processed and integrated with sensory information from other areas of the body.

The somatosensory system uses a number of different types of biosensors to enable the CNS to interpret sensory modalities such as touch, temperature, proprioception, and pain. The sensors encompassing the somatosensory system can be found all over the body, from cutaneous mechanoreceptors in glabrous skin which detect mechanical interactions with the environment via pressure to or stretching of skin, to muscle spindles and Golgi tendon organs that sample changes in muscle length and force. A key role of the somatosensory system is to provide the sensory information required for proprioception, the sense of the relative position of the body parts with respect to each other and the surrounding environment. In addition to providing higher levels of the CNS with proprioceptive information, muscle spindles also contribute to the

maintenance of balance by providing the afferent signals required for extremely fast spinal reflexes, such as the monosynaptic stretch reflex that shortens muscles in response to stretch.

Lastly, the visual system enables humans to detect and process visible wavelengths of light, allowing the relay of information regarding the immediate and adjacent environments, as well as our relation to it. This information help the CNS build internal representations of our surroundings, which can be integrated with other sensory information to determine, prepare for, and modulate appropriate responses to external perturbations.

2.3.1 Balance assessment in concussion testing

Literature states that human standing balance is an important unbiased indicator of concussion severity, and its assessment should therefore be used in conjunction with other tests to aid sports-related concussion evaluations (Davis et al. 2009, Guskiewicz 2011, Riemann and Guskiewicz 2000). One advantage balance assessment possesses over other indicators of concussion is that because balance is normally controlled subconsciously in the brain it cannot be easily cheated by athletes eager to return-to-play. Balance assessment, as well as other concussion evaluation indicators, requires a pre-injury (e.g. baseline) score to compare post-injury values due to intrinsic inter-individual differences (McCrory et al., 2013). Regardless, balance assessment is currently incorporated into many prominent sports-related concussion evaluation protocols (Guskiewicz et al. 2001, Johnson et al. 2011, Cavanaugh et al. 2005).

Human standing balance assessment is also used by medical professionals to diagnose and monitor many other conditions such as stroke, Parkinson's disease, multiple sclerosis, ataxia, and

aging (Berg et al. 1995, Jacobs et al. 2006, Cattaneo et al. 2006, Horak 1997). Yet, alarmingly, limited accessibility to high-end technology and unreliability of simple balance assessment tools have rendered many medical professionals unable to accurately and reliably assess human standing balance outside of a research laboratory setting (Clark et al. 2010, Bell et al. 2011). There are numerous methods available to assess human standing balance: from complex methodology such as 3D motion tracking to very simple techniques such as visually observing individuals as they stand. Though each technique has its own advantages and disadvantages, there is currently no available tool to perform accurate, reliable, cost-effective balance assessment on the sideline (Clark et al., 2010).

2.3.2 High-end balance assessment systems

Optical 3D motion tracking systems such as Optotrak® (Northern Digital, Waterloo, Ontario, Canada) can provide precise measurements (±0.1 mm) regarding kinematics of the human body during standing balance. To quantify balance, these systems simultaneously track infrared light-emitting diode (IRED) markers placed on the body using high resolution cameras, and measure movement of these markers with respect to each other in a user-defined coordinate system. While there is currently no literature using optical 3D motion tracking systems to investigate the effects of sport-related concussion on standing balance, studies have used these methods to show impaired locomotion, delayed reaction times, and reduced capacity to perform simple motor tasks following concussive injury (Fait et al. 2009, Eckner et al. 2011). Unfortunately, these systems are currently restricted to state-of-the-art research laboratories due to costs (US$80,000+), space requirements, and lengthy setup protocols (Paloski et al. 2006, Barela et al. 2011).

Researchers have attempted to simplify the procedure of assessing human standing balance by using derivatives of ground reaction forces (GRF; F_x, F_y, F_z) and moments (M_x, M_y, M_z) produced by individuals standing on a force plate (King and Zatsiorsky 1997, Lafond et al. 2004, Caron et al. 1997, Schmitt et al. 2004). This can be done using force plates in conjunction with custom software, such as Swaywin (AMTI Inc., Watertown, MA, USA), or by sampling raw signals and performing post-collection analysis of balance using estimates of postural stability, such as the displacement of the center of pressure (COP; the point of application of the total ground reaction vector; Winter 1995). The horizontal position of the center of mass (COM) can be estimated using a number of methods, such as the zero-point-to-zero-point double integration technique, which takes advantage of the fact that when the horizontal force produced by a subject is zero, the horizontal position of the gravity line (GLP) of the body passes through the COP. This allows for the determination of the GLP and its velocity fairly accurately by integrating the force in the x-direction from one zero point to another (King and Zatsiorsky, 1997). Alternative ways to use force plate measures have emerged such as the dynamic postural stability index, which may provide improved quantification of standing balance and associated sway (Wikstrom et al. 2005). A number of studies have used force plate technology to document deficits in postural stability following sports-related concussion, showing persistently lower COP oscillation randomness and increased sway, even in athletes whose neurocognitive deficits had resolved (De Beaumont et al., 2011, Ingersoll and Armstrong 1992, Geurts et al. 1996). However, as with other high-end systems, high costs ($US10,000+), large equipment, and technological challenges associated with force plates have limited their use to research laboratories. Researchers have attempted to address a number of the limitations of force plates by

using the simpler and more cost-effective Nintendo Wii Fit Balance Board (WBB; Nintendo, Kyoto, Japan) to quantify balance (Clark et al., 2010). Mixed results indicate that further research is required to validate the reliability and efficacy of the WBB system (Wikstrom, 2012). Mainly because sideline use would be impractical with current technology, there is presently little clinical relevance for the use of force plates in the evaluation and management of sport-related concussions on the sideline.

Computerized dynamic posturography (CDP) systems such as the EquiTest® (NeuroCom International, Clackama, OR, USA) use sophisticated moving force plate systems and visual surrounding equipment to test for abnormalities in postural control by altering information sent to the various sensory systems that contribute to standing balance (vestibular, somatosensory, visual). Specifically, the EquiTest® system provides assessment of balance by running the Sensory Organization Test (SOT). This test is designed to tease out the contribution and functionality of the each sensory system by altering the availability of somatosensory and/or visual information while subjects attempt to maintain static equilibrium. The test involves subjects performing six increasingly difficult 20-second balance trials with variations of visual (eyes open, eyes closed, sway referenced vision) and surface-oriented conditions (fixed, sway referenced). Sway referencing refers to the tilting of the support surface and/or visual surround to match the subject's sway. In sway-referenced support conditions, the force plate platform will rotate about 2 axes as the subject sways, thereby maintaining a relatively constant ankle angle with respect to the surface. In sway-referenced visual surround conditions, the surround will not move with respect to the subject's gaze so that the subject experiences minimal optic flow (Riley and Clark, 2003). Following completion of the SOT, an overall composite equilibrium score is

produced, descriptive of the subject's ability to minimize sway and maintain equilibrium under varied conditions. As with traditional force plate methodology, a subject with less sway is considered to have better "balance". Literature investigating the use of the SOT in evaluating the balance of concussed athletes have showed decreased postural stability which returns to baseline levels approximately 3 days following the concussive injury, and to levels associated with similar matched controls after about 10 days post-injury (Peterson et al., 2003). These results indicate that the balance deficits associated with sports-related concussion may be caused due to a sensory integration problem, where concussed individuals are unable to properly use their visual, vestibular, or somatosensory systems (Guskiewicz, 2001). However, as with other high-end technology, the use of CDP technology such as the EquiTest® for the assessment of sport-related concussions on the sideline is limited due to high costs ($US100,000+), large equipment, and technical challenges.

A deterrent for medical professionals to use high-end systems such as 3D motion tracking, force plates, or CDP to assess balance in sports-related concussion evaluations is that there is no general consensus among researchers on which metric should be considered the gold standard to quantify balance. This lack of a gold standard questions the need to perform elaborate, costly, and time-intensive tests in the first place (Panjan and Sarabon, 2010). In response to the issues with high-end balance assessment technology, simpler tools have been developed to assess balance in potentially concussed athletes on the sideline.

2.3.3 Simple balance assessment tools

Simple balance assessment tools provide convenient balance assessment at the expense of rudimentary measures of sway. Systems such as the Swaymeter, Berg Balance Scale (BBS), and Balance Error Scoring System (BESS) are generally more user-friendly than high-end tools: they require minimal training to use and allow users to quickly gather information regarding standing balance. However, the use of simple tools has been called into question due to reports of marginal validity, biased or unreliable scores, practice effects, or dependency upon the environment (Valovich-McLeod et al. 2003, Onate et al. 2007, Barlow et al. 2011, Clark et al. 2010, Bell et al. 2011).

The "Swaymeter", developed by Stephen Lord and colleges, is a simple balance quantification method which uses a metal rod attached to the back of the waist by a firm belt and a pencil on the end of the rod to trace sway patterns on a sheet of millimeter graph paper (see Figure 2.3). The Swaymeter requires a fairly time-intensive analysis, where experimenters will quantify balance using a number of methods such as counting the total number of graph squares traversed, or measuring the maximal sway in anteroposterior (A-P) and mediolateral (M-L) directions and total length of the sway path traversed by the pencil. To address the labour-intensive analysis of the Swaymeter, recent enhancements have replaced the graph paper with a computerized pad, allowing for quick generation of balance measures. While some researchers have suggested that the Swaymeter is a reliable tool for assessing postural stability, others have discounted the use of this system due its basic methodology (Sturnieks et al. 2011, Lord et al. 2003, Hinman et al. 2002). To date, there is no published literature investigating the use of the Swaymeter for sports-

related concussion assessment. As such, it is difficult to gauge the usefulness of this simple system in the sideline concussion evaluations.

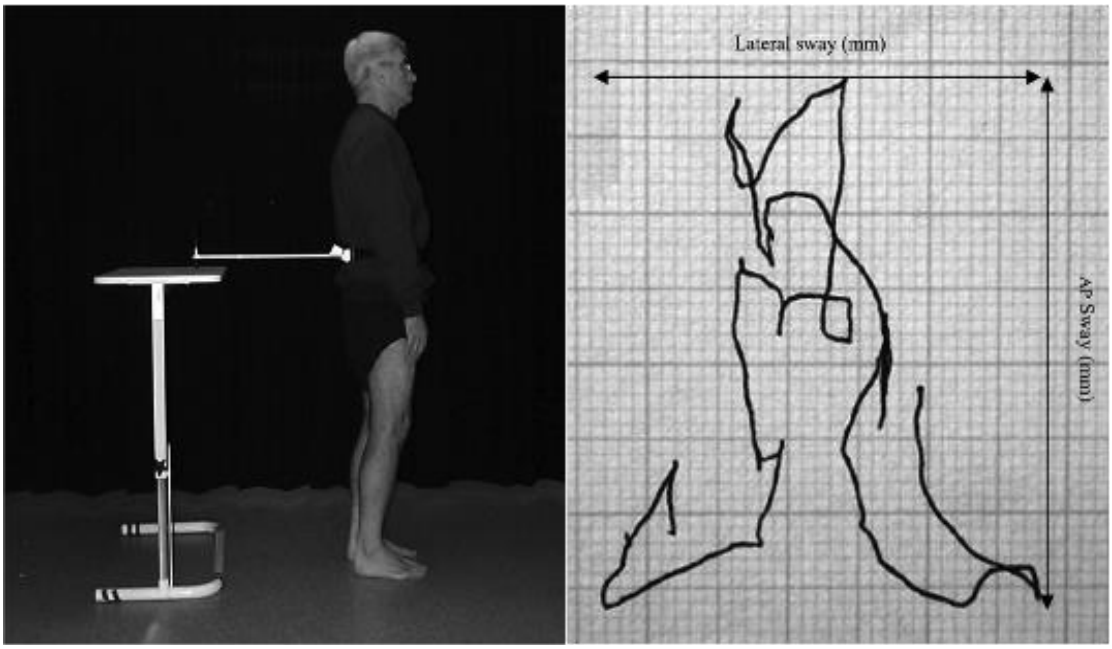


Figure 2.3. Experimental setup of the Swaymeter. The subject has a metal rod attached to the back of his waist by a firm belt, and a pencil on the end of the rod which traces his sway patterns on a sheet of millimeter graph paper. The graph paper shows the sway pattern traced by the pencil during a 30-second eyes closed feet together balance trial, as well as the maximal sway in anteriopostrual (AP) and mediolateral (ML; lateral) directions. Figure 1 and 2 from © Hinman RS, Bennell KL, Metcalf BR, Crossley KM. Balance impairments in individuals with symptomatic knee osteoarthritis: a comparison with matched controls using clinical tests. *Rheumatology*. 2002;41(12):1388-1394. Page 1390. Reproduced with permission from publisher.

Commonly used in stroke patients, the Berg Balance Scale (BBS) is a human scored functional balance test comprising 14 simple balance related tasks (Berg, 1989). The test takes approximately 15-20 minutes to complete and requires a ruler, two standard chairs (one with arm

rests, one without), a footstool, a stopwatch, and a 15 foot walkway. The subject is scored from 0-4 by the experimenter depending on how well, and what degree of assistance that they require to complete each task. The final score is calculated as a sum of the 14 tasks, which categorizes the level of risk the subject is at to experience impairment-related falls (Berg, 1989). To date there is no literature investigating the use of the BBS in sport-related concussion evaluations. While the BBS has shown high test-retest reliability (intra-class correlation (ICC) = 0.97; Conradsson et al. 2007), hindering space (large walkway) and equipment requirements would limit its usability in sideline concussion evaluations.

To provide a more cost-effective and quantifiable method of assessing balance in athletes on the sideline, the Balance Error Scoring System (BESS) was developed by researchers at the University of North Carolina (Guskiewicz, 2001). The BESS involves three balance-testing stances (double-leg, single-leg, tandem) on both a firm and foam surface (see Figure 2.4), and is the current standard for assessing static postural stability in concussed athletes (Valovich McLeod et al. 2012, Hunt et al. 2009). Unrelated clinical research studies have even begun to use the BESS when looking for a simple test to quantify balance (Valovich McLeod et al. 2009, Zammit and Herrington 2005). The BESS has many benefits over high-end balance assessment tools: it takes less than five minutes to complete, is free-to-use, scientifically validated, and can be conducted in any environment. The BESS requires a pen, stopwatch, set of directions, and a scoring sheet. Yet, there is one critical limitation of the BESS: it is scored by the human judgment of a pre-defined set of "balance errors". In general, tests involving the subjective judgment of humans possess a high degree of measurement error, which can lead to inaccurate and unreliable scores. One way measurement error can manifest is scoring differences between

scorers, such as different interpretations and strictness of scoring criteria. Some error criteria of the BESS are easy to implement, such as if the subject opens their eyes during a trial, however other criteria are difficult to judge, such as if the subject flexes their hip beyond a 30 degree angle. While there is literature reporting high reliability of the BESS (ICC = 0.98; Valovich-McLeod et al., 2004), other literature has reported poor reliability both between scorers (ICC = 0.74), and even within scorers judging the same test twice (ICC = 0.57; Finoff et al., 2009). Results have lead researchers to suggest that BESS scores must change by almost 50% for the change to be attributed to a change in balance and not scorer judgment error (minimum detectable change (MDC); Finoff et al., 2009). In fact, it has been suggested that it may be beneficial to create a simpler BESS test that eliminates the subjective scoring criteria, and thus increasing the reliability of the test (Finoff et al., 2009). Another documented limitation of the BESS is strong practice effects, where a study by Valovich McLeod et al. (2003) showed that subjects performing the BESS multiple times over a 7 day period committed less balance errors each session, and committed significantly less errors on day 5 (10.94 ± 2.17) and day 7 (9.44 ± 3.32) than the baseline test (day 1: 12.88 ± 3.34).

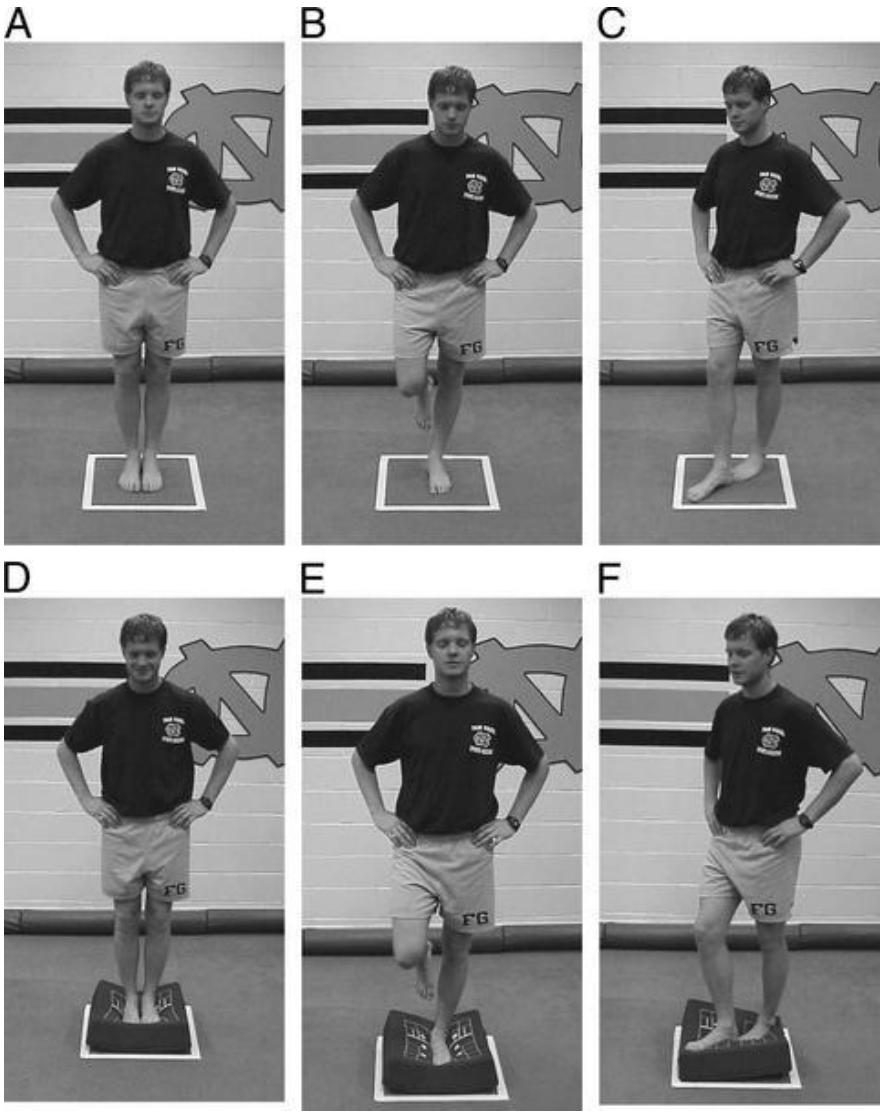


Figure 2.4. The six balance testing conditions of the Balance Error Scoring System (BESS). Figure 3 from © Guskiewicz KM, Ross SE, Marshall SW. Postural Stability and Neuropsychological Deficits After Concussion in Collegiate Athletes. *Journal of Athletic Training.* 2001;36(3):263-273. Page 265. Reproduced with permission from publisher.

A modified version of BESS (mBESS) using only the three stance conditions on the firm surface is currently included in the Sport Concussion Assessment Tool 3 (SCAT3) and the Official NFL Sideline Tool (McCrory et al. 2013, Herring et al. 2011). In addition to providing quicker

assessments of balance, the mBESS removes the requirement of access to a medium-density block of foam, which is needed to perform the full BESS protocol. However, in addition to having no clinical validation in literature, researchers have suggested that the firm surface conditions may not challenge postural stability as well as those performed on the foam surface, leaving the mBESS unable to differentiate between concussed and non-concussed athletes (Valovich McLeod et al. 2005, Hunt et al. 2009).

Simple balance assessment tools offer users quicker and more cost-effective alternatives than high-end technology to assess balance in sideline concussion evaluations. However, while high-end systems suffer from issues such as cost, portability, and questionable outcome measures, simple systems suffer from lack of clinical validation, rudimentary measures, and reliability issues (Valovich McLeod et al. 2003, Hinman et al. 2002, Finoff et al. 2009). While simple balance assessment tools currently offer the most usable options to quantify balance in sports-related concussion evaluations, further research is required to determine the efficacy and reliability of these methods.

Broadly speaking, further research is also required to determine the relationship between balance assessment and sway, as most tools function on the assumption that less sway correlates with higher postural stability (and thus better balance), a theory that is currently debated in literature (Carpenter et al., 2010). BESS, the current standard for assessment of balance in concussed athletes, quantifies the number of "balance errors" rather than sway, and as a result may overcome issues associated with this presumption.

3 Objectives and Hypotheses

The objective of this thesis was to develop and validate a simple objective balance assessment tool that can provide an accurate, reliable, and affordable alternative to the currently available sideline methods. To achieve this, the system will also aim to be portable and easy to adopt by users in the field.

When placed on the body, kinematic sensors offer a rich and objective source of information regarding standing balance and postural stability. As such, the balance assessment system developed in this thesis aimed to use kinematic data collected from the body, rather than the human judgement, to objectively quantify the Balance Error Scoring System (BESS). It is hypothesized that the proposed system will be able to quantify the BESS with greater reliability than current users employing the standardized human-based scoring methods. Additionally, the use of kinematic sensors to produce objective BESS scores (oBESS) will allow this system to quantify BESS at a level of accuracy commonly associated with high clinical reliability (ICC > 0.75).

4 Experiment 1 - Development and Validation of an Objective Balance Error Scoring System (oBESS) in Healthy Subjects

We first performed a feasibility study to assess whether the BESS, the current standard for sideline quantification of balance in concussed athletes, could be quantified using kinematic data collected from the body while subjects performed the standard testing protocol. We chose not to develop our own balance testing protocol due to the inability of previous researchers doing so to gain popularity over the universally accepted BESS. Instead, we opted to allow users to perform the pre-existing BESS protocol and gear our system towards providing an objective quantification alternative to human-based grading methods. While there were a number of laboratory tools available to collect kinematic data during standing balance, such as force plates or wired accelerometer arrays, we determined that in order to achieve our goal of developing a portable sideline tool we needed to perform the first experiment using methods that could operate outside of a research laboratory setting. Customized wireless inertial measurement units (IMUs) housing tri-axial linear accelerometers and angular gyroscopes offered an ideal solution. While rarely used health sciences research, IMUs placed on the body would allow for accurate and cost-effective collection of kinematic information during the BESS. In order to translate IMU data to BESS scores we developed a custom algorithm to identify patterns in the data associated with balance errors, however this method required the true BESS score for each subject in order to properly train the algorithm. To obtain true BESS scores, our "gold standard", we employed experienced BESS raters, a practice common in literature.

4.1 Abstract

Introduction: Limited accessibility to high-end technology and unreliability of simple balance assessment tools has rendered many users unable to accurately and reliably assess human standing balance outside of a research laboratory setting. The goal of this study was to develop and validate a simple objective balance assessment tool that can provide an accurate, reliable, and affordable alternative to the currently available laboratory and clinical methods. Methods: Thirty healthy subjects were filmed performing the six balance testing conditions of the Balance Error Scoring System (BESS). Subjects wore inertial measurement units (IMUs) that measured the linear accelerations and angular velocities from seven landmarks on the body: forehead, sternum, waist, right & left wrist, and right & left shin. Each video was scored by four experienced BESS raters, and mean scores for each subject were used along with IMU data to develop an algorithm allowing the computation of objective BESS (oBESS) scores solely from IMU data. Inter-rater reliability of experienced BESS rater scores, fit of the algorithm to mean experienced BESS rater scores, and the accuracy of algorithm-generated oBESS scores were assessed using intra-class correlations (ICC). Results: Experienced raters displayed low variability in scoring ($ICC_{3,1}$ = 0.91). The oBESS was able to accurately fit mean experienced BESS rater scores ($ICC_{3,1}$ = 0.92) and predict individual BESS scores ($ICC_{3,1}$ = 0.90) using data from only one IMU placed at the forehead. However, using IMU data from the subset of conditions used in popular sport-related concussion protocols, the oBESS was unable to produce scores that accurately fit mean experienced BESS raters ($ICC_{3,1}$ = 0.68). Conclusion: The oBESS can reliably predict total BESS scores in normal subjects. Pending further validation, the oBESS could represent a valid tool to assess balance by offering an objective and reliable alternative to the current scoring methods of the BESS.

4.2 Introduction

Human standing balance is an unbiased indicator of concussion severity (Davis et al. 2009, Guskiewicz 2011, Riemann and Guskiewicz 2000). It has been incorporated into sports-related concussion identification and management protocols used to guide clinical decisions such as return-to-play (Guskiewicz et al. 2001, Johnson et al. 2011, Cavanaugh et al. 2005). However, the limited on-field accessibility to sophisticated equipment and the unreliability of simple sideline tests undermine the clinical utility of balance assessments performed on field (Clark et al. 2010, Bell et al. 2011).

Human standing balance can be assessed in numerous ways, ranging from complex techniques like 3D motion tracking to simple techniques like visually observing individuals as they stand. Optical 3D motion tracking systems and force plates provide precise measurements of the kinematics and kinetics of the human body during standing balance, but are typically restricted to research laboratories due to costs, space requirements, and lengthy setup protocols (Paloski et al. 2006, Barela et al. 2011, Lafond et al. 2004). Simple balance assessment tools such as the Swaymeter, Berg Balance Scale (BBS), and Balance Error Scoring System (BESS) are less precise, but are portable, require minimal training and allow users to quickly gather information regarding standing balance. The use of these simpler balance assessment tools, however, has been called into question due to their marginal validity, biased/unreliable scores, practice effects, and dependency upon the environment (Valovich McLeod et al. 2003, Onate et al. 2007, Barlow et al. 2011, Clark et al. 2010, Bell et al. 2011).

BESS is the current standard for assessing standing balance in concussed athletes on the sideline (Guskiewicz 2011, Valovich McLeod et al. 2012, Hunt et al. 2009). The simplicity of BESS has also led researchers to adopt it when performing studies unrelated to concussion (Valovich McLeod et al. 2009, Zammit and Herrington 2005, Macinnis et al. 2012). BESS consists of counting the total number of pre-defined "errors" a subject makes while balancing using three different stances (two-foot, one-foot, tandem) on two different surfaces (firm, foam). Human judgement of these balance errors introduces variability between raters due to different interpretations and strictness of the scoring criteria. While some literature has reported high reliability for BESS (intraclass correlation coefficient ICC = 0.98; Valovich McLeod et al., 2004), others literature has reported poor reliability both between raters (ICC = 0.74) and within raters grading the same test twice (ICC = 0.57; Finoff et al., 2009). These latter results suggest that BESS scores must change by almost 50% before the difference can be attributed to balance alterations rather than rater judgment error (Finoff et al., 2009).

A modified version of BESS (mBESS) using only the three stance conditions on the firm surface is currently included in the Sport Concussion Assessment Tool 3 (SCAT3) and the Official NFL Sideline Tool (McCrory et al. 2013, Herring et al. 2011). In addition to having no clinical validation in literature, researchers have suggested that using only the firm surface may not challenge postural stability as much as using the foam surface, and as a result mBESS may not differentiate between concussed and non-concussed athletes as well as BESS (Valovich McLeod et al. 2005, Hunt et al. 2009).

The goal of this study was to develop and validate a simple, objective, on-field balance assessment tool that can provide an accurate, reliable and affordable alternative to the currently available laboratory and clinical methods. To achieve this goal, we developed an algorithm to calculate objective BESS scores (oBESS) from kinematic data collected by small wireless sensors worn by players while they perform a regular BESS protocol. We hypothesized that this system would predict BESS scores with a level of accuracy associated with good clinical reliability (ICC > 0.75; Portney and Watkins, 1993) when using sensor data from all six BESS conditions. We also hypothesized that oBESS scores produced using data from the foam surface conditions only would display high correlation with BESS (ICC > 0.75), while those produced using data from the firm surface conditions used by mBESS would not (ICC < 0.75). To optimize the oBESS for sideline use, we selected the "best" algorithm as the one requiring the fewest sensors to accurately predict BESS scores.

4.3 Methods

Subjects

Thirty healthy subjects (15F, 15M) aged 20 to 37 (25.4 years, ± 4.2) participated in the study. Exclusion criteria included neurological or musculoskeletal conditions, respiratory or cardiovascular problems, pregnancy, and the inability to provide informed consent. All subjects gave written informed consent and the study was approved by the University of British Columbia Clinical Research Ethics Board and conformed to the Declaration of Helsinki.

Instrumentation

Inertial measurement units (IMUs) (Shimmer, Realtime Technologies Ltd., Dublin, Ireland) wirelessly collected 6 degree-of-freedom kinematic data (tri-axial linear accelerations and tri-axial angular velocities) sampled at 102.4 Hz, and streamed these data in real time to a desktop computer using a custom LabVIEW program (National Instruments, Austin, TX, USA). IMUs were secured using elastic straps to seven different landmarks: forehead, sternum, anterior waist (below navel), right & left wrist, and right & left shin (see Figure 4.1).

Figure 4.1. Subject performing the two-foot firm surface balance condition of the Balance Error Scoring System (BESS) while wearing inertial measurement units (IMUs) secured to seven landmarks of the body: forehead, sternum, waist, right & left wrist, and right & left shin.

Procedures

Subjects were filmed from the front performing the six standard BESS conditions, i.e., three stances (feet together, one foot, tandem) on two surfaces (firm, foam). The foam pad was medium density and measured 43cm x 43cm x 10cm thick (SunMate foam, Columbia Foam Inc, BC, Canada). Subjects placed their hands on their iliac crests and closed their eyes for all tests (see Figure 4.1). The video clips for each condition were 20s long and the conditions were separated by 30s rest periods to minimize fatigue. Video clips were scored by four experienced raters (15, 20, 25, 150 hours grading experience) from the University of North Carolina at Chapel Hill by counting the number of pre-defined balance errors subjects made during each condition. The balance errors consisted of the following:

Moving the hands off the hips,

Opening the eyes,

Step, stumble, or fall,

Abduction or flexion of the hip beyond 30 degrees,

Lifting the forefoot or heel off the testing surface,

Remaining out of the proper testing position for greater than 5s.

The maximum number of errors per condition was limited to 10, and the total BESS score was the sum of errors committed during all six conditions. If a subject did not maintain the proper stance for at least 5s, or did not otherwise complete the condition, they were given the maximum score of 10. The three stances were performed in order (feet together, one foot, tandem) on the firm surface followed by the foam surface. Prior to each condition, subjects were instructed on

how to perform the stance and verbally given the criteria for each balance error. Once subjects were in the correct stance and comfortably balanced, an auditory tone (750Hz, 100ms duration) signalled the start and end of each 20s condition. This auditory tone was also used to synchronize the IMU data to the video recordings of the balance tests.

Algorithm development

An algorithm was developed to compute objective BESS (oBESS) scores from the IMU data. The algorithm was designed to sum the total number of balance errors committed by the athlete over the duration of the conditions being analyzed, allowing for easy interpretation by users with previous BESS experience. From a general perspective, the IMU data were first sectioned into windows and then the number of windows in which the data exceeded a specified threshold value was summed to generate an oBESS score. Various window lengths, threshold values, number of IMUs and different combinations of data (linear acceleration + angular velocity, $a+\omega$; linear acceleration only, a; and angular velocity data only, ω) were explored to find combinations that yielded the highest ICC values.

All IMU data were first low-pass filtered (5 Hz, $4t^h$ order dual-pass Butterworth). For each condition's 20-s data segment, two resultants were calculated from the tri-axial signals (a_x, a_y, a_z, ω_x, ω_y, ω_z) to yield a linear acceleration signal and an angular velocity signal for each IMU in each condition. Resultant signals were normalized by removing their mean, and were then split into non-overlapping windows varying between one window (20 s long) and 40 windows (each 0.5 s long). Eight thresholds varying from 0.25 × standard deviation (SD) to 2.0×SD in increments of 0.25×SD were considered. Four IMU combinations were investigated: all seven

IMUs, five IMUs (forehead, chest, waist, R & L wrist only), three IMUs (forehead, chest, waist only), and one IMU (forehead only). A raw error score *R* was then defined as the number of windows in which the threshold was exceeded by any IMU included in the analysis during the conditions being analyzed (all, firm only, foam only). A window's error score was binary (1 or 0) and was counted only once even if multiple IMUs exceeded their thresholds within the window.

When subjects could not maintain the testing stance for a minimum of 5 seconds, or otherwise could not complete the condition (an automatic 10 in the standard BESS scoring system), a value of 5 was added to the resultant *R* score for that given condition. Analysis indicated that adding a value of 10 was not needed since part of the balancing behaviour was already incorporated into the data. The oBESS score for a series of conditions was then calculated using the raw error score R and the following equation:

$$\text{oBESS} = c_1R^3 + c_2R^2 + c_3R + c_4 \quad [1]$$

The coefficients (c_1, c_2, c_3, c_4) were calculated using a least squares fit between the mean of the four raters' BESS scores and the raw error scores R for all included IMUs and conditions.

Analysis

Inter-rater reliability was assessed using the intraclass correlation coefficient (ICC) as described by Shrout and Fleiss (1979). The optimal model was selected by maximizing the intraclass correlation coefficient between the oBESS scores and the mean rater BESS scores for every combination (n=3,840) of the four parameters: number of windows (1-40), eight error thresholds

(0.25 to 2.00×SD), three groups of data ($a+\omega$, a, ω), and four combinations of IMUs (7, 5, 3, 1). This entire process was repeated for each combination of conditions (all, firm only, foam only).

The predictive ability of the optimal algorithm using IMU data from all six conditions was then assessed by generating coefficients for equation [1] using data from all but one subjects, and then using these coefficients to predict the missing subject's oBESS score. This "one-by-one" method was repeated for each subject and then the predicted oBESS scores were compared to the mean rater BESS scores using ICC.

For all ICC analyses, the comparisons were considered good if the ICC values were greater than 0.75 and moderate to poor if less than 0.75 (Portney and Watkins, 1993). All analyses were conducted using MATLAB (Version R2012a, The MathWorks Inc, Natick, MA) and, where needed, statistical significance was set to $p = 0.05$

4.4 Results

Data from one subject was removed from the study because the subject balanced on the incorrect foot during one of the six conditions. Analyses were performed using the remaining 29 subjects. The four raters showed little variance in their total BESS scores across all subjects ($ICC_{3,1}$ = 0.91), however they were less consistent when grading conditions performed on the firm surface ($ICC_{3,1}$ = 0.82) than the foam surface ($ICC_{3,1}$ = 0.95; mBESS). Subjects committed an average of 9.78 ± 7.11 balance errors: 3.17 ± 3.55 on the firm surface, and 6.62 ± 4.13 on the foam surface.

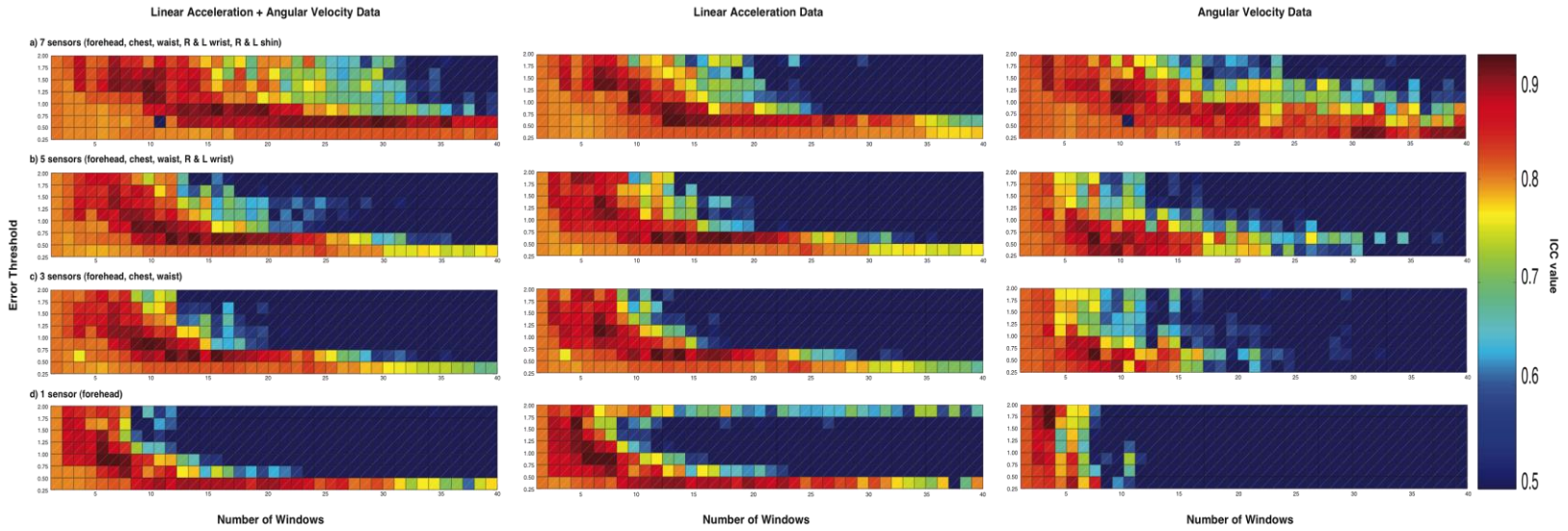


Figure 4.2. Colorbar charts representing the intraclass correlation coefficient of the fit between the mean rater Balance Error Scoring System (BESS) scores and the objective Balance Error Scoring System scores (oBESS) scores generated using IMU data from all six balance conditions and every possible combination of model parameters (n=3,840). Each square represents a combination of four algorithm parameters; number of windows (1-40), error threshold (0.25-2.00 x SD), type of data ($a + \omega$, a, ω), and number of sensors (7, 5, 3, 1). All ICC values less than or equal to 0.50 are shown as the same dark blue colour.

Using all six BESS conditions, many different combinations of parameters produced oBESS scores with a good fit to the mean rater scores ($ICC_{3,1}$ > 0.75; see Figure 2). The graded pattern of ICC values within each panel of Figure 2 showed that the algorithms were relatively insensitive to small changes in the number of windows and the threshold values used. The similarity in the pattern of ICC values between the different panels of Figure 2 showed that the algorithms were similarly insensitive to the groups of data and number of IMUs used.

The algorithm with the best fit to the mean rater scores ($ICC_{3,1}$ = 0.94) had 11 windows (1.8 s long), a 0.50×SD error threshold, and used both linear acceleration and angular velocity data from five IMUs (forehead, chest, waist, L & R wrist). Using only one IMU at the forehead, the algorithm that best fit the mean rater scores ($ICC_{3,1}$ = 0.92) had 4 windows (5 s long), a 1.50×SD error threshold, and relied on linear acceleration data only (Table 1A). This latter simpler algorithm was able to accurately predict individual BESS scores using the "one-by-one" validation method ($ICC_{3,1}$ = 0.90; see Table 1B).

A.

Threshold (xSD)	Number of Windows: 2	3	4	5	6
1	0.8844	0.8890	0.9048	0.9162	0.8891
1.25	0.8861	0.8869	0.9065	0.8698	0.8225
1.5	0.8897	0.8688	0.9243	0.8060	0.7371
1.75	0.8402	0.8199	0.8697	0.8463	0.7652
2	0.8041	0.8709	0.8061	0.7323	0.7909

B.

Threshold (xSD)	Number of Windows: 2	3	4	5	6
1	0.8269	0.8172	0.8391	0.8846	0.8535
1.25	0.8301	0.8323	0.8667	0.8269	0.7675
1.5	0.8268	0.7643	0.8973	0.7512	0.6575
1.75	0.7357	0.6575	0.8192	0.7585	0.6767
2	0.6765	0.8066	0.7136	0.6028	0.6957

Table 4.1. A: Intraclass correlation (ICC3,1) values for the fit of oBESS scores generated using data from a single forehead-mounted inertial measurement unit (IMU) and all six balance conditions to the mean rater BESS scores. The 'best" algorithm (white square) and surrounding adjacent window numbers and error thresholds are shown. B: ICC3,1 values for the comparison between the one-by-one predictions and the mean rater BESS score using the same single-IMU algorithm.

The best algorithm for using IMU data from the subset of conditions performed on the firm surface to calculate total BESS was unable to produce oBESS scores that accurately fit the mean

experienced raters ($ICC_{3,1}$ = 0.68), while the best algorithm for using IMU data from the subset of conditions performed on foam was ($ICC_{3,1}$ = 0.89).

4.5 Discussion

This study showed that it is possible to objectively predict BESS scores from kinematic information collected from the body while subjects perform the standard BESS test. Reliability values suggest that when using data from all six standard BESS conditions, the oBESS can produce scores that accurately fit mean rater BESS scores ($ICC_{3,1}$ = 0.92), and also accurately predict individual BESS scores in normal healthy subjects ($ICC_{3,1}$ = 0.90). These results indicate that oBESS is a valid measure of balance and may offer an objective alternative to the current laboratory and clinical balance assessment methods.

The oBESS required linear acceleration data from only one IMU placed at the forehead to accurately and reliably quantify balance errors. This finding provides the basis for a simple (one sensor), inexpensive (no angular rate sensors required) and portable balance assessment tool. Combined with a blue-tooth enabled smart phone, the resulting system would be convenient to use to objectively measure BESS scores on the sideline.

Sensors at locations other than the forehead did not greatly improve the fit to mean rater scores or predictive ability of the algorithm. This finding suggests that the extra sensors contained either redundant balance information or information that did not correlate with the BESS scores. From a biomechanics perspective, a single sensor located at the forehead makes sense during standing

balance since the head is at the distal end of the kinematic chain and would therefore capture—and perhaps even amplify—the balance error motions of the intermediate body segments.

Interpretations of ICC values as a measure of test-retest reliability vary considerably in a clinical setting. While some literature suggests that a value of 0.60 is the minimum acceptable value for a reliable clinical test (Anastasi, 1988), others argue that values must be greater than 0.90 to make decisions regarding an athlete's cognitive status following a sports-related concussion (Randolph et al., 2005). Literature investigating the reliability of the BESS have suggested that 0.90 is probably too stringent for the BESS, and recommended an ICC of at least 0.75 be considered reliable (Finoff et al. 2009, Portney and Watkins 1993, Broglio et al. 2007). Further work is needed to evaluate the test-retest reliability of the oBESS algorithm developed here.

The inter-rater reliability of BESS raters varied with previous values reported in literature. Our four raters showed little variance in their BESS scores ($ICC_{3,1} = 0.91$), which differed from a recent study that reported an ICC value between experienced BESS raters of 0.57 (Finoff et al., 2009). This difference between our results and those of Finoff et al (2009) may be due to our use of normal healthy subjects instead of athletes, or the lower number of total balance errors in our study (9.78) compared to this prior work (15.1). Normal healthy subjects tend to have worse postural control than athletes, likely resulting in balance errors being more obvious and easier to detect by raters (Davlin, 2004). Lower total balance errors may also minimize inter-rater grading differences due to less reliance on subjective judgement of balance error criteria strictness, including subjective decisions to assign a maximum score (10) for conditions when the subject cannot properly complete the trial. While some balance error criteria of the BESS are easy to

implement, such as if the subject opens their eyes during a stance, others are vague or difficult to judge, such as if the subject flexes their hip beyond a 30 degree angle. In fact, Finoff et al. (2009) have suggested that it may be beneficial to create a simpler BESS test that eliminates subjective errors, and thus increasing the reliability of the test.

The modified version of the BESS (mBESS) used in the SCAT3 protocol relies only the three BESS conditions performed on a firm surface. Our subjects generated fewer balance errors (BESS score = 3.17) in the three firm surface conditions than in the three foam surface conditions (BESS score = 6.62). This finding suggests that mBESS may not challenge standing balance enough to differentiate balance behaviour between subjects (Hunt et al. 2009). Based on similar logic, Valovich McLeod et al. (2012) suggested that the BESS conditions performed on the foam surface (instead of the firm surface) should be considered for future versions of the mBESS. This consideration is supported by our analysis, which showed that oBESS scores generated using data from the foam trials only fit the mean rater BESS scores ($ICC_{3,1}$ = 0.89) better than those generated using data from the firm trials only ($ICC_{3,1}$ = 0.68).

Our study was limited to healthy normal subjects and resulted in relatively low balance error scores (mean = 9.78 ± 7.11). Higher scores have been reported in individuals with sport-related concussion (5.81 ± 6.49 errors above baseline; McCrea et al., 2003) and therefore additional work is needed to validate the current algorithm using subjects with higher BESS scores. Another potential limitation of our study is the inability of the algorithm to objectively identify when subjects opened their eyes. In our data, this type of balance error often occurred simultaneously with other errors (such as putting the elevated foot down during two-legged

stance) and in these instances was potentially captured in the sensor data. Even though our sensors could not detect eye opening, the good correlation with rater BESS scores suggests that the algorithm may have made up for this deficiency by relying on more detailed kinematic data that would be ignored by the raters. And finally, our algorithm relied on a manual addition of 5 error points for non-completion of a trial. This manual input could be incorporated through a handheld device used to report the results of the balance assessment.

4.6 Conclusion

In summary, we have developed and validated an algorithm to objectively measure BESS using a single inertial sensor worn on the forehead. Objectifying the BESS test minimizes the variability introduced by human judgement and generates the same oBESS score regardless of who is administering the test (athletic trainer, team doctor, clinician, coach, or parent). Our findings also suggest that a modified BESS protocol of only three BESS conditions can be used, but that these three conditions should be on the foam surface rather than a firm surface. Further research is required to optimize to oBESS for use in clinical populations, but the present results indicate that the oBESS has the potential to replace the current human-scored BESS test.

5 Experiment 2 – Reliability of the Objective BESS (oBESS) in Subjects with Induced Postural Instability

Although Experiment 1 suggested that oBESS may present a valid and reliable tool to assess BESS scores objectively, we were only able to comment on its ability to do so in normal healthy subjects with low number of balance errors. The potential of this system, and its value to sideline concussion assessment, provided the motivation to perform a second study aimed at evaluating the reliability of the oBESS when individuals commit a greater number of balance errors. Rather than testing patients with documented balance deficits (such as concussed athletes), we opted to create artificial postural instability in normal healthy subjects, which would allow us to capture the required data corresponding to a greater number of balance errors.

To artificially increase the number of balance errors, we chose to expose subjects to simulated high altitude in a hypoxic chamber, a documented stressor to postural stability (Holness et al. 1982, Wagner et al. 2001, Cymerman et al. 2004). Hypoxia is thought to induce balance deficits by reducing the availability of oxygen to the central nervous system (CNS), thereby affecting the balance centers. When the supply of oxygen is inadequate, improper functioning of the CNS leads to impaired neuromuscular coordination and ataxia, thus resulting in a decrease in postural stability (Wagner et al. 2001, Cymerman et al. 2001, Fraser et al. 1987, Holness et al. 1982). It has even been suggested that balance assessment tests, such as the BESS, could be a useful adjunct for the diagnosis of acute mountain sickness (Macinnis et al., 2012).

5.1 Abstract

Introduction: Unreliability of human-generated scores prevents the Balance Error Scoring System (BESS) from providing users with accurate information regarding standing balance. In Experiment 1, accurate and reliable objective BESS (oBESS) scores were generated in healthy subjects using linear accelerations collected from a single kinematic sensor placed at the forehead. The goal of the present study was to evaluate the reliability of oBESS at higher total error scores characteristic of concussed populations. Methods: Twenty healthy subjects wore a network of inertial measurement units (IMUs) and were filmed serially performing twelve BESS tests in a hypoxic altitude chamber, aimed at increasing the number of balance errors. Peripheral blood-oxygen saturation (SpO_2), heart rate (HR), and Lake Louise Score (LLS) were collected following each test to assess acclimation to altitude. All BESS tests were scored by three experienced raters and two athletic trainers. Similarly to Experiment 1, experienced rater scores were used along with IMU data to develop an algorithm to compute objective BESS (oBESS) scores. Results: Experienced raters displayed low inter-rater ($ICC_{3,1}$ = 0.75) and intra-rater reliability ($ICC_{3,1}$ = 0.77, 0.25, 0.75). As such, analyses were performed only using trials where raters displayed marginal scoring differences. While SpO_2, HR, and LLS displayed characteristic responses, mean BESS did not increase in response to simulated altitude. Athletic trainers displayed low inter-rater ($ICC_{3,1}$ = 0.59) and intra-rater reliability ($ICC_{3,1}$ = 0.68, 0.59), and their scores were unable to accurately fit mean experienced BESS scores ($ICC_{3,1}$ = 0.06, 0.06). Using all data, the oBESS was able to fit mean experienced BESS scores with greater accuracy than the two athletic trainers ($ICC_{3,1}$ = 57), but not at a level commonly associated with high clinical reliability (ICC ≥ 0.75). However, if using data where raters displayed a consensus on the number of errors (n=60), the oBESS was able to produce scores with good fit to mean

experienced rater scores ($ICC_{3,1}$ = 0.84). Conclusion: Human raters of the BESS, even if considered experienced, may be unable to produce reliable BESS scores or a suitable gold standard to train other quantification methods.

5.2 Introduction

Human standing balance is an important unbiased indicator of concussion severity (Davis et al. 2009, Guskiewicz 2011, Riemann and Guskiewicz 2000). Sensitivity of standing balance to concussive injuries has led experts to incorporate balance testing into sideline concussion evaluation protocols used to guide medical decisions, such as return-to-play (Guskiewicz et al. 2001, Johnson et al. 2011, Cavanaugh et al. 2005). However, the unreliable scoring methods employed by the Balance Error Scoring System (BESS), the current standard for assessing balance in concussed athletes on the sideline, may undermine the clinical utility of sideline balance assessments (Finoff et al., 2009).

The BESS is a simple human-graded balance test that entails graders summing the number of pre-defined "balance errors" a subject commits while they perform three balance testing stances (two-foot, one-foot, tandem) on two different surfaces (firm, foam). The human judgement involved in detecting these balance errors introduces variability between raters due to different interpretations and strictness of the scoring criteria. While some studies have reported high reliability for the BESS (intraclass correlation coefficient ICC = 0.98; Valovich McLeod et al. 2004, 0.91; Experiment 1), others have reported poor inter-rater (ICC = 0.57) and intra-rater

(ICC = 0.74) reliability (Finoff et al., 2009). In Experiment 1, we demonstrated that the oBESS system was able to generate scores with both an accurate fit to mean experienced BESS raters (ICC = 0.92) also the ability to predict individual BESS scores (ICC = 0.90) using linear accelerations collected from a single inertial measurement unit (IMU) placed at the forehead. However, the use of exclusively normal healthy subjects with low postural instability minimized the mean number of balance errors committed (9.8 ± 7.1). Consequently, we were unable to comment on the accuracy of the oBESS in individuals that commit a higher number of balance errors, such as normally seen in athletes following a sports-related concussion (5.8 ± 6.5 errors above baseline; McCrea et al., 2003). The use of strictly experienced raters also limited comparisons with actual real-world users of the BESS, such as athletic trainers, leaving us unable to gauge the true clinical utility of the oBESS.

The goal of this study was to evaluate the ability of the oBESS to produce accurate scores in subjects who commit a greater number of errors during the BESS than those observed in Experiment 1. To achieve this goal, we attempted to increase the number of balance errors subjects commit while they perform the BESS by exposing them to an environment of reduced oxygen (i.e., hypoxia), a documented stressor to postural stability (Wagner et al. 2011, MacInnis et al. 2012). Using similar methodology to the first study, we developed an algorithm to generate oBESS scores using data from IMUs worn by subjects as they performed the BESS. We hypothesized that the oBESS would be able to produce scores at a level of accuracy associated with good clinical reliability (ICC $\geq$ 0.75; Portney and Watkins, 1993). In addition to our gold standard, experienced raters, we employed two athletic trainers with previous scoring experience

to rate each test allowing for comparison between oBESS with actual field users of the test. We hypothesized that the oBESS would be able to produce scores with greater accuracy than athletic trainers (ICC oBESS > ICC trainer) due to limitations in human-scoring of the BESS.

5.3 Methods

Subjects

Twenty healthy subjects (10F, 10M) aged 19 to 31 (23.3 years, ± 3.2) participated in the study. Exclusion criteria included neurological or musculoskeletal conditions, respiratory or cardiovascular problems, pregnancy, and the inability to provide informed consent. All subjects gave written informed consent and the study was approved by the University of British Columbia Clinical Research Ethics Board and conformed to the Declaration of Helsinki.

Instrumentation

Inertial measurement units (IMUs) (Shimmer, Realtime Technologies Ltd., Dublin, Ireland) wirelessly collected 6 degree-of-freedom kinematic data (tri-axial linear accelerations and tri-axial angular velocities) sampled at 102.4 Hz, and streamed these data in real time to a Lenovo ThinkPad tablet (Lenovo Group Ltd, Beijing, China) running a custom Android application. IMUs were secured to the body of each subject using elastic straps on seven different landmarks: forehead, sternum, waist, right & left wrist, and right & left shin.

Procedures

A practice BESS test prior to beginning the experimental protocol was performed by subjects to familiarize them with test conditions and procedures. Subjects were then filmed from the front performing a series of twelve BESS tests. Each BESS test consisted of the six standard BESS conditions, i.e., three stances (feet together, one foot, tandem) on two surfaces (firm, foam). Conditions were performed for 20s and were separated by 30s rest periods to minimize fatigue. Foam conditions were performed on a balance pad (10" x 10" x 2.5", Airex Balance Pad 81000, Power Systems Inc, Knoxville, TN, USA) aiming to create a more challenging balance task. Subjects placed their hands on their iliac crests and closed their eyes for all conditions.

The first and last BESS tests were performed in a normoxic controlled laboratory setting, while tests 2-11 were performed at a simulated high altitude (up to 4,500 m) inside a hypoxic altitude chamber in the UBC Environmental Physiology Laboratory. Simulated high altitude was aimed to increase the mean number of balance errors normally seen in healthy subjects. Subjects were exposed to 5 hours of simulated altitude in total: 4500m (3 hours), 3000m (1 hour), and 1500m (1 hour; see Figure 5.1). BESS tests were separated by 30 minute rest periods during which experimenters collected measures to assess the acclimation of the subjects to altitude: peripheral blood-oxygen saturation (SpO_2), heart rate (HR), and Lake Louise Score (LLS). SpO2 (%) and HR (beats per minute; bpm) were assessed with the subject sitting in an upright chair via pulse oximetry using a clip placed over the right index finger (Near Infra-Red Spectroscopy; Nonin GO_1 LCD, Nonin Medical Inc., USA), while LLS was assessed using a standardized verbal questionnaire regarding perceived physical changes (see Appendix D).

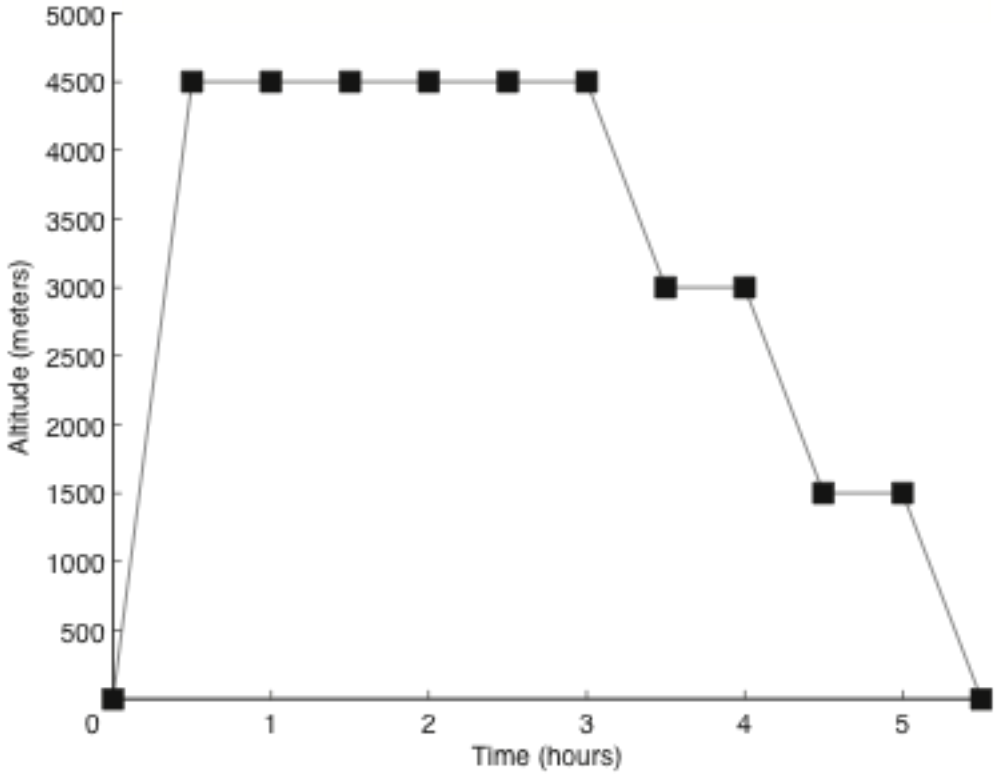


Figure 5.1. Experimental protocol. Subjects were exposed to 5 hours of simulated altitude: 4500m (3 hours), 3000m (1 hour), and 1500m (1 hour). Each black box represents a serially performed Balance Error Scoring System (BESS) test by subjects, which were separated by 30 minutes rest periods.

Video clips were scored by three experienced raters (18.5, 25 & 60 hours grading experience) from the University of North Carolina at Chapel Hill and two experienced athletic trainers (100 & 120 on-field concussion evaluations) from the UBC Department of Athletics. Experienced rater values were screened for consistency and their mean was used as the gold standard, indicating the actual BESS score for each test. Accordingly, to maximize the accuracy of our gold standard, experienced raters were told they were allowed to stop, rewind, or re-watch videos as many times as they would like to obtain the correct BESS score for each video. Athletic trainers, however, watched each video once from beginning to end without stopping, to simulate real-world sideline

administrations of the BESS. Using the standardized scoring methods for the BESS, the pre-defined balance errors consisted of the following:

a. Moving the hands off the hips,
b. Opening the eyes,
c. Step, stumble, or fall,
d. Abduction or flexion of the hip beyond 30 degrees,
e. Lifting the forefoot or heel off the testing surface,
f. Remaining out of the proper testing position for greater than 5s.

The maximum number of errors per condition was limited to 10, and the total BESS score was the sum of errors committed during all six conditions. If a subject did not maintain the proper stance for at least 5s, or did not otherwise complete the condition, they were given the maximum score of 10. The three stances were always performed in the same order (feet together, one foot, tandem) on the firm surface followed by the foam surface. Prior to each condition, subjects were instructed on how to perform the stance and verbally given the criteria for each balance error. Once subjects were in the correct stance and maintained balance comfortably, an auditory tone (750Hz, 100ms duration) triggered by the experimenter signaled the start and end of each 20s condition. This auditory tone was also used to synchronize IMU data with the video recordings, where tones in the video files coincided with the beginning and end of IMU data steams.

Two videos from each subject were randomly chosen to be repeated, where all raters (experienced raters and athletic trainers) unknowingly graded the same video twice. These

repeats were randomly arranged in the sequence of videos allowing for the investigation of intra-rater reliability by comparing scores raters provided when re-grading the same videos.

Algorithm development

Similar to methods in Experiment 1, an algorithm was developed to compute objective BESS (oBESS) scores from linear accelerations and angular velocities collected by the IMUs while subjects performed the balance test. The algorithm was designed to sum the total number of balance errors committed by the subject during the six conditions of the BESS, allowing for easy interpretation by users with previous BESS scoring experience. Simply put, IMU data were first sectioned into windows and then the number of windows in which the data exceeded a specified threshold value was summed to generate an oBESS score. Various window lengths, threshold values, number of IMUs and different combinations of data (linear acceleration + angular velocity, $a+\omega$; linear acceleration only, a; and angular velocity data only, ω) were explored to find combinations that yielded the highest intraclass correlation (ICC) values with mean experienced rater scores.

All IMU data were first low-pass filtered (5 Hz, 4th order dual-pass Butterworth). For each 20s data segment, two resultant vectors were calculated from the tri-axial signals (a_x, a_y, a_z, ω_x, ω_y, ω_z) to yield a linear acceleration magnitude signal and an angular velocity magnitude signal for each IMU. Resultant signals were normalized by removing their mean, and were then split into non-overlapping windows varying between one window (20s long) and 40 windows (each 0.5s long).

Eight thresholds varying from 0.25×standard deviation(SD) to 2.0×SD in increments of 0.25×SD were considered. Four IMU combinations were investigated: all seven IMUs, five IMUs (forehead, chest, waist, R & L wrist only), three IMUs (forehead, chest, waist only), and one IMU (forehead only). A raw error score *R* was then defined as the number of windows in which the threshold was exceeded by any IMU included in the analysis during the six conditions of the BESS. A window's error score was binary (1 or 0) and was counted only once even if multiple IMUs exceeded their thresholds within the window.

When subjects could not maintain the testing stance for a minimum of 5 seconds, or otherwise could not complete the condition (an automatic 10 in the human-scored BESS), a value of 5 was added to the resultant *R* score for that given condition if any of the raters involved in the analysis scored a 10. Analysis indicated that adding a value of 10 to the *R* score was not needed since part of the balancing behaviour was already incorporated into the data. The oBESS score for the total BESS was then calculated using the raw error score *R* and the following equation:

$$\text{oBESS} = c_1R^3 + c_2R^2 + c_3R + c_4 \qquad [1]$$

The coefficients (c_1, c_2, c_3, c_4) were calculated using a least squares fit between the mean BESS scores given by the raters included in the analysis and the raw error scores *R* for all included IMUs and conditions.

Analysis

Inter-rater and intra-rater reliability of raters was assessed using the intraclass correlation coefficient (ICC) as described by Shrout and Fleiss (1979). This statistical test is a modified intraclass (Pearson) correlation that is commonly used to assess the reproducibility of quantitative measures, such as BESS scores, made by different observers measuring the same quantity (Finoff et al. 2009, Valovich McLeod et al. 2004). For all ICC analyses, comparisons were considered good if the ICC values were greater than 0.75 and moderate to poor if less than 0.75 (Portney and Watkins, 1993). All analyses were conducted using MATLAB (Version R2012a, The MathWorks Inc, Natick, MA) and, where needed, statistical significance was set to $p = 0.05$.

The first step was to assess the reliability of our gold standard, the mean experienced rater scores. To minimize uncertainty in the system during algorithm training and reliability analyses, experienced BESS raters were removed from the analysis if they displayed moderate to low intra-rater reliability during the repeated videos ($ICC \leq 0.75$), indicating inconsistent scoring. Following this, individual BESS tests were removed from the analysis if scores fell outside of the 95% confidence bounds (1.96 standard deviations) for the line of best fit between the experienced raters, indicating different interpretations of the number of balance errors committed by the subject during that test. Using the remaining data, the mean number of balance errors committed by all subjects was first compared with Experiment 1 using a t-test. Significant difference in mean experienced rater BESS, SpO2, LLS, and HR between the 12 tests were assessed using one-way

repeated measures ANOVAs. Decomposition of the main effect (to identify differences between tests) was evaluated using post-hoc Tukey HSD methods. Statistical significant was set at $p < 0.05$.

The optimal model to generate oBESS scores was selected by maximizing the intraclass correlation coefficient between the oBESS scores and mean experienced rater BESS scores for every combination (n=3,840) of the four parameters: number of windows (1-40), eight error thresholds (0.25 to 2.00×SD), three datasets ($a+\omega$, a, ω), and four combinations of IMUs (7, 5, 3, 1). Accuracy of fit between oBESS scores generated using the optimal model to our gold standard (mean experienced rater scores), and individual athletic trainer scores to this gold standard were investigated using ICC comparisons. Accuracy of fit between oBESS scores and BESS scores using data where experts displayed consensus or marginal differences in scoring (max ±5 errors difference) were also investigated using ICC comparisons.

5.4 Results

A single BESS test from four subjects was removed from the analysis due to either a video recording from one the six conditions being corrupt, or a data collection issue with the IMUs. One subject withdrew during the experimental protocol, and therefore only BESS1-5 could be used during analysis. Another subject's data was completely removed due to incorrect placement of IMUs. In total, 33 BESS tests were removed, and analyses proceeded using the remaining 217 BESS tests.

Over the repeated videos, two experienced raters displayed good intra-rater reliability ($ICC_{3,1}$ = 0.77, 0.75), while one displayed low intra-rater reliability ($ICC_{3,1}$ = 0.25). Data from this unreliable experienced rater were removed from further analyses (see Figure 5.2). Of the 217 valid BESS tests, 7 were removed as individual experienced rater scores fell outside of the 95% confidence interval bounds calculated using the fit of the data between the two remaining experienced raters (see Figure 5.3). Analyses proceeded with the remaining 210 BESS tests (inter-rater $ICC_{3,1}$ = 0.75).

Figure 5.2. Correlation (intraclass correlation = 0.77, 0.25, 0.75, respectively) between scores given by experienced raters for repeated videos (n=40) of subjects performing the Balance Error Scoring System (BESS). Each dot represents a single BESS test (n=40), and the solid line represents the line of best fit.

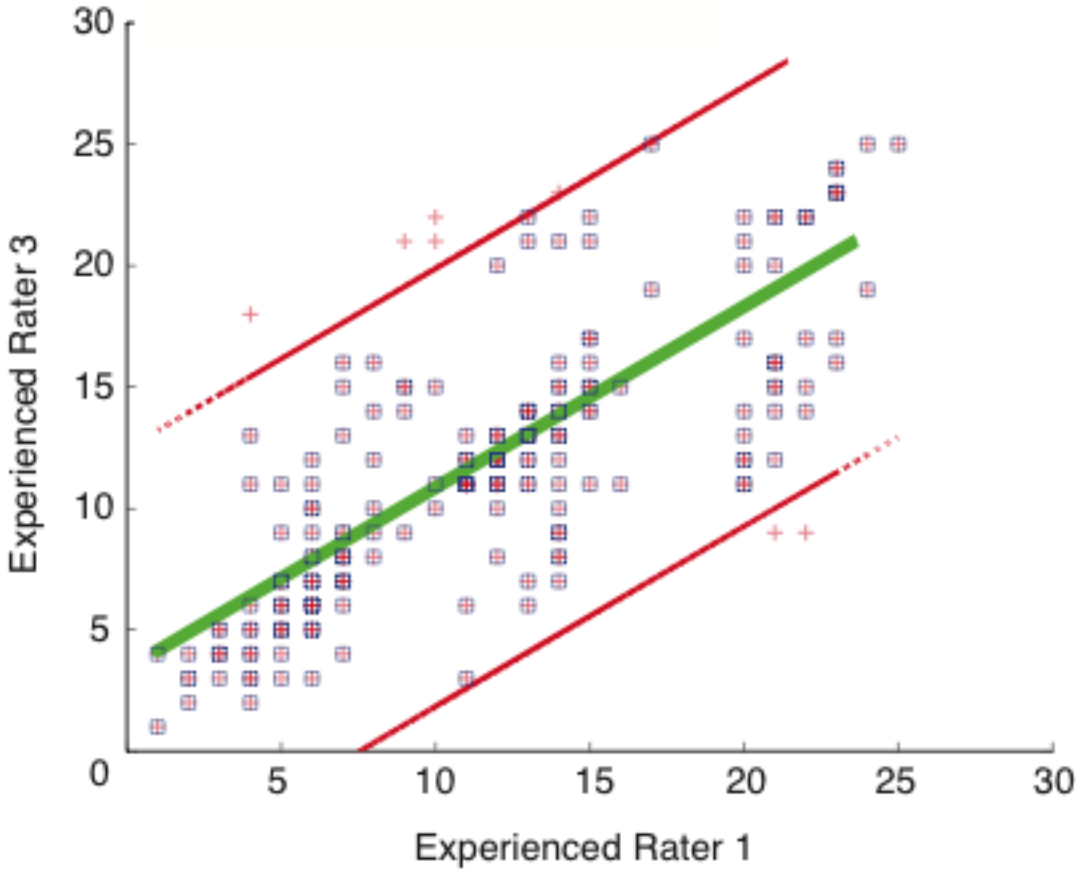


Figure 5.3. Balance Error Scoring System (BESS) scores of experienced raters with good intra-rater reliability (intraclass correlation ICC ≥ 0.75). Red crosses represent individual BESS trials (n=217) from all subjects. The green line represents the line of best fit between the two experienced raters. The red lines represent the 95% confidence interval bounds about this line of best fit, and trials outside of the bounds (n=7) were removed from the analysis. All trials selected for analysis (n=210) are surrounded with a blue box.

SpO_2 significantly decreased when subjects were exposed to simulated altitude ($F = 126.66$, $p < 0.001$, see Figure 5.4) and was at least 21% below baseline for all 6 measures collected at 4500m. Conversely, LLS significantly increased at altitude ($F = 0.944$, $p = 0.499$, see Figure 5.5), however significant differences were not observed between baseline and any of the 12

measures collected. HR also significantly increased at altitude ($F = 9.40$, $p < 0.001$, see Figure 5.6), with 5 of the 6 measures collected at 4500m displaying significant differences to baseline. Mean BESS was consistent over tests performed at simulated altitude, however significant difference was seen between the first (baseline) and final BESS tests, which were both performed in normoxia ($p = 0.003$, see Figure 5.7). Overall, subjects committed significantly more balance errors during the BESS (11.8 ± 5.8) than seen in Experiment 1 (9.78 ± 7.11, $p = 0.046$), with a range of 1 to 37.5 errors.

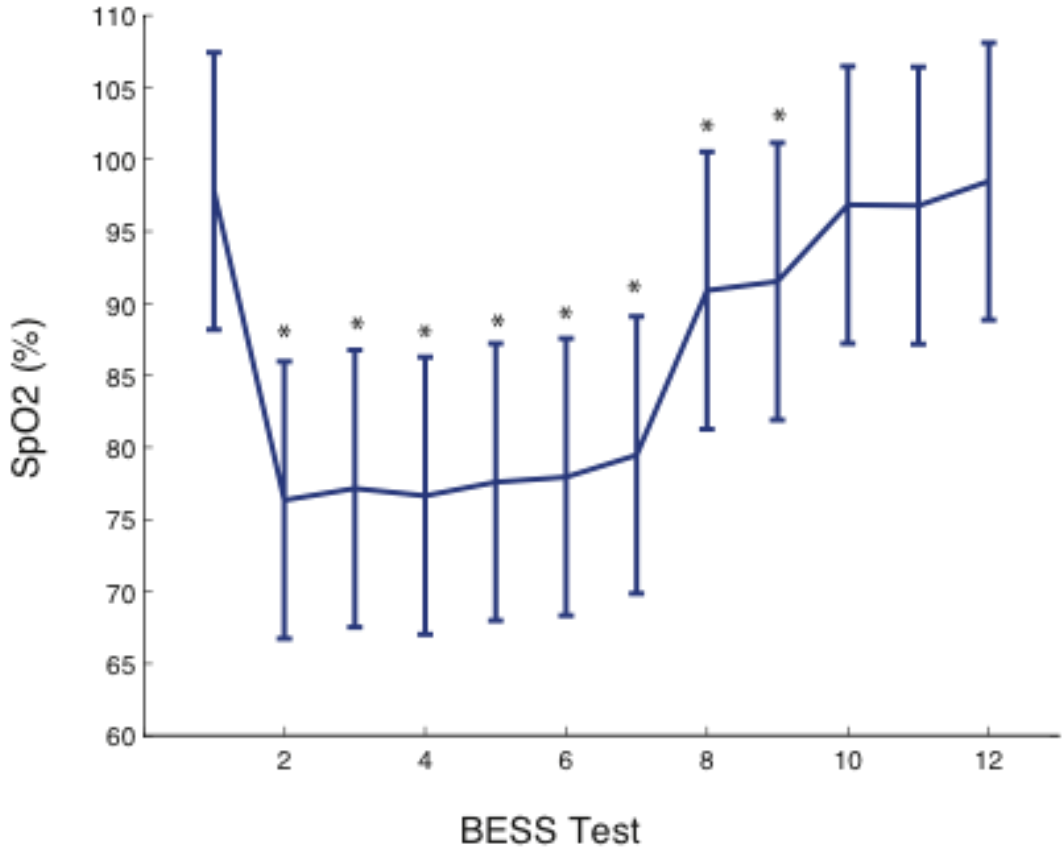


Figure 5.4. Mean peripheral blood-oxygen concentration (percent; SpO_2) in subjects over 12 Balance Error Scoring System (BESS) tests performed serially in a hypoxic altitude chamber. Asterisks indicate significant difference from the first reading (baseline; BESS 1) taken in normoxia. Vertical lines represent error bars.

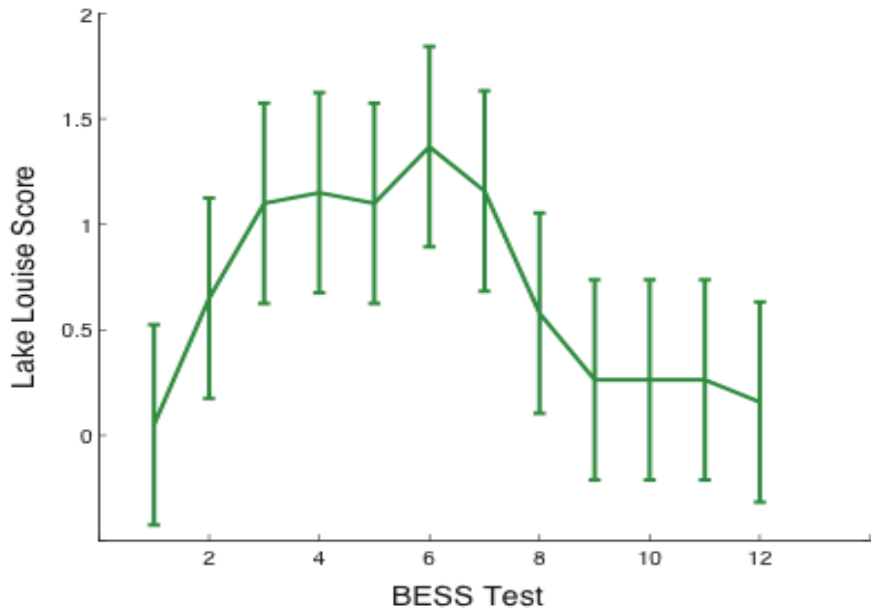


Figure 5.5. Mean Lake Louise Score (LLS) in subjects over 12 Balance Error Scoring System (BESS) tests performed serially in a hypoxic altitude chamber. Vertical lines represent error bars.

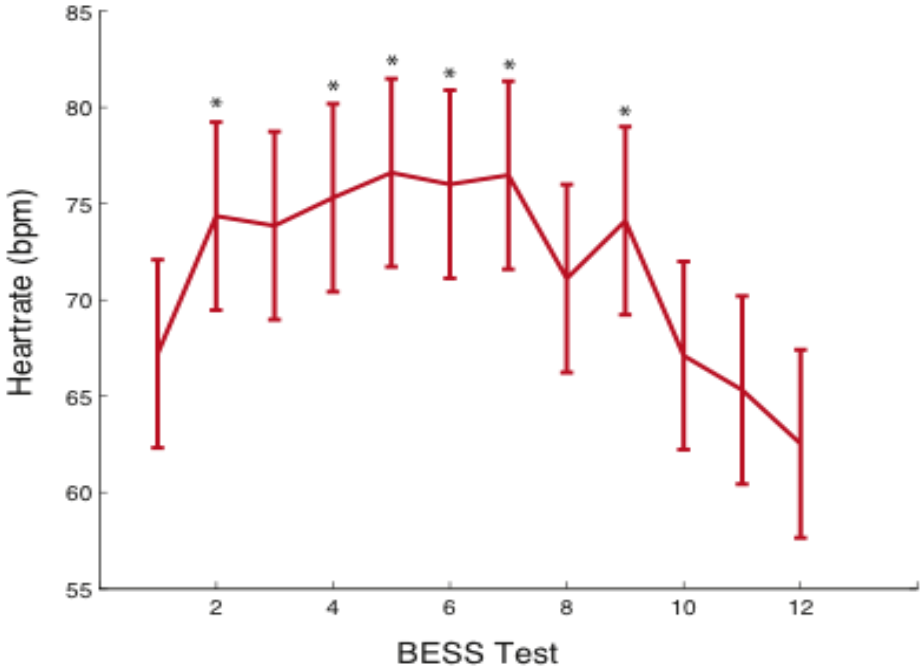


Figure 5.6. Mean heart rate (beats per minute; bmp) in subjects over 12 Balance Error Scoring System (BESS) tests performed serially in a hypoxic altitude chamber. Asterisks indicate significant difference from the first reading (baseline; BESS 1) taken in normoxia. Vertical lines represent error bars.

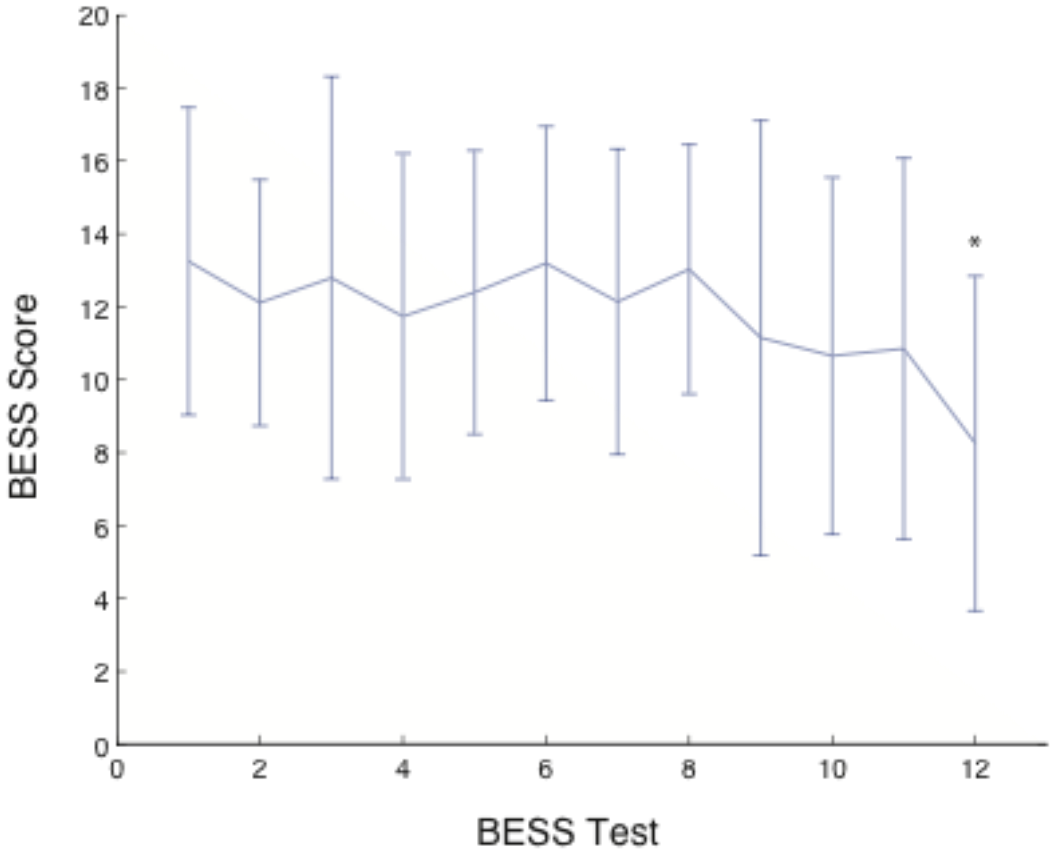


Figure 5.7. Mean number of balance errors committed by subjects (n=19) over 12 Balance Error Scoring System (BESS) tests performed serially in a hypoxic altitude chamber. Asterisks indicate significant difference from the first reading (BESS 1) taken in normoxia. Vertical lines represent error bars.

Athletic trainers displayed moderate inter-rater ($ICC_{3,1}$ = 0.59, see Figure 5.8) and intra-rater reliability ($ICC_{3,1}$ = 0.68, 0.53, see Figure 5.9), while individual trainer scores displayed poor correlation with mean experienced rater BESS ($ICC_{3,1}$ = 0.06, 0.06, see Figure 5.10). Mean athletic trainer scores indicated an average of 9.4 ± 2.9 balance errors per test.

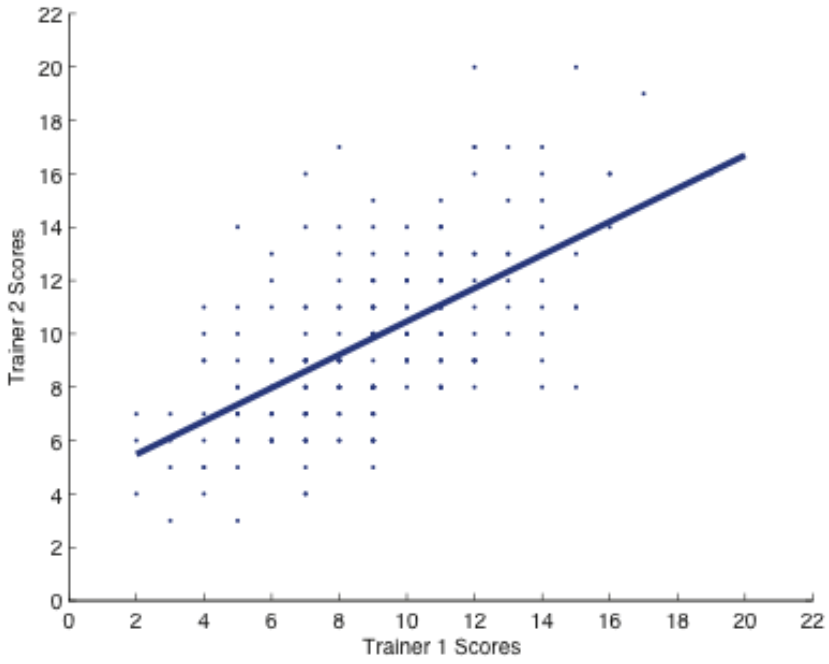


Figure 5.8. Inter-rater reliability of athletic trainers grading the Balance Error Scoring System (BESS). Each dot represents a single BESS test (n=210), and the solid line represents the line of best fit.

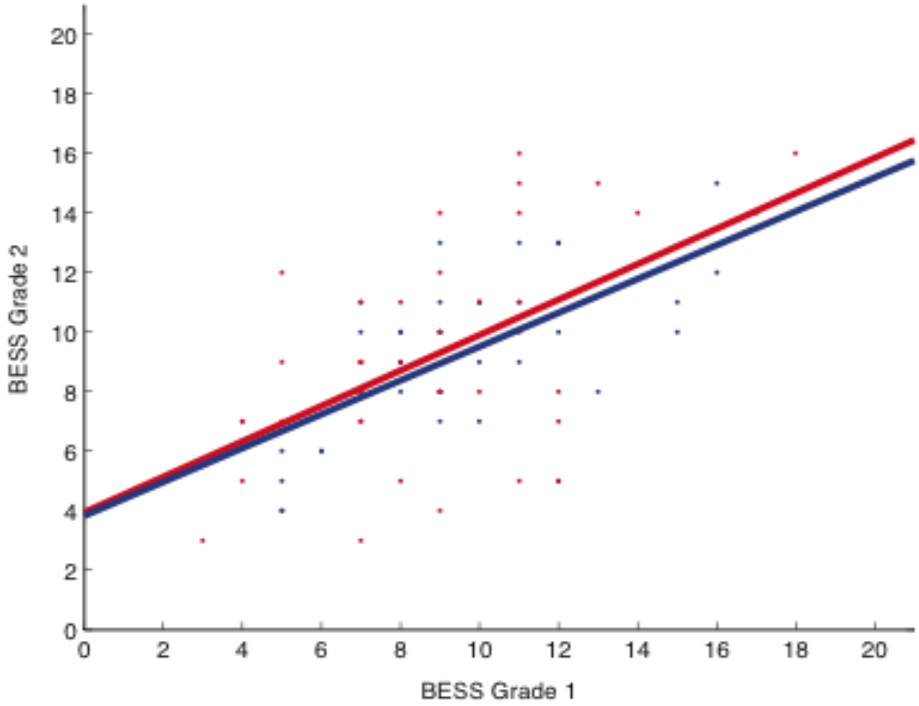


Figure 5.9. Intra-rater reliability of athletic trainers grading the Balance Error Scoring System (BESS). Each dot represents a single re-graded BESS test (n=40), and the solid lines represent lines of best fit.

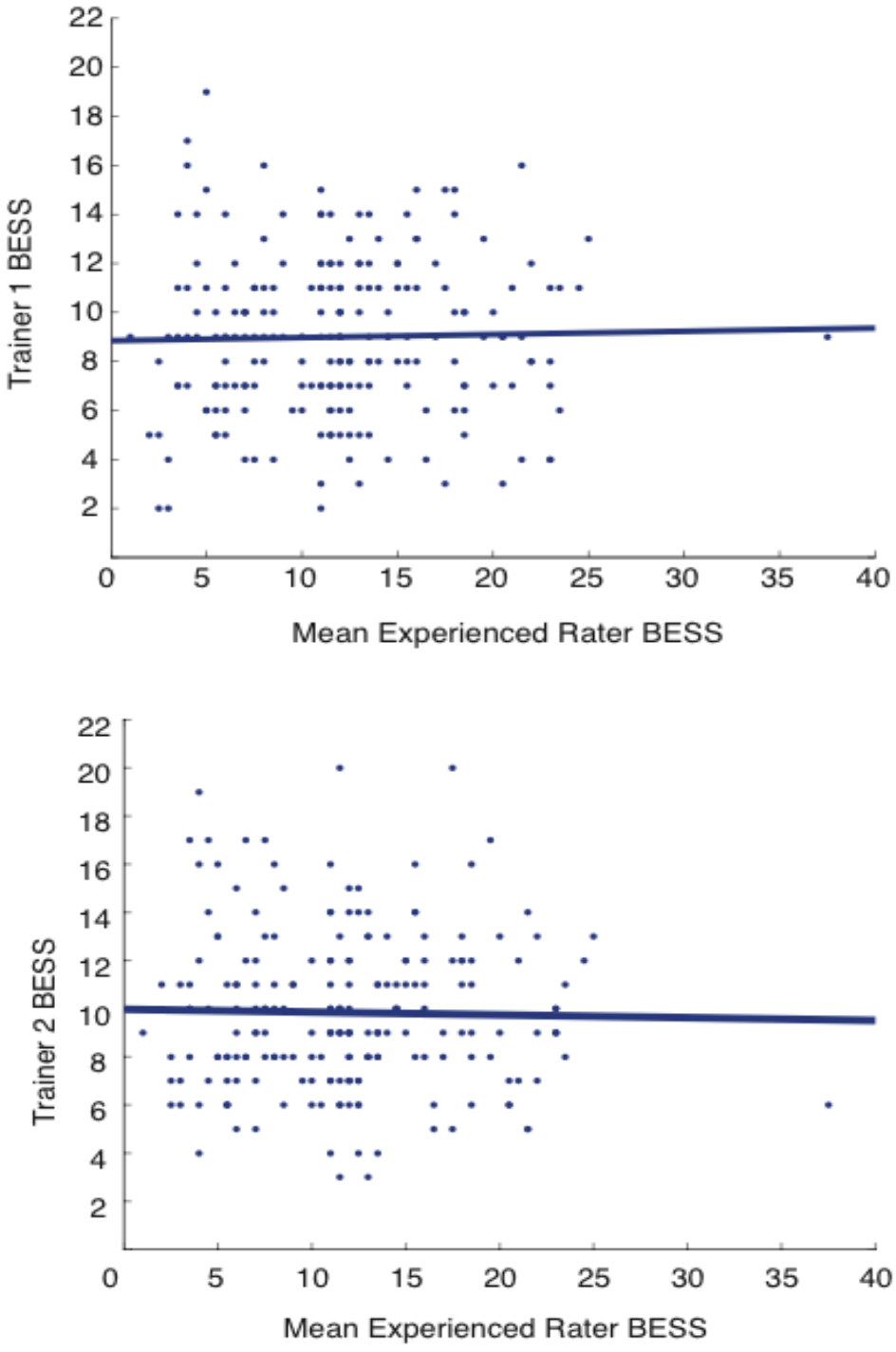


Figure 5.10. Correlation (intraclass correlation coefficients $ICC_{3,1}$ = 0.06, 0.06) between individual athletic trainer and mean experienced rater Balance Error Scoring System (BESS) scores for all tests included in the analysis (n=210). Each dot represents a single BESS test, and the solid line represents the line of best fit.

Using data from all 210 BESS tests, oBESS was unable to produce scores with good fit to mean experienced rater BESS regardless of the combination of model parameters selected (max $ICC_{3,1}$ = 0.57). However, oBESS was able to produces scores with good fit to mean experienced rater BESS when using data from tests where the experienced raters displayed a consensus (±0; n=60; $ICC_{3,1}$ = 0.84) or near consensus (±1; n=119; $ICC_{3,1}$ = 0.82, ±2; n=143 $ICC_{3,1}$ = 0.76, see Figure 5.11) on BESS scores. The graded pattern of ICC values within each panel of Figure 5.12 show that the algorithms were relatively insensitive to small changes in the number of windows and the threshold values used for error detection. The similarity in the pattern of ICC values between the different panels of Figure 5.12 also shows that the algorithms were similarly insensitive to the type of kinematic data and number of IMUs used.

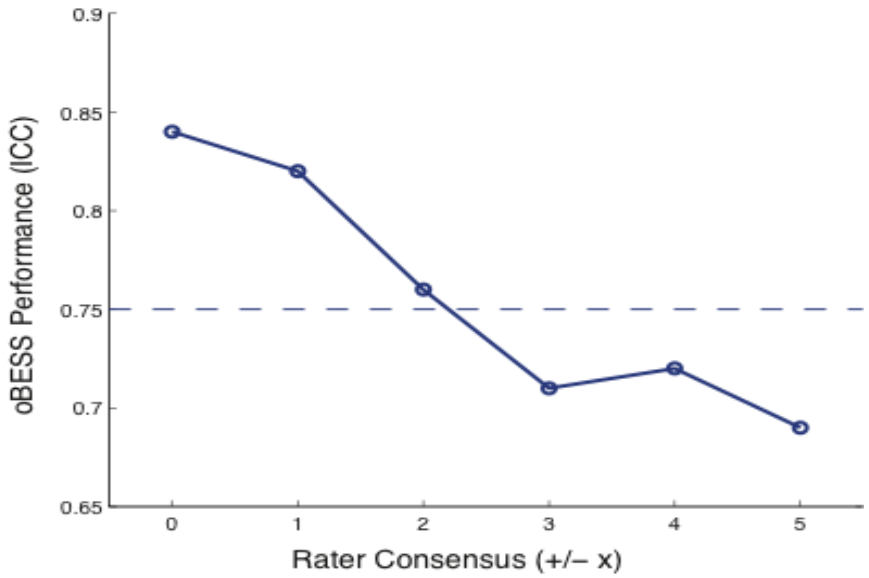


Figure 5.11. Performance of the objective Balance Error Scoring System (oBESS) to produce scores with fit to mean experienced rater BESS scores. Performance is assessed using intraclass correlation coefficients (ICCs) with values equal to or above 0.75 indicating good clinical reliability. Performance is compared between different sets of data including tests where experienced raters varied by ±0 (n=60), ±1 (n=119), ±2 (n=143), ±3 (n=149), ±4 (n=156), and ±5 (n=169).

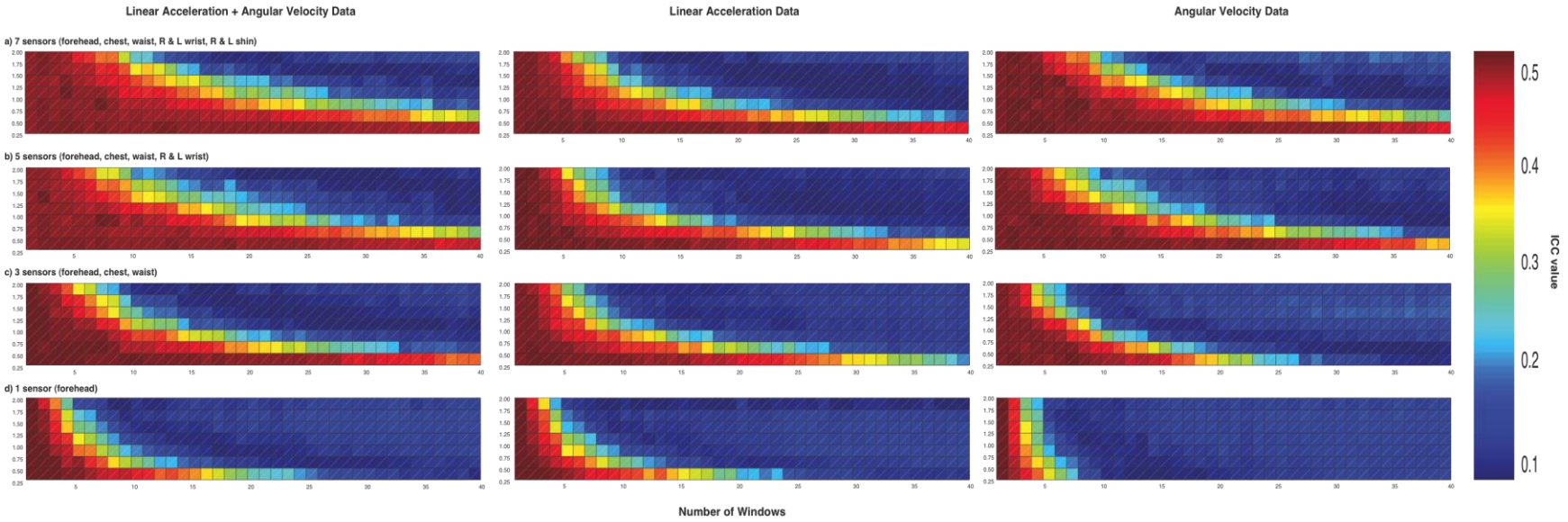


Figure 5.12. Colorbar charts representing the intraclass correlation coefficient of the fit between the mean rater Balance Error Scoring System (BESS) scores using all BESS tests (n=210) and the objective Balance Error Scoring System scores (oBESS) scores generated using IMU data from all six balance conditions and every possible combination of model parameters (n=3,840). Each square represents a combination of four algorithm parameters; number of windows (1-40), error threshold (0.25-2.00 × SD), type of data ($a + \omega$, a, ω), and number of sensors (7, 5, 3, 1).

5.5 Discussion

While subjects performing the BESS in a hypoxic altitude chamber displayed an increased number of balance errors in comparison to Experiment 1, BESS appeared to be insensitive to acute mild hypoxia even though responses of SpO_2, LLS, and HR suggest physiological changes did occur in the body. These results may have been influenced by practice effects, evidenced by subjects committing significantly less errors in normoxic conditions at the end of the protocol than during the first baseline BESS test (also performed in normoxia; p=0.003). ICC values investigating performance of the algorithm demonstrate that while oBESS scores fit mean experienced rater BESS with greater accuracy than the athletic trainers ($ICC_{3,1}$ = 0.57, 0.06, respectively), neither did so at a level commonly associated with good clinical reliability (ICC ≥ 0.75; Portney and Watkins, 1993). However, if using only data where the experienced raters displayed a consensus on the number of errors committed (n=60), the oBESS was able to produce scores with good fit to mean experienced rater BESS ($ICC_{3,1}$ = 0.84).

Inter-rater reliability of the experienced BESS raters ($ICC_{3,1}$ = 0.75) was lower than seen in our previous experiment ($ICC_{3,1}$ = 0.91) and right at the boundary of good clinical reliability (ICC ≥ 0.75; Portney and Watkins, 1993). These results, in addition to the intra-rater reliability measures ($ICC_{3,1}$ = 0.77, 0.75, 0.25), suggest that the experienced raters used in the present study were not consistent, and therefore were unable to provide a precise and reliable gold standard to train the oBESS algorithm. The greater average number of balance errors committed per test than the previous study may help explain the poor inter-rater reliability of experienced raters, as higher total balance errors may amplify inter-rater grading differences. A higher number of errors

means there is more reliance by scorers on the subjective judgement of balance error criteria and strictness, including subjective decisions such as when to assign a maximum score of 10 for conditions when the subject cannot properly complete the trial. While experienced raters were recruited from the research laboratory where the BESS originated and had extensive grading experience (18.5, 25, 60 hours), there is a possibility that those chosen for this study (who were different from Experiment 1) may have been poor raters of the BESS. This may be especially true for the rater that displayed low intra-rater reliability during the repeated videos ($ICC_{3,1}$ = 0.25). However, the inconsistency of scores may also be due to the inherent limitations of using human judgement to grade the BESS. This is supported by the similarity of intra-rater reliability scores between the two more reliable experienced raters (intra-rater $ICC_{3,1}$ = 0.77, 0.75) and previously published values for experienced BESS raters ($ICC_{3,1}$ = 0.74; Finoff et al., 2009).

Athletic trainers exhibited moderate inter-rater reliability ($ICC_{3,1}$ = 0.59). The lack of correlation between individual trainer scores to mean experienced rater BESS ($ICC_{3,1}$ = 0.06, 0.06) suggests trainers were unable to reliably assess the BESS. This finding is significant as at the time of the study both athletic trainers employed a modified version of the BESS (mBESS) for actual sideline concussion evaluations. Mean trainer BESS (9.3) also differed from the experienced raters (13.7), suggesting differences in scoring methods or interpretation of balance error criteria between laboratory and field-based users of the test. One factor leading to this difference was the failure of both trainers to assign subjects a maximum score of 10 for conditions where subjects cannot perform the stance for 5 seconds or otherwise complete the trial. During conditions where subjects clearly could not perform the stance, trainers incorrectly attempted to count the number

of balance errors, once even assigning a score greater than the maximum of 10. Poor inter-rater reliability seen between the two trainers also suggests they have significant differences in their interpretation of scoring methods, which could lead them to different conclusions regarding the status of a potentially concussed athlete. These differences might result in trainers making inappropriate decisions, such as allowing a concussed athlete to return-to-play, where literature suggests a second concussive injury before complete recovery can lead to permanent brain damage or death (Laurer et al. 2001, Guskiewicz et al. 2003).

The poor fit of oBESS scores to mean experienced rater BESS ($ICC_{3,1} = 0.57$) differed with results from Experiment 1 ($ICC_{3,1} = 0.92$). The high predictive ability seen in Experiment 1 suggests that while the oBESS was unable to accurately quantify BESS in the present study, this may be a result of training the algorithm with an unreliable gold standard rather than insufficient ability of the system. The reliability of the two experienced raters who scores were used in the present study (inter-rater $ICC_{3,1} = 0.75$) were right at the boundary of clinically reliable, whereas the four experienced raters in Experiment 1 expressed a much stronger consensus on scores, resulting in a precise gold standard to train the algorithm ($ICC_{3,1} = 0.91$). Uncertainty between raters is likely passed on and incorporated into the predictive algorithm, and therefore the use of mean experienced rater BESS as our gold standard in the present study may have limited the accuracy of oBESS scores. This is supported by the ability of the oBESS to produce accurate scores ($ICC_{3,1} = 0.84$) when using only data where the experienced raters displayed a consensus on the number of errors committed. To address this limitation, future investigations of the oBESS should focus on obtaining a more reliable gold standard to assess the utility of the system

to objectively quantify the BESS. This could be achieved by employing a panel of raters to collectively watch videos of subjects performing the BESS and then discuss, debate, and conclude on a consensus score for each trial. Alternatively, investigators could employ precise laboratory methods such as 3D motion tracking or simple electrical switches to convert subjective scoring criteria into precise computerized analyses.

Although a significant effect of hypoxia on BESS was not realized, the goal of increasing number of balance errors was achieved. Characteristic responses of the altitude acclimation measures (decreased SpO_2, increased LLS and HR) suggests that the hypoxic environment did cause physiological changes in subjects, though results prevent us from specifically commenting on the effect on balance (Holness et al. 1982, Fraser et al. 1987). The actual relationship between hypoxia and BESS may have been obscured by a number of factors including the unreliable gold standard, inter-individual differences in acclimation to altitude, or practice effects. While deficits to postural stability and balance characteristic of high altitude may have occurred (Wagner et al. 2001, Cymerman et al. 2001, Fraser et al. 1987, Holness et al. 1982), strong practice effects associated with serially performing the BESS may have concealed this relationship as subjects gradually improved over the course of the protocol (Valovich McLeod et al., 2003). This is supported by our results showing that statistically less errors were committed during the final BESS test than the first baseline test, both of which were performed in normoxia ($p=0.003$). For this final BESS test, subjects were likely no longer impaired due to hypoxia (evidenced by return to normal SpO_2, LLS, and HR) and thus practice effects could help explain why subjects committed the least number of balance errors of all BESS tests.

5.6 Conclusion

If trained using an unreliable gold standard the oBESS is unable to generate accurate BESS scores. Users of the BESS, even if considered experienced, may be unable to produce reliable BESS scores, and therefore may not represent a sufficient gold standard to train the oBESS algorithm. Athletic trainers were unable to reliably score BESS, suggesting that they may be evaluating sideline concussion assessments with unreliable information regarding standing balance. Future investigations should focus on obtaining a more reliable gold standard for the BESS.

6 Discussion

While results from Experiment 1 suggest the oBESS may offer an accurate and reliable method of assessing balance in concussed athletes on the sideline, results from Experiment 2 indicate that further investigation is required to affirm this finding. The inability of our gold standard, experienced BESS raters, to produce consistent scores in Experiment 2 may have prevented correct training of the algorithm and proper assessment of the reliability of the scores it generated. In contrast to the results from Experiment 1, Experiment 2 also suggested that although the oBESS could not reliably predict BESS at a level commonly associated with high clinical reliability unless using data where raters displayed a consensus, it was still able to do so better than athletic trainers who were real-world users of the BESS. Future investigations of the oBESS, and other methods to quantify balance, must first identify an appropriate gold standard to compare with. The assumption that experienced raters would provide a reliable gold standard is a key limitation of Experiment 2. However the results of this experiment do provide other useful information regarding the BESS and its use as a sideline tool to assess balance.

Experienced raters are commonly used as the gold standard to assess the reliability of the BESS, and their scores are also commonly used compare BESS with other balance quantification methodology (Finoff et al. 2009, Hunt et al. 2009, Valovich McLeod et al. 2006). As such, it was presumed that experienced BESS raters would produce accurate and reliable scores with which to train and validate our algorithm. While this was true for Experiment 1 ($ICC_{3,1}$ = 0.91), the experienced raters in Experiment 2 were unable to do so, evidenced by their inter-rater ($ICC_{3,1}$ = 0.75) and intra-rater ($ICC_{3,1}$ = 0.77, 0.75) reliability measures being right at the boundary of

good clinical reliability (ICC ≥ 0.75; Portney and Watkins, 1993). One limitation that may have led to the reduction in reliability of the experienced raters in Experiment 2 could have been the larger amount of data they were assigned to grade. While Experiment 1 required experienced raters to grade 30 individual tests, Experiment 2 involved 273 individual tests, possibly leading raters to lose interest and motivation with this larger dataset. There are several other possibilities that may explain the low reliability seen by these experienced raters, such as increased mean error scores or the insufficient ability of raters, however it is possible that our findings were merely a product of the inherent limitations of the human-scored BESS. This is supported by low inter-rater reliability measures reported from other investigations of the BESS using experienced raters (ICC = 0.57; Finoff et al., 2009), suggesting that regardless of their level of experience, human raters will always be subject to differences in scoring. Scoring of errors such as if a subject has flexed their hip beyond 30 degrees, or decisions to assign the maximum score because the subject was unable to maintain the stance for at least 5 seconds, require subjective interpretation by the rater and therefore are a likely source of variability between raters. In Experiment 2, while some raters appeared to be lenient identifying errors, especially if subjects were able to quickly regain stability, others appeared to be strict as if they considered exact angles between limbs or the precise periods of time subjects remained in the correct stance. Although that latter strategy would be difficult to implement for BESS scoring on the sideline, perhaps it would be a way to improve the reliability of experienced raters, and thus allow this methodology to generate a more accurate gold standard. Additional clarification by the creators of the BESS on the appropriate strictness of scoring procedures would improve the reliability of test measures produced by experienced raters, but perhaps more importantly, by field users of the BESS such as athletic trainers.

Future investigations into the reliability of the oBESS, or other objective scoring alternatives of the BESS, will first require an accurate and consistent gold standard. Rather than collecting scores from individual experienced BESS raters, a superior method could be to get a number of individuals to watch each video together and debate, discuss, and conclude on a consensus score for each test. If presented with a number of repeated videos this method would similarly be able to determine the intra-rater reliability of this consensus panel of raters. Investigators could even isolate two separate groups of raters to perform this task, allowing further validation of the gold standard through assessment of inter-reliability between the two groups. Although the experiments presented in this thesis used only one video camera located in front of the subject to record each test, perhaps the reliability of a consensus rating panel could be further improved by providing access to a number of different camera angles. While the former method was aimed at providing a source of video that would mimic a specific view point at the time the test was administered, a number of camera angles (e.g. front, side, close-up of feet) would likely result in more consistent scores by eliminating the need to judge scenarios where a subject may have lifted their foot off the support surface. If human rating is consistent and reliable, this consensus panel methodology should result in both high inter-rater and intra-rater reliability, indicating that individuals from different groups agreed upon the same score for each subject. However, if these values prove to be low, investigators will be able to conclude that the inherent limitations of human judgement not only result in unreliable BESS scores, but also that human raters are unable to provide a sufficient gold standard to adequately develop and assess alternative grading methods.

Another method to obtain a more reliable gold standard of the BESS could be to employ instrumented laboratory methods. 3D optical motion tracking would allow for accurate sampling of body motion during the test, such as changes in the joint angles about the body that are required to identify errors such as hip flexion beyond 30 degrees. Other laboratory instruments like simple electrical switches could be used to assess whether subjects remove their hands from their iliac crests or their foot touched the ground. While these methods would allow for precise assessments of definite errors, other vague criteria will still require a degree of interpretation. For example, "Lifting the forefoot or heel off the testing surface" can be interpreted a number of ways. An error could be assigned if a single toe is lifted, or alternatively only if the entire forefoot is removed from the testing surface. Again, this presents a scenario where clarification is needed by the creators of the BESS to limit incorrect interpretations of the scoring procedures. While instrumented laboratory equipment may have the potential to generate a reliable gold standard for the BESS, it presents a much more laborious, technical, and equipment-intensive method to do so than employing experienced BESS raters.

Once an adequate gold standard has been established for the BESS, methods aiming to objectify its scoring procedures will inevitably need to be validated in the target population: concussed athletes. While our approach in Experiment 2 aimed to address the higher error scores characteristic of this population, we were unable to comment on the effect of hypoxia on BESS. In addition to the unreliability of our gold standard, it is likely that the strong practice effects of the BESS affected the desired increase of scores due to artificially induced postural instability (Valovich McLeod et al., 2003). This is supported by the significant difference in errors committed between the first and last BESS tests performed in normoxia. Due to the experimental

protocol in Experiment 2, the experienced raters were blind to the degree of potentially induced postural instability caused by high altitude. This could have also affected the expected relationship between altitude and BESS as raters were unable to take environmental effects into account when grading, a occurrence that may be present in other literature (MacInnis et al., 2012). Future investigations using similar methodology should consider using a sample of control subjects, as running the experimental protocol with subjects exposed to a nominal altitude (e.g. 150 meters) would allow analysis of the effect of practice during serially performed BESS tests. These methods could then provide important information regarding the interpretation of serially performed BESS tests in concussed athletes on the sideline, in addition to field assessments of acute mild hypoxia (Macinnis et al., 2012).

Although it is difficult to conclude on the accuracy and reliability of the oBESS, this thesis does achieve a number of objectives. The use of small kinematic sensors with the ability to connect wirelessly to an electronic device suggests that the oBESS, if proven reliable in future investigations, presents a highly portable method to objectively quantify balance. This is furthered by our effort to produce a custom mobile application that was used for data collection in Experiment 2. As results in Experiment 1 suggest a single sensor sampling accelerations from the forehead would allow the oBESS to reliably predict BESS, an appropriate device could easily be produced for less than US$200, offering a much more affordable alternative to other available instrumented technology. The use of the same protocol and scoring outcome of the BESS also means that the oBESS would be easy to adopt for current users of the BESS, while offering an objective method to quantify it.

The similarity of the colorbar figures representing the fit of oBESS scores to experienced rater means from both experiments (see Figures 4.2, 5.12) indicates that the oBESS algorithm is robust to small changes in model parameters and accompanying data. This suggests that if the reliability of the oBESS can be affirmed, the oBESS algorithm presents a favourable way to predict BESS scores from kinematic data. A number of methods were used to attempt to improve this algorithm, such as principal component analysis (PCA) of acceleration and velocity vectors to reduce the dimensionality of the data. However, analyses indicated that dimensionality could only reduced marginally (14 vectors total; 7 acceleration vectors, 7 velocity vectors→11 vectors total) while still maintaining 95% of the variance in the data, suggesting that the use of PCA was unlikely to improve the oBESS algorithm.

An important result from Experiment 2 was the unreliability of the athletic trainers, who were using a modified version of the BESS (mBESS) at the time of the study during actual sideline evaluations of concussion. This finding suggests that athletic trainers may be incorrectly administering the BESS, and thus making decisions using inaccurate information regarding balance. If so, these faulty scores may lead athletic trainers to make inappropriate decisions on the sideline, such as allowing concussed athletes to return-to-play in the presence of a concussive injury. This result is alarming given the permanent or fatal consequences associated with repetitive concussive injuries (Guskiewicz et al., 2003). One possible explanation for the unreliability of trainers could be unfamiliarity with the scoring criteria that assigns subjects the maximum error score (10) if they cannot correctly perform a stance for a minimum of five seconds, or otherwise properly complete the trial. Because both trainers employed the mBESS for balance assessment in concussed athletes, which incorporates only the conditions performed

on the firm surface that generally result in fewer errors, trainers may not have had experience implementing this "maximum score" rule as athlete balance is never sufficiently challenged (Valovich McLeod et al. 2005, Hunt et al. 2009). With this said, trainers were briefed on the procedures and given a scorecard that clearly states this rule before grading. This unreliability of athletic trainers suggests that concussion evaluation protocols employing the BESS should either look for clarification regarding human-based scoring criteria, or find more objective methods to benefit from the utility balance assessment presents in the sideline evaluation of sports-related concussions.

7 Conclusion

The oBESS may offer an affordable, accurate, and reliable method to quantify balance in potentially concussed athletes on the sideline. Presently, the oBESS can accurately and reliably predict BESS scores in healthy subjects with low total balance errors using acceleration data collected from an IMU located at the forehead. While further research is required to affirm these results, the oBESS can also quantify BESS scores of subjects with artificially-induced postural instability more reliably than athletic trainers employing the standard human-scoring methods, but not at a level commonly associated with high clinical reliability. If scored using the standard human-scoring methods, the BESS may provide users with inaccurate information regarding the balance of potentially concussed athletes. The potentially permanent or fatal consequences associated with inappropriately allowing concussed athletes to return-to-play suggest that further development of the oBESS, or other objective scoring alternatives of the BESS, are required to take full advantage of the utility balance assessment offers to sideline evaluations of concussion.

References

Anastasi A. *Psychological Testing.* 6th ed. New York (NY): Macmillan Publishing Company; 1988.

Alla S, Sullivan SJ, Hale L, McCrory P. Self-report scales/checklists for the measurement of concussion symptoms: a systematic review. *British Journal of Sports Medicine.* 2009;43:i3-i12.

Barela JA, Dias JL, Godoi D, Viana AR, de Freitas PB. Postural control and automaticity in dyslexic children: The relationship between visual information and body sway. *Research in Developmental Disabilities.* 2011;32(5):1814-1821.

Barlow M, Schlabach D, Peiffer J, Cook C. Differences in change scores and the predictive validity of three commonly used measures following concussion in the middle school and high school aged population. *The International Journal of Sports Physical Therapy.* 2011;6(3):150-157.

Bell DR, Guskiewicz KM, Clark MA, Padua DA. Systematic Review of the Balance Error Scoring System. *Sports Health.* 2011;3(3):287-295.

Benson BW, Hamilton GM, Meeuwisse WH, McCrory P, Dvorak J. Is protective equipment useful in preventing concussion? A systematic review of the literature. *British Journal of Sports Medicine.* 2009;43:i56-i67.

Berg, K. Measuring balance in the elderly: preliminary development of an instrument. *Physiotherapy Canada.* 1989;41(6):304-311.

Berg K, Wood-Dauphinee S, Williams JI. The Balance Scale: reliability assessment with elderly residents and patients with an acute stroke. *Scandinavian Journal of Rehabilitation Medicine.* 1995;27(1):27.

Broglio SP, Ferrara MS, Macciocchi SN, Baumgartner TA, Elliott R. Test-Retest Reliability of Computerized Concussion Assessment Programs. *Journal of Athletic Training.* 2007;42(4):509-514.

Cattaneo, D, Alberto R, Matteo M. Validity of six balance disorders scales in persons with multiple sclerosis. *Disability & Rehabilitation.* 2006;28(12):789-795.

Caron O, Faure B, Breniere Y. Estimating the centre of gravity of the body on the basis of the centre of pressure in standing posture. *Journal of Biomechanics.* 1997;30:1169-1171.

Carpenter MG, Murnaghan CD, Inglis JT. Shifting the balance: evidence of an exploratory role for postural sway. *Neuroscience.* 2010;171(1):196-204.

Cavanaugh JT, Guskiewicz KM, Stergiou N. A nonlinear dynamic approach for evaluating postural control: new directions for the management of sport-related cerebral concussion. *Sport Med.* 2005;35(11):935-950.

Cavanaugh JT, Guskiewicz KM, Giuliani C, Marshall S, Mercer VS, Stergiou N. Recovery of Postural Control After Cerebral Concussion: New Insights Using Approximate Entropy. *Journal of Athletic Training.* 2006;41(3):305-313.

Cavanaugh JT, Guskiewicz KM, Giuliani C, Mercer V, Stergiou N. Detecting altered postural control after cerebral concussion in athletes with normal postural stability. *British Journal of Sports Medicine.* 2005;39:805-811.

Chen JK, Johnston KM, Frey S, Petrides M, Worsley K, Ptito A. Functional abnormalities in symptomatic concussed athletes: an fMRI study. *Neuroimage.* 2004;22(1):68-82.

Clark RA, Bryant AL, Pua Y, McCrory P, Bennell K, Hunt M. Validity and reliability of the Nintendo Wii Balance Board for assessment of standing balance. *Gait & Posture.* 2010;31(3):307-310.

Conradsson M, Lundin-Olsson L, Lindelof N, Littbrand H, Malmqvist L, Gustafson T, Rosendahl Erik. Berg Balance Scale: Intrarater test-retest reliability among older people dependent in activities of daily living and living in residential care facilities. *Physical Therapy.* 1997;78(9):1155-1163.

Cymerman A, Muza SR, Beidleman BA, Ditzler DT, Fulco CS. Postural Instability and Acute Mountain Sickness During Exposure to 24 Hours of Simulated Altitude (4300 m). *High Altitude Medicine & Biology.* 2001;2(4):509-514.

Davis GA, Iverson GL, Guskiewicz KM, Ptito A, Johnston KM. Contributions of neuroimaging, balance testing, electrophysiology, and blood markers to the assessment of sport-related concussion. *British Journal of Sports Medicine.* 2009;43:136-145.

Davlin CD. Dynamic Balance in High Level Athletes. *Perceptual and Motor Skills.* 2004;98(3c):1171-1176.

De Beaumont L, Mongeon D, Tremplay S, Messier J, Prince F, Leclerc S, Lassonde M, Théoret. Persistent Motor System Abnormalities in Formerly Concussed Athletes. *Journal of Athletic Training.* 2011;46(3):234-240.

Denny-Brown D, Russell WR. Experimental cerebral concussion. *Journal of Physiology.* 1940;99:153.

Donaldson L, Asbridge M, Cusimano MD. Bodychecking Rules and Concussion in Elite Hockey. *PLoS ONE.* 2013; 8(7):e69122.

Dziemianowicz MS, Kirschen MP, Pukenas BA, Laudano E, Balcer LJ, Galetta SL. Sport-Related Concussion Testing. *Current Neurology and Neuroscience Reports.* 2012;12(5):547-559.

Eckner JT, Lipps DB, Kim H, Richardson JK, Ashton-Miller JA. Can a Clinical Test of Reaction Time Predict a Functional Head-Protective Response? *Medicine & Science in Sport & Exercise.* 2011;43(3):382-387.

Fait P, McFadyen BJ, Swaine B, Cantin JF. Alterations to locomotor navigation in a complex environment at 7 and 30 days following a concussion in an elite athlete. *Brain Injury.* 2009; 23(4):362-369.

Faul M, Xu L, Wald MM, Coronado VG. Traumatic Brain Injury in the United States: Emergency Department Visits, Hospitalizations and Deaths 2002-2006. US Department of Health and Human Services. Centers for Disease Control and Prevention; Mar. 2010.

Finkelstein EA, Corso PS, Miller TR. *The incidence and economic burden of injury in the United States.* New York (NY): Oxford University Press; 2006.

Finnoff JT, Peterson VJ, Hollman JH, Smith J. Intrarater and Interrater Reliability of the Balance Error Scoring System (BESS). *PM&R.* 2009;1(1):50-54.

Foltz EL, Schmidt RP. The role of the reticular formation in the coma of head injury. *Journal of Neurosurgery.* 1956;13(2):145-154.

Fraser WD, Eastman DE, Paul MA, Porlier JA. Decrement in postural control during mild hypobaric hypoxia. *Aviation, Space, and Environmental Medicine.* 1987;58:768-772.

Gennarelli TA, Thibault LE, Adams JH, Graham DI, Thompson CJ, Marcincin RP. Diffuse axonal injury and traumatic coma in the primate. *Annals of Neurology.* 1982;12(6):564-574.

Gennarelli TA, Adams JH, Graham DI. Acceleration induced head injury in the monkey, I: the model, its mechanical and physiological correlates. *Experimental and Clinical Neuropathy.* 1981;7:23–25.

Gessel LM, Fields SK, Collins CL, Dick RW, Comstock RD. Concussions Among United States High School and Collegiate Athletes. *Journal of Athletic Training.* 2007;42(4):495-503.

Geurts ACH, Ribbers GM, Knoop JA, van Limbeek J. Identification of static and dynamic postural instability following traumatic brain injury. *Archives of Physical Medicine and Rehabilitation.* 1996;77(7):639-644.

Gilchrist J, Thomas KE, Wald M, Langlois J. Nonfatal Traumatic Brain Injuries from Sports and Recreation Activities – United States, 2001-2005. *Morbidity & Mortality Weekly Report.* 2007;56(29):733-737.

Giza CC, Hovda DA. The Neurometabolic Cascade of Concussion. *Journal of Athletic Training.* 2001; 36(3):228-235.

Gordon KE, Dooley JM, Wood EP. Descriptive Epidemiology of Concussion. *Pediatric Neurology.* 2006;34:376-378.

Guskiewicz KM. Postural Stability Assessment Following Concussion: One Piece of the Puzzle. *Clinical Journal of Sport Medicine.* 2001;11(3):182-189.

Guskiewicz KM. Balance Assessment in the Management of Sport-Related Concussion. *Clinical Journal of Sports Medicine.* 2011;30:89-102.

Guskiewicz KM, McCrea M, Marshall SW, Marshall SW, Cantu RC, Randolph C, Barr W, Onate JA, Kelly JP. Cumulative Effects Associated with Recurrent Concussion in Collegiate Football Players *Journal of the American Medical Association.* 2003;290(19):2549-2555.

Guskiewicz KM, Ross SE, Marshall SW. Postural Stability and Neuropsychological Deficits After Concussion in Collegiate Athletes. *Journal of Athletic Training.* 2001;36(3):263-273.

Herring SA, Cantu RC, Kuskiewicz KM, Putukian M, Kibler WB. Concussion (Mild Traumatic Brain Injury) and the Team Physician: A Consensus Statement-2011 Update. *Medicine & Science in Sports & Exercise.* 2011;43(12):2412-2422.

Hinman RS, Bennell KL, Metcalf BR, Crossley KM. Balance impairments in individuals with symptomatic knee osteoarthritis: a comparison with matched controls using clinical tests. *Rheumatology.* 2002;41(12):1388-1394.

Holness DE, Fraser WD, Eastman DE, Porlier JA, Paul MA. Postural stability during slow-onset and rapid-onset hypoxia. *Aviation, Space, and Environmental Medicine.* 1982;53(7):647-651.

Horak, FB. Clinical assessment of balance disorders. *Gait & Posture.* 1997;6(1):76-84.

Hunt TN, Ferrara MS, Bornstein RA, Baumgartner TA. The reliability of the Modified Balance Error Scoring System. *Clinical Journal of Sports Medicine.* 2009;19:471-475.

Ingersoll CD, Armstrong CW. The effects of closed-head injury on postural sway. *Medicine and Science in Sports Exercise.* 1992;24(7):739-743.

Iverson G, Lovell M, Collins, MW. Interpreting Change on ImPACT Following Concussion. *The Clinical Neuropsychologist.* 2003;17(4):460-467.

Iverson GL, Lovell M, Collins M. Validity of ImPACT for measuring processing speed following sports-related concussion. *Journal of the International Neuropsychology Society.* 2002;10:1-3.

Jacobs JV, Horak FB, Tran VK, Nutt JG. Multiple balance tests improve the assessment of postural stability in subjects with Parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry.* 2006;77:322-326.

Jantzen KJ, Anderson B, Steinberg FL, Kelso JAS. A prospective functional MR imaging study of mild traumatic brain injury in college football players. *American Journal of Neuroradiology* 2004;25:738-745.

Johnson EW, Kegel NE, Collins MW. Neuropsychological Assessment of Sport-Related Concussion. *Clinical Journal of Sports Medicine.* 2011;30:73-88.

Johnston KM, Ptito A, Chankowsky J, Chen JK. New Frontiers in Diagnostic Imaging in Concussive Head Injury. *Clinical Journal of Sport Medicine.* 2001;11(3):166-175.

King DL, Zatsiorsky VM. Extracting gravity line displacement from stabilographic recordings. *Gait & Posture.* 1997;6:27-38.

Lafond D, Duarte M, Prince F. Comparison of three methods to estimate the center of mass during balance assessment. *Journal of Biomechanics.* 2004;37:1421-1426.

Laurer HL, Bareyre FM, Lee VM, Trojanowki JQ, Longhi L, Hoover R, Saatman KR, Rangupathi R, Hoshino S, Grady S, McIntosh TK. Mild head injury increasing the brain's vulnerability to a second concussive impact. *Journal of Neurosurgery.* 2001;95(5):859-870.

Macciocchi SN, Barth JT, Alves W, Rimmel RW, Jane JA. Neuropsychological Functioning and Recovery after Mild Head Injury in Collegiate Athletes. *Neurosurgery.* 1996;39(3):510-514.

Macinnis MJ, Rupert JL, Koehle MS. Evaluation of the Balance Error Scoring System (BESS) in the diagnosis of acute mountain sickness at 4380m. *High Altitude Medicine & Biology* 2012;13(2):93-97.

Maerlender A, Flashman L, Kessler A, Kumbhani S, Greenwald R, Tostenson T, McAllister T. Examination of the construct validity of ImPACT computerized test, traditional, and experimental neuropsychological measures. *The Clinical Neuropsychologist.* 2010;24(8):1309-1325.

McCrea M, Barr WB, Guskiewicz K, Randolph C, Marshall SW, Cantu R, Onate JA, Kelly JP. Standard regression-based methods for measuring recover after sport-related concussion. *Journal of the International Neuropsychological Society.* 2005;11(1):58-69.

McCrea M, Guskiewicz KM, Marshall SW, Barr W, Randolph C, Cantu RC, Onate JA, Yang J, Kelly JP. Acute effects and recovery time following concussion in collegiate football players. *Journal of the American Medical Association.* 2003:290(19):2556-2563.

McCrea M, Hammeke T, Olsen G, Leo G, Guskiewicz K. Unreported concussion in high school football players: implications for prevention. *Clinical Journal of Sports Medicine.* 2004;14(1):13-17.

McCrory P, Meeuwisse WH, Aubry M, Cantu B, Dvorak J, Echemendia RJ, Engebresten L, Johnston K, Kutcher JS, Raftery M, Sills A, Benson BW, Davis GA, Ellenbogen RG, Guskiewicz, Herring SA, Iverson GL, Jordan BD, Kissick J, McCrea M, McIntosh AS, Maddocks D, Makdissi M, Purcell L, Putukian M, Schneider K, Tator CH, Turner M. Consensus statement on concussion in sport: the 4th International Conference on Concussion in Sport held in Zurich, November 2012. *British Journal of Sports Medicine.* 2013;47:250-258.

Meaney DF, Smith DH. Biomechanics of Concussion. *Clinical Sports Medicine.* 2011;30:19-31.

Meehan WP, Bachur RG. Sport-Related Concussion. *Pediatrics.* 2009;123:114-123.

Onate JA, Beck BC, Van Lunen BL. On-Field Testing Environment and Balance Error Scoring System Performance During Preseason Screening of Healthy Collegiate Baseball Players. *Journal of Athletic Training.* 2007;42(4):445-451.

Paloski WH, Wood SJ, Feiveson AH, Black FO, Hwang EY, Reschke MF. Destabilization of human balance control by static and dynamic head tilts. *Gait & Posture*. 2006;23(3):315-323.

Panjan A, Sarabon N. Review of methods for the evaluation of human body balance. *Sport Science Review.* 2010;(5-6):131-163.

Partridge B, Hall W. Conflicts of Interest in Recommendations to Use Computerized Neuropsychological Tests to Manage Concussion Professional Football Codes. *Neuroethics.* 2013:1-12.

Patricios JS, Collins R, Robers C. Zurich 2012: our cohort of 'concussionologists' – conveying consensus. *British Journal of Sports Medicine.* 2013;47:9-11.

Peterson CL, Ferrara MS, Mrazik M, Scott P, Ronald E. Evaluation of Neruopsychological Domain Scores and Postural Stability Following Cerebral Concussion in Sports. *Clinical Journal of Sport Medicine*. 2003;13(4):230-237.

Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice.* Norwalk (CT): Appleton & Lange; 1993.

Pulsipher DT, Campbell RA, Thoma R, King JH. A critical review of neuroimaging applications in sports concussion. *Current Sports Medicine Reports.* 2011;10(1):14-20.

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

Randolph C, McCrea M. Barr WB. Is neuropsychological testing useful in the management of sport-related concussion? *Journal of Athletic Training.* 2005;40:139-154.

Resch JE, May B, Tomporowski PD, Ferrara MS. Balance Performance With a Cognitive Task: A Continuation of the Dual-Task Testing Paradigm. *Journal of Athletic Training.* 2011;46(2):170-175.

Riemann B, Guskiewicz KM, Shields EW. Relationship between clinical and force plate measures of postural stability. *Journal of Sports Rehabilitation.* 1999;8:71-82.

Riemann BL, Guskiewicz KM. Effects of Mild Head Injury on Postural Stability as Measured Through Clinical Balance Testing. *Journal of Athletic Training.* 2000;35(1):19-25.

Riley MA, Clark S. Recurrence analysis of human postural sway during the sensory organization test. *Neuroscience Letters.* 2003;342:45-48.

Roozenbeek B, Maas AIR, Menon DK. Changing patterns in the epidemiology of traumatic brain injury. *Nature Reviews Neurology.* 2013;9:231-236.

Ropper AH, Brown RH. *Adams and Victor's Principles of Neurology.* 8th ed. McGraw-Hill Professional; 2005. Chapter 35.

Schatz P, Pardini JE, Lovell MR, Collins MW, Podell K. Sensitivity and specificity of the ImPACT Test Battery for concussion in athletes. *Clinical Neuropsychology.* 2006;21(1):91-99.

Schmitt DM, Hertel J, Evans TA, Olmsted LC, Putukian M. Effect of an Acute Bout of Soccer Heading on Postural Control and Self-Reported Concussion Symptoms. *International Journal of Sports Medicine.* 2004;25(5):326-331.

Shehata N, Wiley JP, Richea S, Benson BW, Duits L, Meeuwisse WH. Sport concussion assessment tool: baseline values for varsity collision sport athletes. *British Journal of Sports Medicine.* 2009;43:730-734.

Shrout PE, Fleiss JL. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin.* 1979;86(2):420-428.

Slobounov SM, Gay M, Zhang K, Johnson B, Pennell D, Sebastianelli W, Horovitz S, Hallett M. Alteration of brain functional network at rest and in response to YMCA physical stress test in concussed athletes: RsFMRI study. *Neuroimage.* 2011;55(4):1716-1727.

Sturnieks DL, Arnold R, Lord SR. Validity and reliability of the Swaymeter device for measuring postural sway. *BMC Geriatrics.* 2011;11:63-70.

Talavage TM, Nauman E, Breedlove EL, Yoruk U, Dye AE, Morigaki KE, Feuer H, Leverenz LJ. Functionally-detected cognitive impairment in high school football players without clinically-diagnosed concussion. *Journal of Neurotrauma.* 2013;30:1-12.

Valovich McLeod TC, Armstrong T, Miller M, Sauers JL. Balance improvements in female highschool basketball players after a 6-week neuromuscular-training program. *Journal of Sport Rehabilitation.* 2009;18(4):465-481.

Valovich McLeod TC, Bay RC, Lam KC, Chhabra A. Representative baseline values on the Sport Concussion Assessment Tool 2 (SCAT2) in adolescent athletes vary by gender, grade, and concussion history. *The American Journal of Sports Medicine.* 2012;40(4):927-933.

Valovich McLeod TC, Perrin DH, Gandsneder BM. Repeat Administration Elicits a Practice Effect with the Balance Error Scoring System but Not With the Standardized Assessment of Concussion in High School Athletes. *Journal of Athletic Training*. 2003;38:51-56.

Valovich McLeod TC, Perrin DH, Guskiewicz KM, Shultz SJ, Diamond R, Gansneder BM. Serial administration of clinical concussion assessments and learning effects in healthy young athletes. *Clinical Journal of Sports Medicine.* 2004;14:287-295.

Van Kampen DA, Lovell MR, Pardini JE, Collins MW, Fu FH. The "value added" of neurocognitive testing after sports-related concussion. *American Journal of Sports Medicine.* 2006;34(10):1630-1635.

Wetjen NM, Pichelmann MA, Atkinson JLD. Second Impact Syndrome: Concussion and Second Injury Brain Complications. *Journal of the American College of Surgeons.* 2010;211(4):553-557.

Wagner LS, Oakley SR, Vang P, Noble BN, Cevette MJ, Stepanek JP. Hypoxia-induced changes in standing balance. *Aviation, Space and Environmental Medicine.* 2011;85:518-822.

Wikstrom EA. Validity and Reliability of Nintendo Wii Fit Balance Scores. *Journal of Athletic Training.* 2012;47(3):306-313.

Wikstrom EA, Tillman MD, Smith AN, Borsa PA. A New Force-Plate Technology Measure of Dynamic Postural Stability: The Dynamic Postural Stability Index. *Journal of Athletic Training.* 2005;40(4):305-309.

Williamson IJS, Goodman D. Converging evidence for the under-reporting of concussion in youth ice hockey. *British Journal of Sports Medicine.* 2006;40:128-132.

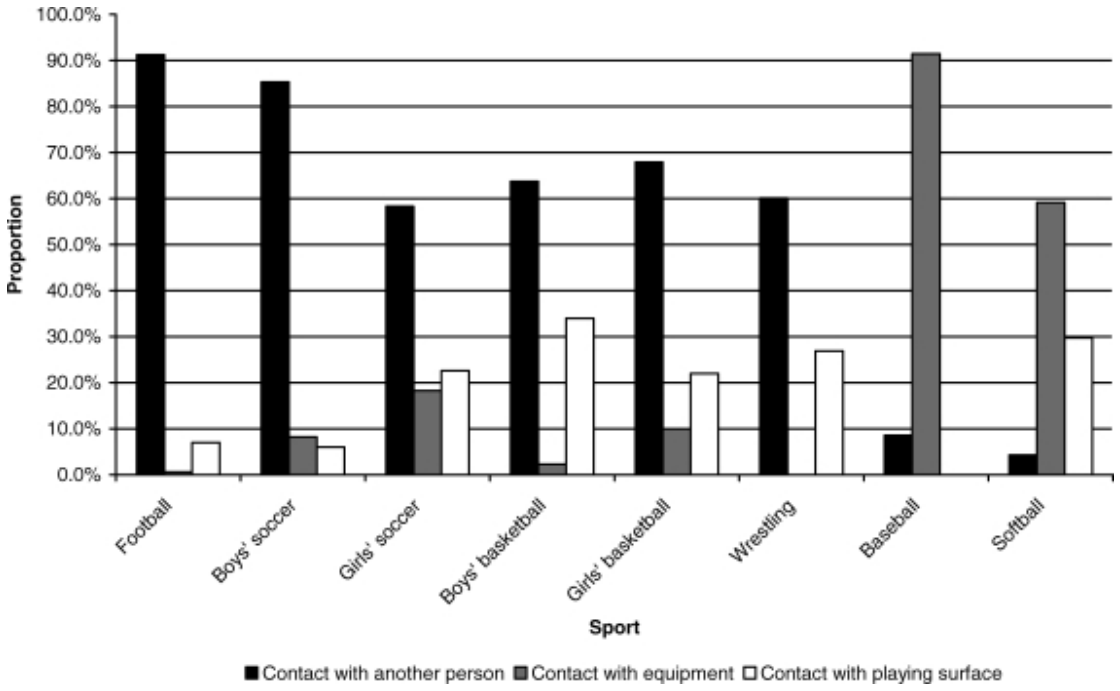
Winter DA. Human balance and posture control during standing and walking. *Gait & Posture.* 1995;3:193-214.

Yarnell PR, Lynch S. Retrograde memory immediately after concussion. *Lancet*. 1970;1:863-864.

Zammit E, Herrington L. Ultrasound therapy in the management of acute lateral ligament sprains of the ankle joint. *Physical Therapy in Sport.* 2005;6(3):116-121.

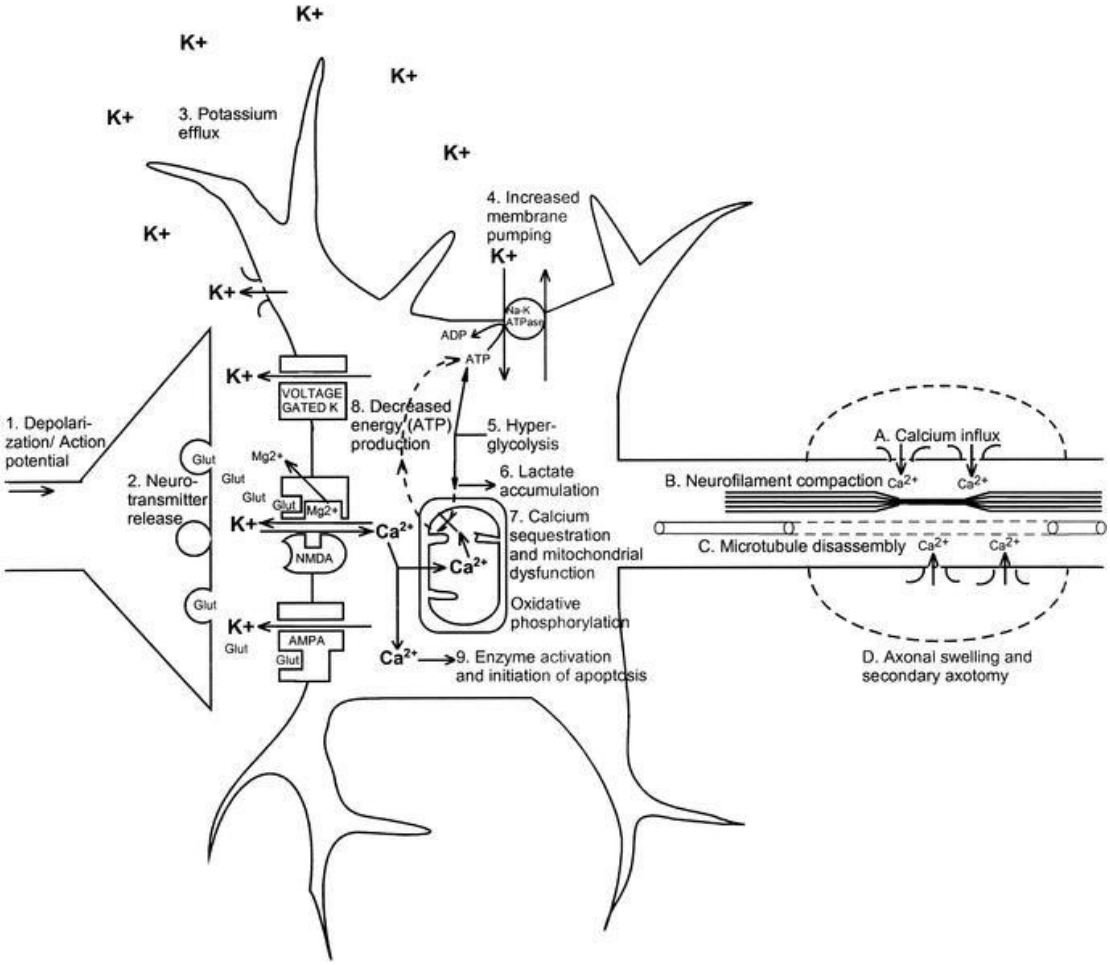
Appendices

Appendix A: National estimates of the mechanism of concussion by sport for high school athletes.



National estimates of the mechanism of concussion by sport for high school athletes, High School Sports-Related Injury Surveillance Study, United States, 2005-2006 School Year.
Figure 3 from © Gessel LM, Fields SK, Collins CL, Dick RW, Comstock RD. Concussions Among United States High School and Collegiate Athletes. *Journal of Athletic Training.* 2007;42(4):495-503. Page 500. Reproduced with permission from publisher.

Appendix B: The neurometabolic cascade of concussion.



The Neurometabolic Cascade of Concussion. Figure 2 from © Giza CC, Hovda DA. The Neurometabolic Cascade of Concussion. *Journal of Athletic Training.* 2001; 36(3):228-235. Page 230. Reproduced with permission from publisher.

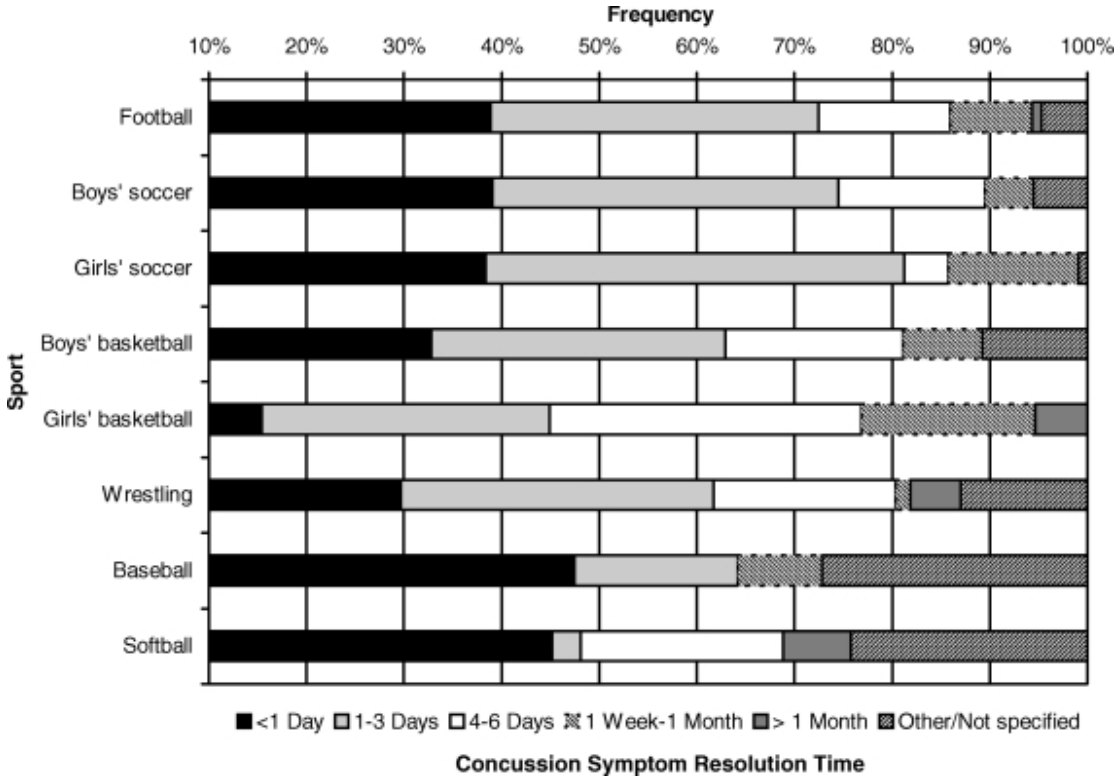
Appendix C: Concussion rates among US high school and collegiate athletes.

Sport	Division	No. of Concussions	National Estimates†	Rates per 1000 Athlete-Exposures			Overall Rate Comparison Collegiate Versus High School		
				Practice	Competition	Overall	Rate Ratio	95% Confidence Interval	*P* Value
Football	High school	201	55 007	0.21	1.55	0.47	N/A‡	N/A	N/A
	Collegiate	245	—	0.39	3.02	0.61	1.31	1.09, 1.58	<.01
Boys' soccer	High school	33	20 929	0.04	0.59	0.22	N/A	N/A	N/A
	Collegiate	42	—	0.24	1.38	0.49	2.26	1.43, 3.57	<.01
Girls' soccer	High school	51	29 167	0.09	0.97	0.36	N/A	N/A	N/A
	Collegiate	57	—	0.25	1.80	0.63	1.76	1.21, 2.57	<.01
Volleyball	High school	6	2568	0.05	0.05	0.05	N/A	N/A	N/A
	Collegiate	14	—	0.21	0.13	0.18	3.63	1.39, 9.44	<.01
Boys' basketball	High school	16	3823	0.06	0.11	0.07	N/A	N/A	N/A
	Collegiate	33	—	0.22	0.45	0.27	3.65	2.01, 6.63	<.01
Girls' basketball	High school	40	12 923	0.06	0.60	0.21	N/A	N/A	N/A
	Collegiate	49	—	0.31	0.85	0.43	1.98	1.31, 3.01	<.01
Wrestling	High school	30	5935	0.13	0.32	0.18	N/A	N/A	N/A
	Collegiate	15	—	0.35	1.00	0.42	2.34	1.26, 4.34	.01
Baseball	High school	9	1991	0.03	0.08	0.05	N/A	N/A	N/A
	Collegiate	12	—	0.03	0.23	0.09	1.88	0.79, 4.46	.22
Softball	High school	10	3558	0.09	0.04	0.07	N/A	N/A	N/A
	Collegiate	15	—	0.07	0.37	0.19	2.61	1.17, 5.82	.03
Boys' sports total	High school	289	87 685	0.13	0.61	0.25	N/A	N/A	N/A
	Collegiate	347	—	0.30	1.26	0.45	1.78	1.52, 2.08	<.01
Girls' sports total	High school	107	48 216	0.07	0.42	0.18	N/A	N/A	N/A
	Collegiate	135	—	0.23	0.74	0.38	2.04	1.59, 2.64	<.01
Overall total	High school	396	135 901	0.11	0.53	0.23	N/A	N/A	N/A
	Collegiate	482	—	0.28	1.02	0.43	1.86	1.63, 2.12	<.01

*Collegiate data for the 2005–2006 school year were provided by the National Collegiate Athletic Association Injury Surveillance System.
†National estimates for the National Collegiate Athletic Association data were not available.
‡Indicates not applicable.

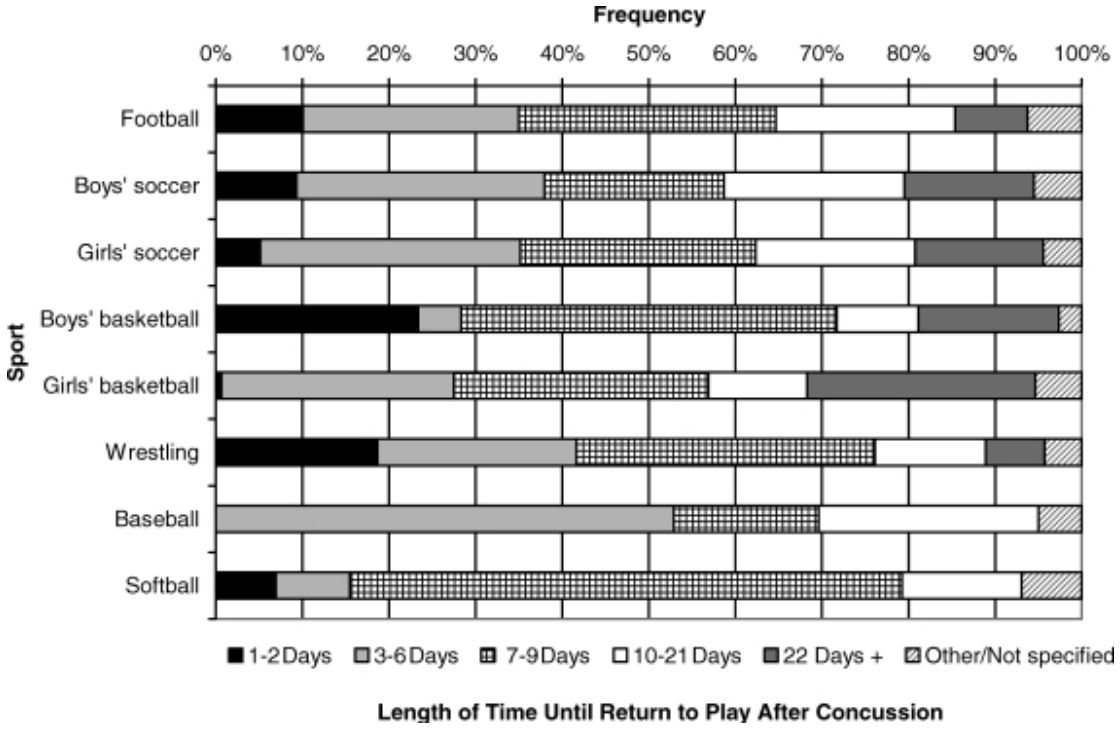
Concussion Rates Among US High School and Collegiate* Athletes, High School Sports-Related Injury Surveillance Study and National Collegiate Athletic Association Injury Surveillance System, United States, 2005-2006 School Year. Table 1 from © Gessel LM, Fields SK, Collins CL, Dick RW, Comstock RD. Concussions Among United States High School and Collegiate Athletes. *Journal of Athletic Training.* 2007;42(4):495-503. Page 497. Reproduced with permission from publisher.

Appendix D: National (US) estimates of concussion symptom resolution time for high school athletes.



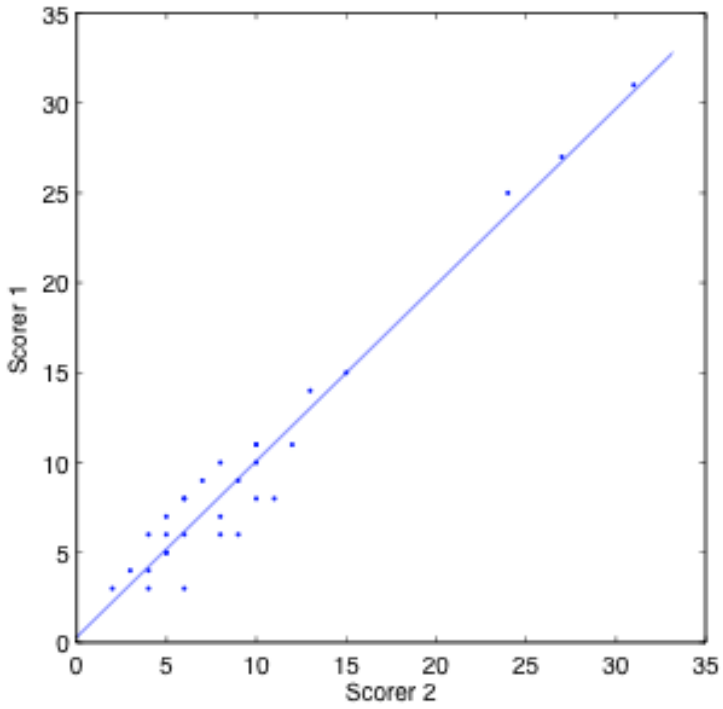
National estimates of concussion symptom resolution time for high school athletes, High School Sports-Related Injury Surveillance Study, United States, 2005-2006 School Year. Figure 1 from © Gessel LM, Fields SK, Collins CL, Dick RW, Comstock RD. Concussions Among United States High School and Collegiate Athletes. *Journal of Athletic Training.* 2007;42(4):495-503. Page 498. Reproduced with permission from publisher.

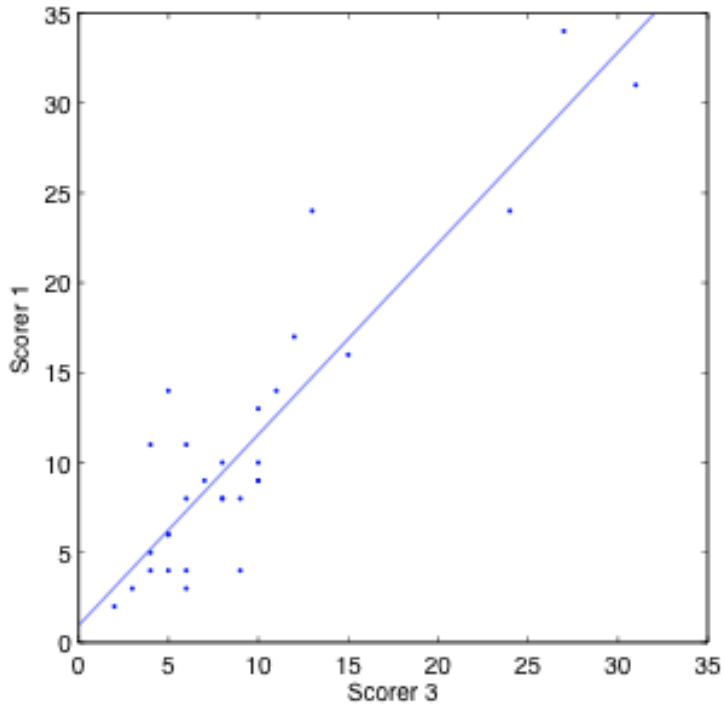
Appendix E: National (US) estimates of length of time until return to play after concussion for high school athletes.

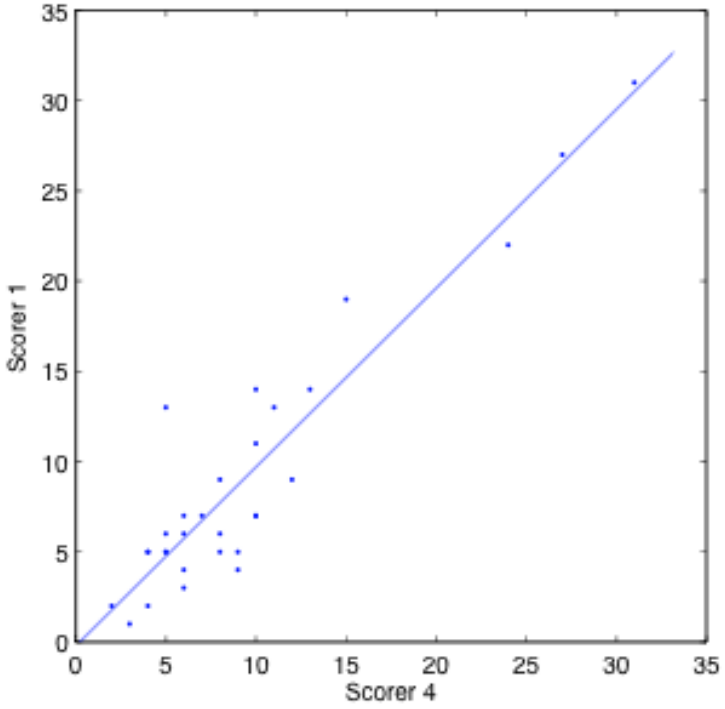


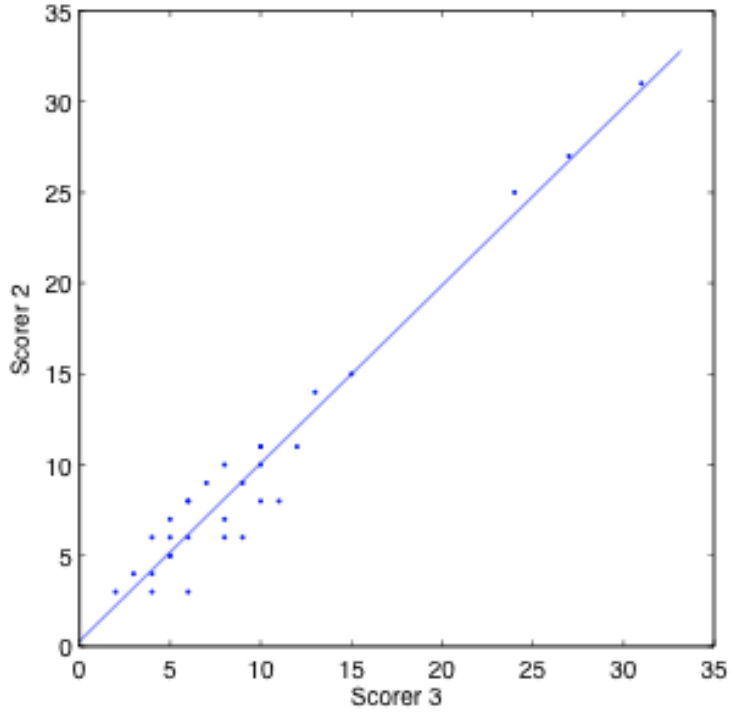
National estimates of length of time until return to play after concussion for high school athletes, High School Sports-Related Injury Surveillance Study, United States, 2005-2006 School Year. Figure 2 from © Gessel LM, Fields SK, Collins CL, Dick RW, Comstock RD. Concussions Among United States High School and Collegiate Athletes. *Journal of Athletic Training.* 2007;42(4):495-503. Page 498. Reproduced with permission from publisher.

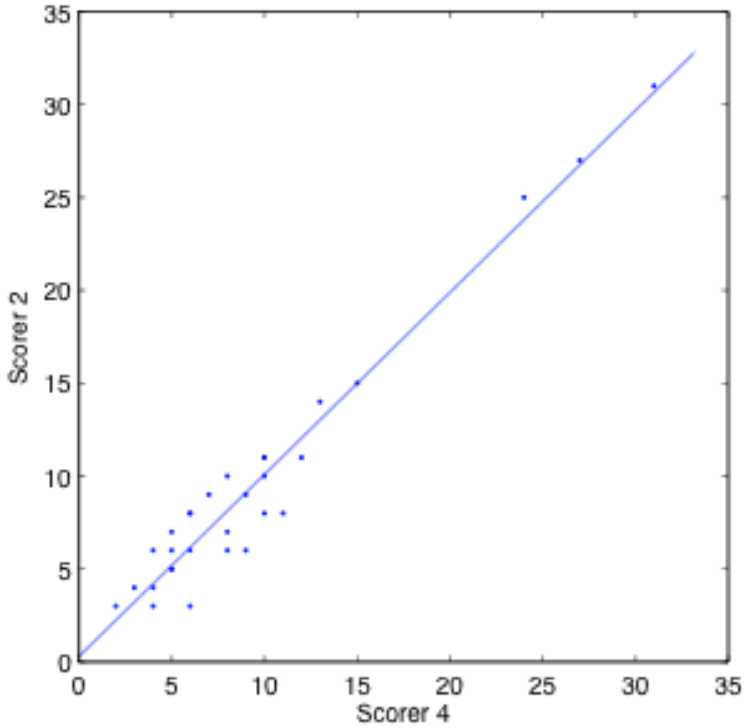
Appendix F: Correlation between experienced BESS raters: Experiment 1

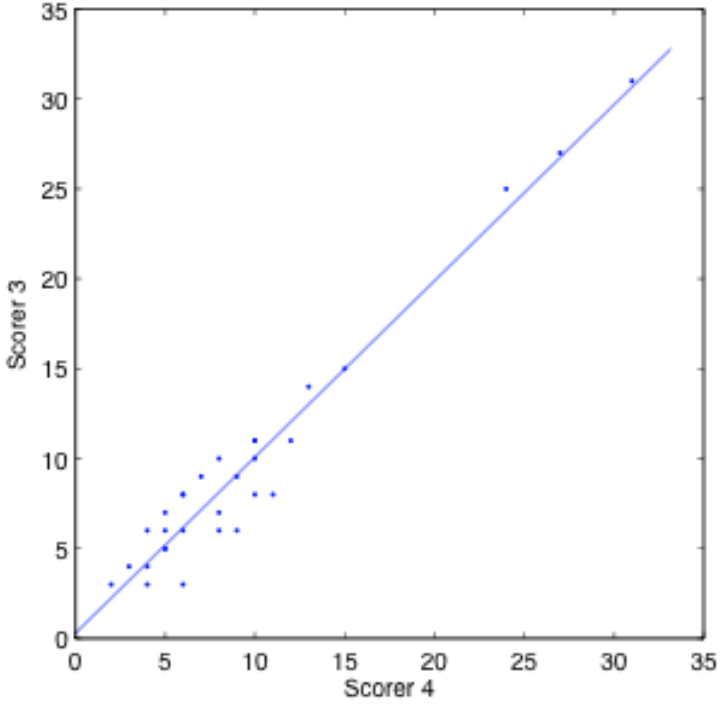




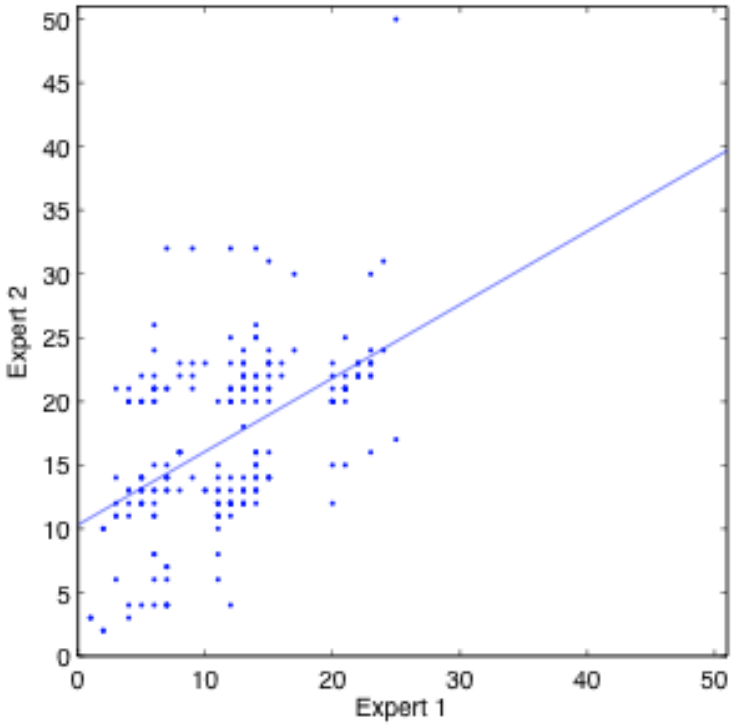

Scorer 1
Scorer 4
0
5
10
15
20
25
30
35

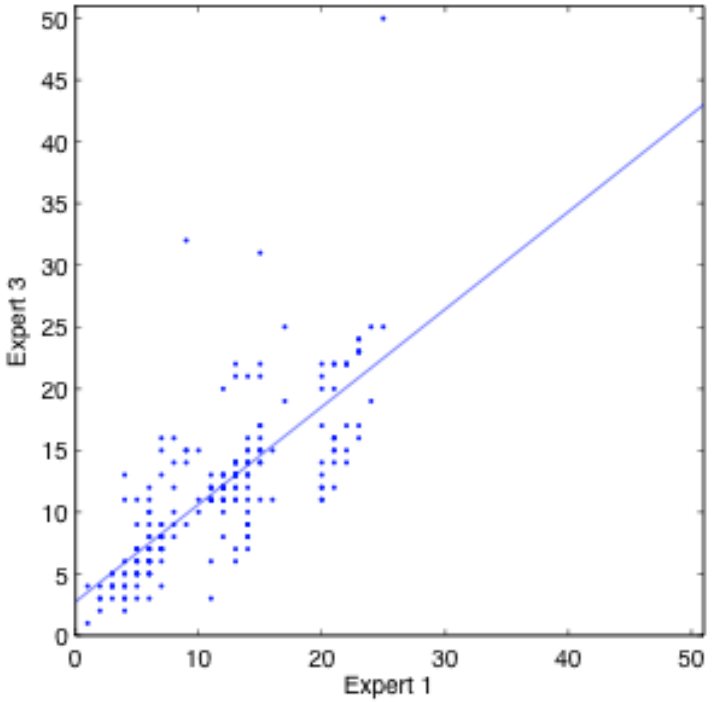

Scorer 2
Scorer 3
0
5
10
15
20
25
30
35

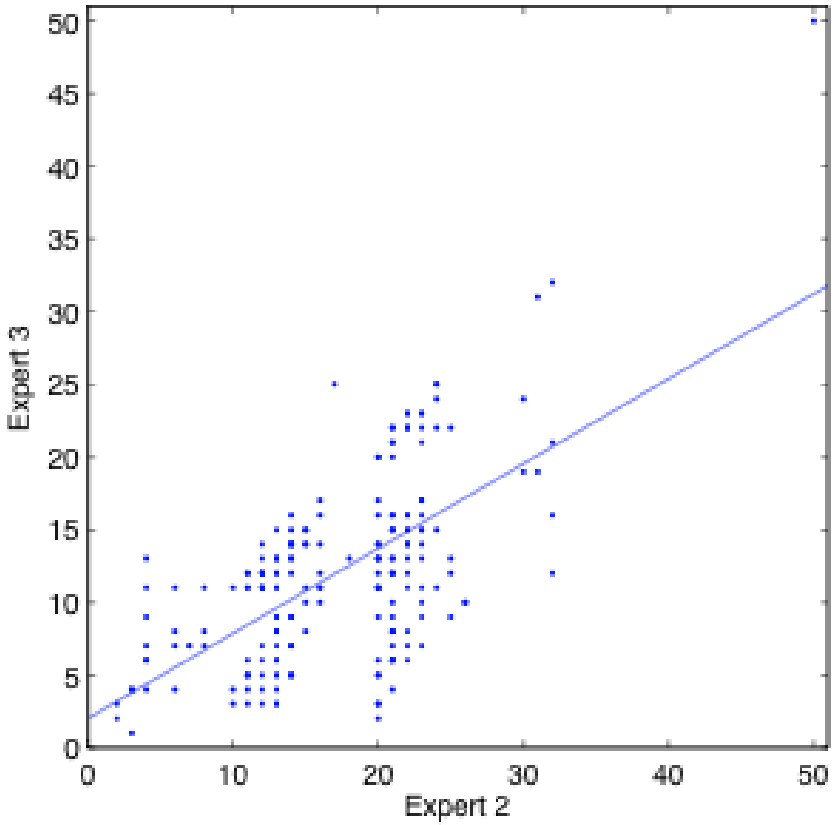

Scorer 2
Scorer 4
0
5
10
15
20
25
30
35


Scorer 3
Scorer 4
0
5
10
15
20
25
30
35

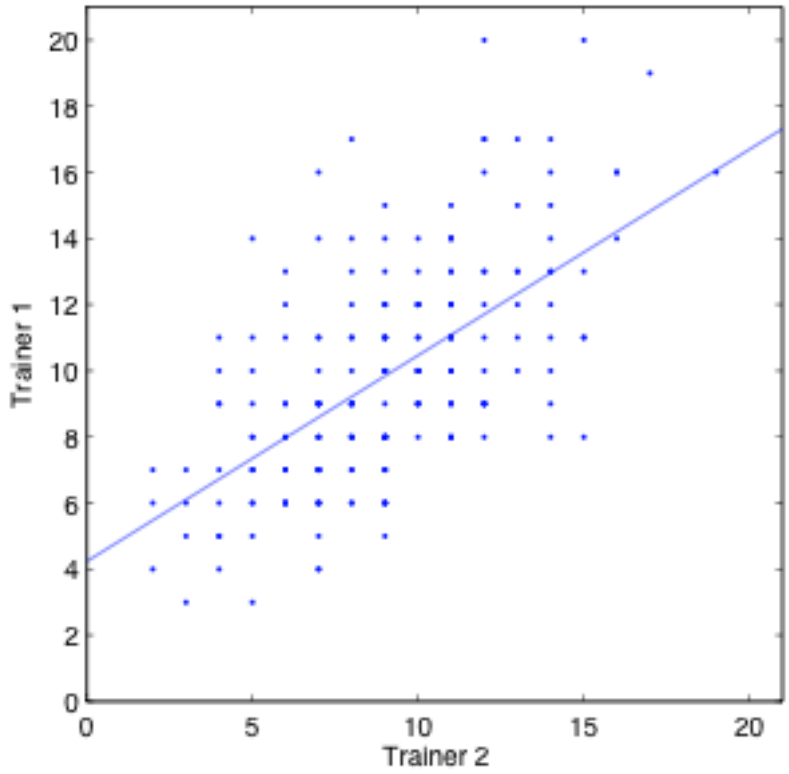
Appendix G: Correlation between experienced BESS raters: Experiment 2






Expert 3
Expert 2
0
5
10
15
20
25
30
35
40
45
50
0
10
20
30
40
50

Appendix H: Correlation between athletic trainer raters: Experiment 2



Appendix I: oBESS scores produced for subjects in Experiment 2 using the optimal algorithm from Experiment 1