2014

DEVELOPMENT AND VALIDATION OF THE ANTERIOR CRUCIATE LIGAMENT INJURY-RISK-ESTIMATION QUIZ (ACL-IQ)

Erich J. Petushek
Michigan Technological University

DEVELOPMENT AND VALIDATION OF THE ANTERIOR CRUCIATE LIGAMENT INJURY-RISK-ESTIMATION QUIZ (ACL-IQ)

By

Erich J. Petushek

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Applied Cognitive Science and Human Factors

MICHIGAN TECHNOLOGICAL UNIVERSITY

2014

This dissertation had been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Applied Cognitive Science and Human Factors.

Department of Cognitive and Learning Sciences

Dissertation Co-Advisor: *Edward T. Cokely*

Dissertation Co-Advisor: *Paul Ward*

Committee Member: *Gregory D. Myer*

Committee Member: *John J. Durocher*

Department Chair: *Susan Amato-Henderson*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

The completion of this dissertation would not have been possible without the help and support from many individuals and institutions. I would first like to thank the National Science Foundation (Grant No. 1108024), the Research Council of Norway (Grant No. 230163), and Michigan Technological University (specifically Dr. Edward Cokely) for their financial support and giving me the opportunity to learn from and collaborate with exceptional researchers around the world.

I would like to give special thanks to my advisor Dr. Edward Cokely, who provided extensive feedback for this work and for his constant support and encouragement throughout this project. I would also like to thank my co-advisor Dr. Paul Ward for his help collecting pilot data, recruiting participants and hosting me during part of my international research experience. This project would not have been possible without Dr. Greg Myer's help recruiting participants as well as providing feedback and the video clips which served as the foundation for this project. Thank you Dr. John Durocher for your comments and feedback throughout this project.

Beyond my committee I would like to thank the international hosts during my National Science Foundation Graduate Research Opportunity Worldwide experience, who helped bring this project to fruition. I would first like to thank Dr. Tron Krosshaug, my main host/supervisor, for his help, feedback, and support during this project as well letting me use his mountain bike to bring joy and balance during the research filled summer. Additionally, I would like to thank the other Oslo Sports Trauma Research Center personnel: Dr. Eirik Kristianslund, Kam Ming Mok, Oliver Faul, Dr. Agnethe Nilstad, Dr. Kathrin Steffen and Dr. Roald Bahr for their support, feedback, and help collecting data for this project. I would also like to acknowledge the support/feedback

from Dr. Mark Williams (Brunel University), Dr. Shenghua Luan, Dr. Lael Schooler, Dr. Jonathan Nelson, and Dr. Gerd Gigerenzer (Adaptive Behavior and Cognition group at the Max Planck Institute for Human Development) as well as Dr. Niklas Keller (Charité Hospital), Dr. Chris Richter and Dr. Kieran Moran (Dublin City University).

I would also like to give special thanks to Sean Wallace for creating the databases and websites for the online data acquisition platform. I would like to thank Dr. William Ebben, Adam Shultz, and Dr. Randy Jensen for reviewing and proofreading. I would like to thank all the individuals who helped recruit/distribute the online survey, especially: Kim Barber Foss, Sarah Leissring, Mitch Stephenson, Alexander Wolfe, Dr. Bryan Dixon, Dr. Christopher Simenz, Dr. Scott Drum, Cheri Baumann, Cora Ohnstad, Kris Rowe, Alison Regal and many other who I may have missed. Additionally, I would like to thank all of those who volunteered their time to participate in these research projects. Without your time and effort, this project would not have been possible.

Finally, I would like to thank my family and friends, who helped me throughout my studies, especially my mother for her constant love and encouragement to let me live out my dream of pursuing this degree. Dr. William Ebben, a mentor, colleague, and friend, who introduced me to research and taught me many skills and values, I thank you. I will also be forever grateful for Breanne Carlson's unconditional support, love, and constant sunshine and warmth in those numerous cold and snow filled days in the Upper Peninsula of Michigan.

ABSTRACT

Over 2 million Anterior Cruciate Ligament (ACL) injuries occur annually worldwide resulting in considerable economic and health burdens (e.g., suffering, surgery, loss of function, risk for re-injury, and osteoarthritis). Current screening methods are effective but they generally rely on expensive and time-consuming biomechanical movement analysis, and thus are impractical solutions. In this dissertation, I report on a series of studies that begins to investigate one potentially efficient alternative to biomechanical screening, namely skilled observational risk assessment (e.g., having experts estimate risk based on observations of athletes movements). Specifically, in Study 1 I discovered that ACL injury risk can be accurately and reliably estimated with nearly instantaneous visual inspection when observed by skilled and knowledgeable professionals. Modern psychometric optimization techniques were then used to develop a robust and efficient 5-item test of ACL injury risk prediction skill—i.e., the ACL Injury-Risk-Estimation Quiz or ACL-IQ. Study 2 cross-validated the results from Study 1 in a larger representative sample of both skilled (Exercise Science/Sports Medicine) and un-skilled (General Population) groups. In accord with research on human expertise, quantitative structural and process modeling of risk estimation indicated that superior performance was largely mediated by specific strategies and skills (e.g., ignoring irrelevant information), independent of domain general cognitive abilities (e.g., metal rotation, general decision skill). These cognitive models suggest that ACL-IQ is a trainable skill, providing a foundation for future research and applications in training, decision support, and ultimately clinical screening investigations. Overall, I present the first evidence that observational ACL injury risk prediction is possible including a robust technology for fast, accurate and reliable

measurement—i.e., the ACL-IQ. Discussion focuses on applications and outreach including a web platform that was developed to house the test, provide a repository for further data collection, and increase public and professional awareness and outreach (www.ACL-IQ.org). Future directions and general applications of the skilled movement analysis approach are also discussed.

CHAPTER 1: INTRODUCTION

Female athletes are approximately three times more likely to tear an anterior cruciate ligament (ACL) compared to their male counterparts (Prodromos, Han, Rogowski, Joyce, & Shi, 2007). Younger (age, 15-25 years) athletes participating in landing and cutting sports such as basketball and soccer are at greatest risk for ACL injury (Griffin et al., 2006). This elevated risk coupled with a nearly two-fold increase in female sports participation over the last 30 years (Irich, 2012; NFSHSA, 2012) has led to a rapid rise in ACL injuries in females (≈ 3 injuries per 10,000 athlete-exposures or around 1 injury per 30 individuals during a sports season). Anterior cruciate ligament surgery cost has been shown to be approximately $5,000, which doesn't include the post-operative rehabilitation or lost time from work/sport (Swenson et al., 2013). In the U.S. alone, the annual cost of ACL injury likely exceeds $3 billion (Kim, Bosque, Meechan, Jamali, & Marder, 2011). Additional consequences of ACL injury include time out of sport/school, scholarship loss, significant risk for re-injury, and osteoarthritis (Ardern, Webster, Taylor, & Feller, 2011; Lohmander, Englund, Dahl, & Roos, 2007; Wright et al., 2007). Interestingly, most ACL injuries occur in a non-contact situation (Agel, Arendt, & Bershadsky, 2005; Krosshaug et al., 2007b) and are likely preventable (Hewett, Myer, Ford, Paterno, & Quatman, 2012).

Neuromuscular training can reduce the relative risk for non-contact ACL injury by 73.4%. Unfortunately, to prevent one injury, 108 individuals must participate in training (Sugimoto, Myer, McKeon, & Hewett, 2012). The time commitment involved in training this number of individuals is non-trivial. Moreover, the prevention techniques including physical training likely only benefit high-risk athletes (Myer, Ford, Brent, & Hewett, 2007). Administering prevention programs to the low-risk likely constitutes an

inefficient use of time and resources. If training instead targeted the high-risk, the number of individuals needed to train to prevent an injury would be reduced.

Anterior cruciate ligament injury risk screening tools have been identified and developed for high school age (15-19 years) female athletes using prospective 3D biomechanical analysis procedures. Specifically, 205 young (age ≈ 16 years) female athletes were screened (using 3D biomechanical motion analysis of the drop vertical jump) and tracked through two sports seasons (13 months) (Hewett et al., 2005). Knee abduction moment (i.e., torque generated rotating the knee inward) and angles (initial contact and peak values) were found to be significant predictors of ACL injury status for the 9 non-contact ACL injuries. Peak knee abduction moment was found to be the best independent predictor of injury status displaying sensitivity of 78% and specificity of 73%. Three-dimensional, biomechanical laboratory screening is not feasible for widespread field-based or clinical use because of the high associated cost, specialized equipment, and implementation times. However, the existing, high-precision tools have provided a biomechanical risk assessment standard that can be used for validation of and comparison with alternative screening approaches.

A clinical based nomogram was developed for identifying individuals at high-risk for ACL injury[1] and involved the use of two standard video cameras, measuring tape, scale, and isokinetic dynamometry (Myer, Ford, Khoury, Succop, & Hewett, 2010b). Analyses of the ACL nomogram screening method revealed considerable advantages over 3D biomechanical laboratory screening techniques. For example, it is relatively accurate and yet considerably quicker and less expensive than 3D biomechanical analysis, requiring only a modest amount of time (≈ 5 to 15 minutes per individual) and more

[1] The ACL nomogram used logistic regression to predict high knee abduction moment (>21.74 Nm), not injury risk.

affordable equipment. The ACL nomogram approach represents major progress in the development of cost-effective, efficient screening tools. Central to the current thesis, theoretically, the success of the ACL nomogram also suggests that other even simpler, less-expensive methods based on observational movement diagnostics may also provide feasible screening methods.

Observational Movement Diagnosis

Visual inspection or observational movement diagnosis is one alternative screening method, which would reduce screening time and cost, while preserving relatively high-risk assessment accuracy (Knudson, 2013). For example, a practitioner (i.e., coach, athletic trainer, physical therapist, etc.) could nearly instantaneously assess ACL injury risk by observing a task where movement patterns would be similar to those that cause ACL injury (i.e., jump landing, cutting, etc.). Unfortunately, the validity and consistency of observational assessment of ACL injury risk is poorly understood. A small body of research has investigated the underlying psychological mechanisms that may give rise to differences in observational movement diagnosis skill. Nevertheless, identifying athletes or patients with abnormal/flawed or inefficient movement is a common task for many coaches (Knudson, 2000), sport judges (Plessner & Haar, 2006), and sports medicine practitioners (i.e., physical therapist, athletic trainer, etc.) (Jensen, Guyer, Shepard, & Hack, 2000). Research in these disciplines can contribute to the understanding of the overall dynamics of observational movement diagnosis. The following review will discuss the quantification methods and initial evidence relating to observational ACL injury risk estimation. Additionally, the various environmental and perceptual-cognitive factors influencing skilled and expert performance will be highlighted.

Measuring Skill and Expert Performance

Various psychometric measurement methods have been used to assess differences in skilled performance for more than a century. Psychometric assessments are commonly used in the selection of employees, awarding of promotion, optimization of training, and have also been extensively validated for use in clinical and educational settings. Specifically, the term "psychometrics" refers to the field of study specializing in theory and measurement techniques in psychology. Within the field of psychometrics, the expert-performance approach provides a systematic framework for assessing high levels of domain-specific expertise, grounded in the use of domain-specific or representative tasks that capture the essence of expertise (Ericsson & Lehmann, 1996). Assessing the level of expert-performance under standardized conditions is the first step in the expert-performance approach. Once objective performance is assessed, cognitive process tracing techniques can be used to investigate the underlying mechanisms mediating superior performance. Additionally, the development of expertise can be investigated through examining practice history or by conducting prospective training studies. The information gained through this systematic approach can then be used to develop training programs or decision support systems that reliably improve performance. Consistent with standards in psychometric theory, (Ericsson & Lehmann, 1996) define expert performance as: "consistently superior performance on a specified set of representative tasks for a domain." This definition is similar to what industrial-organizational (I-O) researchers would call a work sample test—defined as "a test in which the applicant performs a selected set of actual tasks that are physically and/or psychologically similar to those performed on the job" (Ployhart, Schneider, & Schmitt, 2006).

Industrial-organizational researchers and human resources professionals often use supervisor ratings of performance as criteria for "performance" as opposed to work sample tests (Viswesvaran, Ones, & Schmidt, 1996). Supervisor ratings are subjective evaluations of job performance, usually across multiple dimensions of a job (Viswesvaran et al., 1996). Meta-analyses of work sample tests and supervisory ratings of performance have revealed correlations coefficients of .33 (Roth, Bobko, & McFarland, 2005) and .32 (Schmitt, Gooding, Noe, & Kirsch, 1984). However, work sample tests are more reliable than supervisory ratings (mean r = .71 vs. r = .60) (Roth et al., 2005). Since observational movement diagnosis is a specific skill and may be part of many jobs (i.e., physical therapy, athletic training, strength & conditioning, and athletic coaching), a more objective evaluation of performance is likely to be preferable to other types of less objective supervisory ratings. Specifically, in the context of observational movement diagnosis, "expert" or "skilled" observers should display accurate and consistent judgments during a task that is representative of the constraints they would encounter while diagnosing a specific movement in representative, ecological conditions.

A Representative Task

Using the expert-performance approach, Ericsson and Ward (2007) describe a representative task as "… experts' real-world performance is scrutinized to identify naturally occurring situations that require immediate action and that capture the experts' superior selection or execution of actions in the associated domain." Brunswik (1956) conception of "representativeness" in the context of task or experimental design is that perceptual variables should be gathered from an organism's natural environment in which they routinely interact, that is, the environments that participants are adapted to. The expert-performance approach, however, requires that tasks also capture the essence of superiority. For example, de Groot (1978) identified critical chess game situations

where players were to generate the next best move, which has been shown to be highly correlated with tournament ratings (van der Maas & Wagenmakers, 2005) and has subsequently been used as a model for studying expertise in other domains using the expert-performance approach (Ericsson & Williams, 2007). The selection of tasks that capture expertise may be difficult when initially capturing expert or skilled performance, as "expert" cannot be ascertained prior to examining performance on a representative task; and for many domains a "gold standard" of performance is not available. The steps for choosing a representative task using both Brunswikian and expert-performance approaches would be:

1.) Identify a task that is often (or should be) performed in the chosen domain.
2.) Randomly sample the stimuli for the given domain/environment.
3.) Select the stimuli that best discriminate between performance levels (Ericsson, 2007).

For ACL injury risk estimation, the stimuli should be the individuals and screening tasks to be assessed. Since young (15-25 years) females participating in landing and cutting sports are at the greatest risk for ACL injury; this population should be the target of the judgment generalizations. Additionally, the screening task (i.e., drop vertical jump) must also be investigated as a representative task. Several ACL injury risk factors have been identified (for reviews see Hewett, Zazulak, Krosshaug, and Bahr (2012); Smith et al. (2012a, 2012b)). Many of the risk factors are difficult to change/modify (i.e., game vs. practice, gender, joint geometry/laxity, or menstrual cycle phase), thus, may not be suitable to serve as risk screening factors. Thus, only the modifiable risk factors related to movement or landing mechanics will be explored because of their modifiability and relative assessment ease.

Various movements have been used to assess high-risk movement strategies. The drop vertical jump (Ekegren, Miller, Celebrini, Eng, & Macintyre, 2009; Hewett et al., 2005; Mizner, Chmielewski, Toepke, & Tofte, 2012; Myer et al., 2010b; Nilstad et al., 2014; Noyes, 2005; Padua et al., 2009; Whatman, Hing, & Hume, 2012; Whatman, Hume, & Hing, 2013a), single leg squat (Ageberg et al., 2010; Stensrud, Myklebust, Kristianslund, Bahr, & Krosshaug, 2010), and tuck jump (Herrington, Myer, & Munro, 2013; Myer, Ford, & Hewett, 2004) have been used to directly or indirectly assess ACL injury risk. The drop vertical jump (and associated biomechanical variables) is, however, the only movement that has been used to successfully predict ACL injury status in the target population (i.e., young females) (Hewett et al., 2005). With superior performance operationally defined and stimuli/judgment task identified, the next section will describe the procedures for developing a valid test aimed to assess observational ACL injury risk estimating skill.

Test Development and Evaluation

In addition to assessing differences in skill, a standardized assessment of observational ACL injury risk estimation skill could serve many purposes including contributing to programs evaluating the efficacy of observational assessment as an ACL injury risk screening method. The test may also provide an efficient assessment of differences that result from various types of training or other interventions. This section will describe the components of test development as well as score interpretation and validation.

Test development can be informed by two general theories or approaches including classical test (CTT) or item response (IRT) theories. These methods can be used to analyze and select appropriate test items based on the assessment needs. Following will

be a summary of the general procedures, benefits, and shortcomings of each approach (see Hambleton and Jones (1993); Lord (1980) for further review).

Classical test theory has historically been the dominant method for test development. In general IRT is described as "item-based" whereas CTT is described as "test-based," although modern approaches in CTT do incorporate individual item analysis. The main differences reside in the modeling assumptions and sample requirements. The CTT framework assumes a linear relationship between test score and ability and that test scores and error scores are unrelated. Additionally for CTT, item statistics such as difficulty and discriminability are dependent on the specific sample and overall test score. This is problematic if the sample used for test development is different than the intended examinees.

Item response analysis assumes a nonlinear relationship between item performance and ability, which requires fitting complex models to the test data. This requires a large heterogeneous sample (e.g., over 500) but allows the specific item characteristics to remain sample independent. Similar item statistics are calculated such as difficulty and discriminability plus an additional "guessing" factor. Individual item modeling in IRT can be used to describe overall test "informativeness" (item information curves), which allows for flexibility when selecting items based on the developers test objectives. For example, if a developer is interested in creating a test to identify high ability individuals, items that display higher discrimination and greater difficulty will provide more information towards the higher "ability" end of the spectrum and should be selected. Once a test is developed, validation is essential to interpret the score meanings and understand the underlying mechanisms of test performance.

Validation

According to the Standards for Educational and Psychological Testing (1999):

> "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself." (p. 9)

> "Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose." (p. 11)

These definitions were influenced by Messick's unitary concept of construct validity which is corroborated by evidence from five sources: content, response processes, internal structure, relations to other variables, and consequences (Messick, 1995). These components of construct validity evidence will be discussed in detail in the context of developing a test to assess the construct: observational ACL injury risk estimation skill.

Content

Content evidence is the "relationship between a test's content and the construct it is intended to measure" (American Educational Research et al., 1999). The test items/tasks should be solely related to the construct. Describing item/task inclusion rationale and the qualifications of the individual(s) who chose the items can assess this degree of relationship. In the context of ACL injury risk estimation, item inclusion

rationale can be described by representative sampling of stimuli, which was discussed, and in accordance with the Brunswikian point of view. Criteria for performance/expertise was addressed and composed of accuracy and consistency of judgment (ACL injury risk estimation). Thus, the items of the test must ensure judgment accuracy and consistency are captured. Another example of content evidence would be to assess if "experts" agree that the content/items measure what the test is intended to measure.

Response Process

Response process evidence is the "fit between the construct and the detailed nature of performance … actually engaged in" (American Educational Research et al., 1999). Are the cognitive processes while taking the test also representative of the construct intended to be measured? Cognitive process tracing methods such as think aloud or verbal protocol analysis, eye-tracking, and response times may be used to provide evidence for this component, as well as computational modeling techniques (e.g., multinomial comparison). To further illustrate, if the skill of ACL injury risk estimation is the central skill to be measured, and this skill is thought to be a function of extensive domain-specific knowledge and elaborate memory structure, a valid test would reveal a tight link between cognitive process-tracing measurements of skilled individuals, domain-specific knowledge, and overall task performance.

Internal Structure

Internal structure refers to the "interrelations among the scored aspects of task and subtask performance" (Messick, 1995). Reliability, overall factor structure, and individual item response characteristics can be used to determine the internal structure.

Various methods can be used to quantify these aforementioned components. Assessment of internal structure is also essential for test refinement to reduce test length.

Relations to Other Variables

Are scores from other tests intended to measure the same/different construct related/unrelated with the present test (convergent/discriminant evidence)? Do the test scores improve following training (based on theoretical enhancement of the construct) or change based on a length of time purported to influence the construct? Are there differences/similarities in test scores between groups or individuals (i.e., athletes, physical therapists, etc.) that the construct would predict? Answering these questions is part of a thorough evaluation of construct validity.

Another facet of this category is predictive evidence (also sometimes called criterion or concurrent validity). An idealized example would be providing evidence that skilled performers (as identified through a test), who later screened individuals and made intervention recommendations (i.e., training), reduced ACL injury rates as compared to control conditions.

Consequences

Anticipated beneficial and detrimental consequences should be addressed. Because unintended consequences cannot be initially determined, they should be assessed in the future. Potential sources of invalidity such as construct underrepresentation or construct-irrelevant variance should be investigated (Messick, 1990). Finally, classification cut-points should be characterized and justified. A specific example for a test of ACL injury risk estimation skill would be if someone was found to be skilled, went on to assess individuals but their assessment was wrong and high-risk individuals were not correctly identified. ACL injuries are complex and multifactoral,

thus knowing risk with 100% certainty is unlikely. Interactions among various risk factors is currently unknown, therefore by only assessing one risk factor (movement abnormalities) there is a possibility of bias and under diagnosis, which are issues that should be clearly conveyed with the test results (e.g., this is a research instrument that is validated for use in research settings—it is not currently validated as a clinical assessment of risk). The utility of screening tests should also be characterized and quantified by understanding the associated costs due to misclassification and any other ethical issues that may be of concern (e.g., what kind of detrimental effect could poor performance have on a participant's career—does the potential benefit outweigh the potential harms).

Foster and Cone (1995) stated, "Science rests on the adequacy of its measurement. Poor measures provide a weak foundation for research and clinical endeavors." The cumulative evidence from psychometric theories and the five factors (i.e., content, response process, internal structure, relations to other variables and consequences) provides a substantial base for assessing the adequacy of the interpretations of test scores, and increasing the likelihood that test scores are thoroughly valid, robust, ethical, and reliable. The next section will describe the current evidence and limitations for expertise in the context of ACL injury risk estimation.

Estimating ACL Injury Risk

Skill Evidence

Currently, no test/method has been developed to directly assess observational ACL injury risk estimating ability. Previous research has, however, compared observational diagnosis performance with 2D and 3D biomechanical variables purported to assess ACL injury risk. Specifically, four studies used various experimental approaches

to examine the accuracy and consistency of observational diagnosis of the drop vertical jump and will be, hereafter, discussed.

Subjective judgments of frontal plane knee control during live drop vertical jumps were compared with 2D frontal plane knee angle (Stensrud et al., 2010). One observer assessed the drop jump performance of 186 athletes on a three-point scale (0 = good performance; 1 = reduced performance; and 2 = poor performance) based on the amount of frontal plane (i.e., medial/lateral) knee motion. Judgment test-retest reliability was also assessed approximately 30 days later. Subjective classification accuracy was compared to biomechanical measurement (2D frontal plane knee valgus angle) using area under the receiver operating characteristic (ROC) curve (AUC). Results revealed an average AUC measure of .83 (95% confidence intervals of .77-.89). The average global accuracy measurement (.83) is considered 'good' according to the traditional academic point system and similar to the current clinical model (ACL nomogram) for ACL injury risk estimation, which displayed an AUC of .85, albeit predicting high knee abduction moment (Myer et al., 2010b). The observers test-retest Kappa value was .90. Kappa is a measure of reliability not agreement, thus the requisite level of intra-rater agreement or judgment consistency is not clear. The criterion, 2D frontal plane knee valgus angle, has been shown to be correlated with knee abduction moment with a large effect size (r = .59) (Mizner et al., 2012). Thus, the results revealed by Stensrud and colleagues (2010) provided initial indirect evidence of skilled performance during observational ACL injury risk estimation. Subjective judgments of frontal plane knee control during live drop vertical jumps were also compared with 3D knee abduction angle and moment (Nilstad et al., 2014).

Using similar stimuli (i.e., live drop vertical jumps) and the same subjective rating system as Stensrud et al. (2010), Nilstad et al. (2014) assessed the judgment

accuracy of three physiotherapists using 3D biomechanical measurement of knee abduction angle and moment as the criteria. The raters exhibited AUC values between .85 and .89 when compared to 3D knee abduction angle and AUC values between .56 and .57 when compared to 3D knee abduction moment. Percent agreement ranged from 70-90% and no statistical differences were found between raters, suggesting good inter-rater agreement. The low judgment accuracy when compared to 3D knee abduction moment (i.e., AUC .56 - .57) is justifiable as the normative relationship between 3D knee abduction angle and moment in this group was r = .04. Moreover, when asked to judge the amount of knee valgus (i.e., abduction) motion, physiotherapists exhibited accurate judgments when compared to 3D knee abduction angle, but inaccurate judgments when compared to 3D knee abduction moment. The raters were not asked to judge the amount of knee abduction moment (which was not related to angle), thus a direct comparison of judgments with knee abduction moment cannot be ascertained. This research does corroborate the evidence that raters can discriminate individuals with various amounts of 3D knee abduction angle with sufficient performance. In the following two studies, subjective judgment of frontal plane medial knee motion during the drop vertical jump was compared with 3D biomechanical and 2D video criteria.

Three physiotherapists with 12 ± 3 years of clinical experience judged 40-frontal plane videotaped females (age ≈ 15 years) perform drop vertical jumps[2] (Ekegren et al., 2009). The observers were given the specific guidelines of: “If the patella moves inward and ends up medial to the first toe, rate the individual as high-risk,” or “If the patella lands in line with the first toe, rate the individual as low-risk.” The same judgments were

[2] 120 total video clips were shown, as the 40 individuals jumped three consecutive times and the observer had to make a summary judgment following the three jumps.

also reassessed two weeks later to determine test-retest reliability. An “expert” who was able to pause, decelerate, and rewind the videos initially assessed the risk level (i.e., was the patella medial to the first toe?). These “expert” ratings were then compared to 3D knee abduction motion (maximum-minimum angle) to determine the optimum cut-off for the “true” risk rating. Mean (across the three jumps) knee valgus motion greater than 10.83 degrees was considered truly high-risk and below truly low-risk.

All three physiotherapists performed similarly displaying a multi-rater kappa value of .90 at time one and .77 at time two. The best physiotherapist exhibited sensitivity and specificity values of 87 and 72%, respectively, indicating sufficient accuracy. This same physiotherapist displayed an agreement or reproducibility value of 88% between sessions (Kappa value of .75), indicating sufficient test-retest reliability. This study provides additional evidence for skilled performance during observational ACL injury risk estimation. There were, however, several limitations that may influence the generalization of results to ACL injury risk estimating ability and will be, hereafter, discussed.

First, the study instructions were to assess if the patella (knee) was medial to the first toe (from a frontal plane vantage point). The dichotomous variable, knee medial to the toe, has not been directly identified as a risk factor for ACL injury and may not be related to current biomechanical ACL risk factors (i.e., 3D knee abduction moment or angle). The assessment of 2D medial knee motion, which is theoretically related to knee medial to the toe, has a low correlation with 3D knee abduction moment (r (18) = .20) and angle (r (18) = .18) (Pilot data). Similarly, Whatman et al. (2012) revealed that knee medial to the toe was unlikely related to 3D knee abduction angle (1.0 degree difference in means between true “patella medial to the second toe” and “patella not medial to the second toe” groups). These results are in contrast to Ekegren et al. (2009) where 3D knee

abduction motion (>10.83 degrees) predicted knee inside toe location with 87% sensitivity and ≈ 68% specificity (visually estimated with ROC curve). Differences in foot alignment and placement during landing, tibial rotation, as well as 3D marker placement may account for this discrepancy.

Second, the observer responses were not directly compared to a criterion that assessed if the knee was actually medial to the toe. The ratings were compared and based upon an optimal cut-off produced by a ROC curve from "expert" rating of knee medial to the toe and 3D knee abduction motion. For instance, an individual could theoretically jump and land with their knee medial to their toe (judged as "high risk") but display a valgus motion of less than 10.83 degrees (categorized as "truly low risk") recording a "false alarm" when truly should have been a "hit."[3]

Third, a twenty-minute training video was presented to the physiotherapists prior to the judgments. This video provided background information about ACL injury risk and rating instructions. Lastly, the physiotherapists were allowed to practice with feedback and were able to discuss their practice judgments with other raters. These attempts to simplify the task (using sub-optimal criteria) and standardize rater training reduced task complexity and thus representativeness. Subjective judgments of knee location, relative to the toes, during videotaped drop vertical jumps were also compared in a larger sample of physiotherapists (Whatman, Hume, & Hing, 2013b)

Whatman et al. (2013b) conducted a similar study to that of Ekegren et al. (2009) but included a greater number of physiotherapist's (N = 66), did not provide the raters with instructions, and used a younger mixed gender population (11 female and 12 male;

[3] Pilot data from 20 demographically similar individuals (≈ 16 year old female athletes) revealed a knee medial to first toe ("high risk") in 37 out of the 40 legs assessed, whereas abduction motion of greater than 10.83 degrees ("truly high risk") was recorded in only 6 out of the 40 legs, indicating apparent disagreement between these two criteria.

age ≈ 11 years). The raters were instructed to perform a dichotomous decision task in which they assessed if the patella moved medial to the second toe (yes = poor and no = good) during the performance of a drop vertical jump (similar to Ekegren et al. (2009), but the second toe was considered, not the first). The observer ratings were compared with the consensus ratings of three "experts" which used video analysis software to slow, pause, or replay the video as well as overlay lines to provide the "true" classification of "poor" or "good" (i.e., if the patella was truly medial to the second toe based on the same video footage). These expert consensus ratings were also compared to 3D and 2D quantitative motion analysis measures (to confirm a valid criterion). Criterion confirmation results indicated that individuals rated by the "expert consensus" as having a patella medial to the second toe were very likely to have increased 3D peak hip adduction (5.2 degree difference in means between truly "good" and "poor" groups), internal rotation (6.3 degree difference in means between truly "good" and "poor" groups) and 2D knee frontal plane projection angle (15.3 degree difference in means between truly "good" and "poor" groups). Expert ratings were likely not related to 3D knee abduction angle (1.0 degree difference in means between truly "good" and "poor" groups).

For the primary decision tasks, the physiotherapist ratings demonstrated sensitivity and specificity interquartile range (IQR) values of 61-81 and 71-96%, respectively. Percent agreement and agreement coefficient (similar to Kappa and described by Gwet (2012)) within raters was 79% and .60, respectively. The authors also analyzed judgment accuracy in relation to the years of experience. The diagnostic odds ratio, a collective indicator of performance (Glas, Lijmer, Prins, Bonsel, & Bossuyt, 2003), revealed that performance was likely not different between physiotherapists with less than 5 years and 10-14 years of experience. However, physiotherapists with greater than

14 years of experience likely attained higher levels of performance compared to physiotherapists with 5-9 years (diagnostic odds ratio 3 times better). Additionally, a postgraduate qualification did not improve rating performance. Overall, this study provided evidence of individual differences in judgment accuracy and also indicates superior performance is attainable (albeit using a non-ideal criterion or judgment task).

Summary

In addition to the limitations associated with criterion choice and judgment task instructions, three of the aforementioned studies used a limited number of raters (one observer in Stensrud et al. (2010), three observers in Nilstad et al. (2014) and Ekegren et al. (2009)) limiting the assessment of individual differences in ability. When a larger sample of observers were studied by Whatman et al. (2013b), initial evidence for skill-based differences in observational movement diagnosis performance emerged, but generalizations to ACL injury risk estimation are limited due to the criterion/judgment task (knee medial to the toe) and representativeness of stimuli (i.e., individuals ≈ 11 years of age are not at greatest risk for ACL injury). Moreover, all of these studies utilized physio- or physical therapists; accordingly, results cannot be generalized to other individuals who would benefit from assessing ACL injury risk including orthopedic doctors, sport coaches, strength & conditioning coaches, athletes, and parents of athletes. In summary, this cumulative body of evidence suggests individuals may have the capacity to accurately assess ACL injury risk by simple observation. However, limitations need to be addressed and a systematic approach for assessing ACL risk estimation skill must be developed.

Despite the lack of direct evidence for superior performance in observational ACL injury risk estimation, it would be beneficial to discuss the possible factors influencing performance. Herbert Simon, a Nobel Laureate, developed an analogy of behavior (or performance): "Human rational behavior (and the rational behavior of all physical symbol systems) is shaped by a scissors whose two blades are the structure of task environments and the computational capabilities of the actor" (Simon (1990), p. 7). Thus, one must understand the interacting system (person, process, and environment) to better understand why and when performance is sufficient.

Environmental Factors

Critical features must first be identified in order for injury risk to be assessed or movement analyzed. For example, the biomechanical risk factors for ACL injury are knee abduction moment and knee abduction angles (Hewett et al., 2005), with knee abduction moment being the best predictor. These variables may be observable (angles, relative position, etc.) or inferred (moments, muscle activation, etc.) from other visible variables. Observer inaccuracy of various walking and running/cutting variables has been documented and varies considerably by variable type (Krosshaug et al., 2007a; Williams, Morris, Schache, & McCrory, 2009). Spatio-temporal variables such as step length, stance duration, and cadence during walking were judged with higher accuracy than kinematic/kinetic variables such as joint angles and power generation (Cohen's d = 1.81) (Williams et al., 2009). Similar results were found during a running/cutting movement where speed variables resulted in lower judgment errors compared to joint angle assessment (Krosshaug et al., 2007a). For these studies, the raters assessed one variable at a time, thus the interactive effect of the number of variables on judgment accuracy could not be assessed. However, if multiple variables are observed, to the extent these variables need to be integrated to make a judgment, there may be a greater demand on short term memory and perceptual dynamics, which could influence performance (Bays & Husain, 2008). Specifically, performance has been shown to decrease in environments' with large number of cues (Karelaia & Hogarth, 2008). In general, these results indicate that the type of variable dictates the observation rating accuracy with a trend towards greater accuracy when considering less abstract variables.

A summary of the variables used to estimate ACL injury risk (knee abduction moment), their respective ecological validities, and cue utilization profiles are located in

Figure 1. The lens model, developed by Egon Brunswik, is a conceptual framework for understanding "achievement" or judgment performance by comparing the relationship between the human and an idealized (normative) judgment process (Brunswik, 1956; Gigerenzer & Kurz, 2001; Wigton, 2008). The judge (left side in Figure 1) uses "proximal" variables or cues in the uncertain environment to infer the current state (in this case, ACL injury risk status). These cues serve as surrogates for the actual state just as variables serve as predictors in a regression model (this metaphor is, however, psychologically implausible as humans do not optimize like statistical models (Gigerenzer, 1991)). The ability of the judge to correctly assess these cues (which are compared to an objective criterion) is considered the utilization coefficient. There often exist cues that are related or correlated with one another (inter-cue redundancy or vicarious functioning). Thus, the judge must choose the cue(s) that relate or correlate best with the current state. The relationships between the cues and the actual state are considered by Brunswik to be the ecological validities. As Figure 1 suggests, achievement of ACL injury risk estimation has not been directly assessed by one single study. Summarizing the literature from this conceptual framework can lead to significant insights.

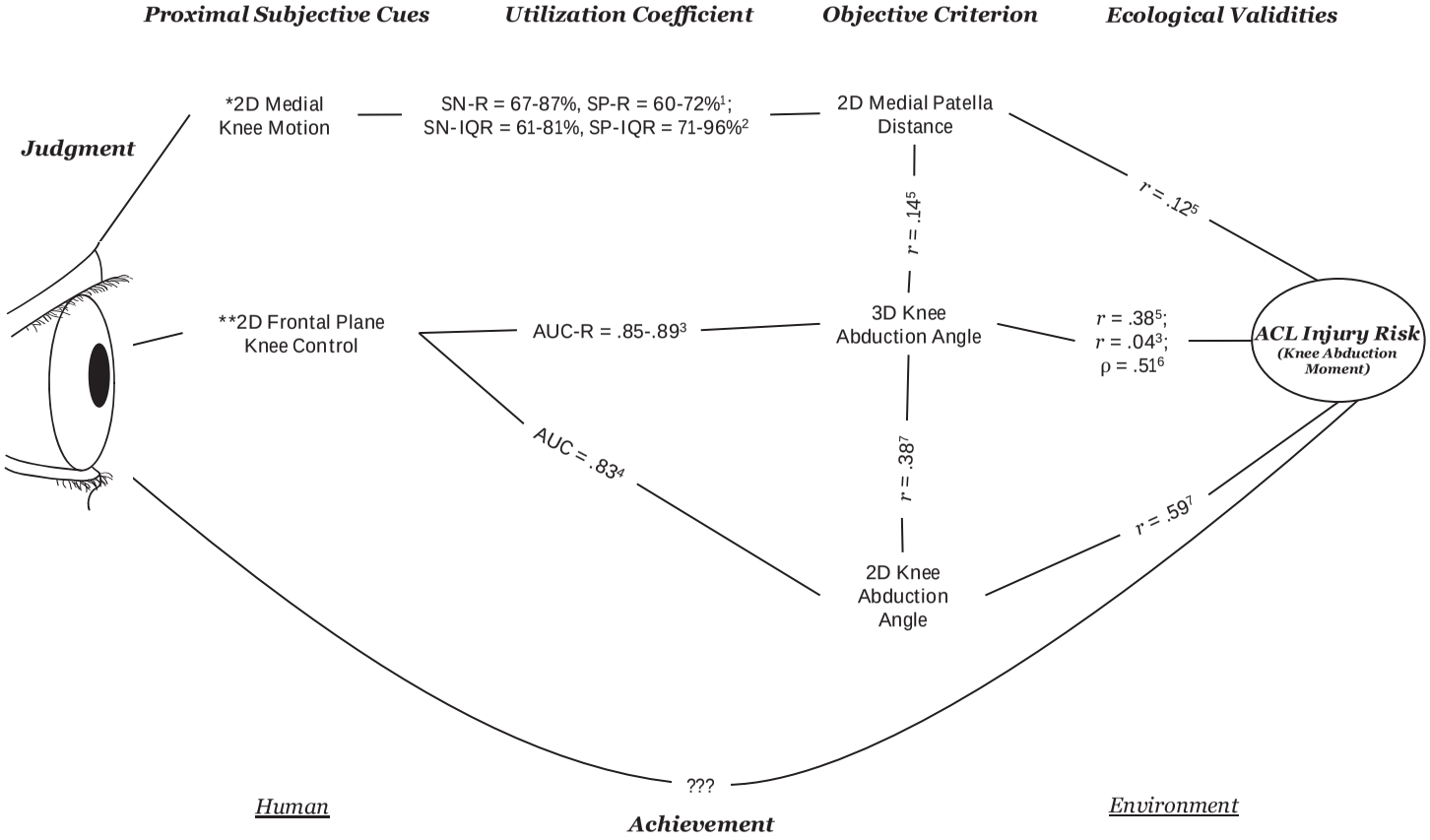

Figure 1. Lens Model for Assessing ACL Injury Risk

Note. SN-R = Sensitivity Range; SP-R = Specificity Range; SN-IQR = Sensitivity Inter-quartile Range; SP-IQR = Specificity Inter-quartile Range; AUC-R = Area Under ROC curve Range; AUC = Area Under ROC; *(Patella medial to the toe; 2pt Scale); **(Knee Valgus Motion/Stability; 3pt Scale); [1]Ekegren et al., 2009 (n = 40 measurements; N = 3 Judges); [2]Whatman, Hume & Hing, 2013 (n = 23 measurements; n = 66 Judges); [3]Nilstad et al., 2014 (n = 60 measurements; N = 3 Judges); [4]Stensrud et al., 2010 (n =186 measurement; n= 1 Judge); [5]Unpublished data (n = 100 measurements); [6]Kristianslund et al., 2013 (n = 120 measurements); [7]Mizner et al., 2012 (n = 36 measurements)

The cue utilizations and ecological validities can provide information regarding the appropriate or optimal cues. Specifically, judging 2D medial knee motion yields high utilization coefficients, but low inter-correlations and ecological validities would suggest that this cue/variable might not be the best estimator of ACL injury risk. Frontal plane knee control (as assessed by Stensrud et al. (2010) and Nilstad et al. (2014)) also yields high utilization coefficients and equivocal ecological validities (as only one study concurrently assessed ecological validity yielding an *r* value of .04 suggesting no relationship with ACL injury risk). Two-dimensional knee abduction angle seems to be a variable with high ecological validity but more studies need to confirm this as only 36 subjects were investigated. By asking participants to rate the ACL injury risk of individuals, achievement can be directly assessed. Also, cognitive tasks analysis methods (verbal reports/eye-tracking) can provide information about the proximal cues used. Thus, there likely exist other cues, which observers will use to determine injury risk status (i.e., landing stiffness, height, weight etc.). This information can provide insight into the development of decision support tools or training systems to enhance performance. Further research is needed to elucidate the effect of the number of observed variables on judgment performance as reliance on fewer cues may provide more robust predictive performance, particularly in the case of co-linear cues (Gigerenzer & Kurz, 2001; Gigerenzer & Todd, 1999). The human visual system has limitations for detecting/recognizing visual stimuli in the environment. Other environmental factors influencing observational movement diagnosis include movement complexity and viewing angle/distance.

Movement speed will influence the perception of various variables/critical features during movement analysis. Observers were more accurate when rating static

knee angle (2-9 degree error) during jumping compared to real time dynamic assessment (20-30 degree error) (Knudson & Morrison, 2000). Additionally, step length estimation error during walking increased at faster walking speeds (Stuberg, Straw, & Devine, 1990). Judgment accuracy was higher for assessing frontal plane knee motion during a slow small knee bend task (Sensitivity = 88%; Specificity = 85%) compared to a fast drop vertical jump (Sensitivity = 70%; Specificity = 79%)(Whatman et al., 2013b). In general, the faster the movement the less accurate one will be in identifying critical features. In addition to movement speed, viewing angle or vantage point influences judgment accuracy.

If the majority of the movement occurs in one plane of motion (i.e., bicep curl), the optimal viewing location for assessment is at a right angle to this plane (i.e., side view). However, out of plane viewing (i.e., perpendicular to the plane of motion) may be more suitable for error detection since relative motion from this vantage point should be minimal. Therefore, movements (i.e., errors) occurring in the perpendicular plane (from which the motion is occurring) may be easier to detect. As the viewing angle deviates from this optimal location, judgment error increases (Plessner & Schallies, 2005). If the movement occurs in many planes of motion the optimal viewing angle remains equivocal and likely leads to judgment errors. Substantial errors (≈1-28 degrees) and inconsistencies between observers were found between visual ratings of joint angles during a multi-planer running/cutting motion (Krosshaug et al., 2007a). Training only resulted in small changes to these visual ratings. In conclusion, movement and critical feature (cue) type must be considered when using observational methods to diagnose movement. Less abstract variables in addition to slow and uni-planer movements seem to provide an environment that fosters more accurate movement diagnosis.

The description of the environment is necessary, as it will dictate judgment performance. Kahneman and Klein (2009) stated: "If an environment provides valid cues and good feedback, skill and expert intuition will eventually develop in individuals of sufficient talent." However, degree of judgment performance can vary with identical environmental structure. Performance differences under similar environments, often studied using the expert-performance approach, describe the perceptual-cognitive characteristics of the observer and are also important when determining expertise mechanisms in judgment and decision-making.

Perceptual-Cognitive Factors

In complex, dynamic environments, such as observational movement diagnosis, cognitive processing must occur to acquire and integrate environmental cues in order to determine the correct diagnosis/decision. Cognitive process theories of diagnosis or decision-making under uncertainty aim to explain how and why individuals make decisions. Theories or models of decision-making include Long-Term Working Memory Theory (Ericsson & Kintsch, 1995), Encapsulation Theory (Boshuizen & Schmidt, 1992), Holistic Model of Image Perception (Kundel, Nodine, Conant, & Weinstein, 2007), Information Reduction Hypothesis (Haider & Frensch, 1999), Recognition Primed Decision Making (Klein, 1997), Adaptive Control of Thought-Rational (Anderson et al., 2004), Cognitive Load Theory (Sweller, 1994), Cognitive Flexibility Theory (Spiro, Feltovich, Jacobson, & Coulson, 1991), Parallel Constraint Satisfaction Models (Herbig & Glöckner, 2009; Simon, Krawczyk, & Holyoak, 2004; Simon, Snow, & Read, 2004) and Cognitive Niches (Gigerenzer & Todd, 1999; Marewski & Schooler, 2011). These theories examine the cues/information utilized, and how this information is integrated in order to make an appropriate decision based on the goals of the task. Thus, the aims of the

following sections are to address theoretical mechanisms that may allow some decision makers to perform better than others given the same environmental constraints. First, the relationship between domain-general perceptual-cognitive factors and skilled performance will be summarized. Second, the relationship between domain-specific perceptual-cognitive processes during decision-making and skilled performance will be summarized.

Domain-General Visual-Spatial Ability

The ability to recognize configurations (i.e., movement patterns) would hypothetically be important for observational movement diagnosis. General tests of spatial abilities have been used to quantify the ability to imagine and mentally transform spatial information or recognize spatial configurations (Uttal et al., 2013). A clear and agreed upon definition of spatial ability has yet to be determined, but recent work by Newcombe and Shipley (2012) have described a typology of spatial skills, which includes a 2 x 2 classification table including intrinsic vs. extrinsic and dynamic vs. static (see Table 1, p.355 of Uttal et al. (2013) for description, measurement methods and relation to defined categories). In short, intrinsic and static describes perceiving objects among distracting background information and could be measured by Embedded Figures tasks, flexibility of closure, and mazes. Intrinsic and dynamic describes visualizing objects into more complex configurations or mentally transforming objects and could be measured by Mental Rotation Test, Purdue Spatial Visualization Test, and the like. Extrinsic and static describes understanding abstract spatial principles and could be measured by Water-Level, Rod and Frame Test, and the like. Finally, extrinsic and dynamic describes visualizing an environment in its entirety from a different position and could be measured by Piaget's Three Mountains Task and Guilford-Zimmerman spatial

orientation. Spatial abilities have been shown to predict achievement in Science, Technology, Engineering, and Mathematics fields even when controlling for math and verbal skills (Wai, Lubinski, & Benbow, 2009). Spatial ability (all types) can be improved with training (Hedge's g = 0.47) and transfers to other tasks (Hedge's g = 0.47) as indicated by a recent meta-analysis (Uttal et al., 2013). Intrinsic-static/dynamic spatial skills have also been postulated to be useful for analysis of movement (Knudson & Morrison, 2000; Morrison & Frederick, 1998).

The relationship between observational movement diagnosis skill and general spatial ability has been investigated. Specifically, movement analysis skill (assessed by the Movement Analysis Test) and two measures of spatial ability (Mental Rotation Test and Group Embedded Figures test) was assessed in 36 undergraduate physical education students (Morrison & Frederick, 1998). Good mental rotation ability was hypothesized to be beneficial for rotating the observed movement to gain relevant information from different vantage points. Disembedding ability was hypothesized to help make decisions about the total movement by examining only parts (i.e., knee or hip movement). The Movement Analysis Test required the respondents to view videos of children performing movements and to indicate on an answer sheet which components were not performed in an adequate fashion (the criterion for the correct responses were based on subjective assessment by domain "experts"). Test-retest reliability was found to be .72. Additionally, the effect of a training intervention was investigated. Training significantly improved movement analysis skill (Cohen's d = 1.07). A linear regression was performed and found that scores on the initial Movement Analysis Test (r = .42) and Group Embedded Figures Test (r = .33) were significant predictors of post-test Movement Analysis Test scores (r = .54), while scores on the Mental Rotation Test were not a significant predictor (r = .05). Interestingly, intrinsic and dynamic spatial ability was not

predictive of movement analysis skill whereas intrinsic and static spatial ability was. The error detection in the Movement Analysis Test may not have required a "skill" for mental rotating but rather simply detecting stimuli among distracters or relations among objects, which is characterized by intrinsic and static (i.e., Group Embedded Figures Test) or extrinsic and dynamic spatial abilities, respectively.

A similar study directly assessed the relationship between the Mental Rotation Test, Group Embedded Figures Test, and visual rating of countermovement jump depth in 43 undergraduate students (Knudson & Morrison, 2000). Students rated the amount of knee flexion during the descent phase of the countermovement jump of 12-videotaped individuals (repeated five times) on a visual analog scale. The mean error in visual rating of the angles was 21.8 degrees with 95% confidence intervals between 14.4 and 29.3 degrees. Video based knee angle assessment ability was not related to Mental Rotation Test (r = -.04) or Group Embedded Figures Test (r = -.13).

In line with the previous study by Morrison and Frederick (1998), the Mental Rotation Test was not related to observational movement analysis performance. However, it was surprising not to see a stronger relationship between the Group Embedded Figures Test and assessment of knee angle. One explanation could be the differences in task demands between the Movement Analysis Test, and assessing knee angle. The former not only requires knowledge of the cues/errors to be detected but also that these cues can be perceived (cue utilization). Assessing knee angle, compared to detecting movement errors, would not require the same cognitive demands because the cue is known and thus seems to be a pure perceptual task. If using the classification system of Newcombe and Shipley (2012), assessing knee angle would require extrinsic and dynamic spatial abilities (which were not assessed). Assessing knee angle appeared to be perceptually difficult according to the high errors (mean of 21.8 degrees). If

assessing knee angle were a predominately visual spatial task it would be expected to have a better relationship between these spatial abilities tests. Interestingly, the Mental Rotation Test and Group Embedded Figures Test displayed a low correlation coefficient of r = .18. In a similar demographic sample (135 female, 83 male, college aged: 22 ± 7.1 yrs) Hegarty, Montello, Richardson, Ishikawa, and Lovelace (2006) found a moderate to large correlation (r = .31) between the Group Embedded Figures Test and the Mental Rotation Test.

This cumulative evidence, albeit with small sample sizes, provides initial evidence suggesting there tends to be small to trivial relations between some measures of spatial ability and observational movement analysis. Future research should investigate the conditions under which these and other measures of spatial abilities (i.e., extrinsic and dynamic/static) are related to movement analysis especially in regards to ACL injury risk estimation.

Domain-General Cognitive Ability

Recent advances in statistical analysis techniques (i.e., Meta- and factor analyses), and studies of expertise mechanisms have lead to a better understanding of the relationship between domain-general cognitive abilities and performance. General cognitive ability is a "construct" with the definition dependent on an adopted theory. According to Schmidt (2011), domain-general cognitive ability is defined and measured as: "the underlying general capacity that causes performance on all mental tasks to be positively intercorrelated, because it is a (partial) cause of every aptitude," and put simply "is the ability to learn."

Domain-general cognitive ability is, however, a theoretical construct that is measured indirectly by various specific domain-general cognitive abilities including verbal, spatial, quantitative, or technical skills. Wechsler's Adult Intelligence Scale,

Wunderlich Intelligence Test, and the Armed Services Vocational Aptitude Battery are often used to assess individual differences in domain-general cognitive abilities. An extensive body of evidence including thousands of participants in a variety of work domains has indicated that domain-general cognitive ability tests predict performance (e.g., supervisory ratings) with validity coefficients ranging from .23 to .58, with strength of validity increasing (but plateauing) with job complexity (Schmidt & Hunter, 2004). The predictive validity coefficients have, however, been shown to decrease over time with an average decrement of -0.45 (corrected for range restriction, reliability and outliers) (Hulin, Henry, & Noon, 1990). Thus, tests of domain-general cognitive abilities likely provide good initial predictors of performance in the absence of domain-specific skill. Additionally, domain-general cognitive abilities have been shown to improve with training (Buschkuehl & Jaeggi, 2010; Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; Jaeggi, Buschkuehl, Jonides, & Shah, 2011). However, as domain-specific skill increases the predictive validity continues to decrease, such that in presence of high-levels of expertise, domain-general cognitive abilities lose all predictive power (Doll & Mayr, 1987; Ericsson, 2013).

Contemporary views on the nature of the relations between domain-general cognitive abilities and expertise present a much more complicated picture. First, general mental ability is a statistical (or psychometric) construct determined from factor analysis of various intelligence tests complicating interpretation. General intelligence is likely not a single underlying cognitive process or capacity, but is a product of mutually reinforcing abilities (Nisbett et al., 2012; van der Maas & Wagenmakers, 2005). Additionally, in a review on giftedness and expert performance, Ericsson, Roring, and Nandagopal (2007) stated: "… we have found no studies that have demonstrated that IQ is predictive of

achievement in domains where reliable, superior performance has been collected meeting our earlier criteria."

Research further shows that rather than primarily reflecting deep-seated differences in cognitive capacities, individual differences in domain-general cognitive abilities are often mediated by differences in simple task strategies and metacognitive dynamics (e.g., thinking about thinking). For example, individuals who score higher on domain-general cognitive fluid intelligence and working memory tests often spend more time preparing for tasks (e.g., reading the instructions), more elaborately encode information, and deliberatively build richer cognitive representations in long-term memory that provide better monitoring and control during subsequent task performance (Baron, 1978, 2005; Cokely & Kelley, 2009; Cokely, Kelley, & Gilchrist, 2006; Ericsson & Kintsch, 1995; Ghazal, Cokely, & Garcia-Retamero, 2014; Hertzog & Robinson, 2005; Mitchum & Kelley, 2010; Sternberg, 1977; Vigneau, Caissie, & Bors, 2006; Ward, Ericsson, & Williams, 2012). Some evidence also indicates that strategy differences in task performance related to domain-general cognitive abilities can be completely eliminated by simple training interventions and modifications of problem representations (Cokely et al., 2006; Garcia-Retamero & Cokely, 2013a, 2013b; Garcia-Retamero & Cokely, 2013; Hertzog & Robinson, 2005; McNamara & Scott, 2001; Nandagopal, Roring, Ericsson, & Taylor, 2010; Stanovich, 2012). Evidence also indicates that individual differences in domain-general cognitive abilities are influenced by differences in motivation and persistence, which are strongly related to differences in overall achievement (Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011; Duckworth & Seligman, 2005).

In summary, domain-general cognitive abilities may predict performance in lower ability individuals or novel tasks settings for a variety of reasons. Nevertheless,

large domain-specific differences in verifiable expertise always reflect differences in acquired cognitive representations (e.g., complex changes in neurology and memory), which are mediated by deliberate practice activities. Therefore, although a comprehensive evaluation of any new test should include comparative testing with some robust domain-general cognitive ability tests, it is unlikely that domain-general cognitive abilities will be related to the anticipated domain-specific differences in expert observational movement diagnosis.

Domain-Specific Cue Acquisition/Integration

During observational movement analysis, where information is abundant, the judge must find information that is most predictive of the actual state. In Brunswikian terms, accurate judgment requires the appropriate utilization of perceptual cues (i.e., high utilization coefficient) with high ecological validities. Time and cognitive characteristics constrain the judge, requiring decision-making based on limited information. Given similar environmental structure and task goals, one open question is: What are the individual differences in type and number of cues searched for and utilized between levels of expertise? What are the individual differences between the interpretation/integration of the acquired cues at different levels of skill?

Verbal protocol analysis during observational movement analysis can provide information regarding cue usage and performance. A process model of motor skill (movement) diagnosis was developed and based on verbal protocol analysis of expert and novice shot put coaches, examining their errors during the shot put movement (Pinheiro & Simon, 1992). Experts detected and diagnosed 40% of the present errors during the first viewing and 44% on the successive viewing, whereas, novices detected 13% of the errors on the first viewing and 7% on the successive viewing. The authors analyzed the verbal report data based on the information-processing framework that

included three stages: cue acquisition, cue interpretation and diagnostic decision-making. Though not discrete stages, this model provided a framework for investigating differences in cognitive processes in skill level of observational movement diagnosis. With regards to cue acquisition, experts mentioned an average of 12 cues for errors in performance, while novices mentioned an average of seven. Experts also made an average of six interpretations/diagnostic decisions, whereas novices made an average of two. When analyzing the motion of a shot putter (to detect errors in movement technique), the superior performance of experts' is mediated by the acquisition of more cues and the use of a greater number of interpretations. Expertise differences, during observational movement analysis have been assessed during other movements such as swimming.

Expertise differences in observational analysis of the freestyle stroke during swimming have also been assessed using verbal report techniques (Leas & Chi, 1993). Two "experts" (recognized by the U.S. Swim Association and had 12 years of coaching experience) and two novices (two years of coaching experience) viewed underwater videotapes of four swimmers. The coaches were asked to view and diagnose/rate the swimmers technique on a scale ranging from 1 (bad) to 10 (good) while "thinking aloud." Experts ratings were more accurate compared to novices when using swimming time as the criterion for comparison (experts: r = .96; novices r = .73); however, these correlations were only based on four data points, thus the results should be interpreted with caution. Large variation existed in the verbalized features during the diagnosis task. Specifically, experts' diagnosis commonly identified "process" type features such as "wide pull" or "stroke unbalanced," whereas, novices' commonly identified specific body parts in a "static" context such as "elbow bent extension." Experts also verbalized a greater number of cause and effect, and prescription type statements. In fact novices' did

not provide any cause/effect or prescriptive statements. In summary, experts identified features which were more "second-ordered" and "dynamic" which was similar to the seminal work of experts solving physics problems (Chi, Feltovich, & Glaser, 1981). If the task is, however, to diagnosis/classify the individual into a dichotomous (good or bad) or a relatively few number of categories, expertise differences in cognitive processes may not reflect these aforementioned differences. For example, experts may require fewer cues to make superior diagnostic decisions.

Shanteau (1992) summarized the relationship between the amount of information used and expertise in a variety of judgment and decision-making tasks. The results of five studies comparing expert and novice information use in auditing, medical diagnosis, and livestock judging consistently revealed that expertise differences were not based on the number of cues but rather the type of cue used. In a majority of these studies, experts tended to use fewer cues than novices (may not have reached statistical significance due to small sample sizes; see Moxley, Ericsson, Charness, and Krampe (2012)). Even in the normative sense, models including fewer cues may perform better when predicting new data than models including a larger number of cues. Specifically, an inverse U-shaped relationship often exists between model complexity (number of parameters/information) and predictive power (Pitt, Myung, & Zhang, 2002). These results suggest that when investigating individual differences in skill and cue usage, studies need to be designed to allow for the assessment of both type (quality) and amount (quantity) of information used by skilled judges.

Eye-tracking can also provide information regarding expertise differences in cue usage. A meta-analysis was conducted to investigate common expertise differences (819 experts, 187 intermediates, and 893 novices) in eye-tracking metrics during the comprehension of visualizations across various domains (sport: $N = 704$; medicine: $N =$

101; transportation: N = 260; Other: N = 110) (Gegenfurtner, Lehtinen, & Säljö, 2011). Overall, experts were more accurate (r = .45) and responded quicker (r = -.38) than novices. Experts compared to non-experts had a similar number of overall fixations (r = -.04), more fixations on task-relevant areas (r = .53) of longer duration (r = .27), less fixations on task-redundant areas (r = -.31) of shorter duration (r = -.43), shorter times to first fixate on task-relevant areas (r = -.31), and longer saccade amplitudes (r = .30). In summary experts displayed efficient allocation of attention to critical/diagnostic information that was related to the level of expertise (for a similar review in the sporting context see Mann, Williams, Ward, and Janelle (2007)). This comprehensive synthesis of data corroborates evidence that supports the importance of information type over amount when assessing expertise differences in decision-making. These eye-tracking measurements, however, do not describe the causal mechanisms or why these expertise differences are present. A closer look at content and organizational differences in knowledge may help elucidate this issue.

Domain-Specific Knowledge

Through the meta-analytic work of Schmidt and Hunter, job knowledge tests have been shown to predict job performance with substantially large correlations (r values ranging from .48 to .61) (Schmidt & Hunter, 2004; Schmidt & Hunter, 1998). Additionally, their path analysis work has shown that job knowledge is largely the reason general mental ability tests predict job performance so well (Schmidt, Hunter, & Outerbridge, 1986). Job type is, however, a moderator for this relationship. This work was based on large amounts of data from various jobs but highlights the importance of job or domain-specific knowledge for performance.

Domain-specific knowledge should be related to observational movement diagnosis skill. Ste-Marie (1999) investigated the ability of gymnastic judges to anticipate

upcoming gymnastic elements using videotaped performances. Additionally, knowledge base was assessed by asking the judges to generate a list of potential gymnastic elements following the stoppage as well as the potential errors using the International Gymnastic Federation Code of Points (IGFCP) book as the standard for comparison. Twelve "experts" (17 years of experience and greater than a Level 5 certification) and twelve novices (two years of experience and less than Level 2 certification) participated in the study. Experts were more accurate at anticipating the gymnastic elements (39.5 vs. 28.3% correct). The correct anticipation was related to better scoring performance of the gymnastic element. Thus, the author's classification of expertise level seems to be justified according to these performance outcomes. Experts were more accurate concerning IGFCP information (error identification, symbol and level of difficulty) than were novices (84 vs. 52% correct). Experts were also able to generate a greater number of alternatives (gymnastic elements and errors therein). These findings suggest domain-specific knowledge (depth and breadth) contribute to expertise differences in judgments.

Using observational movement analysis of the freestyle swimming stroke, Leas and Chi (1993) also examined the expertise differences in domain-specific knowledge and structure. In addition to the mentioned diagnosis tasks (see paragraph three in cue acquisition/interpretation section), the coaches were asked to describe the "ideal" or "prototype" freestyle stroke technique, which was then analyzed based on content and connectedness. The analysis of the verbalizations of the ideal stroke technique revealed that experts identified the four main stroke components, which were described in swimming technical documents, whereas novices only identified two. Additionally, experts verbalized an average of 29 components, which composed 70% of the known feature categories, while novices verbalized 11, which composed 34%. Connectedness was assessed by analyzing the prototype data in regards to the number and length of their

reasoning chains (causal utterances). Experts' chain number and length were significantly larger than novices'. Despite the low sample size in participants and stimuli, experts' knowledge about the critical features for technique analysis seemed to have greater breadth and connectedness compared to the novices.

Studies investigating the mechanisms of the expertise difference in observational movement analysis highlight the importance of possessing adequate domain-specific knowledge but this knowledge is also structured or organized in a way which enhances decision-making accuracy during complex, time constrained tasks. Elaborate memory structures likely allow an expert movement analyst to create an adaptable prototype model of expected performance that facilitates accurate detection, interpretation, and diagnosis of movement abnormalities (Pinheiro & Simon, 1992). The classic study by Chase and Simon (1973) revealed that perceptual expertise in chess and many other domains (Ericsson & Lehmann, 1996) is mediated by superior encoding of representative structured information and not larger short-term memory capacity. These perceived patterns or "chunks" are similar in number to that of novices but are larger and more detailed/complex (Chase & Simon, 1973). This difference in the structure and complexity of memory representations allows experts to attend to the task relevant information while ignoring task irrelevant cues/areas, building and evaluating more diagnostic hypotheses, and better anticipating actions/events (Balslev, 2011; Ericsson & Lehmann, 1996; Gegenfurtner et al., 2011; Mann et al., 2007). These characteristics of expert performance have been shown to be developed through deliberate practice (see Ericsson, Krampe, and Tesch-Romer (1993) for in depth discussion). One must also be aware of potential biases, as previous experience may negatively influence subsequent judgments even in highly skilled individuals.

Biases

Two candidate biases that may affect movement judgment/diagnosis are sequential and/or prior processing effects. In sequential effects, a preceding judgment influences the actual judgment in a systematic way, and have been observed among experienced gymnastics judges (Damisch, Mussweiler, & Plessner, 2006). Specifically, a correlation of r = .30 was found between the target athletes' judgments and the judgments of the previous athletes' performance using data from the 2004 Olympic Games (N = 1,307). Judgments of the second (r = .26) and third (r = .18) judged athletes were also correlated to the target judgment. These results suggest that the magnitude of the sequential effect is lessened as the number of judgments between target assessment increases. Prior processing effects or previous knowledge of the performance have also been documented in gymnastic judging. These effects can be beneficial or detrimental depending on the perceived performance. For example, if the judges saw the same move, they scored the move more accurately (Mean accuracy = 76.2%) than when scoring a new move (Mean accuracy = 72.2%) or if a movement was different to the target performance (Mean accuracy = 68.4%) (Ste-Marie & Lee, 1991). Expertise (as assessed by accuracy of form error detection) did not appear to reduce this bias. These somewhat acute perceptual-cognitive biases may have consequences when diagnosing injury risk status (via observational movement diagnosis), and therefore should be considered.

Problem Statements

This review has highlighted the characteristics of and assessment methods for skilled or expert performance and potential cognitive mechanisms that may give rise to expertise in the context of observational movement analysis. There is, however, a paucity of research directly investigating observational assessment of ACL injury risk estimation ability. Previous research has demonstrated that physiotherapists possess the ability to

observationally assess specific variables purported to be associated with ACL injury risk. However, these variables have not been supported by evidence from longitudinal prospective studies of ACL injury risk. Additionally, the skill of observational injury risk estimation has not been assessed in other populations that may benefit from or commonly use this skill (i.e., coaches, athletes, parents, medical doctors). Similar to other "skills", theoretically, one's ACL injury risk estimation skill is likely a function of adequate domain-specific knowledge (i.e., knowing the function of the ACL and risk factors for injury), and deliberate practice assessing injury risk. Sports medicine practitioners (physical therapists, physiotherapists, athletic trainers, orthopedic doctors) likely possess the requisite domain-specific knowledge for ACL injury risk but may not obtain accurate feedback regarding their current level of skill. To date, ACL injury risk estimation feedback is unlikely, as one would have to see athletes drop jump performance prior to injury, follow-up on the injured athletes, and differentiate movement patterns between the injured and non-injured individuals (while correctly updating their "mental" representation of at risk individuals). Additionally, feedback can also be specified through performance assessment but is currently not formally available. It seems likely that practitioners may often screen for ACL injury risk. Yet, despite their wealth of knowledge about ACL injury risk factors, professionals may not be able to observe differences in risk level.

CHAPTER 2: STUDY 1- DEVELOPMENT OF THE ANTERIOR CRUCIATE LIGAMENT INJURY-RISK-ESTIMATION QUIZ (ACL-IQ)

Introduction

Study 1 was designed to begin to address two questions. First, can some individuals more accurately estimate ACL injury risk based on observations of athletes' drop vertical jump performance? Second, given evidence of skill-based differences in observational diagnosis performance, can we improve assessment of skill differences using modern psychometric methods? For ease of explication, the introduction and analysis of Study 1 will be divided into two parts. Study 1A focuses on testing hypotheses about individual differences in risk estimation skill. Study 1B focuses on refining psychometric assessment of risk estimation skill.

Study 1A: Individual Differences in Risk Estimation Skill

Hypotheses

1.) People working in field of exercise science will have greater understanding of the ACL (i.e., location, function, and risk factors) compared to less experienced individuals.

2.) People working in field of exercise science will make more accurate and consistent ACL injury risk estimates compared to less experienced individuals.

Methods/Procedures

Risk Estimation

Participants viewed brief videotaped clips of athletes performing a drop vertical jump and were asked to estimate the risk for future ACL injury on a 10-point scale (see Figure 2). Actual ACL injury risk was calculated using biomechanical analysis of peak

knee abduction moment and peak knee abduction angle (Myer, Ford, Khoury, & Hewett, 2011). The greater of the two values for both legs was used as the criterion.[4]

	Very Low									Very High
Risk for ACL Injury	O 1	O 2	O 3	O 4	O 5	O 6	O 7	O 8	O 9	O 10

Figure 2. Example Decision Task (Snapshots of Video Sequence)
Note. The figure was created by Erich J. Petushek

Stimuli

Young females participating in landing and cutting sports are at the greatest risk for ACL injury (Hewett, Zazulak, et al., 2012). Accordingly, stimuli consisted of a sample of 20 video clips of female athletes performing a drop vertical jump. The athletes featured in the videos participated in landing and cutting sports and served as the participants for the development and validation of the clinical ACL nomogram (Myer, Ford, & Hewett, 2011) (*M* ± *SD;* age: 15.9 ± 1.3 years; height: 163.6 ± 9.9 cm; body mass: 57 ± 12.1 kg).[5] The athletes featured in the video stimuli were also demographically similar to individuals investigated in the initial prospective injury risk factor study (*M* ± *SD;* age: 16.0 ± 1.35 years; height: 165.9 ± 6.4; body mass: 60.3 ± 8.2 kg) (Hewett et al.,

[4] In the current study, analysis of test items shows that peak knee abduction moment and angle have a marginal, moderate correlation of r (18) = .35, p = .13.
[5] All individuals in the video clips signed a photo release form indicating use of their photo/video in mass media publications, internet, television or movie presentations. Despite the photo release, faces were pixilated to maintain anonymity.

2005). The 20 candidate video stimuli items also depict athletes who have a wide and representative range of injury risk values (from very low to very high).

Participants

A convenience sample of 213 individuals completed the study. The sample included a group of 40-exercise science professionals (e.g., physical/physiotherapists, exercise science students, sports medicine researchers, and orthopedic doctors) recruited from the Norwegian School of Sport Sciences or nearby Physiotherapy clinics (i.e., the Exercise Science group). Two hundred additional participants from Amazon's Mechanical Turk were also recruited as Mechanical Turk participants tend to be demographically diverse and closer to the demographics of the general U.S. population (i.e., the "General Population" group; see Paolacci, Chandler, and Ipeirotis (2010)). Twenty-seven Mechanical Turk participants did not complete the test or were missing a substantial amount of data. One individual from the Exercise Science group was an anesthesiologist and thus was moved into the General Population group. Three Mechanical Turk participants reported working in physical therapy or exercise science and thus were moved to the Exercise Science group.

Study Procedures

The study was fully computerized and hosted online. The testing of the convenience sample of potential exercise science professionals took place in a quiet room using a 61cm wide screen monitor with a resolution of 1920 x 1200 (Model: U2412M, Dell Computer Corporation, USA). The specific study setting for the Mechanical Turk participants cannot be discerned as the study material was distributed online and could be taken on any computer with Internet access. The study was not compatible with mobile devices thus limited to laptop or desktop computers. Once recruited, all

participants completed the entire study on the computer with no initial verbal instructions. Prior to the decision tasks participants were provided with brief written instructions. The instructions stated:

> Video clips will be presented following a 3 second countdown. You will only be able to view the clips once. After viewing the video clips you will be asked to answer 2 questions: First you will rate the athlete's degree of risk for a future anterior cruciate ligament (ACL) injury. You will then be asked to rate your confidence in your chosen risk rating.

Following two practice trials, 25 decision tasks (five of which were repeated items) were presented in randomized order. Following the decision tasks, domain-specific knowledge was measured using three questions related to ACL location, function, and injury risk factors. Finally, demographic information including age and profession were recorded.

Results

Group Demographics

The Exercise Science group consisted of 17-college exercise science students, 13-physio/physical-therapists, 4-sport medicine Ph.D.'s, 3-orthopeadic medical doctors, 3-strength & conditioning coaches, and 2-administrators. The General Population group consisted of individuals with diverse occupations. The Exercise Science group was significantly younger compared to the General Population group (see Table 1).

Domain-Specific Knowledge

Domain-specific knowledge was assessed with three questions (i.e., Where is the ACL located? What is the function of the ACL? and What are the risk factors for ACL injury?). The location of the ACL was coded as one for correct and zero for incorrect. The ACL has two primary functions thus one point was awarded for each correct function.

Various risk factors exist for ACL injury that can be placed into four general categories including biomechanical, anatomic, intrinsic (hormone, genetics, cognitive function and previous injury), and extrinsic (Smith et al., 2012a, 2012b), thus one point was awarded for risk factors reported from any of these four categories. Overall, the Exercise Science group displayed greater domain-specific knowledge compared to the General Population group (see Table 1).

Table 1. Age and Domain-Specific Knowledge of the Exercise Science and General Population Groups

	Exercise Science			General Population				
Variable	*n*	*M*	*SD*	*n*	*M*	*SD*	*p*	Cohen's *d*
Age	42	30.29	7.41	160	35.04	11.23	0.01	0.45
ACL Knowledge	42	3.69	1.35	171	1.49	0.84	< .001	2.31

Note. Maximum ACL knowledge points = 7.

Relative Judgment Accuracy

Relative accuracy metrics were used to assess agreement between estimated and actual ACL injury risk judgments. Analyses of relative accuracy provide information about the resolution of one's judgment, emphasizing participants' ability to discern the relative difference between higher and lower levels of risk, regardless of the absolute risk level (e.g., can participants correctly rank which test items depict higher versus lower levels of risk). Because the judgment data were ordinal, a Spearman Rho rank correlation was calculated between risk estimates on the 10-point scale and biomechanical knee abduction moment criteria (RelAccMom) and angle criteria (RelAccAng) for the 20 video clips. This analysis assessed participants' relative accuracy and describes the overall relationship between known ACL injury risk and subjective assessment, without providing a cut-off value or categorizing the video clips.

The Exercise Science group displayed greater relative accuracy compared to General Population when either knee abduction moment (RelAccMom) or angle

(RelAccAng) was used as criterion (Note: the greater negative value indicates higher relative accuracy) (see Table 2 for descriptive statistics and effect sizes). In aggregate, the General Population relative accuracy measures (using abduction moment and angle) were near 0, indicating a widespread inability to accurately estimate relative changes in injury risk status. On an individual level, the largest rank correlation between a single individual's observational rating and knee abduction moment (RelAccMom) was ρ (18) = -.58, a strong relationship. Four-percent in the General Population group and 7% in the Exercise Science group displayed a statistically or marginally significant RelAccMom (ρ (18) < -.38, $p < .10$). On an individual level, the largest rank correlation between a single individual's observational rating and knee abduction angle was ρ (18) = -.69, a strong relationship. Twelve-percent in the General Population group and 57% in the Exercise Science group displayed statistically or marginally significant RelAccAng performance (ρ (18) < -.38, $p < .10$). The magnitude of relative accuracy was, however, larger for RelAccAng (Mean ρ (18) = -.41) compared to RelAccMom (Mean ρ (18) = -.18), for the Exercise Science group ($t(42) = 13.25$, $p < .001$, $d = 0.77$) but not different for the General Population group ($t(171) = -1.14$, $p = .26$, $d = -0.14$). Exercise Science individuals who displayed greater RelAccMom also displayed greater RelAccAng (r (40) = .66), providing convergent evidence that professionals working or studying in exercise science have some superior ability to visually estimate ACL injury risk.

Table 2. Performance Comparison Between Exercise Science and General Population Groups

	Exercise Science			General Population				
Variable	*n*	*M*	*SD*	*n*	*M*	*SD*	*p*	Cohen's *d*
RelAccMom	42	-0.18	0.13	171	-0.10	0.19	0.01	0.47
AbAccMom	42	2.42	0.35	171	2.40	0.51	0.79	0.05
RelAccAng	42	-0.41	0.15	171	-0.08	0.25	< .001	1.46
AbAccAng	42	2.35	0.40	171	2.50	0.57	0.10	0.28
AbCon	42	0.83	0.45	171	1.15	0.78	0.01	0.45

Note. RelAccMom = relative accuracy using knee abduction moment as criterion (ρ); AbAccMom = mean absolute error using knee abduction moment as criterion; RelAccAng = relative accuracy using knee abduction angle as criterion (ρ); AbAccAng = mean absolute error using knee abduction angle as criterion; AbCon = absolute consistency error

Absolute Judgment Accuracy

Absolute accuracy metrics (i.e., judgment calibration) were also used to assess agreement between estimated and actual ACL injury risk. Individuals could in theory have high associations with the criterion (resolution) without having high-levels of judgment calibration (e.g., participants could overestimate risk across all trials). A measure of absolute agreement would potentially eliminate this limitation and may best serve as an efficient and understandable scoring metric for testing purposes (e.g., those who are perfectly calibrated also have perfect resolution). The video clips were categorized based on the 1-10 scale (according to the biomechanical measures) in order to quantify the absolute accuracy for individual judges. The range of knee abduction moment and angle values were used to linearly transform the continuous biomechanical values into their respective category on a 1-10 scale. The correlation between the raw and categorized video clips was r (18) = -.99 for both knee abduction moment and angle. Average absolute error was used to calculate the absolute accuracy between subjective ratings and criteria for knee abduction moment (AbAccMom) and angle (AbAccAng) over the 20-items. Average absolute error was defined as the average of the 20 absolute value difference scores (i.e., |subjective rating category – criterion category|). This

measure can be interpreted as the average judgment error (i.e., without direction) compared to the criterion, in the actual scale units (1-10).

The Exercise Science group did not differ in terms of absolute accuracy compared to General Population for either knee abduction moment (AbAccMom) or angle (AbAccAng) (see Table 2 for descriptives and effect sizes). Theoretically, the Exercise Science group should have displayed lower AbAccMom and AbAccAng scores compared to General Population because the overall correlation between relative and absolute accuracy was r (211) = .27 (p < .01) for abduction moment and r (211) = .32 for abduction angle across all individuals (p < .01). The absolute accuracy is mathematically lowest when the subjective judgment category is the same as the criterion category. Post-hoc analyses revealed that the current 20-items with the majority of the items located near the mid-range of the scale (mean category for moment criterion (SD) = 4.75 (2.20); mean category for angle criterion (SD) = 5.20 (2.04)) created an opportunity to bias test performance and skill estimates. An individual could "perform" well by anchoring their judgments near the middle of the scale (i.e., 5), a common behavioral tendency observed in test-taking. If an individual chose the risk rating "five" for all 20 video clips, they would display an AbAccMom of 1.65 and AbAccAng of 1.40, which is substantially lower than the aggregate measures. This score would incorrectly indicate that the "rater" showed substantial judgment calibration when in fact they were guessing. Overall, the General Population group's average rating (with standard deviation) across all clips was 4.09 (0.67) compared to 5.22 (2.02) for the Exercise Science group. The standard deviation indicates the Exercise Science group utilized more of the scale whereas the General Population individuals showed a strong tendency to anchor their judgments tightly around the midpoint category 4, resulting in an AbAccMom of 1.55 and AbAccAng of 1.70 across the 20 clips. This evidence paired with the aforementioned correlation

between absolute and relative accuracy suggests the presence of a scale-use anchoring bias. This test design limitation needs to be corrected before absolute accuracy data can be further interpreted (see Study 1B below).

Consistency

A consistency analysis was also performed comparing the consistency of risk estimates on a subset of five video clips that were repeated. Judgment consistency was assessed using a similar approach to that of absolute accuracy with performance calculated as the average of the five absolute value difference scores (i.e., |subjective rating one category – subjective rating two category|), termed AbCon. This measure can be interpreted as the average judgment error (without direction) on a repeated assessment in the actual scale units (1-10).

Exercise science professionals were more consistent (lower AbCon) than the General Population group (see Table 2 for descriptives and effect sizes). Specifically, for a repeated trial, the average expected deviation on the second rating was 0.83 (on a 10-point rating scale) for the Exercise Science group. An AbCon value at or below 1.00 was deemed sufficient for consistency of expert performers as this value will likely not influence injury risk estimation if fewer categories were used (i.e., from 10 to 3) which may best represent decision or intervention points. Moreover, 76% of the Exercise Science raters displayed AbCon values at or below 1.00, compared to 54% of the General Population raters[6]. Overall, results provide evidence that the majority of individuals showed some consistency in their judgments, although Exercise Science rater's showed significantly higher levels of consistency.

[6] No significant relationship was found between consistency and any of the performance metrics (r (211) < .083). This was likely due to the restriction of range for the consistency values.

Study 1A: Discussion

Overall, results provide evidence that participants who work/study in exercise science domains tend to be more knowledgeable about the ACL and injury risk and also tend to show some evidence of superior, reproducible observational movement diagnosis skill. Specifically, in the observational risk estimating task, the group of exercise science professionals showed more consistent judgments and also more accurately estimated the relative differences in ACL injury risk compared to less knowledgeable members of the general population. An analysis of the top performers indicates that some individuals showed relatively high-levels of risk estimation accuracy (using both biomechanical predictors), with relatively good consistency. Four individuals (two from General Population (≈ 1%) and two from Exercise Science (≈ 5%)) displayed a significant association with both knee abduction moment and angle (ρ (18) < -.38) and had consistency error values less than or equal to 1.00. This is noteworthy considering that knee abduction moment should not be directly viewable (and cannot be calculated) from a single frontal plane viewpoint. Estimates of abduction moment are thought to require multiple cameras, a force platform, and many calculations or else involve the use of the ACL nomogram, which requires two cameras and an isokinetic dynamometer. Unfortunately, results also revealed some limitations of the current materials that must be addressed in order to achieve higher levels of psychometric validity and test performance (e.g., artificial inflation of absolute judgment accuracy scores due to anchoring of judgments in the middle of the scale).

Study 1B: Psychometric Test Development

The goal of Study 1B was to use psychometrically robust analytical methods to develop a high-performing test of observational ACL injury risk estimation ability. This

test is intended to be an efficient research tool used for identifying individual differences in an essential ability that may contribute to future clinical applications (e.g., developing screening and training programs). As noted, there are two dominant approaches to modern psychometric test development, i.e., Classical Testing Theory (CTT) emphasizing overall test performance optimization (i.e., test-level) and Item Response theory (IRT) emphasizing theoretical probability distributions modeled on test-item performance (i.e., item level). It is important to note that standard analytic methods need to be modified to address various constraints present in the current data. For example, in Study 1A, absolute measures of accuracy were confounded by an anchoring strategy widely used by less skilled individuals. Classical test theory (CTT) item parameter calculations would be less meaningful under these conditions because they are influenced by overall test performance. The bi-modal distribution of these accuracy metrics would also make CTT item parameter calculations less informative. Unfortunately, an item response theory (IRT) approach would require a substantially greater number of heterogeneous responses and complex model fitting across items, under assumptions that do not hold (e.g., homogeneity of item discriminability indices). Accordingly, the current analysis employed a hybrid approach, using an IRT inspired, modified CTT item-analysis (for a related approach using decision trees see Cokely, Galesic, Schulz, Ghazal, and Garcia-Retamero (2012)). Specifically, analyses examined individual item performance across three dimensions: difficulty, discriminability and guessing. Two key assumptions for this analysis were (a) exercise science professionals have higher ACL injury risk estimating ability as compared to the general population and (b) peak knee abduction moment has a linear relationship with ACL injury risk.

Item analysis difficulty and discrimination values were calculated for each individual test item (i.e., each video clip). Theoretically, the goal was to select items to

represent a wide range of difficulty with a maximum degree of discriminability, while equally sampling from the full range of the scale to reduce artificial test score inflation resulting from any anchoring effect. Item difficulty can be described as the mean absolute error for each item across all individuals. The higher the absolute error, the more difficult the item. Each item should also discriminate between group members who have distinctive differences in general ACL knowledge (i.e., Exercise Science and General Population). Discriminability was assessed as the overall effect size of the mean absolute error differences between the Exercise Science and General Population groups for each item. Items that display large positive effect sizes (Cohen's *d*) are considered more discriminating (i.e., Exercise Science was more accurate than General Population). In order to identify items that optimize both difficulty and discriminability, the product of these two parameters was calculated.

The guessing parameter is directly related to the item's known location on the scale (i.e., the criterion injury risk category). The likelihood of obtaining a low absolute accuracy error by guessing is greater for items near the middle of the scale (i.e., five). Thus, selecting items toward the ends of the scale should reduce the potential benefit of anchoring on any single rating. Guessing performance was analyzed using Monte Carlo simulations of 10,000 "pseudorandom" test performances. For each item in the test, an integer (1-10) was drawn from a standard normal distribution. Test performance was computed and averaged across the 10,000 iterations to estimate "chance" or guessing performance for each individual item and for overall test performance.

The individual item characteristics are presented in Table 3. Based on consideration of available variables with emphasis on the product of discriminability and difficulty indices, seven candidate items were chosen representing the full range of risk

levels[7]. One test was constructed from these initial seven candidate items. Three 6-item tests were constructed by removing one item (either Clip 2, 7 or 8) from the middle range to further reduce the benefit of anchoring. Two additional 5-item tests were also constructed by removing two items with mid-range risk characteristics (Clip 8, which was a 5-risk, and either Clip 2 or 7, which were both a 6-risk). The test scores were determined by the sum of the deviation or error points (using knee abduction moment as the criterion). The greater the error points the poorer the performance. Zero would represent perfect performance. Mathematically the test score and AbAccMom (for each test length) should have a perfect relationship if all items are answered, as the AbAccMom is the sum of the deviations points divided by the number of items (i.e., the mean)[8]. Moreover, six candidate tests were constructed and evaluated based on their various psychometric properties (see Table 4).

[7] Clip 7 was chosen over Clip 11 and 5 because of its greater difficulty. Higher difficulty was prioritized because this test aims to identify high ability individuals.

[8] The average absolute accuracy (i.e. AbAccMom and AbAccAng) for individuals with missing data was calculated based on the number of answered trials, which may not have equaled the total number of items in that test. Individuals were, however, initially excluded from the analysis if they did not answer at least 17 out of the 20 items. Correlations between AbAccMom and all associated test scores (i.e. deviation or error points) were r (211) > .96.

Table 3. Individual Item Analysis

	Criterion (Raw Nm)	Discriminability Ranking (Cohen's *d*)	Difficulty Ranking (Mean absolute error)	Difficulty & Discriminability Ranking (Product[a])	Guessing Mean Absolute Error	Consistency Error (*SD*)
Clip 3*	***10 (-85.20)***	***1 (1.78)***	***2 (4.53)***	***1 (8.06)***	***4.52***	***1.02 (1.19)***
Clip 20*	***10 (-77.58)***	***3 (0.76)***	***1 (5.65)***	***2 (4.29)***	***4.52***	-
Clip 2*	***6 (-41.33)***	***2 (0.94)***	***5 (2.59)***	***3 (2.43)***	***2.51***	-
Clip 17*	***3 (-17.01)***	***4 (0.71)***	***8 (2.24)***	***4 (1.59)***	***3.11***	-
Clip 11	6 (-47.25)	5 (0.57)	14 (1.94)	5 (1.11)	2.51	-
Clip 5	6 (-41.56)	6 (0.46)	15 (1.88)	6 (0.86)	2.51	-
Clip 8*	***5 (-37.70)***	***8 (0.32)***	***12 (2.03)***	***7 (0.65)***	***2.49***	-
Clip 15*	***2 (-10.89)***	***7 (0.39)***	***20 (1.54)***	***8 (0.60)***	***3.74***	***0.80 (1.02)***
Clip 7*	***6 (-48.48)***	***10 (0.19)***	***3 (2.89)***	***9 (0.55)***	***2.51***	-
Clip 13	4 (-29.78)	9 (0.26)	16 (1.83)	10 (0.48)	2.71	1.15 (1.43)
Clip 14	3 (-21.77)	11 (0.16)	19 (1.55)	11 (0.25)	3.11	-
Clip 19	4 (-23.59)	12 (0.12)	18 (1.64)	12 (0.20)	2.71	-
Clip 4	5 (-41.16)	13 (-0.4)	13 (2.02)	13 (-0.81)	2.49	-
Clip 10	3 (-15.76)	15 (-0.79)	17 (1.73)	14 (-1.37)	3.11	-
Clip 16	3 (-18.91)	14 (-0.77)	11 (2.18)	15 (-1.68)	3.11	-
Clip 18	5 (-34.19)	16 (-0.8)	9 (2.23)	16 (-1.78)	2.49	-
Clip 9	4 (-23.68)	17 (-0.95)	10 (2.19)	17 (-2.08)	2.71	1.28 (1.27)
Clip 6	2 (-13.63)	18 (-1.14)	7 (2.26)	18 (-2.58)	3.74	-
Clip 1	4 (-32.2)	19 (-1.58)	6 (2.48)	19 (-3.92)	2.71	1.17 (1.21)
Clip 12	4 (-25.22)	20 (-1.60)	4 (2.72)	20 (-4.35)	2.71	-

Note. *Selected Items for Candidate Tests; [a]Cohen's *d* x Mean absolute error; Guessing was based on Monte Carlo simulation with 10,000 iterations

Table 4. Psychometric Properties of the Candidate Test Instruments

	20-Item	7-Item	6A-Item	6B-Item	6C-Item	5A-Item	5B-Item
Scale Attributes^							
Score Range	0 - 127	0 - 48	0 - 43	0 - 43	0 - 43	0 - 38	0 - 38
Achieved Range (%)	27 - 78 (39 - 79)	8 - 34 (29 - 83)	6 - 30 (30 - 86)	5 - 29 (33 - 88)	7 - 29 (33 - 84)	4 - 26 (32 - 89)	5 - 26 (32 - 87)
Average score							
Overall *M* (%)	47.39 (63)	21.08 (56)	19.1 (56)	18.29 (57)	18.5 (57)	16.31 (57)	16.52 (57)
Overall *Mdn* (%)	47 (63)	21 (56)	19 (56)	18 (58)	19 (56)	17 (55)	17 (55)
Overall *SD* (%)	9.58 (8)	6.06 (13)	5.59 (13)	5.51 (13)	5.15 (12)	5.11 (13)	4.75 (12)
Anchoring 5 (%)	33 (74)	17 (65)	17 (60)	16 (63)	16 (63)	16 (58)	16 (58)
Guessing *M* (%)	59.81 (53)	23.31 (51)	20.7 (52)	20.72 (52)	20.75 (52)	18.22 (52)	18.19 (52)
Discriminability							
Exercise Science *M* (%)	48.41 (62)	14.1 (71)	12.36 (71)	11.43 (73)	12.55 (71)	9.69 (74)	10.81 (72)
Exercise Science *SD* (%)	7.01 (6)	4.34 (9)	3.96 (9)	4.13 (10)	3.83 (9)	3.80 (10)	3.49 (9)
General Population *M* (%)	47.14 (63)	22.8 (53)	20.75 (52)	19.97 (54)	19.97 (54)	17.93 (53)	17.92 (53)
General Population *SD* (%)	10.11 (8)	5.13 (11)	4.61 (11)	4.39 (10)	4.32 (10)	3.96 (10)	3.89 (10)
Cohen's *d* (SD_w)	0.13 (9.50)	1.75 (4.97)**	1.88 (4.48)**	1.98 (4.34)**	1.76 (4.22)**	2.11 (3.92)**	1.87 (3.81)**

Note. ^All numerical values indicate deviation or error points (i.e., Higher numerical values indicate greater judgment error and lower overall test performance); % = Percent correct (100 - percent maximum error); Guessing was based on Monte Carlo simulation with 10,000 iterations; *p < .05; **p < .01; 7-Item (Clips 2, 3, 7, 8, 15, 17 and 20); 6A-Item (Clips 2, 3, 7, 15, 17 and 20); 6B-Item (Clips 2, 3, 8, 15, 17 and 20); 6C-Item (Clips 3, 7, 8, 15, 17 and 20); 5A-Item (Clips 2, 3, 15, 17 and 20); 5B-Item (Clips 3, 7, 15, 17 and 20)

Selecting the Final ACL-IQ Test

All candidate tests were found to be sufficiently difficult as no individual attained perfect performance (i.e., peak-performance was roughly 85% of maximum). Mean and median scores showed converging central tendencies (see Table 4). All candidate tests successfully discriminated between Exercise Science and General Population groups, dramatically improving psychometric discriminability compared to the 20-item test (i.e., 1,623% improvement). All candidate tests displayed significant correlations with relative measures of accuracy using knee abduction moment and angle. All tests showed robust correlations with the ACL knowledge test—an additional index of convergent validity (see Table 5). However, the 7 and 6-item tests continued to display an unacceptable, moderate potential for anchoring bias such that an anchoring strategy would result in better than chance performance. Thus, all 7- and 6-item tests were excluded from further consideration. Both 5-item tests reduced the test bias associated with the anchoring strategy (anchoring and chance were within two raw points or 6%). Both 5-item tests showed considerable agreement across guessing, anchoring, and central tendency variables (i.e., guessing = 52%, anchoring = 58%, mean = 55%, median = 57%). Both 5-item tests also displayed similar difficulty values; however, the 5A-item test exhibited a 13% improvement in discriminability values as compared to the 5B-item test (i.e., a difference of 0.24 *SD*s). The 5A-item test also exhibited a wider range of score dispersion (i.e., actual range) and a greater range of score variance despite maintaining similar discriminability variance. Overall, results based on the hybrid item-analysis indicate that the 5A-item test offers a highly desirable psychometric profile, matching or exceeding performance of all other candidate tests on all essential test-performance variables.

Table 5. Inter-test and Convergent Validity Coefficients for the Candidate Test Scores

	20-Item	7-Item	6A-Item	6B-Item	6C-Item	5A-Item	5B-Item
20-Item							
7-Item	.31**						
6A-Item	.25**	.98**					
6B-Item	.28**	.97**	.96**				
6C-Item	.31**	.98**	.97**	.95**			
5A-Item	.20**	.94**	.96**	.98**	.92**		
5B-Item	.22**	.94**	.97**	.92**	.98**	.94**	
20-Item AbAccMom	.99** [a]	.31**	.24**	.28**	.31**	.20**	.22**
7-Item AbAccMom	.30**	.97** [a]	.96**	.95**	.95**	.92**	.92**
6A-Item AbAccMom	.23**	.96**	.97** [a]	.94**	.94**	.94**	.94**
6B-Item AbAccMom	.27**	.95**	.94**	.98** [a]	.93**	.96**	.90**
6C-Item AbAccMom	.29**	.94**	.93**	.93**	.96** [a]	.90**	.94**
5A-Item AbAccMom	.19**	.92**	.94**	.96**	.90**	.97** [a]	.92**
5B-Item AbAccMom	.20**	.91**	.94**	.90**	.94**	.92**	.96** [a]
20-Item RelAccMom	.23**	.31**	.35**	.34**	.34**	.37**	.38**
7-Item RelAccMom	.09	.62**	.67**	.67**	.67**	.72**	.73**
6A-Item RelAccMom	.11	.61**	.65**	.66**	.67**	.70**	.72**
6B-Item RelAccMom	.10	.64**	.68**	.71**	.68**	.75**	.72**
6C-Item RelAccMom	.09	.63**	.67**	.68**	.68**	.73**	.72**
5A-Item RelAccMom	.10	.62**	.66**	.69**	.67**	.72**	.70**
5B-Item RelAccMom	.10	.62**	.66**	.69**	.67**	.72**	.71**
20-Item AbAccAng	.78**	.52**	.44**	.45**	.48**	.35**	.37**
7-Item AbAccAng	.42**	.74**	.69**	.65**	.68**	.58**	.60**
6A-Item AbAccAng	.53**	.69**	.65**	.60**	.64**	.54**	.58**
6B-Item AbAccAng	.42**	.71**	.66**	.68**	.65**	.61**	.58**
6C-Item AbAccAng	.49**	.63**	.57**	.52**	.61**	.44**	.54**
5A-Item AbAccAng	.54**	.62**	.59**	.60**	.58**	.56**	.53**
5B-Item AbAccAng	.60**	.47**	.44**	.37**	.50**	.32**	.45**
20-Item RelAccAng	.11	.55**	.56**	.60**	.54**	.61**	.55**
7-Item RelAccAng	.10	.53**	.53**	.59**	.51**	.57**	.49**
6A-Item RelAccAng	.06	.58**	.60**	.63**	.54**	.66**	.56**
6B-Item RelAccAng	.12	.54**	.53**	.59**	.52**	.57**	.50**
6C-Item RelAccAng	.12	.53**	.53**	.57**	.56**	.56**	.55**
5A-Item RelAccAng	.08	.60**	.62**	.66**	.56**	.69**	.58**
5B-Item RelAccAng	.10	.59**	.62**	.61**	.63**	.64**	.66**
ACL Knowledge	.08	-.33**	-.35**	-.36**	-.34**	-.38**	-.36**

Note. AbAccMom = mean absolute error using knee abduction moment as criterion; RelAccMom = relative accuracy using knee abduction moment as criterion (ρ); AbAccAng = mean absolute error using knee abduction angle as criterion; RelAccAng = relative accuracy using knee abduction angle as criterion (ρ); [a]Theoretically should be 1.00, see Footnote 7; *p < .05; **p < .01

Brief Study 1 Discussion

This investigation provided some of the first evidence of reliable skill differences in observational assessment of ACL injury risk. Initial analyses indicated that superior and reproducible performance was both attainable and likely among knowledgeable professionals, although a number of constraints limited the interpretability of one essential aspect of judgment performance. To address this limit, and in order to develop a more robust test of relevant observational assessment skills, a hybrid psychometric item-analysis was conducted yielding a number of optimized candidate test structures. Comparative analysis revealed a psychometrically dominant 5-item test structure optimized for sensitivity, discriminability, difficulty, and guessing. Beyond these desirable psychometric properties, analyses provided evidence of convergent validity, indicating that the new ACL-IQ test is also a robust predictor of other measures of judgment accuracy and general ACL knowledge.

Although the current analyses indicate that the new candidate ACL-IQ is likely to be a strong and robust instrument, several potential limitations need to be considered. For example, the shortened test was created by identifying items that in part maximized psychometric performance differences based on group differences (Exercise Science and General Population). The initial Norwegian Exercise Science sample was relatively small (n = 42), and thus there is some risk of sample and thus test bias. Moreover, the test contained individuals from various professions, which may not be representative of the skilled population (e.g., potential restrictions of range). Although steps were taken to mitigate potential risks (e.g., selection of an extra-wide range of difficulty to avoid ceiling effects), out-of-sample cross-validation is needed. To the extent that future studies include large samples of participants working in various professions within the exercise

sciences (incl. sport coaches and parents of athletes), higher-fidelity analysis of theoretically interesting variations in test-performance would also be possible (e.g., are coaches better than physical therapists; are differences in skilled performance mediated by differences in knowledge variables). Furthermore, because education and age differed between groups in Study 1 it is possible that other ability measures could in part influence performance differences (e.g., is test performance related to mental rotation ability or decision-making skill).

CHAPTER 3: STUDY 2- ACL-IQ CROSS-VALIDATION AND PERFORMANCE MODELING

Introduction

In Study 1, observational ACL injury risk estimation performance was empirically described and superior performance evidence established. A 5-item test was developed providing a suitable range of test difficulty with robust skill group discriminability. This test was developed by selecting items that optimized group differences (Exercise Science and General Population) and difficulty based on a relatively small sample of individuals with diverse backgrounds. In Study 2, the goal was to cross-validate the new 5-item ACL-IQ in a larger and more representative sample, assessing out-of-sample psychometric sensitivity, test-retest reliability, and developing a model of some essential cognitive mechanisms and biases (e.g., cue weighting and utilization).

Preliminary analysis of data from Study 1B indicated ACL knowledge (using a 3-item test) was related to professional domain (r (211) = .68, p < .001) and 5-item test performance (r (211) = .38, p < .001). However, hierarchical linear regression revealed adding ACL knowledge scores to professional domain only improved 5-item ACL-IQ performance estimation by ΔR^2 = .005. Sobel test results also indicated ACL knowledge did not necessarily mediate the effect of group on 5-item test performance, indicating a non-significant trend in the correct direction (Sobel statistic = 1.33, p = .18). The low number of ACL knowledge items and unequal variance/bimodal distribution of the group's score likely influenced this lack of mediation. A more extensive knowledge assessment (11-items) was developed to try to explain the performance results (above profession), which parallels current theoretical models of expertise (cue usage). Additionally, assessing cue utilization/importance may better capture domain-specific knowledge, which should also theoretically be related to performance (e.g., Lens

models). Various strategies may be effective for estimating ACL injury risk. However, the strategies that include cues with higher ecological validity (quantified through biomechanical analysis) should be more effective. Finally, demographic factors and domain-general measures of perceptual/cognitive skill will be used to better understand the mechanisms and boundary conditions for ACL-IQ performance models. Overall, the specific aims of the present study were to cross-validate the skill group discriminability/sensitivity results of Study 1, assess test-retest reliability, understand the underlying cognitive mechanisms of ACL-IQ expertise and assess cross-profession differences.

Hypotheses

1.) Consistent with Study 1B, exercise science professionals (i.e., physical therapists, athletic trainers, orthopedic doctors, exercise science academics/students and strength & conditioning coaches) will make more accurate ACL injury risk estimates compared to general population individuals (i.e., non-exercise science professionals, sport coaches, parents/athletes).

2.) The 5-item ACL-IQ will demonstrate robust test-retest reliability by displaying no mean differences and a high correlation coefficient ($r > .70$) between test performances.

3.) ACL knowledge and cue utility will be the best predictors of test performance.

    a. ACL knowledge will be indirectly related to performance through ratings of cue utility. ACL knowledge should thus allow individuals to focus on or ignore task-relevant or irrelevant cues, which in

turn will determine ACL-IQ performance. The path model depicted in Figure 3 is a conceptual diagram of the hypothesis regarding ACL-IQ performance determining factors.

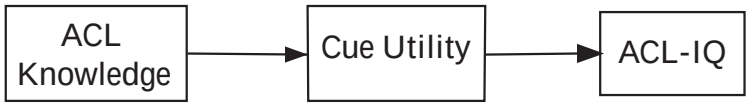

Figure 3. Hypothesized Path Model for Relationship between ACL Knowledge, Cue Utility and ACL-IQ Performance.

4.) Age, gender, education level, risk estimation experience, history of ACL injury, previous/current sport participation, or domain-general perceptual/cognitive skill, if significantly associated with performance, will be mediated by ACL knowledge and cue utility.

5.) Performance differences between groups (i.e., exercise science and general population) will be mediated by ACL knowledge and cue utility.

6.) Professionals within the exercise sciences (i.e., physical therapists, athletic trainers, strength & conditioning coaches, and physicians) who likely encompass the greatest ACL knowledge will perform better than sport coaches, parents, athletes and general population individuals. Additionally, groups hypothesized to have some experience and knowledge regarding ACL injury/prevention (i.e., sport coaches and athletes), will perform better than parents and other general population individuals.

Methods

Participant Characteristics

To enable representative sampling, a variety of exercise science professionals/students (i.e., physicians, sports medicine staff, strength & conditioning coaches, academics, and students) who would benefit from identifying ACL injury risk of athletes, which is similar to the Study 1 Exercise Science group, were recruited. Additionally, a sample of potential non-exercise science or general population individuals, who may also benefit from assessing ACL injury risk, were recruited. All participants were recruited via email through personal networks and list-serv/blog/social media posts or from a paid web panel (i.e., Amazon's Mechanical Turk). Overall, 428 participants completed the study, in which 214 were classified into the Exercise Science group and 214 classified into the General Population group (see Tables 6 and 7 for occupational/subgroup details).

Table 6. Participant Occupation/Subgroup (n = 428)

Occupation	Frequency	Percentage (of Total)
Exercise Science		
Athletic Trainer	50	11.7
Physical Therapist	46	10.7
Physician^	36	8.4
Exercise Science Student	27	6.3
Exercise Science Academic	21	4.9
S&C Coach	34	7.9
Exercise Science Total	214	50.0
General Population		
Other	145	33.9
Parent of Athlete	26	6.1
Young Female Athlete#	11	2.6
Sport Coach	32	7.5
General Population Total	214	50.0

Note. S&C = Strength and Conditioning; ^81% of Physicians Specialized in Orthopedics/Sports Medicine and 19% in Family Medicine; #≤ 25 years old.

Table 7. Demographic Information as a Percentage Within Each Group or Other Specified

Characteristic	General Population (n = 214)	Exercise Science (n = 214)
Highest Degree**		
High School	36.0%	5.6%
Associates	9.3%	0.0%
Bachelors	39.3%	24.8%
Masters	14.0%	32.7%
Doctorate	1.4%	36.9%
Age		
M [95% *CI*]	35.92 [34.22, 37.68]	34.20 [32.70, 35.62]
Mdn [95% *CI*]	33 [30.5, 34]	32 [29, 34]
Gender**		
Female	59.8%	35.5%
Male	40.2%	64.5%
Sport Participation**		
No	43.5%	7.9%
Yes	56.5%	92.1%
Diagnosed with ACL Injury**		
No	92.5%	88.3%
Yes	7.5%	11.7%

Note. *CI* = Confidence Interval (Bootstrap using 1000 samples); **Significant difference (assessed by χ^2) between groups, $p < .01$.

Study Procedures

The study was fully computerized and available online. The ACL-IQ was not compatible with mobile devices and was thus limited to laptop or desktop computers. The participants completed the new 5-item ACL-IQ. Following this test, domain-specific knowledge was assessed with 11 ACL knowledge questions related to location, function, and risk factors for injury (Appendix A). Additionally, cue utility was elicited through a brief survey in which participants indicated the importance of available visual cues (e.g., knee motion, hip motion, trunk motion, landing stiffness, height, weight, etc.) on a 1-10 scale (Appendix B). The analysis of self-reported judgments are likely to have some

limitations (e.g., potential self-serving biases in strategy reporting); however, given the need for quick assessment in an online study, and based on pilot data, the potential costs and benefits of this approach seemed well balanced. Pilot data provided evidence that there was good reason to expect that participants do accurately report key aspects of their strategies, which in turn can be related to more objectively verifiable performance metrics. Additionally, subject rating of cue utility has been used in other expertise studies such as in the Feature Discrimination Task (Loveday, Wiggins, Harris, O'Hare, & Smith, 2013; Loveday, Wiggins, & Searle, 2013; Loveday, Wiggins, Searle, Festa, & Schell, 2012; Wiggins, Brouwers, Davies, & Loveday, 2014).

Domain-general cognitive abilities have been shown to be related to a wide range of differences in superior judgment and decision-making performance; and they often influence differences in strategic task behavior (Cokely et al., 2012; Cokely & Kelley, 2009; Cokely et al., 2006; Ghazal et al., 2014). General cognitive ability may be related to education level, which was potentially different in the groups of Study 1. In order to estimate potential contributions of general cognitive abilities to overall test performance, participants completed the Berlin Numeracy Test (BNT). The BNT has been extensively validated for assessment of statistical numeracy and risk literacy, which is the ability to accurately interpret and make good decisions based on information about risk (Cokely et al., 2012). Theoretically, it is also possible that domain-general mental rotation abilities help determine observational movement analysis performance (i.e., ACL-IQ performance). Previous research indicates that spatial ability (intrinsic and dynamic) is likely a minimal predictor of qualitative movement analysis performance. Nevertheless, to estimate potential contributions of domain-general spatial abilities the 24-item Mental Rotation Test (MRT-A) was administered (Peters et al., 1995; Vandenberg & Kuse, 1978). Personality was assessed using the 10-item Big Five to examine potential

test biases related to ACL-IQ scores (Gosling, Rentfrow, & Swann Jr, 2003). For example, are more conscientious people better at detecting differences in ACL risk level?

Study 2A: Cross-validation and Test-Retest Reliability

Results/Discussion

Basic Attributes and Cross-validation

Range and average measures of ACL-IQ scores are presented in Table 8 (with results from Study 1B presented alongside). No statistically significant difference in effect size (d) was displayed between the two studies (z = 1.17; p = .24). The average score of Study 2A was 63%, which was statistically different from Study 1B, potentially due to the larger sample size of the Exercise Science group (172 more than Study 1B). Similar to Study 1B, no individual scored 100% correct in Study 2A. Additionally, group means between Studies 1B and 2A were nearly identical, corroborating the discriminability evidence for the ACL-IQ, supporting Hypothesis 1.

Table 8. Study 1B and 2A Cross-Validation Comparison

Scale Attributes	Study 1B	Study 2A
M Time in Min:Sec (*SD*)	2:24 (0:47)	
Score Range	0-38 (0-100)	
Achieved Range (%)	12-34 (32-89)	10-36 (26-95)
n	211	428
Overall *M* (%)	21.69 (57)	24.00 (63)*
Overall *Mdn* (%)	21 (55)	24 (63)
Overall *SD* (%)	5.11 (13)	5.86 (15)
Discriminability		
Exercise Science *n*	42	214
Exercise Science *M* (%)	28.31 (74)#	27.97 (74)#
Exercise Science *SD* (%)	3.80 (10)	3.97 (10)
General Population *n*	171	214
General Population *M* (%)	20.07 (53)	20.04 (53)
General Population *SD* (%)	3.96 (10)	4.63 (12)
Cohen's *d*	2.11	1.84
Weighted *SD* (%)	3.92 (10)	4.30 (11)
95% CI	[1.70, 2.48]	[1.60, 2.05]

Note. % = % Correct; *Significantly different from Study 1, $p < .01$; #Significantly different from General Population group, $p < .01$.

Test-Retest Reliability

Internal consistency has been previously assessed (see consistency analysis in Study 1A and results in Table 3). Two of the five repeated video clips, which displayed the highest internal consistency values, were included in the new 5-item ACL-IQ. The 5-item ACL-IQ was administered to a subset of 19 individuals (13 Exercise Science and 6 General Population) on two occasions separated by approximately nine days. Table 9 describes the test-retest characteristics of the ACL-IQ.

Table 9. ACL-IQ Test-Retest Reliability Characteristics (n = 19)

	Test	Retest
Descriptives		
M (%)	28.47 (75)	26.95 (71)
SD (%)	3.82 (10)	4.10 (11)
Range (%)	20-34 (53-89)	18-33 (47-87)
Time Between Tests in Days (*SD*)	9.42 (2.78)	
Reliability Metrics		
Retest Correlation [95% *CI*]	.90 [.74, .96]	
Typical Error (%)	1.28 (3)	
Mean Difference (%)	-1.53 (-4)*	
Cohen's *d*	0.39	

Note. % = % Correct; *CI* = Confidence Interval (Bootstrap using 1000 samples);
*Significant mean difference, $t(18) = -3.68$, $p = .002$.

Despite the high test-retest correlation coefficient (r = .90) a small mean difference was displayed between test sessions[9]. The mean difference is small, within the typical error range and in agreement with Study 1A internal consistency estimates. The typical error represents the amount the score may vary on a repeated performance. For example, if someone scored a 30 (i.e., 8 error points) on the first ACL-IQ there is a 90% probability that their score on a repeated performance will be between 28 and 32 (typical error x 1.65 = 2.11). Thus, based on this typical error profile it is highly unlikely that this statistically significant difference of 1.53 is practically meaningful. Overall, the test-retest results are in support of Hypothesis 2 with the note of a small and likely clinically insignificant mean difference.

[9] Performance level (average across testing sessions), differences in: cue utility ratings or ACL knowledge test score did not independently correlate with test-retest difference scores. Additionally, stepwise linear regression did not reveal any significant predictors of test-retest score differences.

Study 2B: Performance Mechanisms

Results/Discussion

Independent Correlations

Initial independent correlations between various factors and ACL-IQ are displayed in Table 10. All domain-specific factors were related to ACL-IQ score strengthening convergent validity evidence. The new 11-item ACL knowledge test predicted performance to a greater degree than the 3-item ACL knowledge test used in study 1B (i.e., R^2 = .14 vs .35). Various cue utility ratings were also related to ACL-IQ performance. Significant independent task-relevant cues included inward/outward knee/thigh motion and lateral trunk motion. Significant task-irrelevant cues included height and weight of the individual as well as jump height and jump alignment. Regressing ACL-IQ on all cue utility ratings revealed an R^2 = .43, where five cues were statistically significant (jump height, knee/thigh motion, weight, trunk, height and foot alignment). The mean cue importance ratings across levels of ACL-IQ score are depicted in Figure 4.

Table 10. Independent Correlation with ACL-IQ (n = 428)

	ACL-IQ	95% *CI*
Domain-Specific		
ACL Knowledge Test (11-items)	.59**	[.54, .65]
ACL Papers & Books Read/Month	.38**	[.31, .44]
ACL Risk Assessment Experiences (yrs)	.19**	[.11, .28]
Estimated Cue Validity (Cue Utility)		
Arm Motion	-.04	[-.13, .05]
Landing Symmetry	.08	[-.03, .18]
Inward/Outward Knee Motion	.40**	[.32, .47]
Inward/Outward Thigh Motion	.34**	[.26, .42]
Knee & Thigh Composite Average#	.40**	[.33, .47]
Lateral Trunk Motion	.19**	[0.1, .29]
Landing Stiffness	.01	[-.09, .09]
Foot Alignment	-.07	[-.16, .02]
Height of Individual	-.19**	[-.28, -.09]
Weight of Individual	-.38**	[-.46, -.29]
Jump Height	-.54**	[-.61, -.46]
Jump Alignment	-.18**	[-.28, -.09]
Domain-General		
Domain General Perceptual/Cognitive Ability		
Mental Rotation Test-A (24-items)^	.24**	[.15, .33]
Berlin Numeracy Test (4-items)^	.14**	[.05, .24]
Personality Traits		
Extraversion	.12*	[.03, .23]
Agreeableness	-.11*	[-.21, -.01]
Conscientiousness	.17**	[.06, .27]
Emotional Stability	.06	[-.03, .15]
Openness to Experience	-.05	[-.14, .05]
Demographic Variables		
Education Level	.40**	[.32, .47]
Age	-.19**	[-.27, -.10]
Gender	.18**	[.09, .27]
Sport Participation	.30**	[.21, .39]
Diagnosed with ACL Injury	.13**	[.03, .22]

Note. *CI* = Confidence Interval (Bootstrap using 1000 samples); #Variable computed to replace both Knee and Thigh Motion to decrease multicollinearity (Knee and Thigh Motion: r (427) = .71); ^Mental Rotation Test was missing 36(8.4%) values and Berlin Numeracy Test 8(1.9%), Little's Missing Completely at Random test was not significant (p = .53) thus Expectation Maximization (implemented in SPSS) was used to interpolate missing values and used to calculate the correlation coefficient which were not statistically different from the coefficients with missing values (missing data MRT: r (391)= .23** [.14, .33] and BNT: r (419) = .14** [.05, .24]), additionally the means of the interpolated and missing datasets were not statistically different (p < .01); *p < .05; **p < .01;

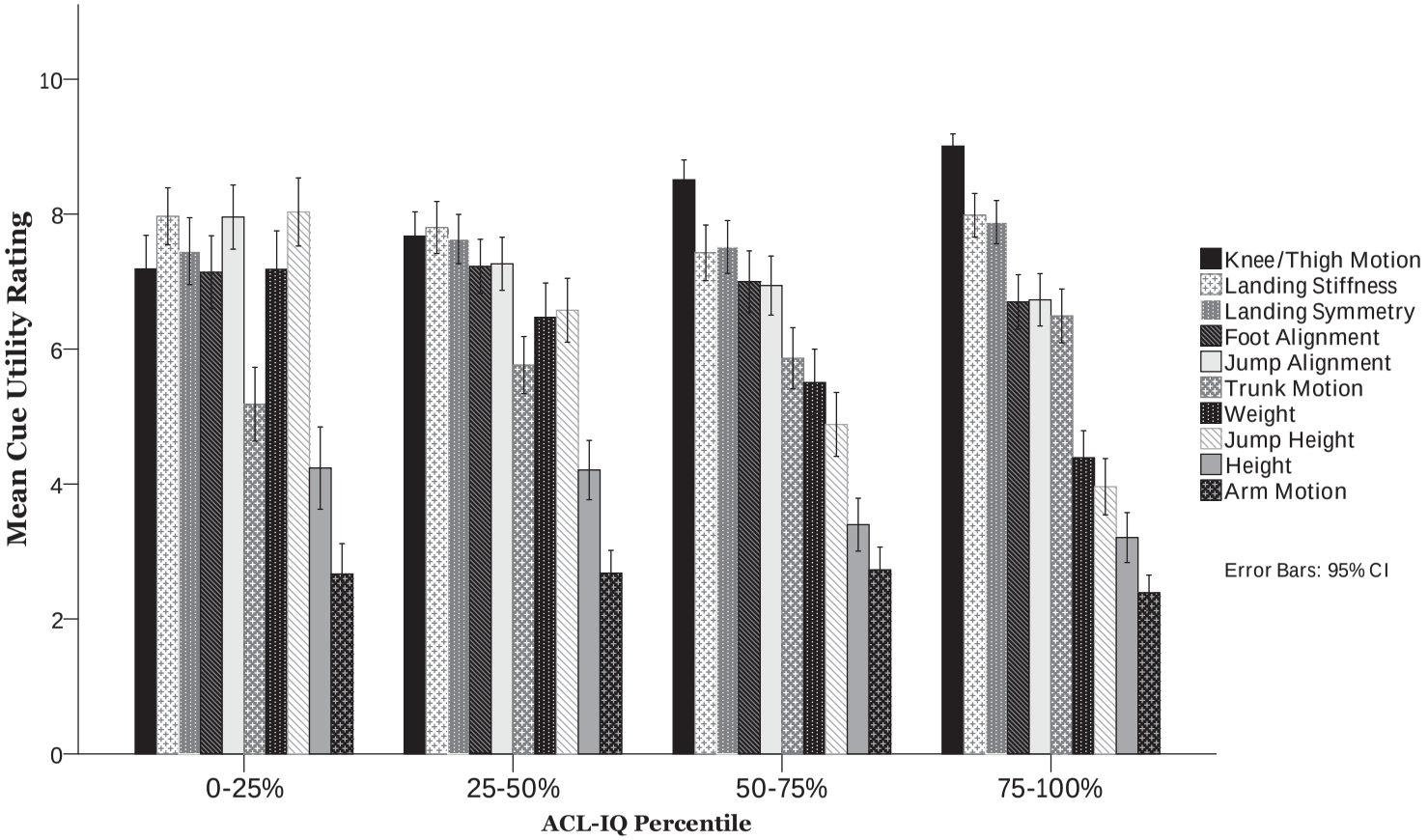

Figure 4. Cue Utility Ratings Across Levels of ACL-IQ Scores (n = 428)

Linear Regression Analysis

To begin to address Hypothesis 3, hierarchical and stepwise regression was performed to assess the independent contributions of ACL knowledge and cue utility ratings for predicting ACL-IQ performance. Additionally, a model was developed to optimize the fit with only the "influential" cue utility variables. The results of these analyses are displayed in Table 11. Together, ACL knowledge and cue utility explained roughly 50% of the variance in ACL-IQ scores. Adding cue importance to ACL knowledge resulted in a 16% increase in variance explained. Alternatively, adding ACL knowledge to cue importance resulted in an 8% increase in variance explained. Through stepwise regression analysis with knowledge and cue utility ratings, of the various cues, four cues were included in the final model (i.e., jump height, knee/thigh motion, weight of individual, and trunk motion).

Table 11. Hierarchical (and Stepwise) Multiple Regression Analysis Predicting ACL-IQ Score from ACL Knowledge and Cue Importance Ratings (n = 428)

	Model 1 (Knowledge Only)		Model 2 (All Cues Only)		Model 3 (Both)		Model 4 (Both/Stepwise)	
Predictor	*B*	β	*B*	β	*B*	β	*B*	β
ACL Knowledge	1.62	.59**			0.94	.34**	0.99	.36**
Cue Utility Ratings								
Jump Height			-0.90	-.44**	-0.61	-.30**	-0.58	-.28**
Knee & Thigh Motion			1.00	.32**	0.65	.21**	0.56	.18**
Weight of Individual			-0.43	-.20**	-0.33	-.16**	-0.28	-.13**
Trunk Motion			0.25	.10*	0.20	.08	0.22	.09*
Height of Individual^			0.33	.13**	0.18	.07		
Foot Alignment			-0.28	-.11*	-0.14	-.05		
Landing Symmetry			-0.14	-.05	-0.06	-.02		
Jump Alignment			0.17	.06	0.05	.02		
Landing Stiffness			0.03	.01	0.01	.004		
Arm Motion			-0.03	-.01	.001	.001		
Constant	14.53		22.43		17.74		17.32	
R^2	.35**		.43**		.51**		.50**	
ΔR^2 Adding Cues	.16**							
ΔR^2 Adding ACL Knowledge			.08**					
ΔR^2 Model 4 to 3							-.005	

Note. Stepwise criteria for entry included F - p < .05 and removal p > .10; B = Unstandardized Coefficient; β = Standardized Coefficient; ^ Weight and Height have a correlation of r (427) = .57 thus the positive regression coefficient for Height may be a product of multicollinearity error (Variance Inflation Factor for Height = 1.82) as the independent relationship with ACL-IQ is negative; *p < .05; **p < .01

To further address Hypothesis 3, hierarchical regression analysis was conducted, using Model 4 as the base, to assess the additive effect of other factors such as domain-general ability, personality and demographics on ACL-IQ performance (Table 12). Although a statistically significant change in F was displayed when adding other domain-specific factors such as the number of ACL books/papers and ACL injury risk assessment experience, the effect was minimal (i.e., $\Delta R^2 = .008$). No significant R^2 change was observed when domain-general factors or personality measures were added to the base model. Demographic measures such as education level, age, and gender significantly improved R^2 by .038. This significant improvement in model fit is relatively small and increases the risk for over-fit. Finally, when other demographic variables such as sports participation and previous ACL injury are added to the base model a significant but minimal change in R^2 resulted ($\Delta R^2 = .009$).

Table 12. Hierarchical Multiple Regression Analysis Predicting ACL-IQ Score (n = 428)

	Model 5		Model 6		Model 7		Model 8		Model 9	
Predictor	*B*	β	*B*	β	*B*	β	*B*	β	*B*	β
Domain-Specific										
ACL Knowledge	0.88	.32**	0.97	.36**	0.97	.36**	0.78	.28**	0.91	.33**
Cue Utility Ratings										
Jump Height	-0.56	-.28**	-0.57	-.28**	-0.55	-.27**	-0.53	-.26**	-0.56	-.28**
Knee & Thigh Motion	0.56	.18**	0.55	.17**	0.52	.16**	0.5	.16**	0.53	.17**
Weight of Individual	-0.27	-.13**	-0.28	-.13**	-0.28	-.13**	-0.28	-.13**	-0.27	-.13**
Trunk Motion	0.17	.07	0.20	.08*	0.21	.08*	0.21	.09*	0.22	.09*
ACL Books/Papers Read	0.26	.10*								
ACL Assessment Experience	0.02	.02								
Domain-General										
Mental Rotation Test			0.07	.07						
Berlin Numeracy Test			-0.17	-.03						
Personality										
Extraversion					0.27	.07*				
Agreeableness					-0.13	-.03				
Conscientiousness					0.24	.05				
Emotional Stability					0.01	.003				
Openness to Experience					-0.35	-.07				
Demographics										
Education							0.29	.13**		
Age							-0.08	-.16**		
Gender							1.09	.09**		
Sport Participation/Injury										
Sport Participation									1.16	.09*
ACL Injury									0.91	.05
Constant	17.68		16.98		17.62		19.82		16.81	
R^2	.51**		.51**		.51**		.54**		.51**	
ΔR^2 (Model 4)	.008*		.004		.009		.038**		.009*	

Note. B = Unstandardized Coefficient; β = Standardized Coefficient; *$p < .05$; **$p < .01$;

Following the hierarchical regression analyses, a stepwise regression analysis was conducted on all the variables to determine the best fitting model without any prior assumptions about predictors (order or type). Furthermore, the stepwise model was assessed for a random 80% of the data in order to cross-validate on a smaller sample to assess model fit/over-fit. The results of two-iterations of training and validating are displayed in Table 13. The demographic variables, number of ACL papers/books read, and trunk motion cue utility were not significant factors in the validation datasets, demonstrating potential over-fitting. Overall, the majority of these performance models (hierarchical and stepwise) included ACL knowledge and three cue utilities (jump height, knee/thigh motion, and weight) to parsimoniously predict ACL-IQ scores while minimizing the potential for over-fit. The hierarchical and stepwise regression analyses provide evidence to support Hypothesis 3, indicating that ACL knowledge and cue utility were the best predictors of ACL-IQ performance.

Table 13. Random 80/20% Cross-validation for Training (Stepwise all Variables) and Validation Datasets for Two Iterations

	80% (n = 342)				20% (n = 86)			
	Training 1		Training 2		Validation 1		Validation 2	
Predictor	B_{1a}	β_{1a}	B_{2a}	β_{2a}	B_{1b}	β_{1b}	B_{2b}	β_{2b}
ACL Knowledge	0.76	.28**	0.74	.28**	**0.74**	**.33****	**0.69**	**.24***
Cue: Jump Height	-0.52	-.26**	-0.43	-.21**	**-0.43**	**-.29****	**-0.85**	**-.40****
Cue: Knee & Thigh Motion	0.41	.13**	0.56	.18**	**0.56**	**.32****	**0.70**	**.20***
Cue: Weight of Individual	-0.26	-.12*	-0.24	-.12**	**-0.24**	**-.20***	-0.22	-.10
Cue: Trunk Motion	0.27	.11*			-0.19	-.07		
ACL Books/Papers Read			0.25	.09*			0.18	.07
Age	-0.09	-.18**	-0.08	-.17**	-0.08	-.09	-0.07	-.12
Education Level^	0.29	.13**	0.31	.14**	0.31	.08	0.06	.03
Gender$	1.35	.12**	1.24	.11**	1.24	0.002	0.17	.01
Constant	20.20		19.39		18.77		22.89	
R^2	.52**		.54**		.64**		.57**	

Note. Stepwise criteria for entry included F - $p < .05$ and removal $p > .10$; B = Unstandardized Coefficient; β = Standardized Coefficient; ^Education Level Coding: High School = 0, Associate = 2, Bachelors = 4, Masters = 6, Doctorate = 8; $ Gender Coding: Male=1, Female = 0; Bold indicates variables which were significant in both training and validation datasets; *$p < .05$; **$p < .01$.

Moreover, the cause of the relationship between the significant domain-general, personality and demographic variables (in Table 10) was likely due to group (Exercise Science and General Population) differences or an indirect effect through ACL knowledge and cue utility. Specifically, when controlling for group membership (Exercise Science and General Population), ACL knowledge, jump height, knee/thigh motion, and weight cue utility ratings were the only significant predictors in a stepwise regression model (see Table 14 for stepwise and cross-validation results). Additionally, mediation analysis was conducted on the significant independent predictors of ACL-IQ score to assess the indirect effect of various predictors on ACL-IQ score through ACL knowledge and cue utility. The total indirect effects of the mediators (ACL knowledge and cues: jump height, knee/thigh motion, and weight), total effects, and direct effects are reported in Table 15. All predictors except age and potentially extraversion had a significant influence on ACL-IQ score through ACL knowledge and cue utility ratings for jump height, knee/thigh motion, and weight (i.e., the total indirect effect 95% *CI* did not contain zero). Domain-general perceptual/cognitive ability, personality factors, and previous ACL injury did not have a significant direct effect on ACL-IQ score (see Direct Effects in Table 15). Factors such as the number of ACL papers/books read, education, and sports participation were partially mediated by ACL knowledge and cue utility factors. Other than age (which has a small influence on ACL-IQ and included in models with likelihood of over-fit), ACL knowledge and cue utility ratings of jump height, knee/thigh motion, and weight, mediated the relationships between significant independent predictors and ACL-IQ, adding confirmatory evidence to Hypothesis 4.

Table 14. Random 80/20% Cross-validation for Training (Stepwise all Variables Controlling Group Membership) and Validation Datasets for Two Iterations

	80% (n = 342)				20% (n = 86)			
	Training 1		Training 2		Validation 1		Validation 2	
Predictor	B_{1a}	β_{1a}	B_{2a}	β_{2a}	B_{1b}	β_{1b}	B_{2b}	β_{2b}
Group^	4.77	.41**	4.63	.40**	**3.30**	**.29****	**3.27**	**.28****
ACL Knowledge	0.49	.18**	0.54	.20**	**0.58**	**.22***	**0.55**	**.21***
Cue: Jump Height	-0.49	-.24**	-0.33	-.16**	**-0.69**	**-.33****	**-0.47**	**-.23****
Cue: Knee & Thigh Motion	0.35	.13**	0.39	.12**	**0.66**	**.21****	**0.77**	**.25****
Cue: Weight			-.18	-.09*			**-0.47**	**-.22****
Constant	18.61		18.00		17.62		18.34	
R^2	.55**		.57**		.63**		.66**	

Note Stepwise criteria for entry included F - $p < .05$ and removal $p > .10$; ^Exercise Science = 1, General Population = 0; B = Unstandardized Coefficient; β = Standardized Coefficient; Bold indicates variables which were significant in both training and validation datasets; *$p < .05$; **$p < .01$.

Table 15. Parallel Mediation Model of Indirect Effects of Various Predictors on ACL-IQ through ACL Knowledge and Cue Utility (Jump Height, Knee/Thigh Motion, and Weight)

Predictors	Total Indirect Effect		Total Effect (*SE*)	Direct Effect (*SE*)
Domain-Specific	Coef (*SE*)	95% CI		
ACL Papers/Books Read	0.70(0.09)	[0.53, 0.89]	1.01(0.12)**	0.30(0.10)**
ACL Risk Assessment Experience	0.20(0.04)	[0.14, 0.31]	0.27(0.07)**	0.06(0.05)
Domain-General				
MRT	0.19(0.03)	[0.13, 0.25]	0.25(0.05)**	0.06(0.04)
BNT	0.83(0.17)	[0.51, 1.19]	0.72(0.24)**	-0.11(0.18)
Personality Traits				
Extraversion	0.22(0.13)	[-0.05, 0.45]	0.44(0.18)*	0.23(0.13)
Agreeableness	-0.41(0.17)	[-0.75, -0.06]	-0.53(0.23)*	-0.12(0.17)
Conscientiousness	0.64(0.16)	[0.32, 0.97]	0.84(0.24)**	0.20(0.18)
Demographics				
Education	0.65(0.06)	[0.54, 0.78]	0.85(0.10)**	0.20(0.09)*
Age	-0.03(0.02)	[-0.07, 0.005]	-0.09(0.02)**	-0.06(0.02)**
Gender	0.86(0.41)	[0.08, 1.76]	2.10(0.56)**	1.23(0.41)**
Sport Participation	2.90(0.43)	[2.11, 3.75]	4.02(0.62)**	1.12(0.49)**
Previous ACL Injury	1.30(0.53)	[0.26, 2.30]	2.33(0.88)*	1.03(0.63)

Note. Coef = Unstandardized coefficient; *SE* = Standard Error; 95% *CI* = Confidence Interval (bias-corrected Bootstrap using 1000 samples)

Why did the Exercise Science group perform better than the General Population group? According to these aforementioned models, Exercise Science professionals would have greater ACL knowledge and rate the importance of jump height and weight lower and knee/thigh motion higher. As previously described, when controlling for group (Exercise Science or General Population), the same four factors, that is, ACL knowledge, jump height, knee/thigh motion, and weight were included in a stepwise regression model and upheld in cross-validation datasets. When these four factors were included as mediators in a model with group predicting ACL-IQ (Figure 5), significant indirect effects were displayed (see Table 16 for specific results).

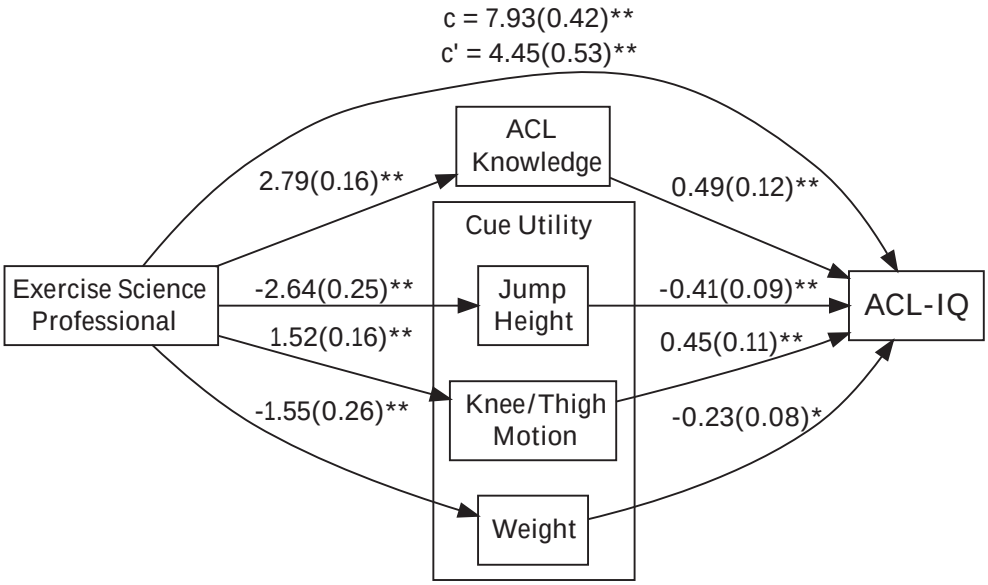

Figure 5. Model 10: Parallel Multiple Mediation Model for the Group Differences in ACL-IQ Scores
Note. Constant = 18.88, n = 428; R^2 = .57). *p < .05; **p < .01

Table 16. Indirect Effects of Group on ACL-IQ though ACL Knowledge and Cue Utility Variables

Predictor	Indirect Effect	Bootstrap *SE*	Bootstrap *CI*
Total	3.48	0.42	[2.71, 4.34]
ACL Knowledge	1.37**	0.35	[0.73, 2.08]
Cue: Jump Height	1.07**	0.28	[0.57, 1.70]
Cue: Knee & Thigh Motion	0.68**	0.19	[0.36, 1.11]
Cue: Weight of Individual Rating	0.35*	0.14	[0.12, 0.67]

Note. Indirect Effect = Unstandardized coefficient; Bootstrap *CI* and *SE* (Standard Error) were bias-corrected with 1000 samples; Normal Theory (Sobel) Test: $*p < .05$; $**p < .01$

This mediation analysis conducted using ordinary least squares path analysis revealed that those in the Exercise Science group displayed higher ACL-IQ scores because they had greater ACL knowledge, rated jump height and weight as less important and knee/thigh motion as more important. The direct effect of group on ACL-IQ, independent of mediators, was still significant but reduced by approximately 44% due to the inclusion of the mediators (difference between c and c' or total and direct effects). Furthermore, the addition of group membership (Exercise Science or General Population) to ACL knowledge and cue utility of jump height, knee/thigh motion, and weight, improved ACL-IQ model fit by ΔR^2 of .072. However, of the explainable variance, ACL knowledge and three cue utilities contributed 87% to the models prediction ability.

To test Hypothesis 3a, the path model: ACL-Knowledge -> Cue Utility -> ACL-IQ, a mediation analysis was conducted. Figure 6 and Table 17 display the results.

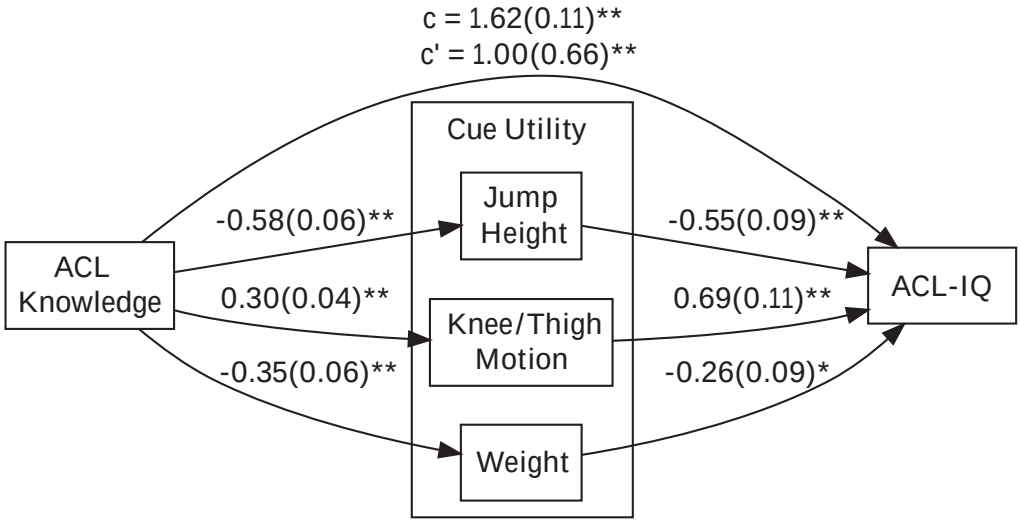

Figure 6. Model 11: Parallel Multiple Mediation Model for the Relationship Between ACL Knowledge and ACL-IQ Scores
Note. Constant = 17.17, n = 428; R^2 = .50). *p < .05; **p < .01

Table 17. Indirect Effects of Knowledge on ACL-IQ though Cue Utility Variables

Predictor	Indirect Effect	Bootstrap *SE*	Bootstrap *CI*
Total	0.62	0.08	[0.48, 0.79]
Cue: Jump Height	0.32**	0.07	[0.20, 0.46]
Cue: Knee & Thigh Motion	0.20**	0.04	[0.13, 0.31]
Cue: Weight of Individual Rating	0.09*	0.03	[0.03, 0.16]

Note. Indirect Effect = Unstandardized coefficient; Bootstrap *CI* and *SE* (Standard Error) were bias-corrected with 1000 samples; Normal Theory (Sobel) Test: *p < .05; **p < .01;

An individual scoring 1.0 point higher on their ACL knowledge test is estimated to have a 0.62 higher ACL-IQ score through their cue utility ratings of the three cues (i.e., total indirect effect). Similarly, independent of cue utility ratings, an individual with 1.0 higher on their ACL knowledge test is estimated to have a 1.0 higher ACL-IQ score (i.e., direct effect or c'). Both cue utility ratings of jump height and knee/thigh motion have statistically larger indirect effects than weight as indicated by the 95% Bootstrap confidence intervals for the contrasts not including zero. If group membership was

included as a covariate, similar results were displayed but the indirect/direct effect magnitudes were smaller (see Figure 7 and Table 18 for specific results).

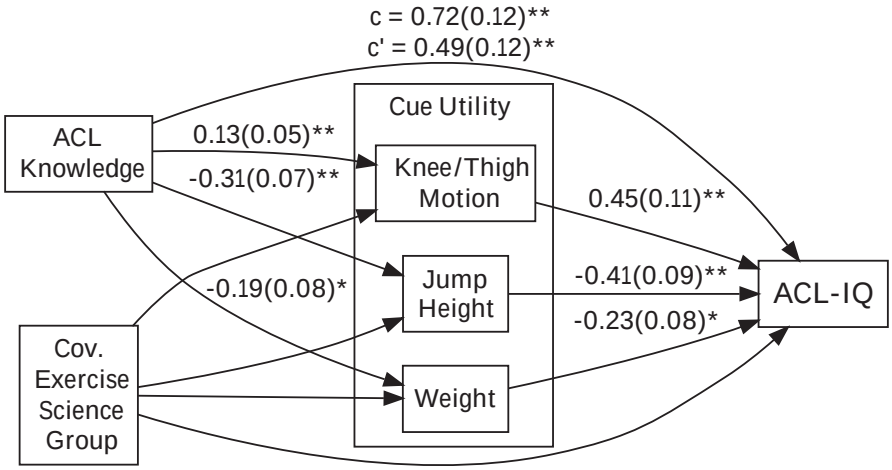

Figure 7. Model 12: Parallel Multiple Mediation Model for the Relationship Between ACL Knowledge and ACL-IQ Scores with Group as a Covariate
Note. Constant = 18.88, n = 428; R^2 = .57). *p < .05; **p < .01

Table 18. Indirect Effects of Knowledge on ACL-IQ though Cue Utility Variables Controlling for Group

Predictor	Indirect Effect	Bootstrap *SE*	Bootstrap *CI*
Total	0.23	0.05	[0.14, 0.35]
Cue: Jump Height	0.13**	0.04	[0.06, 0.23]
Cue: Knee & Thigh Motion	0.06*	0.03	[0.01, 0.13]
Cue: Weight of Individual Rating	0.04#	0.02	[0.01, 0.11]

Note. Indirect Effect = Unstandardized coefficient; Bootstrap *CI* and *SE* (Standard Error) were bias-corrected with 1000 samples; Normal Theory (Sobel) Test: *p < .05; **p < .01; #p = .08.

Conditional Effects

How robust are the relationships in process Model 11? Do various degrees of mental rotation ability change the relationship between cue utility and ACL-IQ? Does playing sports change the relationship between ACL knowledge and cue utility? These

and other theoretically plausible conditional effects were tested (see Figure 8 for conceptual diagram of the various conditional effects).

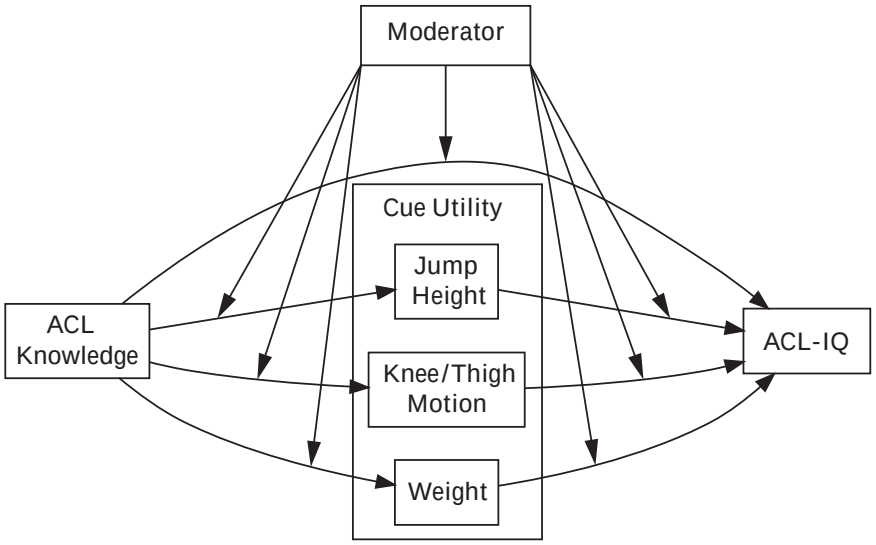

Figure 8. Conceptual Diagram of Potential Moderation Effect on Process Model 11

The moderation effects of domain-general and demographic variables did not significantly alter the performance of, or relationships within, Model 11. Various interaction or moderation terms were statistically significant ($p < .05$), however marginal change in R^2 resulted. Specifically, the inclusion of various moderators in Model 11 resulted in R^2 changes below .017. Furthermore, using the Johnson-Neyman technique to probe the interaction to determine where the values of a moderator change the relationships to become non-significant, no statistical significant transition point was observed for any of the tested moderator (e.g., BNT, MRT, age, education). Additionally, the inclusion of a potential moderator did not alter the indirect effects of ACL knowledge on ACL-IQ through the various cue utilities. Overall, Model 11 appears to be robust against the conditional effects of domain-general and demographic variables. The one boundary condition for this model includes using it to differentiate/explain performance in highly skilled individuals (i.e., Exercise Science professionals).

When group membership (Exercise Science or General Population) is included as a moderator, the interaction term with jump height cue utility and group becomes significant (p = .037). When this independent interaction term is added to the Model 11 predicting ACL-IQ, the ΔR^2 = .005 (p = .02). Moreover, the relationship between jump height and ACL-IQ is significant for the General Population group but not for the Exercise Science group. Additionally, the indirect effects of ACL knowledge on ACL-IQ through the cue utility mediators are conditioned upon group membership. Specifically, the indirect effects of ACL knowledge on ACL-IQ through cue utility mediators becomes non-significant (all cue utility indirect effects include zero in the 95% Bootstrap *CI*) for Exercise Science group whereas the indirect effects remain significant in the General Population group. These between-group differences in model performance/structure can be further exemplified by regressing ACL-IQ on the predictors (ACL knowledge, cue utility ratings of: jump height, knee/thigh motion, and weight), for each group separately. The General Population group R^2 = .28, where all variables were significant (p < .05) and Exercise Science R^2 = .12, where all variables except jump height were significant (p < .05). Thus, as the level of expertise increases, it becomes more difficult to predict performance with these simplified and moderate reliability assessment methods for cue utility.[10] Additionally, for groups with high levels of expertise, the indirect effects of knowledge on ACL-IQ through the cue utility mediators become trivial.

The moderation effect of group membership (Exercise Science or General Population) accords with expertise research where higher fidelity/reliability process level data such as eye-tracking and verbal protocol analysis are better at identifying expertise differences when performers have similar or high abilities. Thus, the error resulting from

[10] All cue utility rating test-retest reliability coefficients (r) ranged between .49 to .78 (n = 19).

restriction of range and relatively imprecise cognitive process level data substantiate the moderation effect of group membership. However, given the ease of using this cue utility assessment, the results indicate these factors were robust against many plausible moderators and useful for describing the overall nature of risk-estimating skill across a wide range of ability levels.

Cross-Profession Discriminability

Subgroup/occupation ACL-IQ scores and ACL knowledge test results are depicted in Figure 9 and cue importance ratings in Figure 10. One-way ANOVA with post-hoc pairwise comparison (Tukey HSD) on ACL-IQ revealed that General Population subgroups such as non-exercise science professionals and parents (i.e., Other and Other/Parent) performed lower than all other subgroups ($p < .05$). Female athletes performed lower than exercise science students ($p < .05$).[11] Sport coaches displayed lower ACL-IQ scores than exercise science students and academics, physicians, strength and conditioning coaches, athletic trainers, and physical therapists ($p < .05$). There was no statistically significant difference in ACL-IQ between exercise science students and academics, physicians, strength and conditioning coaches, athletic trainers, or physical therapists ($p > .05$). When grouped according to Hypothesis 6 predictions, that is, High: physical therapists, athletic trainers, strength & conditioning coaches, and physicians; Medium: sport coaches and athletes; and Low: parents and other non-exercise science individuals (i.e., Other) results are displayed in Figure 11. All three groups displayed significant mean differences ($p < .05$) in ACL-IQ score.

All General Population subgroups displayed lower ACL knowledge compared to the Exercise Science subgroups ($p < .05$). There were no significant differences in ACL

[11] Only 11 non-exercise science female athletes were included in the sample thus mean estimates are imprecise.

knowledge between exercise science students and academics, physicians, strength and conditioning coaches, athletic trainers, or physical therapists ($p > .05$).

The mean cue utility differences for the cues significantly related to performance were analyzed using a one-way ANOVA with post-hoc pairwise comparisons (Tukey HSD). Exercise science students and academics, physicians, strength and conditioning coaches, athletic trainers, and physical therapists rated knee/thigh motion higher than the Other subgroup (which was not statistically different from Other/Parent, female athletes, or sport coaches) ($p < .05$). Strength and conditioning coaches and athletic trainers rated trunk motion higher than the Other subgroup ($p < .05$). No statistically significant difference between subgroups was displayed for the cue utility rating of height. The Other subgroup rated weight higher than athletic trainers ($p < .05$). Other/Parent subgroup rated weight higher than exercise science students and academics, physicians, athletic trainers, and physical therapists. Sport coaches rated weight higher than athletic trainers ($p < .05$). Other and Other/Parent subgroups rated jump height higher than exercise science students and academics, physicians, strength and conditioning coaches, athletic trainers, and physical therapists. Sport coaches rated jump height higher than exercise science students & academics, physicians, athletic trainers and physical therapists. The Other subgroup rated jump alignment greater than exercise science academics ($p < .05$). Other/Parent subgroup rated jump alignment greater than exercise science students and academics ($p < .05$).

Overall, the conclusions from the subgroup analysis parallel the aforementioned performance modeling results. That is, the General Population subgroups (i.e., Other, Other/Parents, Sport Coaches) have lower ACL-IQ, lower ACL knowledge, rate the importance of knee/thigh motion lower, and weight and jump height higher. The slightly higher ACL-IQ of sport coaches over the Other subgroup is likely due to the slightly

higher ACL knowledge and higher knee/thigh importance rating. Finally, Hypothesis 6 was supported as significant mean differences in ACL-IQ between the three groups were displayed.

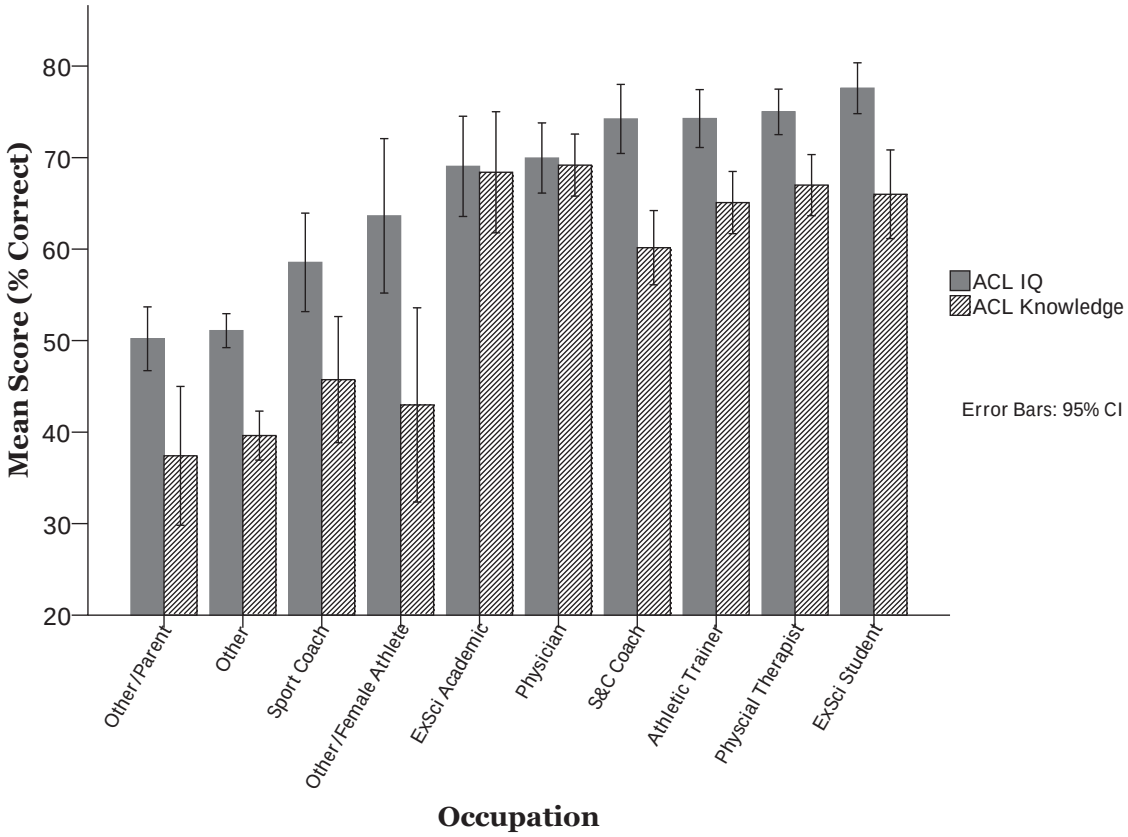

Figure 9. ACL-IQ and ACL Knowledge Scores of Various Subgroups.
Note. ExSci = Exercise Science; S&C = Strength and Conditioning

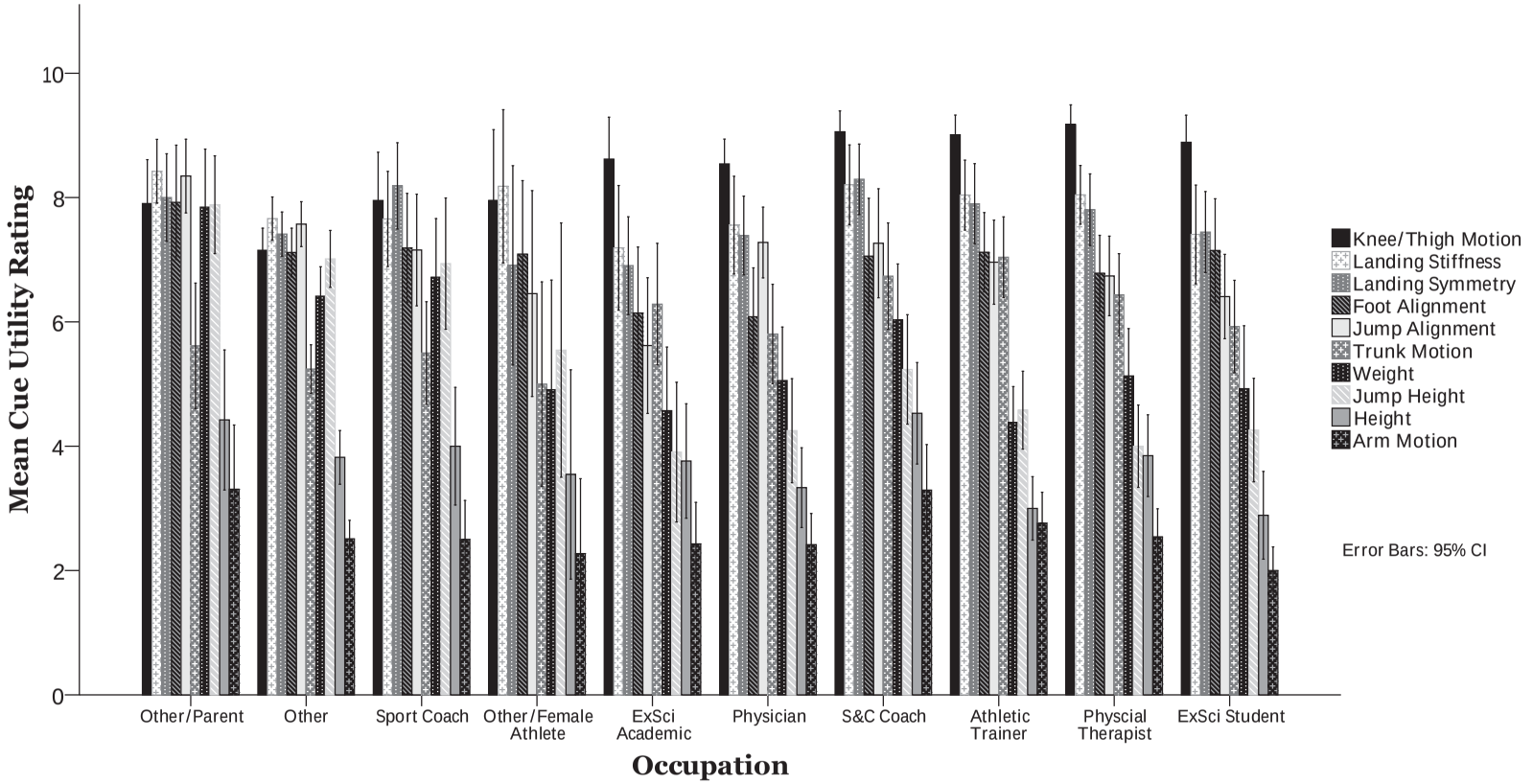

Figure 10. Cue Utility Ratings of Various Subgroups.
Note. ExSci = Exercise Science; S&C = Strength and Conditioning

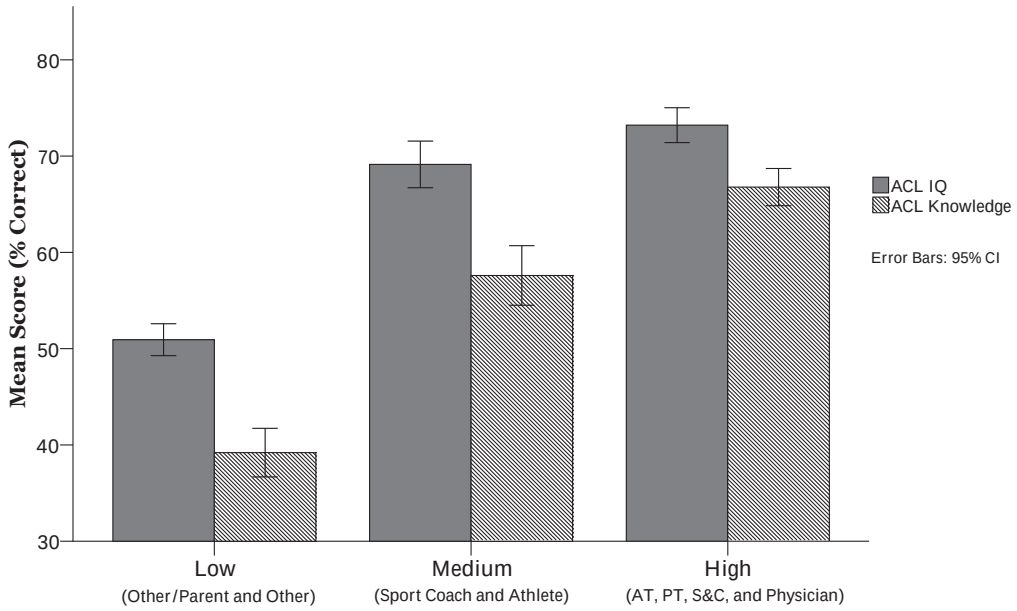

Figure 11. ACL-IQ and ACL Knowledge Scores of Subgroups in Three Hypothesized Skill Levels.
Note. AT = Athletic Trainer; PT = Physical Therapist; S&C = Strength and Conditioning Coach.

Brief Study 2 Discussion

The results from Study 2A corroborate evidence from Study 1B indicating the ACL-IQ is a reliable and sensitive tool for assessing ACL injury risk estimating expertise. Additionally, further validity evidence was established demonstrating the ACL-IQ works well because it conforms with current theories of expertise where domain-specific factors, and importantly, judgment processes, contribute highly to describe performance mechanisms. The boundary conditions for the ACL-IQ process model demonstrated sufficient robustness and parallels contemporary theories of expert performance. Furthermore, occupational differences in ACL-IQ reflect differences in domain-specific knowledge and cue utility. Finally, results from Study 2B document the individuals/groups who would likely need to improve performance, and the established process Model 11 provides a foundation for developing these training tools/programs.

CHAPTER 4: CONCLUSIONS

The initial results of Study 1A provided the first evidence that ACL injury risk can be reliably and accurately assessed by visual inspection. Additionally, in Study 1B, a short test was developed to assess this skill using modern psychometric techniques. The results of Study 2A replicated the results of Study 1B providing additional converging evidence demonstrating the 5-item ACL-IQ was a psychometrically robust and reliable research tool for examining individual differences in ACL injury risk estimating ability. An additional aim of Study 2 was to extend results of Study 1B by examining the potential cognitive mechanisms underlying performance. Domain-specific knowledge (including measures of cue usage/importance) was the best predictor of performance. Additionally, a process model (Figure 6) was developed describing the indirect effect of ACL knowledge on ACL-IQ through cue utility ratings of three cues (mediators). This process model was robust against potential moderators and is essential for both strengthening theories of skilled performance as well as the development of training or decision support tools. Finally, this framework for systematically assessing observational movement analysis skill can be extended to other clinical situations where identifying biomechanical abnormalities is important for assessing injury risk, performance enhancement, or rehabilitation progress. Overall, this dissertation developed the first systematic approach and technology for assessing individual differences in observational movement analysis skill.

Theoretical Contribution

The results from Study 1B and 2B provide information regarding the nature of expertise, which parallels contemporary theories of expert performance. Specifically, results of study 2B demonstrated that the best factors that influenced performance were

domain-specific knowledge and cue utility. Previous expertise research in occupational (Schmidt & Hunter, 2004; Schmidt & Hunter, 1998) and movement analysis (Leas & Chi, 1993; Ste-Marie, 1999) disciplines also acknowledged the importance of domain-specific knowledge as a factor influencing expertise. Similarly, in Study 2B, domain-general abilities were related to performance but only because of or through domain-specific factors, which is also consistent with previous findings (Schmidt et al., 1986). Furthermore, expertise differences in the current study were related to cognitive process measures such as cue utility.

Previous meta-analyses using process level data revealed that experts focus on task-relevant while ignoring task-irrelevant cues (Gegenfurtner et al., 2011; Mann et al., 2007). These findings are consistent with ACL-IQ expertise as cue importance ratings, lower fidelity process level data, were related to expertise. Specifically, individuals who rated knee/thigh motion as more important and jump height/weight as less important, performed better. Cue utility ratings have been included in an extensive battery of expertise assessment methods (e.g., Expert Intensive Skills Evaluation or EXPERTise), which have been shown to differentiate levels of expertise in a variety of tasks (Loveday, Wiggins, Harris, et al., 2013; Loveday, Wiggins, & Searle, 2013; Loveday et al., 2012; Wiggins et al., 2014).[12]

The current results also specified the relationship between domain-specific knowledge and cue utility. Specifically, as theory would suggest, greater domain-specific ACL knowledge leads to better performance but this is (in part) due to the use of task-

[12] Interestingly, for the high level performers (i.e., Exercise Science group), individuals with greater variance in their cue utility ratings (representing greater discrimination and similar to the Cochran-Weiss-Shanteau index) performed better. The cue utility variability metric was the strongest independent correlate with ACL-IQ r (213) = .29 (p < .001) in the high ability (Exercise Science) group. When combined with the General Population the correlation between ACL-IQ and this cue utility variability index was marginal r (427) = .17 (p < .01).

relevant information while ignoring irrelevant information. These characteristics of expert performance are related to differences in structure and complexity of memory representations and are developed through deliberate practice (Ericsson et al., 1993; Ericsson & Lehmann, 1996).

However, the results from the current study also suggest the skill of observational ACL injury risk assessment may not be the type of skill which requires extensive deliberate practice over many years as seen in chess, surgery, athletics, and many other skills. For instance, regression results revealed that high ACL-IQ could result from appropriate cue usage even if ACL knowledge is "low" and the individual is not an exercise science professional.[13] Furthermore, results from the development of the ACL nomogram by Myer and colleagues (2011; 2011; 2010a; 2010b; 2011) indicated that specific cues such as tibia length, medial knee motion and knee flexion predict ACL injury risk, and in theory anybody could use this decision support tool to become an "expert." Accordingly, it may be possible to become proficient in ACL injury risk estimation by only assessing the appropriate cues which may be learned in a short period of time with simple instructions or with decision support tools (e.g., nomogram, decision tree, etc.).

Decision Support/Training Applications

As previously discussed, cue usage/utility is important for successfully estimating ACL injury risk. Thus, in order to improve performance, attention should be allocated to the specific task relevant cues such as knee/thigh motion (as well as landing stiffness and symmetry) while ignoring task irrelevant cues such as jump height and weight. The

[13] As an example, 10 General Population individuals under the 50th percentile in ACL knowledge, had an ACL-IQ higher than the mean of the Exercise Science group (i.e. ACL-IQ >74% correct).

previously developed ACL nomogram provides a decision support tool for assessing ACL injury risk but requires two vantage points and assessment of other factors/cues not observable through motion information (i.e. hamstring to quadriceps strength ratio). Additionally, the weighting of the various cues in the ACL nomogram represents a different strategy compared to skilled individuals in the current study. For example, the cue with the greatest relative importance for assessing ACL injury risk, according to the ACL nomogram, is height,[14] whereas the observation "experts" in this study (i.e., 75-100 percentile) rated knee/thigh motion as most important. Therefore, a different decision support tool using evidence from psychological process data of skilled individuals may provide a better alternative approach to improve performance (see Figure 12 for an example). A simple decision tree may also serve as an effective learning aid. For instance, following many uses of the decision tree, an individual may learn these simple heuristics and not need the decision tree. Furthermore, cognitive process tracing methods such as verbal protocol analysis or eye-tracking can be used to reverse engineer superior performance in order to optimize training and decision support tools (Cokely et al., 2012; Cokely & Kelley, 2009; Ericsson, Charness, Feltovich, & Hoffman, 2006; Ericsson et al., 1993; Hoffman et al., 2013; Ward, Suss, & Basevitch, 2009).

[14] The actual factor is tibial length but is highly correlated to height $r(19) = .96$.

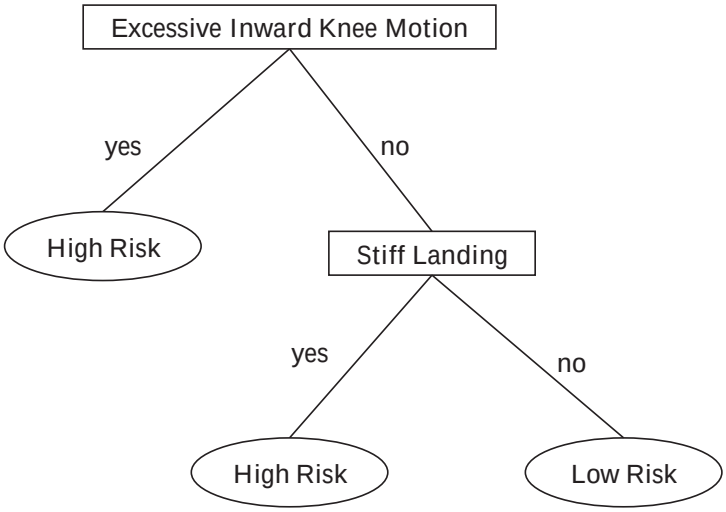

Figure 12. Example Decision Tree for Estimating ACL Injury Risk

Clinical Application/Score Meaning

What is the clinical significance of the ACL-IQ score? The ACL-IQ can be transformed into practical meaning by simply subtracting the score from 38 (maximum points) and dividing by 5 (number of items/video clips). This value represents the average deviation from the criterion on any given video clip or test item. For example, if an individual scored a 28, their average absolute error would be 2.00 ([38-28]/5), meaning if a video clip was presented with an actual risk value of 5 (on the 1-10 point scale: see Figure 2) this individual would, on average, be within ± 2 or between 3 and 7 (see Figure 13 for depiction of mean error across occupations).

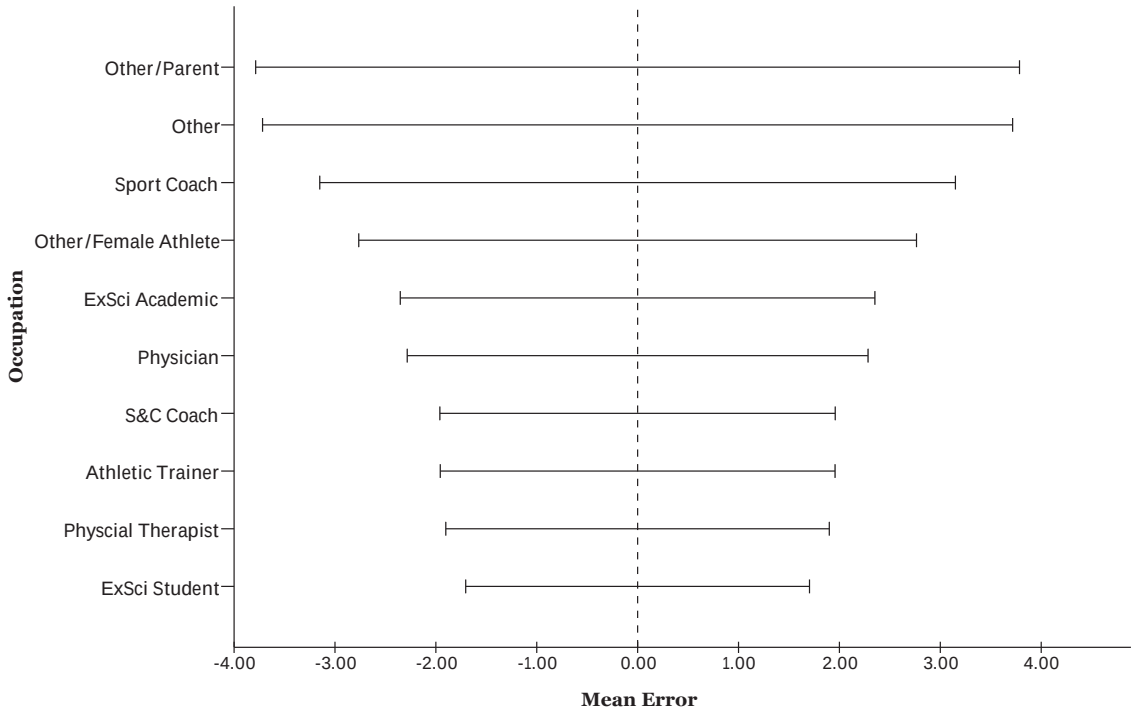

Figure 13. Mean Error Across Occupation
Note. ExSci = Exercise Science; S&C = Strength and Conditioning

A mean error of 2 may seem unacceptable to some, but if the purpose of identifying the ACL injury risk level of an athlete (i.e., screening) is to decide an appropriate intervention (e.g., feedback, training, etc.) the athlete may only need to be classified into a "high" or "low" risk group (a classical signal detection task). Unfortunately, with only 5 items/trials, a reliable signal detection analysis cannot be conducted. However, if we classify the risk level of an athlete at greater than 5 (on the 10 point scale) as "high" risk we can determine the number of judges who correctly classified all 5 athletes (video clips) into either "high" (i.e., above 5) or "low" (i.e., below 5) risk categories.[15] Overall, 20% of the total sample classified all 5 video clips into

[15] Three of the five ACL-IQ items/video clips were "high" risk (knee abduction moments above 41 Nm) and two were "low" risk (knee abduction moments below 17 Nm). Previous

correct “high” and “low” risk categories. Group-wise, 35% of Exercise Science individuals and 4% of the General Population classified all 5 video clips into correct “high” and “low” risk categories. The average ACL-IQ for these individuals with 100% two-category classification accuracy was 31.18 (6.81 error points, mean error of 1.34, or 82% correct).

ACL-IQ scores can also be compared to the ACL nomogram performance.[16] When transformed into 1-10 categories, the ACL nomogram demonstrated 8 error points (ACL-IQ score of 30 or 79% correct). Overall, 23% of the total sample performed better than or equal to the ACL nomogram. Group-wise, 40% of Exercise Science individuals and 6% of the General Population performed better than or equal to the ACL nomogram. Conducting a one-sample *t* test with ACL nomogram performance across various occupations revealed that, on average, the ACL nomogram performed better than all professional subgroups except Exercise Science Students (t (26) = -1.01, p = .32). A summary of the subgroup proportions above these specific clinical thresholds (i.e., two-category classification and nomogram) is located in Table 19.

---

research has used a knee abduction moment of 25 and 22 Nm as a cut-point for “high” and “low” risk.

[16] The current video clips used in this study had concurrent ACL nomogram assessment for only the left leg. Right and left leg knee abduction moment (actual risk criterion), demonstrated a correlation coefficient of r(19) = .62.

Table 19. Summary of Proportions Within Each Subgroup Above Specific Criteria (n = 428)

	100% Two-Category Accuracy		Greater than or Equal to Nomogram	
Occupation	Frequency	Percentage (within Subgroup)	Frequency	Percentage (within Subgroup)
Exercise Science				
Athletic Trainer (n = 50)	22	44.0	24	48.0
Physical Therapist (n = 46)	21	45.7	20	43.5
Physician (n = 36)	6	16.7	9	25.0
Exercise Science Student (n = 27)	11	40.7	15	55.6
Exercise Science Academic (n = 21)	4	19.0	4	19.0
S&C Coach (n = 34)	11	32.4	14	41.2
Exercise Science Total (n = 214)	75	35.0	86	40.2
General Population				
Other (n = 145)	5	3.4	5	3.4
Parent of Athlete (n = 26)	0	0	0	0
Young Female Athlete (n = 11)	1	9.1	2	18.2
Sport Coach (n = 32)	3	9.4	6	18.8
General Population Total (n = 214)	9	4.2	13	6.1

Note. S&C = Strength and Conditioning.

The collective clinical criteria suggest an ACL-IQ score of around 80% correct may be suitable for justifying the use of observation as a suitable screening method for ACL injury risk. However, more data is needed to support this claim and in particular, an appropriate signal detection analysis would reveal estimates of sensitivity and specificity, which can be used to assess the efficacy of a screening approach (in addition to cost and time associated with the screening method and misses/false alarms). Additional studies, which are prospective in design, are also needed to assess the predictive validity for using the ACL-IQ to identify individuals who are exceptional at predicting actual injury risk.

Prospective studies should be conducted to assess if ACL injuries can be reduced by observational screening. A prospective injury risk study could be conducted by incorporating observational screening with appropriate training intervention and comparing injury rates with no screening or training everyone. Prospective studies are resource intensive and often require many years of data collection. To begin to understand if observation can be used to assess ACL injury risk, a pseudo-prospective study could be conducted using video-taped individuals (i.e., drop vertical jump) who later went on to injure their ACL in a non-contact situation. Specifically, video clips of a representative sample of athletes could be shown/rated by observers with various levels of ACL-IQ. Classification accuracy could then be established by comparing the observer ratings to actual outcomes (no injury/ injury). This design would significantly reduce time and any ethical dilemmas associated with identifying injury risk level by unskilled individuals as well any confounding effects due to training. The goal would be to establish evidence that ACL-IQ scores would be correlated with observers' classification accuracy (ROC area, sensitivity, specificity, etc.) with actual injurious events.

Broader Applications

The ACL-IQ represents an interdisciplinary contribution to longstanding research programs aimed at preventing injuries and understanding human performance. This work provides a foundation for future research investigating the degree to which simple observational screening can prevent ACL injuries. Additionally, the web-based nature of test and online platform: www.ACL-IQ.org enhances outreach and awareness, maximizing research impact. The website not only houses the ACL-IQ, which provides individualized feedback to individuals regarding their ACL-IQ performance, but is also a repository for data collection, a means of informing the public and other researchers about ACL injury/prevention, as well as place to house future training programs or decision support tools.

Beyond ACL-specific implications and applications, the current dissertation provides a framework that can be applied to other problems related to movement analysis/musculoskeletal injury risk assessment. For example, biomechanical (movement) analysis has been used to predict several costly and debilitating musculoskeletal injuries including lower back disorders (Marras et al., 1995), concussive head impacts (Rowson et al., 2012), tibial stress fractures (Pohl, Mullineaux, Milner, Hamill, & Davis, 2008), fall risk in the elderly (Callisaya et al., 2011), and second ACL injury (Paterno et al., 2010) to name a few. Consequently, current biomechanical approaches are costly, time intensive, and require specialized training to operate/use.

An alternative framework focuses on *skilled movement analysis*. Simple visual inspection, as opposed to biomechanical instruments, significantly reduces cost and time. Specifically, specialized equipment is not needed and measurement is nearly instantaneous. Theoretically, many domains could benefit from using observation as a

movement analysis method. However, it is not known if individuals have the ability to accurately and reliably visually detect movement patterns that place individuals at risk for injury in these other domains or situations. Going forward I predict a *skilled movement analysis* approach will be useful when we can:

1.) Establish a normative benchmark
2.) Efficiently assess observational skill
3.) Construct a model of skill mechanisms
4.) Develop skill training and decision support

Efficient systems improve health and human performance.

REFERENCES

Ageberg, E., Bennell, K. L., Hunt, M. A., Simic, M., Roos, E. M., & Creaby, M. W. (2010). Validity and inter-rater reliability of medio-lateral knee motion observed during a single-limb mini squat. *BMC Musculoskeletal Disorders, 11*, 265-273.

Agel, J., Arendt, E. A., & Bershadsky, B. (2005). Anterior cruciate ligament injury in national collegiate athletic association basketball and soccer: a 13-year review. *American Journal of Sports Medicine, 33*(4), 524-530.

American Educational Research, A., American Psychological, A., National Council on Measurement in, E., Joint Committee on Standards for, E., & Psychological, T. (1999). *Standards for educational and psychological testing*: American Educational Research Association.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*(4), 1036-1060.

Ardern, C. L., Webster, K. E., Taylor, N. F., & Feller, J. A. (2011). Return to the preinjury level of competitive sport after anterior cruciate ligament reconstruction surgery: two-thirds of patients have not returned by 12 months after surgery. *American Journal of Sports Medicine, 39*(3), 538-543.

Balslev, T. (2011). *Learning to diagnose using patient video cases in paediatrics: perceptive and cognitive processes.* (Doctor of Philosophy), Maastricht University, Denmark.

Baron, J. (1978). Intelligence and general strategies. *Strategies of information processing*, 403-450.

Baron, J. (2005). *Rationality and intelligence*: Cambridge University Press.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science, 321*(5890), 851-854.

Boshuizen, H., & Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cognitive Science, 16*(2), 153-184.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*: University of California Press.

Buschkuehl, M., & Jaeggi, S. M. (2010). Improving intelligence: A literature review. *Swiss Medical Weekly, 140*(19-20), 266-272.

Callisaya, M. L., Blizzard, L., Schmidt, M. D., Martin, K. L., McGinley, J. L., Sanders, L. M., & Srikanth, V. K. (2011). Gait, gait variability and the risk of multiple incident falls in older people: a population-based study. *Age & Ageing, 40*(4), 481-487.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*(1), 55-81.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121-152.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making, 7*(1), 25-47.

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making, 4*(1), 20-33.

Cokely, E. T., Kelley, C. M., & Gilchrist, A. L. (2006). Sources of individual differences in working memory: Contributions of strategy to capacity. *Psychonomic Bulletin & Review, 13*(6), 991-997.

Damisch, L., Mussweiler, T., & Plessner, H. (2006). Olympic medals as fruits of comparison? Assimilation and contrast in sequential performance judgments. *Journal of Experimental Psychology: Applied, 12*(3), 166-178.

de Groot, A. (1978). *Thought and choice in chess*: The Hague: Mouton.

Doll, J., & Mayr, U. (1987). Intelligence and success in chess playing: An examination of chess experts. *Psychologische Beitrage, 29*(2-3), 270-289.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences, 108*(19), 7716-7720.

Duckworth, A. L., & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science, 16*(12), 939-944.

Ekegren, C. L., Miller, W. C., Celebrini, R. G., Eng, J. J., & Macintyre, D. L. (2009). Reliability and validity of observational risk screening in evaluating dynamic knee valgus. *Journal of Orthopaedic & Sports Physical Therapy, 39*(9), 665-674.

Ericsson, K. A. (2007). An expert-performance perspective of research on medical expertise: the study of clinical performance. *Medical Education, 41*(12), 1124-1130.

Ericsson, K. A. (2013). Why expert performance is special and cannot be extrapolated from studies of performance in the general population: A response to criticisms. *Intelligence, 45*, 81-103.

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (2006). *The Cambridge handbook of expertise and expert performance*: Cambridge University Press.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*(2), 211-245.

Ericsson, K. A., Krampe, R. T., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363-406.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: evidence of maximal adaptation to task constraints. *Annual Review of Psychology, 47*, 273-305.

Ericsson, K. A., Roring, R. W., & Nandagopal, K. (2007). Giftedness and evidence for reproducibly superior performance: An account based on the expert performance framework. *High Ability Studies, 18*(1), 3-56.

Ericsson, K. A., & Ward, P. (2007). Capturing the naturally occurring superior performance of experts in the laboratory. *Current Directions in Psychological Science, 16*(6), 346-350.

Ericsson, K. A., & Williams, A. M. (2007). Capturing naturally occurring superior performance in the laboratory: translational research on expert performance. *Journal of Experimental Psychology: Applied, 13*(3), 115-123.

Foster, S. L., & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment, 7*(3), 248-260.

Garcia-Retamero, R., & Cokely, E. T. (2013a). Communicating health risks with visual aids. *Current Directions in Psychological Science, 22*(5), 392-399.

Garcia-Retamero, R., & Cokely, E. T. (2013b). Simple but powerful health messages for increasing condom use in young adults. *Journal of Sex Research*(Epub ahead-of-print), 1-13.

Garcia-Retamero, R., & Cokely, E. T. (2013). The influence of skills, message frame, and visual aids on prevention of sexually transmitted diseases. *Journal of Behavioral Decision Making, 27*(2), 179-189.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review, 23*(4), 523-552.

Ghazal, S., Cokely, E. T., & Garcia-Retamero, R. (2014). Predicting biases in very highly educated samples: Numeracy and metacognition. *Judgment and Decision Making, 9*(1), 15-34.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review, 98*(2), 254-267.

Gigerenzer, G., & Kurz, E. M. (2001). Vicarious functioning reconsidered: A fast and frugal lens model. In K. R. Hammond & T. R. Stewart (Eds.), *The Essential Brunswik: Beginnings, Explications, Applications*: Oxford University Press.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*: Oxford University Press, USA.

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. M. (2003). The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology, 56*(11), 1129-1135.

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*(6), 504-528.

Griffin, L. Y., Albohm, M. J., Arendt, E. A., Bahr, R., Beynnon, B. D., Demaio, M., . . . Yu, B. (2006). Understanding and preventing noncontact anterior cruciate ligament injuries: a review of the Hunt Valley II meeting, January 2005. *American Journal of Sports Medicine, 34*(9), 1512-1532.

Gwet, K. L. (2012). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters*: Advanced Analytics Press.

Haider, H., & Frensch, P. A. (1999). Eye movement during skill acquisition: more evidence for the information-reduction hypothesis. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25*(1), 172-190.

Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional Module on. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hegarty, M., Montello, D. R., Richardson, A. E., Ishikawa, T., & Lovelace, K. (2006). Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence, 34*(2), 151-176.

Herbig, B., & Glöckner, A. (2009). Experts and decision making: First steps towards a unifying theory of decision making in novices, intermediates and experts. *Preprints of the Max Planck Institute for Research on Collective Goods, 2009*(2), 1-29.

Herrington, L., Myer, G. D., & Munro, A. (2013). Intra and inter-tester reliability of the tuck jump assessment. *Physical Therapy in Sport, 14*(3), 152-155.

Hertzog, C., & Robinson, A. (2005). *Metacognition and Intelligence*: Sage Pubplications.

Hewett, T. E., Myer, G. D., Ford, K. R., Heidt, R. S., Jr., Colosimo, A. J., McLean, S. G., . . . Succop, P. (2005). Biomechanical measures of neuromuscular control and valgus loading of the knee predict anterior cruciate ligament injury risk in female athletes: a prospective study. *American Journal of Sports Medicine, 33*(4), 492-501.

Hewett, T. E., Myer, G. D., Ford, K. R., Paterno, M. V., & Quatman, C. E. (2012). The 2012 ABJS Nicolas Andry Award: The sequence of prevention: a systematic

approach to prevent anterior cruciate ligament injury. *Clinical Orthopaedics and Related Research, 470*(10), 2930-2940.

Hewett, T. E., Zazulak, B. T., Krosshaug, T., & Bahr, R. (2012). Clinical basis: epidemology, risk factors, mechanisms of injury, and prevention of ligament injuries of the knee. In M. Bonnin, N. A. Amendola, J. Bellemans, S. J. MacDonald & J. Menetrey (Eds.), *The Knee Joint: Sugical Techniques and Strategies*. Paris, France: Springer-Verlag.

Hoffman, R. R., Ward, P., Feltovich, P. J., DiBello, L., Fiore, S. M., & Andrews, D. H. (2013). *Accelerated Expertise: Training for High Proficiency in a Complex World*: Guilford Press.

Hulin, C. L., Henry, R. A., & Noon, S. L. (1990). Adding a dimension: Time as a factor in the generalizability of predictive relationships In T. R. Carretta (Ed.), *AFHRL Technical Paper* (pp. 89-67). Brooks Air Force Base: Manpower and Personnel Division.

Irich, E. (2012). NCAA Sports Sponsorship and Participation Rates Report 1981-82 - 2011-12.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences, 105*(19), 6829-6833.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short-and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences, 108*(25), 10081-10086.

Jensen, G. M., Guyer, J., Shepard, K. F., & Hack, L. M. (2000). Expert Practice in Physical Therapy. *Physical Therapy, 80*, 28-43.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *American Psychologist, 64*(6), 515-526.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: a meta-analysis of lens model studies. *Psychological Bulletin, 134*(3), 404-426.

Kim, S., Bosque, J., Meechan, J. P., Jamali, A., & Marder, R. (2011). Increase in outpatient knee arthroscopy in the united states: a comparison of national surveys of ambulatory surgery, 1996 and 2006. *Journal of Bone and Joint Surgery, 93*, 994-1000.

Klein, G. (1997). *The recognition-primed decision (RPD) model: Looking back, looking forward*: Lawrence Erlbaum Associates, Inc.

Knudson, D. V. (2000). What can professionals qualitatively analyze? *Journal of Physical Education, Recreation and Dance, 71*(2), 19-23.

Knudson, D. V. (2013). *Qualitative diagnosis of human movement: improving performance in sport and exercise*: Human Kinetics.

Knudson, D. V., & Morrison, C. (2000). Visual ratings of the vertical jump are weakly correlated with perceptual style. *Journal of Human Movement Studies, 39*, 33-44.

Krosshaug, T., Nakamae, A., Boden, B., Engebretsen, L., Smith, G., Slauterbeck, J., . . . Bahr, R. (2007a). Estimating 3D joint kinematics from video sequences of running and cutting maneuvers-assessing the accuracy of simple visual inspection. *Gait & Posture, 26*(3), 378-385.

Krosshaug, T., Nakamae, A., Boden, B. P., Engebretsen, L., Smith, G., Slauterbeck, J. R., . . . Bahr, R. (2007b). Mechanisms of anterior cruciate ligament injury in basketball: video analysis of 39 cases. *American Journal of Sports Medicine, 35*(3), 359-367.

Kundel, H. L., Nodine, C. F., Conant, E. F., & Weinstein, S. P. (2007). Hollistic component of image perception in mammogram interpretation: gaze-tracking study. *Radiology, 242*(2), 396-402.

Leas, R. R., & Chi, M. T. (1993). Analyzing diagnostic expertise of competitive swimming coaches. In J. L. Starkes & F. Allard (Eds.), *Cognitive Issues in Motor Expertise*: Elsevier Science.

Lohmander, S., Englund, M., Dahl, L., & Roos, E. (2007). The long-term consequence of anterior cruciate ligament and meniscus injuries: osteoarthritis. *American Journal of Sports Medicine, 35*(10), 1756-1769.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*: Routledge.

Loveday, T., Wiggins, M. W., Harris, J. M., O'Hare, D., & Smith, N. (2013). An objective approach to identifying diagnostic expertise among power system controllers. *Human Factors, 55*(1), 90-107.

Loveday, T., Wiggins, M. W., & Searle, B. J. (2013). Cue Utilization and Broad Indicators of Workplace Expertise. *Journal of Cognitive Engineering and Decision Making, 8*(1), 98-113.

Loveday, T., Wiggins, M. W., Searle, B. J., Festa, M., & Schell, D. (2012). The Capability of Static and Dynamic Features to Distinguish Competent From Genuinely Expert Practitioners in Pediatric Diagnosis. *Human Factors, 55*(1), 125-137.

Mann, D. T. Y., Williams, A. M., Ward, P., & Janelle, C. M. (2007). Perceptual-cognitive expertise in sport: A meta analysis. *Journal of Sport and Exercise Psychology, 29*, 457-478.

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review, 118*(3), 393-437.

Marras, W. S., Lavender, S. A., Leurgans, S. E., Fathallah, F. A., Ferguson, S. A., Gary Allread, W., & Rajulu, S. L. (1995). Biomechanical risk factors for occupationally related low back disorders. *Ergonomics, 38*(2), 377-410.

McNamara, D. S., & Scott, J. L. (2001). Working memory capacity and strategy use. *Memory & Cognition, 29*(1), 10-17.

Messick, S. (1990). Standards of validity and the validity of standards in performance asessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Mitchum, A. L., & Kelley, C. M. (2010). Solve the problem first: constructive solution strategies can influence the accuracy of retrospective confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 699-710.

Mizner, R. L., Chmielewski, T. L., Toepke, J. J., & Tofte, K. B. (2012). Comparison of 2-dimensional measurement techniques for predicting knee angle and moment during a drop vertical jump. *Clinical Journal of Sport Medicine, 22*, 221-227.

Morrison, C., & Frederick, C. M. (1998). Relationship of initial and final scores on a qualitative analysis of movement test. *Perceptual and Motor Skills, 87*, 651-655.

Moxley, J. H., Ericsson, K. A., Charness, N., & Krampe, R. T. (2012). The role of intuition and deliberative thinking in experts' superior tactical decision-making. *Cognition, 124*(1), 72-78.

Myer, G. D., Ford, K. R., Brent, J. L., & Hewett, T. E. (2007). Differential neuromuscular training effects on ACL injury risk factors in"high-risk" versus "low-risk" athletes. *BMC Musculoskeletal Disordorders, 8*, 39-46.

Myer, G. D., Ford, K. R., & Hewett, T. E. (2004). Rationale and Clinical Techniques for Anterior Cruciate Ligament Injury Prevention Among Female Athletes. *Journal of Athletic Training, 39*(4), 352-364.

Myer, G. D., Ford, K. R., & Hewett, T. E. (2011). New method to identify athletes at high risk of ACL injury using clinic-based measurements and freeware computer analysis. *British Journal of Sports Medicine, 45*(4), 238-244.

Myer, G. D., Ford, K. R., Khoury, J., & Hewett, T. E. (2011). Three-dimensional motion analysis validation of a clinic-based nomogram designed to identify high ACL injury risk in female athletes. *The Physician and Sportsmedicine, 39*(1), 19-28.

Myer, G. D., Ford, K. R., Khoury, J., Succop, P., & Hewett, T. E. (2010a). Clinical correlates to laboratory measures for use in non-contact anterior cruciate ligament injury risk prediction algorithm. *Clinical Biomechanics, 25*(7), 693-699.

Myer, G. D., Ford, K. R., Khoury, J., Succop, P., & Hewett, T. E. (2010b). Development and validation of a clinic-based prediction tool to identify female athletes at high risk for anterior cruciate ligament injury. *Amercan Journal of Sports Medicine, 38*(10), 2025-2033.

Myer, G. D., Ford, K. R., Khoury, J., Succop, P., & Hewett, T. E. (2011). Biomechanics laboratory-based prediction algorithm to identify female athletes with high knee loads that increase risk of ACL injury. *British Journal of Sports Medicine, 45*(4), 245-252.

Nandagopal, K., Roring, R. W., Ericsson, K. A., & Taylor, J. (2010). Strategies may mediate heritable aspects of memory performance: a twin study. *Cognitive and Behavioral Neurology, 23*(4), 224-230.

Newcombe, N. S., & Shipley, T. F. (2012). Thinking about spatial thinking: New typology, new assessments *Studying visual and spatial reasoning for design creativity*. New York, NY: Springer.

NFSHSA. (2012). 2011-12 High School Athletics Participation Survey.

Nilstad, A., Anderson, T. E., Kristianslund, E., Bahr, R., Myklebust, G., Steffen, K., & Krosshaug, T. (2014). Physiotherapists an identify female football players with high knee valgus angles during vertical drop jumps using real-time observational screening. *Journal of Orthopaedic and Sports Physical Therapy, 44*(5), 358-365.

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. *American Psychologist, 67*(2), 130-159.

Noyes, F. R. (2005). The Drop-jump screening test: Difference in lower limb control by gender and effect of neuromuscular training in female athletes. *American Journal of Sports Medicine, 33*(2), 197-207.

Padua, D. A., Marshall, S. W., Boling, M. C., Thigpen, C. A., Garrett, W. E., Jr., & Beutler, A. I. (2009). The Landing Error Scoring System (LESS) Is a valid and reliable clinical assessment tool of jump-landing biomechanics: The JUMP-ACL study. *American Journal of Sports Medicine, 37*(10), 1996-2002.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making, 5*(5), 411-419.

Paterno, M. V., Schmitt, L. C., Ford, K. R., Rauh, M. J., Myer, G. D., Huang, B., & Hewett, T. E. (2010). Biomechanical measures during landing and postural stability predict second anterior cruciate ligament injury after anterior cruciate ligament reconstruction and return to sport. *Amercan Journal of Sports Medicine, 38*(10), 1968-1978.

Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse mental rotations test-different versions and factors that affect performance. *Brain and Cognition, 28*(1), 39-58.

Pinheiro, V. E. D., & Simon, H. A. (1992). An operational model of motor skill diagnosis. *Journal of Teaching in Physical Education, 11*, 288-302.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*(3), 472-491.

Plessner, H., & Haar, T. (2006). Sports performance judgments from a social cognitive perspective. *Psychology of Sport and Exercise, 7*(6), 555-575.

Plessner, H., & Schallies, E. (2005). Judging the cross on rings: a matter of achieving shape constancy. *Applied Cognitive Psychology, 19*(9), 1145-1156.

Ployhart, R., Schneider, B., & Schmitt, N. (2006). Staffing organizations: Contemporary practice and research: Mahwah, NJ: Lawrence Erlbaum Associates.

Pohl, M. B., Mullineaux, D. R., Milner, C. E., Hamill, J., & Davis, I. S. (2008). Biomechanical predictors of retrospective tibial stress fractures in runners. *Journal of Biomechanics, 41*(6), 1160-1165.

Prodromos, C. C., Han, Y., Rogowski, J., Joyce, B., & Shi, K. (2007). A meta-analysis of the incidence of anterior cruciate ligament tears as a function of gender, sport, and a knee injury-reduction regimen. *Arthroscopy, 23*(12), 1320-1325.

Roth, P. L., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology, 58*, 1009-1037.

Rowson, S., Duma, S. M., Beckwith, J. G., Chu, J. J., Greenwald, R. M., Crisco, J. J., . . . Maerlender, A. C. (2012). Rotational head kinematics in football impacts: an injury risk function for concussion. *Annals of Biomedical Engineering, 40*(1), 1-13.

Schmidt, F. L. (2011). A Theory of Sex Differences in Technical Aptitude and Some Supporting Evidence. *Perspectives on Psychological Science, 6*(6), 560-573.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: occupational attainment and job performance. *Journal of Personality and Social Psychology, 86*(1), 162-173.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of reseach findings. *Psychological Bulletin, 124*(2), 262-274.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71*(3), 432-439.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407-422.

Shanteau, J. (1992). How much information does an expert use? Is it relevant? *Acta Psychologica, 81*, 75-86.

Simon, D., Krawczyk, D. C., & Holyoak, K. J. (2004). Construction of preferences by constraint satisfaction. *Psychological Science 15*(5), 331-336.

Simon, D., Snow, C. J., & Read, S. J. (2004). The redux of cognitive consistency theories: evidence judgments by constraint satisfaction. *Journal of Personality and Social Psychology, 86*(6), 814-837.

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41*(1), 1-20.

Smith, H. C., Vacek, P., Johnson, R. J., Slauterbeck, J. R., Hashemi, J., Shultz, S., & Beynnon, B. D. (2012a). Risk factors for anterior cruciate ligament injury a review of the literature—Part 2: Hormonal, genetic, cognitive function, previous injury, and extrinsic risk factors. *Sports Health, 4*(2), 155-161.

Smith, H. C., Vacek, P., Johnson, R. J., Slauterbeck, J. R., Hashemi, J., Shultz, S., & Beynnon, B. D. (2012b). Risk factors for anterior cruciate ligament injury: a review of the literature - part 1: Neuromuscular and anatomic risk. *Sports Health, 4*(1), 69-78.

Spiro, R. J., Feltovich, P. J., Jacobson, M., & Coulson, R. L. (1991). Cognitive flexibility, constructivism, and hypertext: advanced knowledge acquisition in ill-structured domains. *Educational Technology, 31*, 24-33.

Stanovich, K. E. (2012). On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. *The Oxford Handbook of Thinking and Reasoning*, 343-365.

Ste-Marie, D. M. (1999). Expert-novice difference in gymnastic judging: an information-processing perspective. *Applied Cognitive Psychology, 13*, 269-281.

Ste-Marie, D. M., & Lee, T. D. (1991). Prior processing effects on gymnastic judging. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(1), 126-136.

Stensrud, S., Myklebust, G., Kristianslund, E., Bahr, R., & Krosshaug, T. (2010). Correlation between two-dimensional video analysis and subjective assessment in evaluating knee control among elite female team handball players. *British Journal of Sports Medicine, 45*(7), 589-595.

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*: Lawrence Erlbaum.

Stuberg, W., Straw, L., & Devine, L. (1990). Validity of visually recorded temporal-distance measures at selected walking velocities for gait analysis. *Perceptual and Motor Skills, 70*(1), 323-333.

Sugimoto, D., Myer, G. D., McKeon, J. M., & Hewett, T. E. (2012). Evaluation of the effectiveness of neuromuscular training to reduce anterior cruciate ligament injury in female athletes: a critical review of relative risk reduction and numbers-needed-to-treat analyses. *British Journal of Sports Medicine, 46*(14), 979-988.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295-312.

Swenson, D. M., Collins, C. L., Best, T. M., Flanigan, D. C., Fields, S. K., & Comstock, R. D. (2013). Epidemiology of Knee Injuries Among US High School Athletes, 2005/06-2010/11. *Medicine and Science in Sports and Exercise, 45*(3), 462-469.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: a meta-analysis of training studies. *Psychological Bulletin, 139*(2), 352-402.

van der Maas, H. L., & Wagenmakers, E.-J. (2005). A psychometric analysis of chess expertise. *The American Journal of Psychology*, 29-60.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills, 47*, 599-604.

Vigneau, F., Caissie, A. F., & Bors, D. A. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence, 34*(3), 261-272.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*(5), 557-574.

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology, 101*(4), 817-835.

Ward, P., Ericsson, K. A., & Williams, A. M. (2012). Complex perceptual-cognitive expertise in a simulated task environment. *Journal of Cognitive Engineering and Decision Making, 7*(3), 231-243.

Ward, P., Suss, J., & Basevitch, I. (2009). Expertise and Expert Performance-based Training (ExPerT) in complex domains. *Technology, Instruction, Cognition & Learning, 7*(2), 121-145.

Whatman, C., Hing, W., & Hume, P. (2012). Physiotherapist agreement when visually rating movement quality during lower extremity functional screening tests. *Physical Therapy in Sport, 13*(2), 87-96.

Whatman, C., Hume, P., & Hing, W. (2013a). Kinematics during lower extremity functional screening tests in young athletes - are they reliable and valid? *Physical Therapy in Sport, 14*(2), 87-93.

Whatman, C., Hume, P., & Hing, W. (2013b). The reliability and validity of physiotherapist visual rating of dynamic pelvis and knee alignment in young athletes. *Physical Therapy in Sport, 14*(3), 168-174.

Wiggins, M., Brouwers, S., Davies, J., & Loveday, T. (2014). Trait-based cue Utilization and initial skill acquisition: implications for models of the progression to expertise. *Cognition, 5*(541), 1-8.

Wigton, R. S. (2008). What do the theories of Egon Brunswik have to say to medical education? *Advances in Health Sciences Education, 13*(1), 109-121.

Williams, G., Morris, M. E., Schache, A., & McCrory, P. (2009). Observational gait analysis in traumatic brain injury: accuracy of clinical judgment. *Gait & Posture, 29*(3), 454-459.

Wright, R. W., Dunn, W. R., Amendola, A., Andrish, J. T., Bergfeld, J., Kaeding, C. C., . . . Spindler, K. P. (2007). Risk of tearing the intact anterior cruciate ligament in the contralateral knee and rupturing the anterior cruciate ligament graft during the first 2 years after aneterior cruciate ligament reconstruction: a prospective MOON cohort study. *American Journal of Sports Medicine, 35*(7), 1131-1134.

APPENDIX A

11-Item ACL Knowledge Test

1.) What does ACL stand for?

- ☐ Anterior cruciate ligament
- ☐ Anterior collateral ligament
- ☐ Anatomical cruciate ligament
- ☐ Anterior condyle ligament
- ☐ Anatomical collateral ligament

2.) What joint is the ACL located within?

- ☐ Hip
- ☐ Ankle
- ☐ Pelvis
- ☐ Knee
- ☐ Shoulder

3.) ACL injuries are most common in:

- ☐ Weightlifting
- ☐ Soccer
- ☐ Running
- ☐ Volleyball
- ☐ Hockey

4.) Risk for ACL injury is _____ in men compared to women when participating in the same sport at the same competition level.

- ☐ Lower
- ☐ Higher
- ☐ Equal

5.) The ACL functions to prevent:

- ☐ A. Anterior tibial translation (shin sliding forward)
- ☐ B. Posterior tibial translation (shin sliding backward)
- ☐ C. Tibial rotation (shin rotating)
- ☐ D. Knee flexion (knee bending)
- ☐ All of the above
- ☐ A. and C. only
- ☐ A. and D. only

6.) The ACL attaches to the:

- ☐ Front of the tibia to the back of the femur
- ☐ Back of the tibia to the front of the femur
- ☐ Medial (inside) side of the tibia and femur
- ☐ Lateral (outside) side of the tibia and femur
- ☐ Front of the tibia to medial side of the femur

7.) Dynamic knee valgus is:

- ☐ Knee abduction (distal tibia or foot moving away from midline in a frontal view)
- ☐ Knee adduction (distal tibia or foot moving towards midline in a frontal view)
- ☐ Rotations of multiple lower body segments resulting in a "knock-knee" position (i.e., inward knee motion)
- ☐ Stiff landing with straight or extended knee
- ☐ Toes rotated outward

8.) What is a knee abduction moment:

- ☐ Torque generated rotating the knee outward
- ☐ Torque generated rotating the knee inward
- ☐ Angle between the tibia and midline
- ☐ Ground reaction force compressing the knee
- ☐ Perpendicular distance from the knee to the ground reaction force vector in the frontal plane

9.) What is the best training intervention to prevent ACL injuries?

- ☐ A. Plyometric (jump) training
- ☐ B. Resistance (strength) training
- ☐ C. Balance training
- ☐ D. Aerobic/endurance training
- ☐ All of the above
- ☐ A., B., and C. only

10.) What factors can influence peak knee abduction moment during a drop vertical jump task (check all that apply)?

- ☐ Tibia length
- ☐ Quadriceps (Q) Angle
- ☐ Medial knee motion
- ☐ Knee flexion angle
- ☐ Jump Height
- ☐ Weight

11.) In young females, low ACL injury risk drop vertical jump technique includes (check all that apply):

- ☐ Landing softly
- ☐ Landing stiff or hard
- ☐ Pointing toes outward
- ☐ Keeping knees in line with toes (avoiding a "knock-knee" position)
- ☐ Keeping trunk or upper body vertical (no flexion)

APPENDIX B

Rate the importance of the following cues for making your ACL injury risk rating.

	Not Important									Very Important
Arm Motion	O 1	O 2	O 3	O 4	O 5	O 6	O 7	O 8	O 9	O 10
Landing Symmetry	O 1	O 2	O 3	O 4	O 5	O 6	O 7	O 8	O 9	O 10
Inward/Outward Knee Motion	O 1	O 2	O 3	O 4	O 5	O 6	O 7	O 8	O 9	O 10
Inward/Outward Thigh Motion	O 1	O 2	O 3	O 4	O 5	O 6	O 7	O 8	O 9	O 10
Lateral Trunk Motion	O 1	O 2	O 3	O 4	O 5	O 6	O 7	O 8	O 9	O 10
Landing Stiffness	O 1	O 2	O 3	O 4	O 5	O 6	O 7	O 8	O 9	O 10
Foot Alignment	O 1	O 2	O 3	O 4	O 5	O 6	O 7	O 8	O 9	O 10
Height of Individual	O 1	O 2	O 3	O 4	O 5	O 6	O 7	O 8	O 9	O 10
Weight of Individual	O 1	O 2	O 3	O 4	O 5	O 6	O 7	O 8	O 9	O 10
Jump Height	O 1	O 2	O 3	O 4	O 5	O 6	O 7	O 8	O 9	O 10
Jump Alignment	O 1	O 2	O 3	O 4	O 5	O 6	O 7	O 8	O 9	O 10

Figure 14. Cue Utility Survey