Accuracy and Reliability of Peer Assessment of Clinical Skills and Professional Behaviors Among Undergraduate Athletic Training Students

Dissertation

Submitted to Northcentral University

Graduate Faculty of the School of Education
in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF EDUCATION

by

JEANINE M. ENGELMANN

Prescott Valley, Arizona
October 2014

Approval Page

Accuracy and Reliability of Peer Assessment of Clinical Skills and Professional Behaviors Among Undergraduate Athletic Training Students

By

Jeanine M. Engelmann

Approved by:

11/11/2014

Chair: Helen Zaikina-Montgomery, Ph.D. Date

Certified by:

Kelley Walters 11/13/2014

Dean of School: Kelley Walters, Ph.D. Date

Abstract

Peer assessment is used by health care professionals as a way to share knowledge and evaluate the performance of colleagues. Peer assessment is used widely in medical education as a preparatory tool for students, but peer assessment research in athletic training education is lacking. Athletic trainers are healthcare providers with a similar skill-set to physicians, thus, athletic training education can benefit from the use of peer assessment. Athletic training educators need to research the use of peer assessment as an evaluation tool in order to better prepare students to practice as healthcare professionals. This study investigated the accuracy and reliability of undergraduate athletic training students in their ability to assess their peers. This quasi-experimental study used between-group and within-group designs to answer the research questions. Junior-level students, senior-level students, and their instructors were enrolled as participants. Each student group's ratings of clinical skills and professional behaviors were compared to instructor ratings to measure accuracy, and each student group's ratings were compared for reliability. Cohen's kappa coefficient measured inter-rater agreement for all statistical analyses. Both groups of students were accurate raters ($p < .05$) of their peers on clinical skills, but only the senior-level students were accurate in rating professional behaviors. Both groups of students were reliable in rating their peers on about half of the clinical skills. The senior-level students were also reliable in evaluating professional behaviors, but the junior-level students were not. The data for this study showed high levels of observed agreement for most clinical skills, subscales and the professional behaviors, but some items had low Cohen's kappa values, most likely due to a known paradox that occurs with the kappa statistic. As the first study in athletic training education to use undergraduate students, live data collection, and rating of professional behaviors, the findings were promising for future research. Future research needs to include training in peer assessment, use of repeated measures, and comparison of instructor scores in order to better understand peer assessment in this population. Additionally, there is a need to establish consistent, quality measures in peer assessment research, including those used in athletic training education.

Acknowledgements

First and foremost, I would like to thank my mentor, Dr. Helen Zaikina-Montgomery for the unwavering support and frank, honest feedback she provided throughout this long, arduous process. Next, my family and friends who have showed their interest and provided support and encouragement during the past six and a half years leading up to this point. Finally, a huge “thank you” to the students, faculty and preceptors who took time out of their busy schedules to volunteer as participants in this study; without whom, I would not have been able to perform this research.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Medical and health care professionals use peer assessment as a valuable method to share knowledge among colleagues, evaluate the performance of fellow professionals in a manner in which the lay person is unable to participate, and foster professional growth (Evans, Elwyn, & Edwards, 2004; Hulsman, Peters, & Fabriek, 2013; Speyer, Walmari, Van Der Kruis, & Brunning, 2011). For both educational and evaluative purposes, peer assessment provides working professionals the opportunity to be judged by others of similar training and expertise, who share a language and commonly used knowledge and skills (Finn & Garner, 2011; Hulsman et al., 2013).

Peer assessment in higher education involves the use of learners of the same academic level in the process of determining the quality, worth, or level of successfulness of the outcomes or products of learning (Casey et al., 2011; Topping, 1998). Through the peer assessment process, professional growth is fostered in students prior to entering the workforce (Garner, McKendree, O'Sullivan, & Taylor, 2010) and may help students develop the necessary self-assessment skills to judge their own abilities when working as independent healthcare practitioners (Hulsman et al., 2013) . Peer assessment, as a formative evaluation tool, has been widely studied with beneficial results in medical education (Garner et al., 2010; Speyer et al., 2011), but only two currently published studies address the use of peer assessment in athletic training education (Marty, Henning, & Willse, 2010; Marty, Henning, & Willse, 2011).

Education programs for both the medical and athletic training professions require graduates to possess and demonstrate knowledge, skills, and affective traits suitable for practice in a modern health care setting (Commission on Accreditation of Athletic

Training Education [CAATE], 2012; Liaison Committee on Medical Education [LCME], 2012; National Athletic Trainers' Association [NATA], 2011). Common educational competencies among the groups are: (a) collaboration with other healthcare providers, (b) performance of physical examinations in order to establish diagnoses and develop treatment plans, (c) providing emergency care for life-threatening conditions, (d) preventing disease through health and wellness promotion, (e) demonstration of professional behaviors and ethical standards, and (f) communicating effectively with patients and families (Association of American Medical Colleges [AAMC], 1998; Accreditation Council for Graduate Medical Education [ACGME], 2013; CAATE, 2012; LCME, 2012; NATA, 2011). Due to the job similarities, students of both disciplines can see similar outcomes from peer assessment. Therefore, athletic training education can benefit from performing peer assessment research similar to the research performed in medical education.

This study sought to identify if undergraduate athletic training students were accurate and reliable assessors of their peers on clinical skills and professional behaviors. Investigation into peer assessment in the undergraduate athletic training population can determine peer assessment's use as a valuable evaluation tool in athletic training education, as its been established in medical education (Finn & Garner, 2011; Garner et al., 2010; Speyer et al., 2011). If athletic training students are accurate and reliable assessors of their peers, athletic training educators can incorporate peer assessment into the evaluation process prior to students entering the workforce.

Background

Traditionally, educational assessment has been used for summative measurement purposes (Falchikov, 2005; Iqbal & Mahmood, 2008), usually in the form of tests (Kubiszyn & Borich, 2007); and has included objective items such as multiple choice and true-false questions, along with higher-order test items such as essay questions. There are multiple negative outcomes for evaluating students through traditional assessment methods. First, these types of evaluations promote extrinsic rewards for students that focus on the product, not the process, of learning (Topping, 2009). Students do not learn the value of intrinsic motivation, which promotes responsibility, autonomy and ownership in the learning process (Falchikov, 2005; Topping, 2009). Second, traditional assessment methods are not reflective of the social and interactive nature of the learning process (Hodges, 2011). This is problematic because interaction and discussion among students helps them learn to justify their position and give and accept criticism and suggestions, all transferrable life skills (Topping, 2009). Students become passive consumers of their education. Passive learning and assessment processes limit students' abilities to build new knowledge on their own, without the guidance of an instructor (Hodges, 2011). Students are also limited in developing skills for lifelong learning through traditional examination formats because these designs do not promote independent critical thinking or creativity; two skills important to creating personal learning opportunities (Elton & Johnson, 2002; Hodges, 2011).

In addition to the negative outcomes associated with traditional assessment methods, limitations exist with use of these assessment tools (Falchikov, 2005). Objectivity and standardization receive priority over student and teacher collaboration.

Students typically have little or no input into the assessment process through which their academic standing is determined. Having little or no control over the assessment process does not provide students with the tools needed to develop such skills as problem solving and reflection for lifelong independent learning (Hodges, 2011; Topping, 2009; Welsh, 2007).

Student-centered activities, such as peer assessment, are proposed as alternatives to traditional assessment methods. Peer assessment, for example, provides opportunities for students to gain autonomy (Casey et al., 2011; Iqbal & Mahmood, 2008) and independence, and increase their sense of responsibility and self-efficacy (Casey et al., 2011; Tillema, Leenknecht, & Segers, 2011). The promotion of autonomy among students increases ownership of the assessment process and can improve motivation (Henning & Marty, 2008; Hulsman et al., 2013; Tiew, 2010; Welsh, 2007).

Marty et al. (2010) asserted the use of peer assessment in athletic training education could enhance the understanding and performance of psychomotor skills. Improved levels of autonomy, responsibility, self-efficacy, and independence can all help athletic training students meet the educational standard that requires they be able to integrate professional knowledge and skills, including decision-making and professional behaviors (CAATE, 2012).

Because they are on the same level as their peers, students often perceive peer feedback as more understandable and useful than instructor feedback (Ploegh, Tillema, & Segers, 2009; Tillema et al., 2011; Topping, 1998). The use of peer assessment allows for immediate one-on-one feedback that is often not available with instructor feedback (Gielen, Dochy, Onghena, Struyven, & Smeets, 2011; Iqbal & Mahmood, 2008; Topping, 1998). Additionally, the inherent collaboration involved in student participation allows students to improve their social skills, co-operation and diplomacy (Casey et al., 2011; Topping, 1998; Welsh, 2007; Yurdabakan, 2011). Nofziger, Naumburg, Davis, Mooney, and Epstein (2010) supported this finding in their investigation of professional development of medical students. Subjects reported improved interpersonal skills and professional growth after engaging in peer assessment practices. These findings were consistent with the previously stated need for athletic training students to develop professional behaviors, including communication skills with patients and other healthcare providers (NATA, 2011).

The movement to involve students in assessment dates to the 1950s with seminal studies continuing through the 1960s, including research on general peer assessment (Kubany, 1957), methodological issues with peer assessment (Falchikov, 2005), and problems with traditional assessment (Falchikov, 2005). Research from the 1970s produced more total investigations and shifted to the benefits of student involvement in the assessment process (Friesen & Dunning, 1973). Pressures in the traditional educational context, such as increased staff time and energy on assessment; increased amount and types of mandatory learning objectives, and greatly differing levels in student ability, spurred the increased interest in alternative assessment methods (Falchikov, 2005).

The benefits to student involvement in assessment in professional education studies began to emerge in the 1980s (Falchikov, 2005). Along with this shift in the peer assessment research, came the recognition that communication skills were very important in many professions (Earl, 1986). Studies comparing self- and peer assessment marks also started to be reported during the 1980s, adding to the amount of literature that investigated reliability and validity of student marks (Moreland et al., 1981). Research about student involvement in the assessment process grew during the 1990s; more studies were conducted than all the preceding decades combined (Falchikov, 2005). Two main themes emerged out of the literature: the benefits of involving students in the assessment process (Edwards & Sutton, 1991; Mathews, 1994) and pressure on the part of teachers in higher education (Dochy & McDowell, 1997; Young, 1999).

There are some common concerns about peer assessment, especially from students. Some authors reported students believed peer assessment to be a stress-inducing activity (Garner et al., 2010; Kaufman & Schunn, 2011; Tiew, 2010). Some students have also reported feeling unprepared to assess their peers (Garner et al., 2010) and perceive assessment to be the responsibility of the instructor (Kaufman & Schunn, 2011; Topping, 1998; Vickerman, 2009). Others have demonstrated concern about the fairness of the peer assessment process and the quality and accuracy of the feedback provided by fellow students (Harris, 2011; Kaufman & Schunn, 2011; Vickerman, 2009). Topping (2009) reported that peer assessment outcomes might be influenced by friendships, enmity, popularity, or collusion among students to submit similar scores. Maintaining anonymity in the assessment process and disallowing friendships to bias results were also reported as concerns by students (Casey et al., 2011; Garner et al., 2010; Topping, 1998; Vickerman, 2009).

In response to many of these concerns, the importance of training and practice with peer assessment among students prior to implementation has been highlighted by many authors in the literature (Falchikov & Goldfinch, 2000; Garner et al., 2010; Marty et al., 2011; van Zundert, Sluijsmans, & van Merriënboer, 2010; Topping, 2010; Vickerman, 2009). Specifically, studies have cited the importance of training students to give and receive quality feedback as an important aspect to the learning process for students engaged in peer assessment activities (Garner et al., 2010; Henning & Marty, 2008; Marty et al., 2011; Topping, 2009).

Peer assessment is the most commonly used method of incorporating students in the assessment process, and is an established valid and reliable tool in both education (Falchikov, 2005; Topping, 1998) and professional practice for medical and healthcare professionals (Finn & Garner, 2011; Speyer et al., 2011; Topping, 1998).

However, outside of medical education, there is a need to produce quality research, including validated measurement instruments, into peer assessment in other healthcare education fields (Speyer et al., 2011). Peer assessment in athletic training education has thus far produced two published studies (Marty et al., 2010; Marty et al., 2011). Both studies used videotaped clinical skill performances for graduate students to evaluate their peers, and neither investigated professional behaviors demonstrated during the clinical skills. While these recent studies provide early information, further investigation regarding how well undergraduate athletic training students evaluate peers compared to instructors during live performance of clinical skills and professional behaviors is needed to determine if peer assessment can be used similarly to its use in medical education. Additionally, research in this population will highlight where additional research and training is necessary to most effectively implement peer assessment within athletic training education.

Statement of the Problem

Participation in peer assessment prior to entering the workforce fosters professional growth (Garner et al., 2010), and has been supported in the medical education literature as a method to prepare students to practice as fully competent professionals within the larger healthcare setting alongside other health professionals (Finn & Garner, 2011; Garner et al., 2010; Speyer et al., 2011). Athletic training and medical education share many skills and traits that are needed by healthcare professionals to provide quality patient care and engage with other healthcare providers, including: (a) performing physical examinations, (b) providing immediate care, (c) preventing disease, (d) communicating effectively with patients and families, (e) collaborating with other healthcare professionals, and (f) portraying professionalism and ethics at all times (CAATE, 2012; LCME, 2012: NATA, 2011).

Initial studies of peer assessment in athletic training education (Marty et al., 2010; Marty et al., 2011) have shown early evidence that athletic training students are valid evaluators of their peers during videotaped clinical skills demonstrations. However, there is a need to examine the validity and reliability of peer assessment using live participation and to compare students' ratings of clinical skills and professional behaviors to those of instructors in the undergraduate athletic training student population (Marty et al., 2010). This type of peer assessment will allow athletic training educators to better identify the quality of peer assessment among pre-professional athletic training students in order to implement peer assessment practices most effectively prior to entering the workforce. Athletic training educators need to research the use of peer assessment as an evaluation tool in order to better prepare students to practice as healthcare professionals, or risk placing students at a clinical disadvantage to their counterparts in the medical field upon graduation.

Purpose of the Study

The purpose of this quantitative study was to investigate the accuracy and reliability of undergraduate athletic training students to assess their peers on clinical skills and professional behaviors. The results of the proposed study further investigated peer assessment as an effective assessment tool for use in athletic training education. In comparison to instructors, if athletic training students accurately assess the clinical skills and professional behaviors of their peers, this initial investigation can be developed into larger, more complex studies.

This was a quasi-experimental study with three, relatively small (n≤11) groups; non-randomly assigned as instructors, senior-level students, and junior-level students. For the purpose of this study, the definition of accuracy was the amount of agreement between student (peer) scores and instructor scores. Peer and instructors scores have been compared through percentage agreement or mean values in multiple studies in order to determine the accuracy or level of agreement between scores (Bucknall et al., 2008; Chenot et al., 2007; Evans, Leeson, & Petrie, 2007; Marty et al., 2010; Marty et al., 2011). Level of accuracy is how closely students score their peers in relation to the scores of the instructors, not how well the evaluated skills are performed. For the clinical skills, Cohen's kappa coefficient was used to measure inter-rater agreement between instructor and student scores for the accuracy measure and among each student group for the reliability measure. Cohen's kappa is the most widely used measure of inter-rater reliability for dichotomously scored data (Howell, 2002; Warner, 2008; von Eye & von

Eye, 2008). For the professional behaviors, a weighted Cohen's kappa coefficient was used to measure inter-rater agreement between instructor and student scores for the accuracy measure and among each student group for the reliability measure.

Participants were recruited from a sample of current students and instructors affiliated with an accredited undergraduate athletic training program at a large university in the mid-Atlantic region of the United States. The measurement tool used for data collection was adapted, with permission, from an athletic training textbook designed for use of clinical skills documentation of athletic training students (Amato, Hawkins, & Cole, 2006). The measurement tool used a simple dichotomous nominal Yes/No scale for assessment of clinical skills and a 5-point Likert-type (5= Always, 1= Never) scale for global ratings of professional behaviors.

Research Questions

The current sought to determine the accuracy and reliability of undergraduate athletic training students to assess their peers on clinical skills and professional behaviors. Student scores were compared to instructor scores for inter-rater agreement to determine accuracy, while within-group inter-rater agreement scores were used to determine reliability among the students. Eight research questions were answered through the analysis of data in this study.

Q1. In relation to instructor scores, how accurate are junior-level students in scoring the clinical skills performance of undergraduate athletic training students for the a) Biceps Femoris Manual Muscle Test, b) Kleiger Test, c) Lachman Test, d) Noble's Compression Test, and e) Thompson's Test?

Q2. In relation to instructor scores, how accurate are senior-level students in scoring the clinical skills performance of undergraduate athletic training students for the a) Biceps Femoris Manual Muscle Test, b) Kleiger Test, c) Lachman Test, d) Noble's Compression Test, and e) Thompson's Test?

Q3. How reliable are junior-level students to each other in their ability to evaluate the clinical skills performance of undergraduate athletic training students for the a) Biceps Femoris Manual Muscle Test, b) Kleiger Test, c) Lachman Test, d) Noble's Compression Test, and e) Thompson's Test?

Q4. How reliable are senior-level students to each other in their ability to evaluate the clinical skills performance of undergraduate athletic training students for the a) Biceps Femoris Manual Muscle Test, b) Kleiger Test, c) Lachman Test, d) Noble's Compression Test, and e) Thompson's Test?

Q5. In relation to instructors, how accurate are junior-level students in scoring professional behaviors during clinical skills performance of undergraduate athletic training students?

Q6. In relation to instructors, how accurate are senior-level students in scoring professional behaviors during clinical skills performance of undergraduate athletic training students?

Q7. How reliable are junior-level students to each other in their ability to evaluate professional behaviors during clinical skill performance of undergraduate athletic training students?

Q8. How reliable are senior-level students to each other in their ability to evaluate professional behaviors during clinical skill performance of undergraduate athletic training students?

Hypotheses

The hypotheses below reflect the research questions in the following way: hypotheses one and three addressed Q1; hypotheses two and four addressed and Q2; hypotheses five and seven addressed Q3; hypotheses six and eight addressed Q4; hypothesis nine addressed Q5; hypothesis ten addressed Q6; hypothesis 11 addressed Q7; and hypothesis 12 addressed Q8.

$H1_0$: There is no statistically significant agreement between instructor and junior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble’s Compression Test, Thompson Test) performance ratings.

$H1_a$: There is statistically significant agreement between instructor and junior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble’s Compression Test, Thompson Test) performance ratings.

$H2_0$: There is no statistically significant agreement between instructor and senior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble’s Compression Test, Thompson Test) performance ratings.

$H2_a$: There is statistically significant agreement between instructor and senior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger

Test, Lachman Test, Noble's Compression Test, Thompson Test) performance ratings.

$H3_0$: There is no statistically significant agreement between instructor and junior-level student scores among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

$H3_a$: There is statistically significant agreement between instructor and junior-level student scores among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

$H4_0$: There is no statistically significant agreement between instructor and senior-level student scores among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

$H4_a$: There is statistically significant agreement between instructor and senior-level student scores among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

$H5_0$: There is no statistically significant agreement between junior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble's Compression Test, Thompson Test) performance ratings.

$H5_a$: There is statistically significant agreement between junior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble's Compression Test, Thompson Test) performance ratings.

$H6_0$: There is no statistically significant agreement between senior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble's Compression Test, Thompson Test) performance ratings.

H6$_a$: There is statistically significant agreement between senior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble’s Compression Test, Thompson Test) performance ratings.

H7$_0$: There is no statistically significant agreement between junior-level students among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

H7$_a$: There is statistically significant agreement between junior-level students among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

H8$_0$: There is no statistically significant agreement between senior-level students among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

H8$_a$: There is statistically significant agreement between senior-level students among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

H9$_0$: There is no statistically significant agreement between instructor and junior-level students in professional behaviors ratings.

H9$_a$: There is statistically significant agreement between instructor and junior-level students in professional behaviors ratings.

H10$_0$: There is no statistically significant agreement between instructor and senior-level students in professional behaviors ratings.

H10$_a$: There is statistically significant agreement between instructor and senior-level students in professional behaviors ratings.

H11$_0$: There is statistically significant agreement between junior-level students in professional behaviors ratings.

H11$_a$: There is no statistically significant agreement between junior-level students in professional behaviors ratings.

H12$_0$: There is statistically significant agreement between senior-level students in professional behaviors ratings.

H12$_a$: There is statistically significant agreement between senior-level students in professional behaviors ratings.

Nature of the Study

This quantitative study of peer assessment among undergraduate athletic training students used a quasi-experimental method. Both between-groups and within-groups designs were used for data analysis. The independent variables were the groups of junior-level students, senior-level students, and instructors for the accuracy measures; and the groups of junior-level students and senior level-students for the reliability measures. Level of accuracy of each student group (junior-level students, senior-level students) was determined through a between-groups design. Similarity in scores between each student group and the instructor group were computed to determine the level of accuracy of the students' scores. The dependent variables for the accuracy measures of clinical skills were the individual components for each of the five clinical skills; and the subscales of patient position, clinician position, and test performance across all clinical skills. The dependent variable for the accuracy measures of professional behaviors was the individual behaviors.

A within-group design addressed the research questions concerned with reliability. Each treatment consisted of two students from either the senior-level student group or the junior-level student group. This allowed the author to compare the student participant scores in order to determine inter-rater reliability (Warner, 2008). The dependent variables were the same for the reliability measures as for the accuracy measures.

The measurement tool used for data collection was adapted, with permission, from an athletic training textbook designed for clinical skills documentation of athletic training students (Amato et al., 2006). A field test was performed on the data collection instrument prior to use in the full study. Participants in the field test included athletic training educators, clinical athletic trainers, and undergraduate athletic training students who were not be part of the full study.

The data collection instrument (Appendix A) contained five clinical skills and eight professional behaviors. The instrument used a simple 2- point nominal scale for assessment of clinical skills and a 5-point Likert scale for global rating of professional behaviors. The 2- point nominal scale (Yes/No) determined completion of the individual tasks needed for each clinical skill, as described on the data collection instrument. The 5-point Likert scale (5 = Always, a score of 4 = Frequently, a score of 3 = Occasionally, a score of 2 = Rarely, and a score of 1 = Never) measured the frequency with which the participants observed the model clinician’s professional behaviors during the performance of the five clinical skills. The professional behaviors were assessed at the conclusion of the performance of the five clinical skills and were used as a global rating.

Differences in scores determined both accuracy and reliability of undergraduate athletic training students in evaluation of clinical skills and professional behaviors. Cohen's kappa coefficient was used to measure inter-rater agreement between instructors and students for the students' accuracy of scoring clinical skills. The same test measured the reliability of each student group to score clinical skills.

A weighted Cohen's kappa coefficient was used to measure inter-rater agreement between instructors and students for the students' accuracy of scoring professional behaviors. The same test measured the reliability of each student group to score professional behaviors.

Significance of the Study

The current study in peer assessment contributed to the literature in a number of ways. This was the first study to use undergraduate athletic training students as participants in formally investigated peer assessment practices. As of May 2014, there were 365 accredited athletic training education programs in the United States (CAATE, 2013b). Of the 365 programs, 338 were undergraduate programs. Since almost 93% of the education programs were undergraduate programs, peer assessment research using this student population can contribute greatly to the literature in an effort to determine its value in athletic training education.

Second, data collection occurred concurrently with the student and instructor participants evaluating clinical skills and professionalism traits at the same time during live skills demonstrations. The previous athletic training education studies performed by Marty et al. (2010; 2011) used videotaped presentations shown to individual participants for data collection, and did not use instructor ratings for comparison to student scores.

The inclusion of live skills demonstrations was significant because it reflects real world clinical practice where students are regularly judged by instructors and peers. The concurrent data collection from peers and instructors allowed for better comparison measures between groups. With live action skills demonstrations, both groups needed to be present during data collection to ensure each participant observed the same presentation.

Third, professionalism traits were assessed for the first time as part of a peer assessment study in athletic training education. CAATE requires athletic training students to demonstrate sound clinical skills and professional behaviors before graduating from an accredited athletic training program (CAATE, 2012). With this in mind, the incorporation of peer assessment of professional behaviors during clinical skills demonstration provided the opportunity to evaluate both components of professional practice athletic training students need in order to function well as healthcare providers after graduation.

Finally, this study used a field tested data collection instrument with athletic training educators and students as participants. There were currently no validated measurement instruments in the field of athletic training education for use in the current study. The data collection instrument was adapted, with permission, from a textbook of commonly taught clinical skills in athletic training education. Lack of validity testing on peer assessment instruments, particularly in healthcare education, is an ongoing concern in the literature (Speyer et al., 2011; Topping, 1998). The data collection instrument was field tested prior to use in the full study, whereas in most peer assessment research in healthcare education, the data collection instruments were designed by the researchers and not tested at all prior to use in published studies (Speyer et al., 2011).

Definition of Key Terms

Accuracy. The amount of agreement between student (peer) scores and instructor scores. Instructor scores are the “expert” mark. Peer and instructor score comparisons determine the amount of correlation between the participants. As the correlation between student and instructor scores improves, the student scores are more accurate. This definition is the work of the author.

Athletic training. A health care profession whose members collaborate with physicians to optimize activity and participation of patients. Athletic training comprises the prevention, diagnosis, and intervention of emergency, acute, and chronic medical conditions involving impairment, functional limitations and disabilities. (National Athletic Trainers’ Association, n.d.).

Athletic Training Education Competencies. Requirements of knowledge, skills, and clinical abilities to be mastered by students enrolled in professional athletic training education programs. Mastery of these Competencies provides the entry-level athletic trainer with the capacity to provide athletic training services to clients and patients. (National Athletic Trainers’ Association, 2011).

Board of Certification (BOC). Certification program for entry-level athletic trainers. The BOC establishes and regularly reviews both the standards for the practice of athletic training and the continuing education requirements for BOC Certified Athletic Trainers. The BOC has the only accredited certification program for athletic trainers in the US. (Board of Certification, Inc., 2013).

Clinical skills. The hands-on techniques required of athletic trainers to perform tasks within their respective work setting. Such techniques include, but are not limited to

(a) bony and soft tissue palpations, (b) functional movement assessments, (c) ligamentous and special tests, and (d) neurological assessments. This definition is the work of the author.

Commission on Accreditation of Athletic Training Education (CAATE). Provides accreditation services to colleges and universities that offer athletic training programs, verifies that all CAATE accredited programs meet Standards for professional athletic training education, and supports continuous improvement in the quality of athletic training education. (Commission on Accreditation of Athletic Training Education, 2013a).

Foundational Behaviors of Professional Practice. Basic behaviors that permeate professional practice. The NATA states these behaviors be infused into instruction and assessment throughout the athletic training education program. (National Athletic Trainers' Association, 2011).

Instructors. Participant group that includes the classroom faculty and clinical preceptors responsible for evaluating the undergraduate athletic training student participants during their academic tenure in the athletic training program where data collection takes place. This definition is the work of the author.

Liaison Committee on Medical Education (LCME). Nationally recognized accrediting authority for medical education programs leading to the M.D. degree in United States and Canada. The LCME is sponsored by the Association of American Medical Colleges and the American Medical Association. (Liaison Committee on Medical Education, 2013).

National Athletic Trainers' Association (NATA). The professional membership association for athletic trainers and others who support the profession of athletic training. (National Athletic Trainers' Association, 2014).

Peer assessment. The process whereby students are involved in grading the work of fellow students. (Tillema et al., 2011).

Professional behaviors. The identified characteristics in the *Foundational Behaviors of Professional Practice* under the sub-heading of professionalism. These characteristics include being an advocate for the profession, demonstrating honesty and integrity, exhibiting compassion and empathy, and demonstrating effective interpersonal communication skills. This definition is adapted from the *Athletic Training Educational Competencies* (5th ed.). (National Athletic Trainers' Association, 2011).

Summary

Healthcare professions recognize the importance of peer assessment in the development of professional skills and behaviors (Evans et al., 2004; Speyer et al., 2011). Students preparing to become healthcare practitioners need to engage in peer assessment activities in order to prepare for their careers (Garner et al., 2010; Henning & Marty, 2008; Hulsman et al., 2013; Topping, 1998). The benefits to incorporating peer assessment in higher education (increased accountability and autonomy, faster feedback, improved communication and assessment skills, etc.) outweigh the potential drawbacks (fairness, potential stress, etc.); especially once students have been trained and have practiced with peer assessment activities (Garner et al., 2010; Rush, Firth, Burke, & Marks-Maran, 2012; Tillema et al., 2011; Topping, 2010). Peer assessment is supportive of a student-centered learning environment reflected in the constructivist learning theory

(Hodges, 2011; Hulsman, 2013; Topping, 1998; Yurdabakan, 2011). Due to the similarities in educational requirements (CAATE, 2012; LCME, 2012: NATA, 2011), peer assessment research in athletic training education needs to be performed to determine if its value is as great to athletic training professional preparation, as it has been to medical professional preparation (Finn & Garner, 2011; Garner et al., 2010; Speyer et al., 2011).

Chapter 2: Literature Review

This study seeks to determine the accuracy and reliability of undergraduate athletic training students to assess their peers on clinical skills and professional behaviors. The following literature review presents an analysis of previous and current peer assessment literature as it pertains to the conceptual underpinnings of the constructivist learning theory, and the validity and reliability of its use in undergraduate, medical, and allied health education. The current medical and athletic training education requirements are presented, along with an analysis of the similarities between the two disciplines. Peer assessment research performed in medicine and allied health education on clinical skills, professional behaviors, and didactic skills are included for perspective on the types of studies presented in the literature.

Multiple education and medical databases were searched using the key terms “peer assessment” and “peer evaluation”. These terms were then searched with the addition of the terms of “medical education”, “athletic training education”, “clinical skills”, and “professionalism”. The majority of the sources used for this literature review are from peer-reviewed journals. The education requirements for medical and athletic training are cited from the accrediting agencies responsible for establishing the standards for each discipline (CAATE, 2012; LCME, 2012).

Constructivism and Peer Assessment

Although it is not consistently referenced in the literature, constructivism is the learning theory that works best within a peer assessment framework (Luxton-Reilly & Denny, 2010; Topping, 1998; Yurdabakan, 2011). Social constructivism, Piaget’s model of cognitive conflict, and Vygotsky’s concept of scaffolded learning are cited as the primary foundations for peer assessment (Topping, 1998). Social constructivism promotes the use of discussion and interaction between learners to develop a joint construction of knowledge. Likewise, Piaget supported cooperative learning through reciprocal interaction among learners. Vygotsky promoted communication as an implicit component to learning. Ranging from activities that require only written feedback through grades, to group activities that allow students to openly discuss and build a learning product, peer assessment requires learners to interact and communicate on some level (Topping, 1998). Through its support of joint construction of knowledge through interaction and discourse among learners, social constructivism is reflected in peer assessment's innately interactive nature in creating a learning environment (Yurdabakan, 2011).

Recent studies have connected collaborative learning and peer assessment to the theory of constructivism. Problem-based learning and other constructivist activities in a collaborative healthcare environment promote critical thinking, adaptability, peer assessment, and team consensus-building (Hodges, 2011). The author believed nurses needed to be proactive problem-solvers and work within an interdisciplinary team environment in order to function well in a professional setting. Through collaborative learning experiences, including peer assessment, nursing students co-constructed their knowledge with instructors, improving their abilities to adapt to changes in their work environment and gain control over their learning.

Peer assessment in group work, and its relationship to social constructivism, has also been recently investigated (Yurdabakan, 2011). The author asserted that learning and assessment activities co-existed through peer assessment. Peer assessment activities during group work also led to improvements in participation and reflection on the learning process due to the students questioning of each other.

Constructivist learning activities, including peer assessment, improved the critical thinking and communication skills of computer science students (Luxton-Reilly & Denny, 2010). The interaction among the students allowed them to observe how their peers solved problems and improved their lifelong learning skills. The authors stated feedback provided by classmates offered valuable learning opportunities and this advice and skill modeling were crucial elements for independent learning.

In a constructivist curriculum, the roles of students and teachers defy tradition. Peer assessment activities reflect this difference. Peer assessment is most effective when students have direct input into the assessment process (Finn & Garner, 2011; Ploegh et al., 2009). Student involvement in the assessment process shifts power away from a traditionally teacher-centered practice, to a more collaborative activity. Student input in the assessment process, as performed through peer assessment practices, gives students some control over their learning and empowers students to take ownership of their education (Henning & Marty, 2008; Hodges, 2011; Schunk, 2012). Outcomes of student-centered classroom techniques include a stronger connection to the content by students than normally occurs within the traditional curriculum (Roloff, 2010; Rush et al., 2012) and improved application of learned material in a professional setting (Hodges, 2011).

Peer assessment engages students in learning and assessment processes (Yurdabakan, 2011). Active student participation is a hallmark of constructivism (Powell & Kalina, 2009; Schunk, 2012). Discussion and dialogue are encouraged in constructivist activities and the opportunity for this discourse is an important component to peer assessment (Powell & Kalina, 2009; Topping, 1998; Yurdabakan, 2011). Active engagement among nursing students was crucial for development of the critical thinking and problem solving techniques needed to be competent health care providers (Hodges, 2011). The collaborative learning found in peer assessment activities will benefit athletic training students as healthcare providers in the same way (CAATE, 2012; NATA, 2011). The social process of learning is supported in peer assessment through its inherent collaborative characteristics (Roloff, 2010; Schunk, 2012; Wright & Grenier, 2009). Peer assessment allows for immediate feedback from one or more individuals, providing an opportunity for students to gain an immediate deeper understanding of content (Gielen et al., 2011; Rush et al., 2012), and the mutual review process of peer assessment has a positive impact on student learning (Harris, 2011).

Constructivism and peer assessment both improve critical thinking and problem solving abilities in students (Blaik-Hourani, 2011; Casey et al., 2011; Falchikov, 2005; Schunk, 2012). Constructivist activities include those that require students to critically evaluate the information presented to them and allow for interpretation of the information by the students for their own learning purposes (Powell & Kalina, 2009). Meta-cognitive skills such as problem solving, mastery, and critical thinking are enhanced through peer assessment activities (Falchikov, 2005; Gielen et al., 2011). Peer assessment encourages critical thinking and problem solving in participants through observation of other students' abilities and the analysis of feedback provided to, and given by, peers (Luxton-Reilly & Denny, 2010; Marty et al., 2010). Athletic training students must be able to use critical thinking and problem-solving skills in the patient evaluation process in order to determine a diagnosis and provide quality treatment to their patients (NATA, 2011).

Peer assessment provides opportunities for students to reflect on their knowledge and skills (Luxton-Reilly & Denny, 2010; Welsh, 2007; Yurdabakan, 2011). Activities that promote reflection in students are also key parts to a constructivist curriculum (Roloff, 2010; Schunk, 2012). Reflection provides students with an important opportunity to develop their skills as lifelong learners (Falchikov, 2005; Luxton-Reilly & Denny, 2010). Assessing the work of a peer and reflecting on their own work through peer assessment activities teaches students the ability to learn on their own in the future (Gielen et al., 2011; Kaufman & Schunn, 2011; Welsh, 2007; Vickerman, 2009). Lifelong learning skills are important to athletic trainers because they are required to advance their professional knowledge in order to remain current with the constantly changing healthcare landscape. Athletic training students are required to begin this process prior to graduation (NATA, 2011).

Historical and Background Studies in Peer Assessment

The first collection of studies regarding peer assessment in higher education was reported by Topping (1998). In addition to determining the nature, quality and extent of the peer assessment literature, the author identified a typology for peer assessment, explored its theoretical underpinnings, clarified the mechanisms through which peer assessment has its effects, and gave recommendations for future peer assessment research.

Journal articles published between 1980 and 1996 were included in the study, which used the Social Science Citation Index, ERIC, and Dissertation Abstracts International databases to complete data collection (Topping, 1998). All articles that focused on peer assessment among students in higher education were included, resulting in 109 total items. Of these, 42 were descriptive in nature and 67 provided outcome data through research processes. The author noted that of these 67 studies there were very few that used rigorous analyses or measures of known validity and reliability (Topping, 1998). Sweeping conclusions regarding peer assessment typology were not made due to the large number of variables (17) and diverse activities found in the literature.

Four mechanisms through which peer assessment creates its effects were identified by Topping (1998). First, peer assessment promotes cognition and meta-cognition among its users. Specifically, students learn by assessing others. Learners must dissect and evaluate a product in order to ask intelligent questions of its originator. Peer assessment also requires students to increase their time on tasks. In order for learners to review, clarify, provide feedback, correct inaccurate or misconceived information, and consider alternate ideas, they must devote more time and effort into their activities. Students can use peer assessment activities as a form of norm-referencing. Peer assessment allows learners to locate their performance in relation to the performances of their peers and designated learning objectives. Additionally, self-assessment is improved. Finally, peer assessment provides learners with faster, more abundant feedback compared to instructors. Higher order and better quality thinking occurs with more immediate feedback, along with less potential for confusion about a concept to linger for learners.

The second peer assessment mechanism identified by Topping (1998) was affect. The qualities that are promoted through peer assessment include ownership, personal responsibility and motivation among students. In addition, interaction among students is increased and students bond over assigned activities.

Next, it was noted that peer assessment allows for social and transferrable skills to be used by learners (Topping, 1998). Teamwork and interactive learning are enhanced. Verbal communication and negotiation skills are improved, along with diplomacy. Topping also believed professional skills can be transferred through peer assessment.

Finally, user insight into the assessment process was cited as a mechanism through which peer assessment works (Topping, 1998). Greater insight into assessment allows students to develop their assessment skills for their own, and their peers', work.

Disadvantages to peer assessment focused on student relationships (Topping, 1998). Students may not accept the responsibility of evaluating their peers. In addition, it is important to oversee the peer assessment process in order to ensure power relationships among learners do not take over the assessment process.

Thirty-one studies investigated validity and reliability of peer assessment among college students (Topping, 1998). Most of these studies compared peer and teacher grades through a variety of mechanisms. The author noted that many studies that cited a reliability measure were actually validity studies by design. Findings of Topping (1998) regarding reliability included 18 of 25 studies (72%) that reported acceptably high reliability in peer assessment marks compared to teacher marks. Reliability measures in these studies were reported by correlation coefficients, percentage agreement, or measures of central tendency and variance. The seven remaining studies that investigated validity and reliability identified the validity and reliability of peer assessment to be unacceptably low compared to teacher assessment (Topping, 1998).

Peer assessment performed through tests, marks, or grades was used in many disciplines, and was the most commonly reported method of peer assessment among college students (Topping, 1998). Students reported peer assessment to be demanding, but anxiety reducing. There were frequent reports of improvements in test and skills performance.

Peer assessment for professional skills among college students was also noted (Topping, 1998). The author drew a parallel to the use of peer appraisal between professionals in the workplace. Nursing and physical therapy studies were grouped with medicine and the main conclusion drawn by Topping (1998) was the issue of acceptability among students.

The author suggested areas that required clarification in future investigations of peer assessment (Topping, 1998). Expectations, objectives, and acceptability were the three concepts identified. Clarification in these areas was needed for all stakeholders to develop trust in peer assessment as a valid assessment tool. Specifically, Topping (1998) mentioned training, practice, and coaching as areas where students should be better prepared for peer assessment participation. In addition, the assessment criteria by which students evaluate each other need to be made clearer in the research. Studies of higher methodological quality that did not vary greatly in type and organization were of high priority to be investigated in order to advance the peer assessment literature.

A meta-analysis of research studies on peer assessment among undergraduate students was performed in 2000 by Falchikov and Goldfinch. The authors investigated level of agreement in marking between peer assessed grades and teacher assessed grades. A database search (BIDS, ERIC, PsychINFO, Socinfo, FirstSearch) located more than 100 articles published from 1959-1999 (Falchikov & Goldfinch, 2000). Of these, 48 studies were review papers and qualitative studies performed in higher education settings.

An additional 48 articles reported the quantitative comparisons of teacher and peer grades in higher education. Inclusion criteria for the meta-analysis were (a) the studies had to be performed in a higher education setting and (b) the articles had to report correlation coefficients or proportions that allowed for level of agreement between teacher and peer marks to be determined (Falchikov & Goldfinch, 2000). Subject areas included business and management, medicine, dentistry and paramedical subjects, science and engineering, and social science and arts.

Findings included an overall average *r* value of 0.69. This signified that peer and teacher marks agreed well across all studies. Additionally, the authors found that teacher and peer marks agreed most similarly when global judgments were evaluated, as opposed to when individual grading scales were used (Falchikov & Goldfinch, 2000).

The second significant finding by the authors was a better correlation between peer and teacher marks when traditional academic process or products were graded, as opposed to items related to professional practice. Academic product items, which included such things as written exams and essays, were found to have a coefficient value of $r = 0.75$. An *r* value of 0.83 was found for academic process items such as oral presentations and participation in group work. Professional practice items, which included studies that investigated clinical skills and teaching performance, were found to have a coefficient value of $r = 0.54$ (Falchikov & Goldfinch, 2000). This finding is interesting in relation to the current study, as clinical skills and professional behaviors can be categorized as professional practice items, as defined by Falchikov and Goldfinch.

The recommendations of the authors at the conclusion of the study included suggestions to further investigate the interaction between variables in peer assessment studies, the effects of multiple experiences with peer assessment on participants, and investigation into friendship bias among students involved in peer assessment activities. These recommendations furthered the previously made recommendations by Topping (1998).

A recent literature review demonstrated the great amount of variability found in the literature about peer assessment in medical and allied health education settings (Speyer et al., 2011). The purpose of the review was to provide an overview of the instruments and questionnaires that have been used for peer assessment in medical and allied health education, and to then present the psychometric properties of these tools, as described in the literature. The authors (Speyer et al., 2011) argued that confidence in, and acceptance of, peer assessment among students can only be achieved when concurrent validity is found in the instruments used for peer assessment.

Independent literature searches using five databases (Pubmed, ERIC, Embase, PsychINFO, and Web of Science) were performed by two reviewers. The searches yielded 2,899 articles, but only 28 met the inclusion criteria of the authors (Speyer et al., 2011). These criteria were (a) articles that described a peer assessment tool in allied health or medical education settings and (b) original articles that described peer assessment tools and articles that presented information about the validity or reliability of any of the peer assessment tools. Within these final 28 articles, the authors identified 22 different assessment instruments used. Most authors developed their own peer assessment instrument and applied it to their individual educational setting, in accordance with their own criteria and scoring system.

Findings included only two articles that did not occur in a medical education setting; one performed in pharmacy education, and one performed on a combined group of medical and dental students. Also, only three articles described the concept of validity on any level and six articles provided no psychometric data at all. The authors concluded that their attempt at a statistical pool of data was not possible for their designed review. The included articles offered too much heterogeneity in their designs, instrument diversity, and restricted data availability of psychometric characteristics to be able to provide a useful review (Speyer et al., 2011).

The limited number of studies found by Speyer et al. (2011) that identified peer assessment tools and presented information regarding validity and reliability measures reiterated the concern over the variability of designs, tools and measurements used in peer assessment research that was presented by Topping (1998).

Another recent study performed via database search analyzed the different conceptualizations of quality of peer assessment in the literature (Gielen et al., 2010). The authors stated a cluttered picture of peer assessment research has been the result of an increased output of studies evaluated against a variety of quality criteria. A database search (ERIC, SSCI, Academic Search Premier, PsychINFO) yielded between 174 and 1,196 studies published between 1952 and 2006. These studies were then examined for their conceptualizations of quality for peer assessment, including their goal for using peer assessment in a particular practice (Gielen et al., 2011).

Five major goals of peer assessment were presented, along with the quality conceptualizations found in the literature for each. The first goal of peer assessment was identified as the use of peer assessment as a social control tool. This form of peer assessment related to the increased external motivation many students obtain when they are aware their work will be judged by a peer, as opposed to a teacher (Gielen et al., 2011). The quality concept found by the authors for this goal was based on behavior change. Studies that used peer assessment as a social control tool reported success as greater occurrences of a desired behavior, or fewer occurrences of an undesired behavior. The authors concluded that such performance changes were only indirectly related to successful peer assessment.

Peer assessment as an evaluation tool was the most apparent practical use of peer assessment (Gielen et al., 2011). The authors found substantial variations in the arrangements of the instruments used in these studies. The quality concept for this goal had three distinct references used for comparison. The first, a validity measure, concerned the comparison of teacher and peer marks. Studies that compared student marks to other students used inter-rater reliability or a generalizability coefficient as their quality concept. The third quality concept identified was a consistency measure in the form of a comparison against self-assessment.

The third goal of peer assessment identified its use as a learning tool (Gielen et al., 2011). The quality concept for this goal was determined by consequential validity, or the effect peer assessment had on the participants and their learning. The articles found that used peer assessment as a learning tool varied in their measurement of learning effects, thus the authors noted that this particular goal had diverse criteria results.

Peer assessment as a "learn-how-to-assess" tool involved learning on a meta-level. Such learning is required to become a life-long learner, something that peer assessment has been identified as promoting in its participants (Gielen et al., 2011).

Consequential validity was also the quality criteria for this goal. However, the authors did not find any studies that directly addressed and measured the effects of peer assessment on lifelong learning. Such a measurement required a longitudinal study, of which none were identified in the authors' searches.

The final goal, use of peer assessment as an active participation tool, had been reported as an empowering practice for students, and one that developed student autonomy (Gielen et al., 2011). The quality concept identified for this goal required a change in classroom culture. It was measured as the accomplishment of creating shared ownership of the assessment process. The authors were unable to find any specific methods or criteria to measure this change, but did note that qualitative studies were the best choice of study design to develop a measurement for this change.

It was concluded that the quality criteria used for peer assessment studies must fall back to the intended goal of the use of peer assessment (Gielen et al., 2010). The authors argued that goals for peer assessment have progressed beyond its use as an assessment tool, introducing new expectations and novel concepts concerning the quality of peer assessment. This change has created a problem. A clear view of the relationship between these new goals and definitions is getting lost (Gielen et al., 2010). The authors further explained that some researchers are not clear about their intended goals for the use of peer assessment, but still draw conclusions regarding its quality.

A 2011 (Tillema et al.) literature search attempted to identify the types of quality criteria that are attended to in the development of peer assessment activities used for learning. Specifically, the authors evaluated the studies with regard to their use of quality criteria under two conditions. The first was their recognition of criteria for educational measurement. The second condition was their consideration of student involvement in learning assessment. The search yielded a total of 1,151 articles, with 132 accepted for inclusion in the study based on the following criteria (Tillema et al., 2011): (a) a publish date after 2000, (b) studies had to be conducted within an educational context, and (c) studies had to analyze the implementation of peer assessment as an assessment tool.

The authors found that specific quality criteria were taken into account in relation to specific steps in the assessment process (Tillema et al., 2011). Seven steps were identified and a total of 129 potential quality criteria could be assigned to each step. The first two steps (purpose and goal setting, construction of tasks) in the assessment process were found to consider the criteria of authenticity most, with the purpose and goal setting step including both criteria. The quality criteria of fairness and transparency were most often considered during the third step, choosing scoring criteria, and the sixth step, appraisal. The fifth step in the assessment process, scoring, considered the criteria of transparency, fairness, and generalizability the most. This step also represented the greatest amount of quality criteria (31 of a potential 129) considered during any step in the assessment process. The fourth and seventh steps, administration of the assessment and providing guidance and feedback, respectively, were the steps where the least consideration towards quality criteria was given.

The authors (Tillema et al., 2011) were also able to determine that each step in the assessment process offers an opportunity to involve students in their own evaluations. The amount and type of quality criteria varied in each step of the assessment process, but student involvement was always an option. The authors did not report information on the outcome measures associated with the quality criteria for each step in the assessment process.

This study's (Tillema et al., 2011) findings supported the use of peer assessment as a way to increase ownership among students for their education, supporting the earlier reported study of Gielen et al. (2011). This finding has relevance for the proposed study. If the study finds that undergraduate athletic training students are accurate and reliable graders of their peers, peer assessment activities may be incorporated more thoroughly into the athletic training curriculum in order to promote ownership among students for their education. Increasing ownership relates to the established requirement in athletic training education for students to continually advance their knowledge through critical self-examination, continuing education, and evidence-based practice (NATA, 2011).

Current State of Peer Assessment Research

Many questions still exist about peer assessment and its outcomes. *Learning and Instruction* dedicated an entire issue in 2010 to peer assessment. The issue evaluated the current literature and outlined areas where improvements in the research should be addressed. Specific areas where there exists a need for improvement include establishing relationships between goals, processes, measures, and outcomes of peer assessment research (Gielen et al., 2011; Strijbos and Sluijsmans, 2010; Topping, 1998; 2010; van Zundert et al., 2010). These concerns are mirrored in the peer assessment research in healthcare education as well, with no significant research outside of medical education into the use of peer assessment (Marty et al., 2010; Speyer et al., 2010).

Kollar and Fischer (2010) provided a commentary that identified peer assessment research as being in its adolescent stage of growth. The authors maintained that peer assessment research needed to form a clear identity and affiliation in order to develop a more cohesive message in the literature. Identity formation can be achieved through the use of shared terminology and joint theory building. Peer assessment has been called by other names, including peer revision and peer feedback. This diversity in terminology among researchers made it difficult to clearly describe the phenomenon of peer assessment in the literature. The lack of a commonly agreed-upon joint process model contributed to an inconsistent identity (Kollar & Fischer, 2010). Such a process model would help identify what activities constitute peer assessment and the processes associated with these activities. The previously mentioned studies by Topping (1998) and Speyer et al. (2011) also presented concerns over the number of designs, tools and measurement variations found in the peer assessment literature.

Peer assessment's lack of clear affiliation within a particular research field was cited as another issue to address in future peer assessment research (Kollar & Fischer, 2010). Peer assessment has close ties to the fields of peer tutoring, help seeking, and collaborative learning. Kollar and Fischer (2010) asserted that peer assessment research could benefit from the information discovered in these other fields, especially collaborative learning, as peer assessment is fundamentally a collaborative activity.

The suggestion made by the authors (Kollar & Fischer, 2010) designed to address the needs of identity formation and affiliation was to develop a process model that was cognitively themed for peer assessment. The authors started by identifying four activities that occur consistently during peer assessment. These activities were (a) task performance, (b) feedback provision, (c) feedback reception, and (d) revision.

Next, the authors (Kollar & Fischer, 2010) designated ways that high level cognitive and discursive process occurred during peer assessment; allowing for a clearer process model for peer assessment activities. For task performance, the cognitive and discursive processes varied by type of task, but the authors believed that interactivity during this step evoked higher order thought processes, resulting in positive learning outcomes. Feedback provided to a learner during peer assessment must support that learner in solving the task if high level learning gains are to be achieved. If feedback was given in the form of a simple response of “correct” or “incorrect”, then higher cognitive processes were not used by either learner in the peer assessment process. Feedback must be of good quality and relevant to the task in order for it to be received and used to facilitate learning. Finally, comparison and integration of information by the learner was needed in the revision stage for higher order thought processes to be reached. The authors encouraged allowing communication during this stage in order to alleviate some of the stress associated with this process by the learner (Kollar & Fischer, 2010).

This commentary provided a method for improving the current state of identity of peer assessment in the literature (Kollar & Fischer, 2010). The authors clearly supported the connection of peer assessment to collaborative learning research and provided a process model that demonstrated that connection. Peer assessment, as part of a more participatory culture of learning, can play an important role in the design of such collaborative learning environments (Kollar & Fischer, 2010).

A literature review of 26 empirical studies performed on peer assessment by van Zundert et al. (2010) resulted in the authors’ development of peer assessment variables and subsequent investigation of their inter-relatedness. The four variables included psychometric qualities; those related to validity and reliability of peer assessment scores compared to instructor scores. Next, the variable that pertained to domain-specific skills included studies that make claims about student learning from the use of peer assessment in a specific discipline. The variable of peer assessment skills related to the quality of feedback and feedback style given and received between students. The final variable developed by the authors was student attitudes towards peer assessment; specifically student confidence in using peer assessment and the perceived learning benefits (van Zundert et al., 2010).

The findings illustrated the gaps in content in the peer assessment literature; including the finding that most of the studies were neither quasi-experimental nor experimental in design (van Zundert et al., 2010). Also, most of the literature on peer assessment had been performed in higher education, with very little information available on peer assessment in lower educational levels. Only 12 of the 26 reviewed studies identified clear relationships between their methods, conditions, and outcomes (van Zundert et al., 2010).

The identified psychometric studies, eight in all, did not provide any conclusive evidence towards what contributed to the psychometric qualities. Also, within those studies, the psychometric qualities were expressed in multiple ways and their findings were quite diverse (van Zundert et al., 2010). The main finding from the domain-specific studies was a lack of longitudinal research in the area. For the third variable, the authors found a lack of studies that attempted to distinguish between the effects of assessing peers and being assessed by peers. The largest number of studies reviewed (15 total) were categorized by the authors in the fourth variable, student attitudes towards peer assessment (van Zundert et al., 2010). The results of these studies showed an overall positive attitude (12 of the 15 studies) of students toward peer assessment. However, the authors made a point to acknowledge that within these studies there was a tremendous amount of variety in both data collection instruments and procedures.

Additional findings by the authors (van Zundert et al., 2010) included the discovery that training and experience of peer assessors had a positive impact on both the psychometric qualities of the related studies and student attitudes towards the use of peer assessment. Also, learning outcomes improved when peer assessed feedback was provided before revisions on student work were performed, thus improving the quality of the work in the discipline. Finally, the authors found that training improved peer assessment skills and development of those skills was related to the students' thinking styles and academic achievement level.

Three areas that are widely regarded as known outcomes of peer assessment in the literature were identified by Topping (2010). The first was the positive impact training and practice in the use of peer assessment has on its validity and reliability. Second, student attitudes towards peer assessment, as both a formal and informal evaluation measure, were positive. Finally, feedback between peers was important to learning outcomes when using peer assessment, but non-directional feedback tended to be received and utilized more often than direct feedback.

A critical evaluation of five empirical studies and one literature review was subsequently performed and identified multiple variables in the literature that required further exploration (Topping, 2010). Specifically, definitions of samples, types of peer assessment, nature of interventions, and the measures used to evaluate change, were areas the author believed were necessary to investigate. Topping (2010) believed identifying the challenges that lay ahead in the peer assessment literature provided the best way to start addressing them in the research.

The author identified a number of demographic issues that arose in the sample populations that needed further exploration in their relation to peer assessment (Topping, 2010). These included age of participants, subject content studied, learning culture of participants (Eastern vs. Western, etc.), and prior experience of participants with peer assessment.

Since peer assessment can be greatly varied in its practice, the author cited that the type of peer assessment studied should be an area of further exploration. Such things as whether the peer assessment was reciprocal among participants, the ability or level of matched participants, anonymous or personal feedback practices, on-line or face-to-face feedback practices, and length of peer assessment activities were all research ideas presented by Topping (2010). The need to study the relationship between multiple variables was also noted by Falchikov and Goldfinch (2000) in their earlier study.

The nature of the peer assessment intervention needed further specifying in the research as well (Topping, 2010). Specifically, the effects of length of training, length of interaction, number of assessments, joint development of assessment criteria, and format of assessment reporting, were all areas where peer assessment intervention needed to be developed further.

Finally, it was noted that research measurements would always reflect the author's theoretical inclinations (Topping, 2010). However, it was suggested that the amount and quality of feedback given during peer assessment be investigated further.

Also, the type of peer assessment task; written, practical skill, presentation, etc., that is studied should impact the type of measurement used. The author also believed the trust between, and psychological safety of, the participants were an area that researchers could benefit from exploring further.

An editorial piece was contributed that focused on the need for development in the areas of the methodology, function and concepts within peer assessment research (Strijbos & Sluijsmans, 2010). This need stemmed from the lack of empirical evidence for the effects of peer assessment in the current literature. Specifically, the authors wanted to address what was known about peer assessment and what had been claimed about the benefits of peer assessment for learners. Strijbos and Sluijsmans (2010) went further and agreed with the previously mentioned studies of Gielen et al. (2011), Topping (1998; 2010), and van Zundert et al. (2010), that there was a clear need to establish relationships between the goals, processes, measures, and outcomes of peer assessment research studies.

The authors stated that most studies included self-reported learning effects of peer assessment by students, but the need existed for studies to investigate specific mechanisms in relation to specific outcomes through experimental and quasi-experimental designs (Strijbos & Sluijsmans, 2010). Methodological developments could occur through the inclusion of a greater variety of designs, instruments and analytic techniques in peer assessment studies. The authors also suggested generalizability of the peer assessment literature could be fostered through the establishment of quality criteria for peer assessment research.

Also, functional development of peer assessment research was needed in order to identify clear definitions of peer assessment purposes and relation to learning outcomes (Strijbos & Sluijsmans, 2010). Most studies that compared student and teacher evaluations did so only once, and subsequent student evaluation performances were not investigated. This summative application of peer assessment had traditionally been a strong focus of the peer assessment literature, but did not provide any information about the effects peer ratings had on subsequent performances.

The final area for research development in peer assessment was concept. The majority of studies performed on peer assessment focused on the evaluation aspect, not the collaborative process involved with its use. Additionally, these studies used peer assessment intervention as a one-time event, not the interactive and cyclical process that continued collaboration provided. The authors suggested further investigation into the reciprocity effect of peer assessment. The reciprocity effect encompasses the social aspects of peer assessment, and the authors believed its effects on the peer assessment process and subsequent student learning needed to be explored (Strijbos & Sluijsmans, 2010).

An on-line questionnaire was used to investigate the criteria used for peer assessment in the classrooms of secondary vocational teachers by Ploegh et al., 2009. Eighty-four questionnaires were administered, with 56 returned for data analysis. Participants were teachers in the vocational areas of health, technology and economics.

Each questionnaire gathered initial exploratory information from the participants that determined how peer assessment was used in their classrooms. Thirty subjects responded peer assessment was used for formative, or learning, assessment. The remaining 26 respondents cited their peer assessment practices as summative, or grading, assessments. The second part of the questionnaire established more thoroughly how the teachers used peer assessment. Specifically, the purpose of the peer assessments, types of learning objectives evaluated, scoring methods used, and privacy of peer assessors, were all identified (Ploegh et al., 2009).

The final part of the questionnaire included the items for criterion use under the assessment process (Ploegh et al., 2009). A Yes/No response format was used regarding the consideration teachers gave to the features of quality criterion in their peer assessment classroom practices. Twenty percent of respondents involved students in the development of scoring criteria for peer assessment. Additionally, 14% allowed students to have total control over decisions regarding scoring criteria. Half of the respondents used peer assessment only once during the length of a course, while 37% used peer assessment practices throughout a course's duration.

Interpretation and valuing of scores by the authors (Ploegh et al., 2009) showed greater emphasis was placed on using peer assessment for grading, rather than learning purposes, contradicting the information gathered in the first part of the survey where more teachers stated they used peer assessment for learning rather than grading purposes. When used for grading, transparency, fairness and reproducibility were the most important aspects of the peer assessment process to the respondents. Finally, two-thirds of respondents provided feedback, along with the student assessors, while only one-third provided feedback on their own, excluding peer feedback from their practice.

Regarding the assessment process, the authors (Ploegh et al., 2009) discussed their findings in relation to four steps of assessment. For the establishment of scoring rules, teachers were the main deciders, but they explained their chosen criteria well and transparently. The appraising and scoring process was directed towards providing feedback and most often occurred face-to-face and at the end of a course. How and when the feedback was given was specified beforehand. Score interpretation was predominantly the job of the teachers. The authors found very little student involvement in this area. Lastly, in order to provide direction for future activities, feedback was delivered by the teachers after collecting all the information from the peer assessment process.

It was concluded that peer assessment practices entailed many of the already-established quality criteria in measurement and evaluation (Ploegh et al., 2009). The authors made note, however, that many of these criteria were adapted to the peer assessment environment from their original generic forms.

Although Ploegh et al. (2009) identified that secondary vocational teachers used quality criteria for measurement and evaluation in their courses that applied peer assessment practices, the participants of their study were not consistent on their uses of peer assessment.

It is clear there is a need to establish peer assessment as a valid and reliable classroom practice. While peer assessment has been used in medical education for a number of years (Finn & Garner, 2011) with positive results (Speyer et al., 2011), other healthcare education programs have not performed significant research into the use of peer assessment (Marty et al., 2010; Speyer et al., 2011).

A clear, coherent process model, as recommended by Kollar and Fischer (2010) is one suggestion that can help move peer assessment research forward. Strijbos and

Sluijsmans (2010), Topping (2010), and van Zundert et al. (2010) all reported a similar need, based on the current reports of peer assessment in the literature. Particularly, the lack of established relationships between methods, conditions and outcomes was cited by van Zundert et al. (2010) as an area where gaps were found in the literature. Topping (2010) agreed and noted that peer assessment studies needed to improve their definitions of samples, types of peer assessment interventions used, and the measures used to determine outcomes. Stijbos and Sluijsmans (2010) echoed this sentiment by pointing to the lack of empirical evidence for the effects of peer assessment. The authors' response called for development of methodology and outcomes in peer assessment research.

Van Zundert et al. (2010) and Topping (2010) also agreed that training, practice and experience all have a positive effect on peer assessment outcomes; of which Topping stated validity and reliability, specifically, were both improved. A third area where agreement was reached with these authors was the importance of feedback in the peer assessment process (Topping, 2010; van Zundert et al., 2010). Van Zundert et al. (2010) reported improved learning outcomes when peer feedback was provided between students prior to submission for grading.

Medical and Athletic Training Education Requirements

The educational competencies for athletic training education are similar in nature to those for medical education (ACGME, 2013; AAMC, 1998; CAATE, 2012; LCME, 2012; NATA, 2011). In the United States, LCME and CAATE are the accrediting bodies for medical education and athletic training education, respectively (CAATE, 2012; LCME, 2012). These accreditation organizations both require education programs to provide quality didactic and clinical education experiences for their students in order to prepare them to practice as qualified healthcare providers after graduation.

The formal standards for medical education (LCME, 2012) require medical schools to provide students with learning opportunities in multidisciplinary academic and clinical environments that allow for interaction with students of other health professions (Standard IS-12). Lifelong learning skills must also be fostered within medical schools through provision of active learning instructional opportunities to their students (Standard ED-5A). A medical education curriculum is required to include preventative, acute, chronic, rehabilitative, and continuing care when covering all organ systems with their students (Standard ED-13). Academic and clinical experiences for medical students must relate to the phases of the human life cycle and prepare the students to recognize signs of health, opportunities for health promotion, signs and symptoms of disease; and establish differential diagnoses and treatment plans, and educate and assist patients in addressing health-related concerns (Standard ED-15).

The LCME (2012) requires medical schools to provide instruction and opportunities for students to develop professional behaviors, separate from academic knowledge and clinical skills. Communication skills must be taught in relation to interactions with colleagues, patients and their families, and other health professionals (Standard ED-19). Medical school faculty and students are required to demonstrate an understanding of diverse cultures and belief systems, and how people of these cultures and belief systems perceive health and illness and respond to their symptoms, pathologies, and treatments (Standard ED-21). Gender and cultural biases must be recognized and addressed by medical students in the process of health care delivery

(Standard ED-22). Medical ethics and human values must be instructed within a medical education program, and students must demonstrate scrupulous ethical principles when caring for patients and communicating with families and others involved in patient care (Standard ED-23). The learning environment of a medical education program must promote the development of specific and appropriate professional attributes through both formal learning activities and informal interactions by individuals who come into contact with medical students (Standard ED-31A).

The medical education program standards established by the LCME (2012) do not dictate how medical schools should address the requirements within their curriculums. Instead, documents written by other organizations are cited as resources for medical programs to refer to in order to meet the established standards. In order to develop the competencies expected by the medical profession and public, Standard ED-1A (LCME, 2012) requires the establishment of outcomes-based learning objectives that include the desired knowledge, skills, behaviors and attitudinal attributes of a physician. The LCME Standards (2012) refer to the first report of the Medical School Objectives Project (MSOP), developed by AAMC (1998), the Common Program Requirements, adopted by the ACGME (2013), and the American Board of Medical Specialists (ABMS) (2006-2012).

Since 1932, the AAMC (1998) has called upon medical schools to develop and institute learning objectives in their curricula. The MSOP fulfilled a recommendation set forth by the AAMC to develop a set of goals and objectives to provide guidance to medical schools in the establishment of learning objectives for their medical education programs (AAMC, 1998). The MSOP reflects a consensus among medical education leaders on four attributes that physicians must possess in order to effectively practice medicine.

Altruism is the first attribute described by the MSOP. Physicians need to be compassionate, empathic, trustworthy, and truthful in caring for patients and all professional duties (AAMC, 1998). Specifically, before graduation, medical students must demonstrate appropriate decision-making based on knowledge of theories and principles that govern ethics and the ethical dilemmas in medicine. Medical students must also portray a commitment to advocate for the interests of their patients at all times, and use honesty and integrity in their interactions with patients, families, and colleagues.

Physicians must be knowledgeable. The MSOP (AAMC, 1998) described the need for medical students to understand and apply the scientific basis of medicine in their practice. This includes appropriate knowledge of the structure and function of all body and major organ systems, and the mechanisms that work to maintain homeostasis. The MSOP states this fundamental knowledge is needed in order to understand disease and use diagnostic and therapeutic modalities wisely in the practice of medicine.

Providing care to patients requires medical school graduates to be highly skilled (AAMC, 1998). Obtaining an accurate medical history, performing complete and organ-specific physical examinations and diagnostic procedures, and constructing appropriate treatment strategies are all clinical skills necessary to being a competent physician. The management of acute and chronic conditions of the medical, psychiatric, and surgical nature, along with those conditions needing short- and long- term rehabilitative care; are all responsibilities of the resident physician. Immediately life threatening illnesses such as cardiac, pulmonary, or neurological conditions must be recognized and an appropriate emergency therapy initiated by medical school graduates. The MSOP (1998) described the need for physicians to understand the etiologies, pathogeneses, and manifestations of diseases and conditions; along with the scientific basis and evidence for the effective use of available therapeutic interventions for appropriate patient care.

It is the physician's duty to collaborate with other healthcare providers and use systematic approaches in the promotion, maintenance, and improvement of the health of patients and populations (MSOP, 1998). Knowledge and understanding of the risk factors and preventative measure for disease and injury must be used in patient care. Medical school graduates must be able to provide counseling to promote healthy behaviors among patients and families and be advocates for improving access to medical care for everyone. This applies particularly to members of traditionally underserved populations.

Building on the initial work of the AAMC, the ACGME (2013) specified six core competencies that must be integrated into a medical education curriculum. Patient care and procedural skills (IV.A.5.a) must be provided by residents in an effective, appropriate, and compassionate manner. Residents must demonstrate medical knowledge (IV.A.5.b) of established and evolving clinical, biomedical, social-behavioral, and epidemiological sciences in order to apply this knowledge to patient care. Practice-based learning and improvement (IV.A.5.c) requires residents to be able to investigate and evaluate their patient care, appraise and use scientific evidence, and continuously improve patient care through self-evaluation and lifelong learning. Goals in the area of practice-based learning and improvement include identification of strengths, deficiencies, and limits in the residents own knowledge and expertise, incorporation of formative feedback into daily practice, and participation in the education of patients and their families, students, residents, and other health professionals. Residents must demonstrate interpersonal and communication skills that allow for effective exchange of information and collaboration between colleagues, patients, families, and other health professionals, regardless of socioeconomic and cultural backgrounds (IV.A.5.d). Professionalism (IV.A.5.e) must be demonstrated through a commitment to carry out responsibilities and adherence to ethical principles that include (a) respect for patient privacy and autonomy, (b) integrity, compassion, and respect for others, and (c) accountability and sensitivity to patients regardless of their socioeconomic or cultural backgrounds. Systems-based practice (IV.A.5.f) within the larger context and system of healthcare must be shown through awareness and responsiveness by the resident. Also, the resident must demonstrate the ability to effectively use other resources within the system in order to provide optimal care to their patients, including working in interprofessional teams.

Athletic trainers are healthcare providers who collaborate with physicians to optimize activity and participation of patients. Recognized by the American Medical Association, athletic training comprises the prevention, diagnosis and intervention of emergency, acute and chronic medical conditions involving impairment, functional limitations and disabilities (CAATE, 2012; NATA, 2011).

Sponsored by the American Academy of Family Physicians, the American Academy of Pediatrics, the American Orthopaedic Society for Sports Medicine, and the NATA, CAATE develops and maintains the minimum education requirements for entry-level athletic training education programs (CAATE, 2012). As part of the educational standards for athletic training education, programs are required to demonstrate their students interact with other healthcare professionals (Standard IV.C), have opportunities to integrate athletic training knowledge and skills, including clinical decision-making and professional behaviors (Standard IV.F), and clinical education places students in a variety of healthcare settings, including primary care, emergency, outpatient, and specialties outside of orthopaedics (Standard IV.I).

In addition to the previously mentioned standards, CAATE requires formal instruction and evaluation of the *Athletic Training Education Competencies* in a structured classroom and laboratory environment (NATA, 2011). These competencies encompass the specific content knowledge and skills deemed necessary for minimal professional ability as an entry-level athletic trainer (NATA, 2011).

Eight content areas have been identified within the *Competencies* (NATA, 2011), all with relevance to the need for athletic training students to acquire similar knowledge and skills set to medical students. Evidence-based practice (EBP) is the first content area introduced in the athletic training competencies. Practicing athletic training in an evidence-based manner is necessary in order to make sound clinical decisions during practice and for critical examination of the practice itself. EBP is used as a systematic approach to answering clinically relevant questions through review and application of current research and practice evidence (NATA, 2011).

Prevention and health promotion is the area of athletic training designed to limit the incidence of injury and illness and to optimize the overall health of patients (NATA, 2011). Included within this content area are the knowledge and skills that pertain to (a) prevention principles, strategies, and procedures, (b) protective equipment, taping, and wrapping, (c) fitness and wellness, (d) general nutrition, performance enhancing drugs and supplements, and (e) weight management, body composition, and disordered eating and eating disorders.

The content area of clinical examination and diagnosis includes all body systems, with an emphasis placed on those systems (musculoskeletal, neurological) most likely to be affected in patients in the clinical settings where athletic trainers work (NATA, 2011). Athletic training students are expected to be able to apply appropriate clinical reasoning skills during the physical examination process in order to use the collected information and form a differential diagnosis (NATA, 2011). Within the knowledge and skills designated to this area, athletic training students must use standard techniques and procedures in the examination of patients for injuries, conditions, illnesses, and diseases that include the following steps (Competency CE-20): (a) history taking, (b) inspection, observation and palpation, (c) musculoskeletal functional assessment, (d) selective tissue (joint) testing, (e) neurological assessment, (f) respiratory, circulatory, and abdominal assessment, and (g) eye, ear, nose, and throat assessment (NATA, 2011). Athletic training students must also be able to determine when their examination findings warrant referral of the patient to another healthcare provider (Competency CE-22) (NATA, 2011).

An athletic trainer is often the first healthcare provider when an acute condition occurs to a patient, therefore, athletic training education programs must provide a significant amount of education to their students in the area of acute care of injuries and illnesses (NATA, 2011). Emergency planning, patient examination, and transportation are all included within the acute care content area. Specific knowledge and skills are required of athletic training students in the immediate emergent management of the following conditions (Competency AC-36): (a) cardiac arrest, (b) brain injury, (c) spine trauma, (d) heat illness, (e) internal hemorrhage, (f) diabetic emergency, (g) asthma, (h) allergic reaction, (i) seizures, and (j) musculoskeletal injuries (NATA, 2011).

Therapeutic interventions are used to enhance patient function through identification, remediation, and prevention of impairment and functional limitations in order to maximize activity level (NATA, 2011). Therapeutic interventions include those things involved with physical rehabilitation, therapeutic modalities and therapeutic medications. Specific items included in the therapeutic interventions content area that athletic training students must gain knowledge and skills include techniques to: (a) reduce pain and limit edema, (b) restore joint mobility, muscle extensibility and neuromuscular function, (c) improve strength, endurance, speed, power, balance, coordination, agility, (d) improve gait, posture and body mechanics, (e) home-based exercise programs, (f) activity-specific exercises, (g) aquatic therapy, (h) thermal, electrical, mechanical, and ultrasonic agents, and (i) prescription and over-the-counter medications (NATA, 2011).

The competency area of psychosocial strategies and referral is the domain that maintains athletic trainers must be able to identify patients who exhibit abnormal social, emotional, or mental behaviors, and intervene and refer these patients as needed (NATA, 2011). Athletic training students must be able to demonstrate they understand the theoretical backgrounds of psychosocial and emotional principles, psychosocial strategies for improving the well-being of their patients, and when referral for mental, social, or emotional reasons is necessary for their patients (NATA, 2011).

The NATA (2011) maintains that athletic trainers work within the context of the larger, complex healthcare system. Integral to their function as healthcare providers is an understanding of risk management, healthcare delivery, insurance, reimbursement, documentation, patient privacy laws, and facility management. The content area of healthcare administration encompasses the organization and administration aspects of the practice of athletic training (NATA, 2011).

The eighth content area established within the *Competencies* (NATA, 2011) is concerned with professional development and responsibility. The knowledge and skills required of athletic training programs to instruct their students includes everything that pertains to maintaining professional competence within the world of healthcare. Such things include (a) practicing within state and national regulations, (b) using sound moral and ethical judgment, and (c) collaborating with other healthcare professionals for the benefit of patients (NATA, 2011).

Also part of the *Competencies*, but not a formal content area, the *Foundational Behaviors of Professional Practice* are intended to be infused throughout the athletic training education curriculum (NATA, 2011). Seven behavioral areas are contained within the *Behaviors*. These areas are primacy of the patient, team approach to practice, legal and ethical practice, advancing knowledge, cultural competence, and professionalism (NATA, 2011). Under the area of primacy of the patient, the athletic trainer is expected to be an advocate for their patients and provide the best care possible, while protecting patients' privacy and avoiding any conflict of interest (NATA, 2011). A team approach to practice is best described as the recognition and use of other healthcare providers when needed, and the incorporation of the patient in the decision-making process (NATA, 2011). Legal practice requires athletic trainers to follow state and national laws while practicing in a competent manner. In contrast, ethical practice pertains to the athletic trainer's responsibility to practice under the NATA's *Code of*

Ethics and the Board of Certification's *Standards of Professional Practice* (NATA, 2011). Advancing knowledge in the field of athletic training is also a necessary behavior in the instruction of athletic training students. The use of critical examination, EBP, and an appreciation for the role of continuing education in the advancement of the profession are all aspects of advancing knowledge (NATA, 2011). Cultural competence is displayed through the awareness that a patient's culture will impact their attitudes towards healthcare and that athletic trainers must demonstrate the knowledge and behaviors to improve the health of diverse patient populations (NATA, 2011). Under the area of professionalism, athletic trainers are expected to be advocates for the profession. Additionally, athletic trainers must practice with honesty, integrity, compassion, and empathy, while using effective communication skills with their patients and other healthcare professionals (NATA, 2011).

The requirements presented for medical and athletic training education programs share many similarities. First, both groups of students are required to interact and collaborate with healthcare professionals outside of their specialties in order to provide better patient care and understand other providers' roles within the healthcare system (ACGME, 2013; CAATE, 2012; LCME, 2012; NATA, 2011). Second, the knowledge and skills to perform physical examinations in order to develop diagnoses and treatment strategies are major necessities for both groups of students (AAMC, 1998; LCME, 2012; NATA, 2011). Third, education programs for both medicine and athletic training must educate their students on the immediate identification and care of life-threatening medical conditions prior to graduation (AAMC, 1998; NATA, 2011). Fourth, health promotion and wellness for the prevention of illness and disease is a responsibility for both groups of healthcare professionals (AAMC, 1998; LCME, 2012; NATA, 2011). Fifth, while the two groups have different names for it (i.e. evidence-based practice, practice-based learning), the ability to critically examine healthcare practice, followed by an appraisal of the practice in order to improve patient care, is a necessity in both medical and athletic training education (ACGME, 2013; NATA, 2011). Sixth, understanding of the healthcare system in a larger context and the roles different healthcare providers play are also required of both groups of students (AGCME, 2013; NATA, 2011). Finally, appropriate professional behavior and adherence to ethical standards are expected to be instilled by medical and athletic training education programs in their students. Common themes of professional behaviors include: (a) patient advocacy, (b) communication skills with patients, families, and colleagues, and (c) appreciation for cultural diversity (AAMC, 1998; ACGME, 2013; LCME, 2012; NATA, 2011). Additionally, medical and athletic training students are expected to conduct themselves with honesty and integrity, along with compassion and empathy towards their patients (AAMC, 1998; ACGME, 2013; LCME, 2012; NATA, 2011).

Peer Assessment of Clinical Skills

Peer assessment in athletic training education.

The first study published of peer assessment in athletic training education was performed by Marty et al. (2010). The authors incorporated video presentation of three commonly used diagnostic tests to determine the accuracy and reliability of first- and second-year entry-level master's athletic training students in assessing their peers' abilities to perform the tests properly. The tests included (a) a manual muscle test to assess the strength of the middle deltoid muscle, (b) the FABER test for hip conditions, and (c) the Slocum drawer test to assess rotational knee instability. Ten presentations were evaluated on two separate occasions by 13 (n= 5 first-year; n= 8 second-year) participants. Data collection occurred through the use of an adapted peer-reviewed check-off sheet from an athletic training textbook. The data collection sheets used a Yes/No format for the nine components to each of the skills assessed. Accuracy scores were determined via comparison to the principal investigator's assessment of the videos, which was reviewed by a panel of five certified athletic trainers to ensure accuracy on the part of the principal investigator. Reliability measures were determined through a generalizability (G) study, followed by a decision (D) study in order to produce a summary coefficient (φ), which is similar to a reliability coefficient used in classical test theory.

The authors (Marty et al., 2010) found high levels of accuracy (middle deltoid manual muscle test = 96.84%, FABER test = 94.83%, and Slocum drawer test = 97.13%) among the participants in evaluation of their peers. There were no differences between the two groups of students in scoring their peers. Thus, academic year did not influence the students' abilities to assess their peers accurately. The athletic training students, however, were not found to be reliable assessors when only one occasion was used for the measure, with one-time reliability values ranging from φ = 0.37- 0.86, with only one-third of measures meeting the minimally accepted value of φ = 0.70. Reliability measures did improve when multiple assessors were used on multiple occasions.

As part of the D study, summary coefficients for multiple participants on multiple occasions were determined. Findings included improved scores of φ = 0.79 when the FABER test was assessed by two participants on three occasions, φ = 0.76 when the

Slocum drawer test was assessment by one participant on two occasions, and $\varphi = 0.72$ when the middle deltoid manual muscle test was assessed by three participants on two occasions. All of these findings exceeded the minimally accepted standard of $\varphi = 0.70$ (Marty et al., 2010).

Findings were compared to multiple studies performed in both medical and allied health education, most of which were from the 1970s and 1980s, which demonstrated inconsistent findings regarding accuracy and reliability of students to assess their peers on clinical skills (Marty et al., 2010). This first study of peer assessment among athletic training students provided a starting point for the discussion to include peer assessment practices in athletic training education (Marty et al., 2010). Peer assessment accuracy and reliability of athletic training students must be investigated further in order to determine whether peer assessment is an appropriate evaluation tool in athletic training education. The proposed study seeks to identify if undergraduate athletic training students are accurate and reliable evaluators of their peers, as compared to instructor evaluations, on clinical skills and professional behaviors.

The authors (Marty et al., 2011) followed up their 2010 article with a study that examined the accuracy and quality of feedback provided by athletic training students during psychomotor skills practice sessions. All participants had prior experience with peer assessment of psychomotor skills in previous athletic training coursework. Eleven students enrolled in an entry-level master's athletic training program (6 first-year students, 5 second-year students) participated. Participants evaluated 10 video presentations of a peer performing the FABER hip pathology test on two separate occasions. The presentations were viewed one week apart from each other. The authors

(Marty et al., 2011) once again used percent correct scores to determine accuracy of the participants in evaluating the clinical skills of their peers. The authors then categorized feedback given by the participants as comments that addressed either correct or incorrect performance items. Quality of the feedback given by participants was categorized as either general or detailed.

Accuracy measures were judged against the evaluations of the principle investigator, then were reviewed by a panel of five certified athletic trainers to ensure accuracy on the part of the principal investigator. Results for the accuracy measures were high, with an average of 97.83% (97.58% for first set of evaluations, 98.08% for second set of evaluations) (Marty et al., 2011). All participants scored greater than 80% accuracy, which was the minimum standard set by the authors. Five assessments reached 100% accuracy during the course of data collection. The authors also found no significant difference in accuracy scores between sessions (F[1,9] = 0.30, p = 0.57). Similar to the 2010 Marty et el. study, the main effect of academic year in the athletic training education program was also not significant (F[1,9] = 1.88, p =0.20). There were no differences between first- and second-year students in either accuracy of peer assessment or quality of feedback provided to classmates.

Feedback results contained 451 total comments provided by participants (Marty et al., 2011). Feedback on incorrect items accounted for 297 of these comments, and 154 comments were given for correct items. Incorrectly performed items were given feedback 90% (n = 330) of the time, while no incorrect feedback was given on incorrectly performed items. Additionally, no feedback was provided for 77.22% of the total items (n = 1,980) and correctly performed items only received feedback 9.33% of the time (n =

1,650). Participants provided feedback for 905 of the items that were performed incorrectly. Detailed feedback was given 54.32% (n = 451) of the time and general feedback was offered 45.68% (n = 451) of the time.

The study by Marty et al. (2011) expanded on the authors' 2010 study by including the feedback provided to students during peer assessment procedures. Conclusions drawn by the authors (Marty et al., 2011) agreed with the accuracy findings of the 2010 Marty et al. study. Athletic training students were accurate peer assessors of clinical skills. The authors further concluded from the 2011 study that athletic training students also provided accurate, if not consistently detailed feedback to their peers.

These two studies are the only currently published articles on peer assessment using athletic training students. Although both had entry-level master's students as participants, the skills investigated are similar to those that will be used to determine the ability of undergraduate athletic training students to accurately and reliably assess the clinical skills of their peers. The proposed study will also investigate undergraduate athletic training students' abilities to accurately and reliably evaluate the professional behaviors of their peers. Neither of the two currently published peer assessment studies in athletic training education investigated professional behaviors among participants (Marty et al., 2010; Marty et al., 2011).

Peer assessment in medical education.

The use of objective structured clinical examinations (OSCEs) is common practice in medical education (Chenot et al., 2007). They are time consuming and require a much higher faculty workload than giving a traditional written examination, thus placing a greater demand on teaching faculty. Chenot et al. (2007) performed a study to investigate the use of peer examiners of third year medical students during an OSCE, and to also determine the acceptance of peer assessment among the students.

Twenty fourth- and fifth-year medical students and 25 teaching doctors were trained prior to data collection. The training included sample video presentations and detailed instruction regarding how to evaluate the performance of the students during an OSCE. Four stations of the OSCE were included in data collection. These stations were cardiovascular risk assessment, electrocardiogram, depression screening, and occupational assessment. The data collection instrument consisted of checklist items for individual skills and a global rating scale for overall performance at each station. Following the OSCE, all 214 participants were asked to respond to a questionnaire regarding their feelings towards the OSCE itself and provide feedback regarding the peer assessment process (Chenot et al., 2007).

Overall, the students were more lenient in their evaluations than teaching doctors were, scoring their peers slightly higher (.02- .20 on a 5-point Likert scale) for checklist and global ratings. Inter-rater agreement, calculated by kappa values and paired *t*-tests showed a range of 0.41- 0.64 for checklist and global ratings for the four stations, resulting in moderate to good agreement levels between groups (Chenot et al., 2007).

The questionnaire had a 90% rate of return from the 214 participants. The majority (90%) of respondents had no prior experience with OSCEs. Feedback from the students on the use of peer assessment on the OSCE was overall positive. Most (69%) students believed their peers would be objective graders and that they themselves would also be objective when participating as a peer assessor (95%). Additionally, 64% of respondents believed that peer assessed evaluations would result in the same grade as teacher evaluations (Chenot et al., 2007).

Distinguishing itself from other research performed on peer assessment in medical education, a study performed by Machado, Machado, Grec, Bollela, and Vieira (2008) tracked peer, self and instructor assessment grades over the course of seven semesters as part of summative assessment practices. All participants were first-year medical students, with the total number reaching 349.

The authors used an ANOVA and post hoc test for statistical analysis. The mean values for peer assessment and self-assessment grades increased six out of seven semesters, but no statistically significant difference was found between the grades of the two groups ($r = 0.806$). The authors determined that peer and self-evaluations were consistently higher than instructor grades over the course of the study, and both sets of grades differed from the instructor assessed grades every semester. Significant differences were found between the instructor group and each of the student groups (Machado et al., 2008). The correlation values between each student group and the instructor group demonstrated lower values. The instructor assessment - peer assessment correlation value was $r = 0.456$, while the instructor assessment - self-assessment correlation value was $r = 0.376$. The low correlation between the two student groups and the instructor group led the authors to conclude that peer and self-assessment were not valid tools in the summative assessment process.

Both Chenot et al. (2007) and Machado et al. (2008) found students to be more lenient evaluators than instructors. However, Chenot et al. (2007) believed medical students had the ability to assess their peers accurately during a practical skills exam.

Machado et al., on the other hand, concluded that peer assessment was not a valid tool for student evaluation. These contradictory conclusions may be due to the difference in the type of peer assessment investigated. Chenot et al. (2007), used peer assessment as a one-time formative evaluation tool, whereas Machado et al. (2008), used it as a summative tool. The longer length of the Machado et al. study provided for seven semesters of data collection, resulting in a more longitudinal study. It can be argued that this difference in data collection lead to a more valid conclusion when compared to the findings of Chenot et al. (2007).

The proposed study into peer assessment of undergraduate athletic training students will be performed within one semester, therefore the reliability findings may be suspect due to the lack of longitudinal data. A longitudinal design is not practical for the current study due to the structure of the athletic training education program from where the participants are going to be recruited. The program is a two-year program; therefore, it consists of four semesters, which is not enough time to conduct a longitudinal study using the same participants throughout.

Peer assessment in allied health care education.

In comparison to the number of studies performed using peer assessment in medical education, there are far fewer in the literature that pertain to the use of peer assessment in other health care education fields. Evans et al. (2007) investigated to determine if peer assessment was more reliable than self-assessment when compared to instructor assessment during oral surgery. Dental students were asked to perform a third molar extraction procedure and were evaluated simultaneously by a classmate and an instructor on both technical skills and affective traits.

A total of 38 participants were observed by 19 peer assessors and five instructors. Data was collected through an itemized checklist and global rating scale. Paired *t* tests were used to identify differences between assessors, followed by Lin concordance coefficient to determine reliability of scores among the groups. The authors found no statistically significant difference between the evaluations of the instructors and peers. Inter-rater reliability between the instructors and peers on the checklists was r= 0.92, and r= 0.91 for the global rating scale. These values suggested an excellent level of agreement between the two groups (Evans et al., 2007). In contrast, when the scores from the self-assessments were compared to the evaluators' scores, there was only a moderate level of agreement for both the checklist and global rating scales (r= 0.55 for both scales). The differences in the mean scores for the self-assessments were significantly different than those of the evaluators. The participants assessed themselves higher than both the instructors and peer assessors.

Similarly positive findings in the use of peer assessment were found by Bucknall et al. (2008) when testing basic life support skills during a final practical exam. In addition to testing the reliability of student scores compared to instructor scores, the authors inquired about students' attitudes towards peer assessment. Participants (n=162) were each evaluated by a faculty member and peer on individual skill items and global pass/fail criteria. All assessors were certified course instructors for basic life support and automated external defibrillation through a national certifying agency.

The authors used percentage agreement between the two groups of evaluators to determine inter-rater reliability. Findings included better than 95% agreement for all individual skill items, except for the skill of chest compressions, which had a 93% level of agreement. Summative global pass/fail agreement level was determined to be 86%. Peer assessors had a lower pass rate (71%) than the instructors (82%), accounting for the difference between global and individual skill agreement levels. Also, for the pass/fail rates, peers gave failing grades when instructors gave passing grades 20 times, whereas peers gave passing grades three times when an instructor gave a failing grade. Using the instructor scores as a gold standard, the authors surmised a sensitivity rate of 85% and a specificity rating of 90% for peer grades. Additionally, a positive predictive value of 97% was found for the probability that a peer assessed passing grade was a true passing grade (Bucknall et al., 2008).

Responses to the questionnaire showed that 76% of students reported preferring peer assessment to instructor assessment. Also, the majority of students believed their peers were competent evaluators of their skills. Finally, anxiety levels of students being assessed by their peers were found to be neutral and of no consequence to student performance (Bucknall et al., 2008).

Both Bucknall et al. (2008) and Evans et al. (2007) had positive results between peer assessed and instructor-assessed evaluations of clinical skills. Evans et al. (2007) had excellent agreement levels for both individual skills and global ratings. Bucknall et al. (2008), on the other hand, had better agreement between groups on individual skills than global passing rate. When compared to the earlier reported findings of Falchikov and Goldfinch (2000) that student grades were more similar to instructor grades when global judgments were made, both Bucknall et al. (2008) and Evans et al. (2007), differed.

The proposed study will use individual skill ratings for the clinical skills assessment and global ratings for professional behaviors. The findings from Falchikov and Goldfinch (2000), Bucknall et al. (2008), and Evans et al. (2007) are interesting to the proposed study because of the conflicting findings of these authors. How well the results of the clinical skills and professional behavior scores correlate between peers and instructors can help support the use of one method of scoring over the other, and add to the findings of these authors.

Peer Assessment of Professional Behaviors in Medical Education

Research performed on peer assessment of professionalism in medical students has covered a broader range of research questions than the clinical skills-related research performed in recent years. To date, there have been no published studies of peer assessment of professionalism among athletic training students. However, professionalism traits of medical students have been investigated often. In addition, students' perceptions of the peer assessment process, and the impact peer assessment may have on the students involved in the research have also been studied within medical education.

Hulsman et al. (2013) investigated peer assessment of medical students' communication skills compared to instructor assessment. Second-year medical students were trained in patient history-taking during their first semester. Included in the training was instruction on the importance of verbal and non-verbal active listening skills, principles of effective feedback, and one video-recorded history-taking session with a simulated patient for personal review.

In the second semester of the year, all students were once again recorded while taking a history from a simulated patient. Students shared their video with classmates and instructors for evaluation. Approximately four weeks after evaluation, students received peer and instructor feedback in a meeting with peers and an instructor. After the meeting, students were asked to complete a questionnaire about the evaluation that included items regarding communication in history-taking skills, personality domains, social and academic reputation, and perceptions of peer assessment.

A total of 320 students participated in the evaluation, with 244 questionnaires providing complete data-sets for analysis (Hulsman et al., 2013). Pearson correlations and *t*-tests were used for bivariate analyses between instructor and peer scores. Two key results emerged. The first, peer ratings were significantly higher than instructor ratings ($t= 6.4$; $p< .001$) for communication skills for global ratings and history-taking subscales. Global ratings between instructors and peer correlated significantly, but weakly ($r= 0.28$; $p<.001$). Second, peer scores were related to academic reputation, but not social reputation. Instructor scores were also related to academic reputation ($r= 0.26$; $p <.001$), but not to social reputation.

The authors (Hulsman et al., 2013) noted that peer assessment scores did not replicate instructor scores for summative assessment in communication skills, as students were more lenient than instructors, despite the significant correlation on the global scale between the groups. Also, the authors expected students to be vulnerable to the social and academic reputations of their peers when scoring their communication skills, but found this to be untrue.

Kovach, Resch, and Verhulst (2009) investigated peer assessment of medical students' professionalism in order to determine if peer scores correlated to traditional performance measures and faculty scores on professionalism. Anonymous student peer assessment of professionalism accounted for 20 percent of a non-cognitive behaviors grade, and seven percent of a final grade for a medical clerkship. Students and faculty used the same 5-point rating scale instrument over the course of five years for data collection.

A total of 349 student peer ratings were compared to faculty ratings. Descriptive measures, correlations, paired *t*-tests, and analysis of variance were used for data analysis. Mean scores for peer ratings (4.18) were lower than faculty mean scores (4.27, $p<0.001$), with a weak correlation ($r = 0.29$, $p<0.001$), for professionalism. There was also a weak, but statistically significant correlation, between peer ratings for professionalism and performance on traditional assessment measures that included faculty ratings of clinical skills ($r = 0.28$), performance on a competency exam ($r= 0.30$), and election to a medical honor society ($r = 0.24$).

Students were allowed to write comments on their peer evaluation forms. The authors (Kovach et al., 2009) noted some striking differences between peer and faculty comments about the same student on a number of occasions. Despite 41% of students commenting that they inflated their peer assessment grades, results showed the students were tougher graders than faculty. The authors believed peer assessment of professionalism provided value to the final grade for the medical clerkship, as students did grade similarly to faculty and were able to provide a unique perspective of their peers.

Hulsman et al. (2013) and Kovach et al. (2009) used quantitative analyses to compare peer-assessed scores to faculty scores with conflicting results. Despite a weak, but significant correlation, medical students were significantly more lenient than instructors on grading communication skills (Hulsman et al., 2013). Alternately, medical students were found to be harsher graders than instructors for professionalism during a clerkship.

Recently, a qualitative study looked at the impact of peer assessment on the professional development of second- and fourth-year medical students (Nofziger et al., 2010). Narratives were used to determine what types of feedback were most memorable and what reactions or transformations were experienced as a result of peer assessment. Peer assessment was a built-in component of the students' formative, comprehensive assessment process, thus all subjects had experience of at least one year with peer assessment practices.

Responses were themed and coded (Nofziger et al., 2010). The results were put into the following categories: (a) content of peer assessment, (b) cognitive reactions to peer assessment, (c) emotional reactions to peer assessment, and (d) personal transformations related to peer assessment. The authors found 73% of second-year students and 63% of fourth-year students remembered specific feedback they received during peer assessments, including both negative and positive forms of feedback. Sixty-seven percent of the 183 respondents believed peer assessment to be reassuring, confirming, and helpful of something they already knew. Additionally, 65% of the total respondents reported they had increased awareness and improved attitudes and behaviors because of peer assessment.

An important conclusion formed by the authors was that peer assessment was a great tool for the formation of professional behaviors, especially interpersonal skills, as reported by the subjects (Nofziger et al., 2010). The authors also recommended providing training for peer assessment participants prior to implementation, particularly in the area of providing high quality constructive feedback.

A 2010 study performed by Garner et al. used focus groups to investigate undergraduate medical students' views towards the use of peer assessment of professional behavior. Two medical schools in England that used peer assessment as part of a problem-based learning curriculum in small groups participated. A total of four focus groups were formed from the two schools (n_{total}= 30). The interview sessions were transcribed, analyzed and coded for themes.

Training and preparation of students to give and receive feedback were found to be key aspects to the successful use of peer assessment among the students (Garner et al., 2010). The authors found discrepancy among the students when the preferred method of feedback was discussed. Some students preferred face-to-face feedback to allow for further discussion and explanation, but others believed this would cause complications and may affect personal relationships. Students also had mixed views about anonymity during the peer assessment process and how the evaluation information was used. It was revealed that paper forms allowed for some students to recognize the assessor through handwriting, thus compromising anonymity. In addition, the authors discovered some participants revealed the names of the students they were responsible for evaluating. The final notable finding of the authors was concern among the students that peer assessment evaluations would permanently affect their school records or become part of their portfolios.

Based on the overall positive attitudes of the participants, the authors were able to conclude that peer assessment can be a valuable feedback tool for formative learning. In particular, feedback on professionalism in undergraduate students could help ensure better clinical practice in the future for the participants (Garner et al., 2010).

The conclusions made by Garner et al. (2010) regarding the value of peer assessment as a feedback tool for the development of professional behaviors echoed the sentiment of Nofziger et al. (2010). In addition, both studies recommended training in the use of peer assessment, especially in the area of giving quality feedback, prior to its implementation. Earlier cited studies performed by Topping (2010) and van Zundert et al. (2010) supported the notion that peer assessment had better outcomes when training was incorporated into the peer assessment process.

A 2006 study by Lurie, Nofziger, Meldrum, Mooney, and Epstein sought to investigate the longitudinal stability of peer assessment ratings, and the differences between multiple groups using the same evaluation tool for peer assessment of professionalism traits. The authors followed two consecutive classes (2003 graduates and 2004 graduates) of medical students for the entirety of their second and third years of medical school. All participants (162) evaluated 6-12 of their classmates towards the end of each year using the school's standard assessment form for professionalism. The outcome measures used by the authors were based on mean numerical ratings on scales of professional work habits (WH) and interpersonal attributes (IA).

Reliability measures for each scale were calculated using Cronbach's alpha (α). Results showed high internal consistency for both groups in both scales (Lurie et al., 2006). Second-year WH data showed an α value of 0.84 that increased to 0.89 in the third-year data. IA data had an α value of 0.94 for second-year students and a value of $\alpha = 0.92$ for third-year students. Also, scores from the second year were found to be predictive of third year scores for both scales. Stability of individual ratings for the two scales between years two and three were determined through the use of Pearson's correlation coefficient. For the graduating class of 2003, the WH scale's *r* value was 0.71 from Year Two to Year Three, and 0.56 for the IA scale. The 2004 graduating class had similar, but slightly lower correlation values. Their WH scale *r* value was 0.55, while the IA scale was *r* = 0.65. The authors found no statistical difference between correlations between years.

Additional findings by the authors (Lurie et al., 2006) included third-year students scored consistently higher than second-year students for both scales, but all scores were highly correlated, even with different students doing the evaluating between years. Also, second-year medical students rated highly by their peers were likely to be rated highly in their third year, and second-year students receiving the lowest peer assessment ratings showed improvement during third-year ratings. One inconsistency noted by the authors was that each group was more discriminating with one scale than the other group. This finding led the authors to note that individual groups may differ in their ability to discriminate specific types of skills.

An investigation into whether or not participation in peer assessment improved professional behavior in medical students was conducted by Schönrock-Adema, Heijne-Penninga, van Duijn, Geertsma, and Cohen-Schotanus (2007). Occurring over a period of two consecutive trimesters, peer evaluators assessed their classmates in the domains of task performance, communication aspects, and personal performance using a Likert-type scale rating system of 1-10. Faculty derived professionalism scores of those students who assessed their peers were then compared to the professionalism scores of the students who were not peer assessors during the study.

None of the participants had prior experience or training in peer assessment. The first trimester had 278 participants and the second trimester had 272 participants. For each trimester all participants were randomly assigned to a group and each group randomly assigned to a condition of either peer assessment or no peer assessment. The students assigned to be peer assessors were given verbal and written instructions prior to data collection in order to improve the reliability of the assessments (Schönrock-Adema et al., 2007).

Results showed that those students who were peer assessors demonstrated greater improvement in their professionalism scores, as rated by faculty, than those students who were not peer assessors (Schönrock-Adema et al., 2007). Scores from both groups in the first trimester were approximately the same. The second trimester, however, did show slightly improved scores for the peer assessment condition. A learning effect was found to exist from the first trimester in the domains of task performance ($z = 3.34$, $P< 0.001$) and personal performance ($z = 1.69$, $P<0.05$). No significant effect was found for the aspects of communication domain.

The results partially supported the authors' hypothesis that students who assessed their peers would score higher than those who did not evaluate their peers on professional behaviors. The first trimester's data did not demonstrate the peer assessors to be more highly rated by faculty on their professionalism. The authors asserted that adjusting to a more complex learning environment and learning the new method of assessment may have contributed to this difference in the data between the two trimesters, but believed the students showed improvement in professional behavior once they were acclimated to the peer assessment process, which occurred in the second trimester (Schönrock-Adema et al., 2007).

The Lurie et al. (2006) and Schönrock-Adema et al. (2007) studies, performed over multiple courses provided information to the effect peer assessment had on professionalism. Although the findings of Schönrock-Adema et al. only partially substantiated their hypothesis that medical students who engaged in peer assessment would have higher professionalism scores than their peers, the authors did see an improvement in this group's scores from one trimester to the next. Also, Lurie et al. demonstrated a high level if consistency from one year to the next in the work habits and interpersonal skills of medical students.

Peer Assessment of Didactic Skills in Medical and Allied Health Education

A method for implementing peer assessment in an undergraduate nursing program and report on their findings was described by Casey et al. (2011). The purpose of the peer assessment was to improve student engagement. Thirty-seven second-year nursing students developed marking criteria and graded two of their peers' assignments anonymously. Qualitative interpretive descriptive design was used through six focus groups. The focus group interviews were conducted two months after completion of the peer assessment process. Each focus group had an average of six participants and was led by two facilitators who used an interview guide. None of the participants had prior experience with peer assessment.

The interview transcripts were coded and themed by three of the authors (Casey et al., 2011). Three themes emerged from the data. First, impact on student engagement lead to enhanced student learning among the participants. Students experienced greater involvement in their learning, had more confidence in evaluating their own work, and were more eager to learn, with the majority of participants reporting they enjoyed the learning experience. The second theme reported by the authors was challenges of peer assessment. Students reported that the developed marking criteria did not work as well as had been anticipated and there were some concerns over feedback interpretation. Students did not want to be viewed as mean by their peers and were reluctant to give a poor grade, even when the grade was merited.

The final theme focused on making peer assessment better, based on the recommendations of the participants. Some students believed the 15% awarded through peer assessment for the assignments was too high, while others believed this amount gave weaker students a chance to improve their grades. Confidentiality should have been emphasized more, especially regarding discussion of the grading process with others. These two points concerned this particular peer assessment investigation and the authors (Casey et al., 2011) noted they should not be projected onto peer assessment as a whole.

The report by Casey et al. (2011) described a successful implementation of peer assessment among undergraduate nursing students. The authors asserted their program was supported by the self-regulation theory of learning which focused on the responsibility and autonomy students have in their learning. The authors' findings demonstrated peer assessment practices to enhance learning and strengthen the capacity to learn. The implemented peer assessment also prompted more critical thinking and reflection among the participants.

In a 2006 study, first-year medical students were found to be only slightly harsher graders of their peers than instructors (English, Brookes, Avery, Blazeby, and Ben-Shlomo, 2006). After being supplied with a model answer and grading criteria, the students assessed an essay-type exam about data interpretation. A total of 289 students participated in the study over the course of two years. Random assignment to groups resulted in 147 peer assessors and 142 control subjects. The treatment occurred blinded and anonymously. The first year of the study yielded a mean difference of 2.2% between faculty and peer grades, with peers grading more harshly than faculty. The second year had a wider gap between faculty and students, with peers grading, on average, 5% lower than faculty.

The authors (English et al., 2006) also sought to determine if peer assessment had any effect on exam performance and determined it did not. Two months after the treatment, all participants took an end-of-year examination. Students who were assigned to the treatment group performed only 1.5% higher than the control group on the final exam. Given the 95% confidence interval the authors used, they concluded there was insufficient evidence to suggest that peer assessment had a time effect on exam performance.

Interviews conducted with the participants after the conclusion of the study proved students gained insight into the evaluation process, but not necessarily a deeper understanding of the content that was tested after the use of peer assessment (English et al., 2006). Only six participants accepted the opportunity to join the focus group at the conclusion of data collection. Two facilitators used a topic schedule for the recorded interviews and transcripts were recorded, coded and themed. Supported by the reports of

Garner et al. (2010), Topping (2010), and van Zundert et al. (2010), the authors found that students believed they needed training in proper marking procedures in order to improve the validity of their grades. Also, even though they were provided with a grading scheme, students were concerned about giving an incorrect grade. Finally, the students expressed the opportunity to assess their peers provided help in preparing for the final exam. Specifically, the provided grading scheme allowed them to understand the grading process better.

A similar study, performed by Langendyk (2006) also required students to peer assess a classmate's essay response, and were given a model answer and grading criteria. In this study, however, each student also performed a self-assessment of their own essay. Faculty evaluated all essays using the same criteria as students following the peer assessment portion of the study.

Participants numbered 175 and all were third-year medical students. Results included mean scores (95% confidence interval) of 56.8 for self-assessment, 58.8 for peer-assessment, and 58.3 for faculty assessment. Paired *t*-tests were used to determine statistical significance between groups and the Pearson correlation coefficient determined the relationship between the three groups. The author found moderately strong correlations between the self-assessment and faculty assessment groups (r= 0.55, P< 0.01) and the peer assessment and faculty assessment groups (r= 0.63, P<0.01). Additionally, peer assessment did not differ significantly from faculty assessment, as demonstrated by a mean difference of -0.5% (P= 0.39) (Langendyk, 2006).

The author (Langendyk, 2006) concluded that the majority of third-year medical students were capable of accurately assessing themselves and their peers. However, additional outcomes showed that the students with a history of high achievement scored themselves more harshly than faculty members, and were accurate peer assessors. Lower achieving students, however, scored themselves and their peers more leniently than faculty members.

Although the purposes of the English et al. (2006) and Langendyk (2006) studies differed, the author's findings regarding students' abilities to assess their peers on an essay were similar. The results from English et al. (2006) showed a difference of, or less than, 5% between peer and faculty grades, with peers grading slightly lower than faculty. Langendyk (2006) found moderately strong correlations between peer and faculty grades, with no statistically significant difference between the two groups.

Peer Assessment in Undergraduate Students

A recent study to determine the effectiveness and reliability of undergraduate physiology students to grade a peer's laboratory report in a large class environment was conducted by Harris (2011). Conducted over two years, two cohorts were formed with approximately 180 students per cohort. The lab reports consisted of both close-ended and open-ended questions. The peer assessment session took place two weeks after the lab was performed and 24 hours after students submitted their final reports to the course instructor. The lab reports were randomly, but not anonymously, distributed to the class for grading. Peer assessors were also known to the original owner of each lab report. After distribution, the course instructor guided the grading process of the students through the use and explanation of correct answers for each question on the lab report.

Following grading, a sample from each cohort (40 of 172 reports and 28 of 185 reports, respectively) was used for data analysis. The author (Harris, 2011) found students to grade higher than faculty on an average of 2.5% during the first year. Second-year data showed a 2.9% difference in grades, with students again grading higher than faculty. It was found that over 80% of the students graded their peers higher than faculty. These differences were found to be statistically significant at a *P* value of 0.001, but had excellent correlation both years ($r = 0.96$ and 0.98, respectively). The correlation values between student and faculty marks were higher than other studies performed in the sciences, which the author mentioned may have been due to the structured format of the grading process. The author noted that the occurrences of over-marking by students occurred predominantly in the sections of the lab reports that were open-ended questions. These were questions where greater amounts of critical judgment needed to be used by the graders (Harris, 2011).

Student feedback on the peer assessment process was overall positive (Harris, 2011). Students noted their understanding of the content improved, as did their understanding of how best to present a lab report for grading. Seventy percent of participants were comfortable with peer assessed grades contributing a small amount (5%) to the final course grade.

The author (Harris, 2011) concluded that peer assessment can be a reliable tool for large classes. Faculty grading time was decreased by 95%, as all the reports were graded by the students in less than one hour. The comparison of mean scores by students and faculty grades was similar to other studies, which demonstrated students to be more generous in grading than faculty (Chenot et al., 2007; Harris, 2011; Machado et al., 2008). These findings are in contrast to the previously reported finding of English et al. (2006) that stated peers graded more harshly than faculty.

Student perceptions and experiences with peer assessment were explored by Vickerman (2009). Ninety students enrolled as undergraduates in sports studies at a university in the United Kingdom participated in the study designed to determine if peer assessment, when used for the first time, effected learning development. As part of their coursework, the participants were required to write four annotated bibliographies based on journal articles about social inclusion in sport. Of these, two assignments were graded using peer assessment and two were graded by the course instructor. At the conclusion of the course, participants were asked to complete a questionnaire that consisted of 12 Likert-type items and four open-ended questions about the peer assessment process. All of the participants completed the questionnaire for data collection.

Results showed that 55% of the students either agreed or strongly agreed that their knowledge and understanding of the course topic was improved through peer assessment (Vickerman, 2009). Next, 58% of the students stated they gained confidence in student-led discussion, independent learning skills, and most enjoyed sharing ideas and concepts with peers. Also, 77% of students agreed or strongly agreed that their referencing skills were more effectively enhanced by peer assessment. Only 3% of students reported they did not like the peer assessment process. Reasons for their displeasure included a greater need for teacher support, rather than self-directed learning. Almost half (43%) of the students agreed or strongly agreed they would like to have greater involvement on the assessment process in future courses (Vickerman, 2009).

Findings from Vickerman's (2009) study supported the use of peer assessment as a formative learning tool. The positive feedback from participants reinforced the notion that peer assessment provided students with the opportunity to increase their confidence and autonomy in their learning. Improvement is these areas correspond well to the requirement of athletic training education programs to ensure their students are able to integrate the knowledge and skills necessary to make sound clinical decisions for patient care, including diagnoses, treatment strategies, and when necessary, referral to other healthcare professionals (CAATE, 2012; NATA, 2011).

Multiple methods of peer assessment for writing assignments in undergraduate history courses were used by Van den Berg, Admiraal, and Pilot (2006) in order to determine effective ways to incorporate peer assessment practices into the curriculum. The authors noted effective peer assessment methods were those that were easily implemented and provided optimal learning outcomes. Seven types of peer assessment activities were incorporated into seven courses based on 10 design features of peer assessment. Nine instructors and 168 students participated. All classroom activities were monitored to ensure peer assessment implementation during the course of the study. Learning outcomes included the revisions students made, the grades assigned to the written products, and the students' perceived progress of their products and writing skills.

Results showed that most students believed their writing improved, as did their ability to process peer feedback (van den Berg et al., 2006). Instructors also reported improvements in student interaction in their classes when peer assessment practices were used. For the three courses that implemented peer assessment as a means to compare student grades to instructor grades, there was no statistically significant difference in the two groups of grades.

The authors (van den Berg et al., 2006) concluded three design features were most beneficial for the use of peer assessment in writing. The first was the allowance of sufficient time for revision by the students before instructor assessment. This allowed for greater correlations between student and instructor grades. Second, peer feedback among students should be two-way and reciprocal in nature. In this situation, each student involved was the assessor and the assessed, which allowed for better exchange of ideas and concepts. Last, the authors determined the ideal number of students in a peer feedback group was between three and four. The use of only two students allowed for feedback from only one peer. Feedback from more than one person provided the opportunity to compare remarks for better processing by the students.

A case study that investigated the use of peer assessment in an undergraduate sociology program and focused on the students' experiences was reported by Vu and Dall'Alba (2007). Nine of 11 second-year students enrolled in a communications and personnel relations course participated in the semester-long study. Seven of the nine participants had prior experience with peer assessment practices. Peer assessment was used to evaluate a videotaped interview of each student with the course instructor. The student in each interview represented a stakeholder for a particular position and was made to defend that position against the course instructor. Peer assessment was used for the videotaped portion of the course in order to promote the students' abilities to evaluate interview performance, and give and receive feedback in a professional manner.

All participants were given instructions on the use of a checklist for evaluating the interviews of their peers. The checklists were completed and students wrote comments on the forms after each interview presentation. Items evaluated through the checklist included accuracy, confidence in presentation, and fluency of argument (Vu & Dall'Alba, 2007).

Data collection was conducted using five methods (Vu & Dall'Alba, 2007). Analysis of the institution's assessment policy, a two-part questionnaire, classroom observations, focus group interviews, and private interviews with the course instructor were all taken into consideration by the authors in determining their findings. The authors concluded that there were several conditions that supported the effectiveness of peer assessment. These included: (a) adequate and appropriate preparation of the students, (b) alignment of the peer assessment, learning objectives, and broader course purposes, (c) availability of the course instructor for assistance with the peer assessment process, and (d) incorporation of constructive discussions following peer assessment practices.

The pre- and post- treatment ratings of the participants demonstrated overall approval for the use of peer assessment (Vu & Dall'Alba, 2007). Students improved their ratings in multiple areas. Students believed peer feedback was as useful as instructor feedback. Peer assessment was also seen to offer ways for students to learn from each other and enhanced their understanding. Students also felt the increased workload of using peer assessment was worthwhile. Although the authors did not compare peer and instructor grades for accuracy measures, their findings supported the use of peer assessment with undergraduate students when practical, observable skills were evaluated (Vu & Dall'Alba, 2007).

Van den Berg (2006), Vickerman (2009), and Vu and Dall'Alba (2007) all had similar results for two effects of peer assessment. First, students in all three studies believed their content knowledge and understanding improved through the incorporation of peer assessment practices in their courses. In addition, participants in all three studies reported greater confidence and improved ability in learner-centered activities used a part of peer assessment. Such things included learning from each other, independent learning skills, and student interaction and feedback processing. These findings support the proposed investigation into peer assessment in undergraduate athletic training education. Athletic training students must be able to incorporate their classroom knowledge into clinical practice. In addition, as healthcare providers, athletic training students must demonstrate the ability to advance their professional knowledge and use good communication skills with patients, families, and other healthcare providers (CAATE, 2012; NATA, 2011).

Students' negative perceptions about an online peer assessment program for undergraduate writing were investigated by Kaufman and Schunn (2011). The authors looked specifically at the nature of the students' resistance, the factors that influenced their resistance, and how their perceptions impacted their revision work on assignments. An initial end-of-course survey was administered to 250 students in 10 classes across six universities that used the online peer assessment program SWoRD for writing assignments. This initial survey demonstrated a low-level of agreement among students when asked how much they agreed that it was reasonable to receive grades for peer assessment (2.3/5). Also, students generally disagreed with the statement that peer assessed feedback was acceptable instead of teacher feedback (2.6/5). The highest level of agreement occurred with a statement that involved improvement in the writing process based on recommended revisions from peers (3.5/5).

As a follow-up to their initial study, Kaufman and Schunn (2011) held semi-structured interviews with 84 students from one class that used SWoRD to peer assess two draft and final versions of two papers in their course. No instructor feedback or grades were used in the class for the assignments submitted through SWoRD. For the students interviewed, there were reservations regarding the fairness of peer assessed grades. Eighty-three of the 84 participants expressed some concern about their own, or their peers', qualifications to evaluate student writing. These reservations were primarily due to the fact that the instructor had no input into the feedback or grades for the papers that were peer assessed.

The negative perceptions developed by this group of students were echoed in the results of their post-course surveys (Kaufman & Schunn, 2011). Based on a 5- point Likert scale, the mean agreement value for the usefulness of peer feedback dropped from 3.9/5.0 to 3.5/5.0. For the validity of peer feedback, student agreement dropped from a mean value of 3.5/5.0 to 3.1/5.0. Finally, the level of agreement for peer assessment being a fair practice dropped from an average of 3.8/5.0 to 3.2/5.0.

The authors (Kaufman & Schunn, 2011) were able to conclude that negative perceptions of peer assessment were most closely associated with classes where the instructor did not participate in the grading portion of the peer assessment process. It was suggested that instructors participate on some level in order to allay the concerns of the students. However, the authors also discovered that the negative perceptions were unrelated to, and did not impact, the extent of the revision work performed by students.

A 2006 study by Liu and Carless examined the rationale for using peer feedback in the college classroom, with emphasis placed on its potential to enhance student learning. A large scale questionnaire and survey (1,740 students, 460 faculty) was used to acquire data about the use of peer feedback and peer assessment in college classrooms in Hong Kong. The authors argued that the preponderance of using grades as a means for peer assessment undermines the potential for student learning through the feedback students received from their peers.

Results showed that a significant number of faculty and students resisted the use of peer assessment practices that used grades, and that the majority of the faculty surveyed (70%) never or rarely had students grade each other. The authors scanned the responses to determine the reasons behind the lack of use of peer assessment in the classrooms. These reasons included the perceived questionable reliability of peer assessment grading, the balance of power over grades, and the perceived increase in workload on the part of the faculty (Liu & Carless, 2006).

Respondents cited students' limited content knowledge and evaluative experience as reasons for believing peer assessed grades were less reliable than faculty grades (Liu & Carless, 2006). Teachers were reluctant to give up their control of the grading system in their classes. Peer assessment practices redistribute power in the classroom. Students and faculty were both uncomfortable with the idea of students having authority over their peers' grades. Responses included the belief that peer assessment practices were generally more complex than teacher evaluation methods. Faculty believed that this required a greater amount of time to be devoted to grading when peer assessment was used.

The authors (Liu & Carless, 2006) conclusions, based on their findings, were that peer feedback practices had greater potential for improved learning, but conceded that a combination of the two practices may be needed for practical purposes. Included in their recommendations on how to achieve this was to imbed peer feedback into coursework.

This would then provide a good starting point to incorporate peer assessment towards the latter part of the course.

Summary

Peer assessment among pre-professional healthcare students has been investigated primarily in medical education, with mixed results concerning clinical skills (Bucknall et al., 2008; Chenot et al., 2007; Evans et al., 2007) and professional behaviors (Garner et al., 2010; Nofziger et al., 2010).Given the many different research designs and tools reported in the peer assessment literature (Strijbos & Sluijsmans, 2010; Topping, 2010), this is no surprise.

Current literature on peer assessment supports its use to improve critical thinking and problem-solving skills (Luxton-Reilly & Denny, 2010; Marty et al., 2010), independent learning and autonomy (Van den Berg, 2006; Vickerman, 2009; Vu & Dall'Alba, 2007), educational ownership (Gielen et al., 2011; Tillema et al., 2011), self-examination and reflection (Gielen et al., 2011; Tillema et al., 2011), and lifelong learning skills (Gielen et al., 2011; Kaufman & Schunn, 2011; Vickerman, 2009). These skills are all important in the development of the knowledge and skills needed by both medical and athletic training students in order to serve as healthcare professionals (ACGME, 2013; CAATE, 2012; LCME, 2012; NATA, 2011).

There is strong support in the literature to use peer assessment activities within medical education in order to prepare future physicians for a career in healthcare (Finn & Garner, 2011; Speyer et al., 2011), but this has not extended to athletic training education. The need exists to explore peer assessment in athletic training education in order to determine its effectiveness as a formative assessment tool that can be used to prepare athletic training students to become competent healthcare professionals.

Chapter 3: Research Method

Participation in peer assessment prior to entering the workforce fosters professional growth (Garner et al., 2010), and has been supported in the medical education literature as a method to prepare students to practice as fully competent professionals within the larger healthcare setting alongside other health professionals (Finn & Garner, 2011; Garner et al., 2010; Speyer et al., 2011). Athletic training and medical education share many skills and traits that are needed by healthcare professionals to provide quality patient care and engage with other healthcare providers, including: (a) performing physical examinations, (b) providing immediate care, (c) preventing disease, (d) communicating effectively with patients and families, (e) collaborating with other healthcare professionals, and (f) portraying professionalism and ethics at all times (CAATE, 2012; LCME, 2012; NATA, 2011).

Initial studies of peer assessment in athletic training education (Marty et al., 2010; Marty et al., 2011) have shown early evidence that athletic training students are valid evaluators of their peers during videotaped clinical skills demonstrations. However, there is a need to examine the validity and reliability of peer assessment using live participation and to compare students' ratings of clinical skills and professional behaviors to those of instructors in the undergraduate athletic training student population (Marty et al., 2010). This type of peer assessment will allow athletic training educators to better identify the quality of peer assessment among pre-professional athletic training students in order to implement peer assessment practices most effectively prior to entering the workforce. Athletic training educators need to research the use of peer assessment as an evaluation tool in order to better prepare students to practice as healthcare professionals, or risk placing students at a clinical disadvantage to their counterparts in the medical field after graduation.

The purpose of this quantitative study was to investigate the accuracy and reliability of undergraduate athletic training students to assess their peers on clinical skills and professional behaviors. The results of the proposed study further investigated peer assessment as an effective assessment tool for use in athletic training education. In comparison to instructors, if athletic training students accurately assess the clinical skills and professional behaviors of their peers, this initial investigation can be developed into larger, more complex studies.

Chapter 3 introduces the reader to the research design and methods used for this study of peer assessment among undergraduate athletic training students. A quasi-experimental, research design was used with a sample of athletic training students and instructors, all of whom were recruited from the same accredited undergraduate athletic training education program. The data collection instrument, processing and analyses are then presented. Finally, the assumptions, limitations, delimitations, and ethical assurances pertinent to the study's design and methods are explained. Eight research questions were answered during this study.

Q1. In relation to instructor scores, how accurate are junior-level students in scoring the clinical skills performance of undergraduate athletic training students for the a) Biceps Femoris Manual Muscle Test, b) Kleiger Test, c) Lachman Test, d) Noble's Compression Test, and e) Thompson Test?

Q2. In relation to instructor scores, how accurate are senior-level students in scoring the clinical skills performance of undergraduate athletic training students for the a) Biceps Femoris Manual Muscle Test, b) Kleiger Test, c) Lachman Test, d) Noble's Compression Test, and e) Thompson Test?

Q3. How reliable are junior-level students to each other in their ability to evaluate the clinical skills performance of undergraduate athletic training students for the a) Biceps Femoris Manual Muscle Test, b) Kleiger Test, c) Lachman Test, d) Noble's Compression Test, and e) Thompson Test?

Q4. How reliable are senior-level students to each other in their ability to evaluate the clinical skills performance of undergraduate athletic training students for the a) Biceps Femoris Manual Muscle Test, b) Kleiger Test, c) Lachman Test, d) Noble's Compression Test, and e) Thompson Test?

Q5. In relation to instructors, how accurate are junior-level students in scoring professional behaviors during clinical skills performance of undergraduate athletic training students?

Q6. In relation to instructors, how accurate are senior-level students in scoring professional behaviors during clinical skills performance of undergraduate athletic training students?

Q7. How reliable are junior-level students to each other in their ability to evaluate professional behaviors during clinical skill performance of undergraduate athletic training students?

Q8. How reliable are senior-level students to each other in their ability to evaluate professional behaviors during clinical skill performance of undergraduate athletic training students?

The following are the hypotheses for this study.

$H1_0$: There is no statistically significant agreement between instructor and junior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble's Compression Test, Thompson Test) performance ratings.

$H1_a$: There is statistically significant agreement between instructor and junior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble's Compression Test, Thompson Test) performance ratings.

$H2_0$: There is no statistically significant agreement between instructor and senior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble's Compression Test, Thompson Test) performance ratings.

$H2_a$: There is statistically significant agreement between instructor and senior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble's Compression Test, Thompson Test) performance ratings.

$H3_0$: There is no statistically significant agreement between instructor and junior-level student scores among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

$H3_a$: There is statistically significant agreement between instructor and junior-level student scores among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

H4$_0$: There is no statistically significant agreement between instructor and senior-level student scores among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

H4$_a$: There is statistically significant agreement between instructor and senior-level student scores among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

H5$_0$: There is no statistically significant agreement between junior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble’s Compression Test, Thompson Test) performance ratings.

H5$_a$: There is statistically significant agreement between junior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble’s Compression Test, Thompson Test) performance ratings.

H6$_0$: There is no statistically significant agreement between senior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble’s Compression Test, Thompson Test) performance ratings.

H6$_a$: There is statistically significant agreement between senior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble’s Compression Test, Thompson Test) performance ratings.

H7$_0$: There is no statistically significant agreement between junior-level students among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

$H7_a$: There is statistically significant agreement between junior-level students among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

$H8_0$: There is no statistically significant agreement between senior-level students among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

$H8_a$: There is statistically significant agreement between senior-level students among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

$H9_0$: There is no statistically significant agreement between instructor and junior-level students in professional behaviors ratings.

$H9_a$: There is statistically significant agreement between instructor and junior-level students in professional behaviors ratings.

$H10_0$: There is no statistically significant agreement between instructor and senior-level students in professional behaviors ratings.

$H10_a$: There is statistically significant agreement between instructor and senior-level students in professional behaviors ratings.

$H11_0$: There is statistically significant agreement between junior-level students in professional behaviors ratings.

$H11_a$: There is no statistically significant agreement between junior-level students in professional behaviors ratings.

$H12_0$: There is statistically significant agreement between senior-level students in professional behaviors ratings.

H12$_a$: There is statistically significant agreement between senior-level students in professional behaviors ratings.

Research Method and Design

This quantitative, quasi-experimental study of peer assessment among undergraduate athletic training students included both between-groups and within-groups designs for data analysis. The study fit this classification of design because the independent variables were selected by the investigator, the groups were not be randomly assigned, there was no control group, and all participants in the proposed study participated in the same treatment (Millsap & Maydeu-Olivares, 2009).

For the purpose of this study, a quantitative quasi-experimental procedure was the best choice to collect and analyze data for the accuracy and reliability measures, but establishment of a causal relationship was not needed to answer the research questions this study posed (Warner, 2008). In a true experimental design, there is a need to control for all extraneous variables in order to rule out other possible explanations for the results. A qualitative design would not have provided the best data to answer the research questions posed in this study.

Degree of accuracy of clinical skills assessment for each student group (junior-level students, senior-level students) was determined through calculation of Cohen’s kappa coefficient. Each data collection session, or triad, consisted of two students from the same group and one instructor viewing a live performance of clinical skills by a student of the same level as the two observing students in the triad. Student group ratings were compared to the instructor group ratings. Instructor scores traditionally represent how well a student performs on evaluations, therefore the instructor scores were used as the measuring stick to compare the student scores against for the accuracy measure. Actual correctness of the evaluated skills did not matter for the accuracy measure. This study measured whether students scored the skills similar to instructors. Reliability of clinical skills ratings was assessed through a measurement of Cohen's kappa coefficient. The scores of the students within each group were used to determine inter-rater reliability.

Accuracy and reliability of professional behaviors were examined through a linear weighted Cohen’s kappa coefficient. A weighted Cohen’s kappa measure is preferred over the standard kappa measure when using a Likert-type scale. The weighted kappa takes into account how much of a difference exists between rater scores (Vanbelle & Albert, 2009; Warrens, 2011). The dichotomous scale used for the clinical skills is best measured by the traditional Cohen’s kappa coefficient because with only two categories to choose from (Yes or No), the difference between scores was always one. The 5-point scale used to measure the professional behaviors needed a measurement tool that took into account the number of points on the scale the two raters may have differed from one another. The weighted kappa measurement essentially penalized the result when the categories chosen by the raters were farther apart than a difference of one, as seen in 2x2 contingency tables (Vanbelle & Albert, 2009; Warrens, 2011).

Population

The problem and purpose statements for this study pertained specifically to undergraduate athletic training students. For this reason, the study’s population from which the sample was chosen from, was undergraduate junior-level students, senior-level students, and the instructors (classroom faculty and clinical preceptors) responsible for grading these students.

Sample

This study of peer assessment was conducted with a convenience sample from an accredited undergraduate athletic training education program at a large university in the mid-Atlantic region of the United States. All students and instructors of the athletic training education program where the study was conducted were recruited for participation through email communication from the primary investigator. Three, relatively small ($n \leq 10$) groups; non-randomly assigned as junior-level students, senior-level students, and instructors served as participants, with a total participant number of 28. The groups were non-randomly assigned in order to best identify differences in peer assessment accuracy and reliability between the two academic levels of the student groups (junior-level, senior-level) and the instructor group, thus answering the research questions most effectively. Additionally, recruitment of participants from only one athletic training education program ensured the clinical skills used in the study were taught and assessed in a similar manner across all student participants.

The athletic training education program where the study was performed had a maximum enrollment of 20 students per class. The athletic training education program was an upper-division undergraduate program, so only two classes of students were enrolled in the program at any given point in time (junior-level and senior-level students). With only two small classes of students, the instructor group was correspondingly low.

Although an a priori power analysis is usually performed to predict sample size, given the known limitations in this area for the current study, power analyses for Cohen's kappa between two raters were performed using the known sample sizes using the PASS Sample Size Software (Version 13) (NCSS, 2014). For the clinical skills scored between students and instructors, the sample size was set at 20, alpha value at 0.05, two categories for ratings, and the value of kappa under the alternate hypothesis was 0.60 to reflect the upper margin of moderate agreement between raters (Landis & Koch, 1977). The resulting power was 0.82, with a Beta value of 0.18; signifying a low risk for committing Type I and Type II errors. For the clinical skills scored between students, the sample size was set at 10 with the remainder of the parameters the same as the power analysis for instructors and students. The resulting power was 0.47, with a Beta value of 0.53; signifying a moderate risk for committing Type I and Type II errors.

For the professional behaviors scored between students and instructors, the sample size was set at 20, alpha value at 0.05, five categories for ratings, and the value of kappa under the alternate hypothesis was 0.60 to reflect the upper margin of moderate agreement between raters (Landis & Koch, 1977). The resulting power was 0.99, with a Beta value of 0.01; signifying a very low risk for committing Type I and Type II errors. For the professional behaviors scored between students, the sample size was set at 10 with the remainder of the parameters the same as the power analysis for instructors and students. The resulting power was 0.84, with a Beta value of 0.16; signifying a low risk for committing Type I and Type II errors.

Additionally, post hoc power analyses were performed to determine the likelihood the true kappa values were calculated through analyses. For the clinical skills scored between students and instructors, the sample size was set at 20, two categories for ratings, alpha value at 0.05, beta value at 0.20, and power was set to 80%. The output confirmed these parameters achieved 80% power to detect a true kappa value of 0.59. For the clinical skills scored between students, the sample size was set at 10 with the remainder of the parameters the same as the power analysis for instructors and students. The output confirmed these parameters achieved 80% power to detect a true kappa value of 0.78. Both of these analyses produced results close to the target kappa value of 0.60 set on the a priori power analyses to signify a high level of moderate agreement between raters.

For the professional behaviors scored between students and instructors, the sample size was set at 20, five categories for ratings, alpha value at 0.05, beta value at 0.20, and power was set to 80%. The output confirmed these parameters achieved 80% power to detect a true kappa value of 0.42. This value did not meet the minimum kappa value of 0.60 in the a priori analysis to signify a high level of moderate agreement, but did fall within the moderate level of agreement range between 0.40- 0.60 (Landis & Koch, 1977). For the professional behaviors scored between students, the sample size was set at 10, with the remainder of the parameters the same as the power analysis for instructors and students. The output confirmed these parameters achieved 80% power to detect a true kappa value of 0.58. This analysis produced a result close to the target kappa value of 0.60 set on the a priori power analysis to signify a high level of moderate agreement between raters.

Inclusion criteria for each participant was they were either a junior- or senior-level student, or instructor, currently enrolled, or employed, within the undergraduate athletic training education program at the university where data was collected. All student subjects completed the coursework in which the clinical skills assessed during data collection were taught and evaluated within the athletic training education program.

Average time from completion of the related coursework to data collection was between six months and 18 months. Senior-level students had one year longer from completion of related coursework than junior-level students. All instructor participants were required to have experience in formal evaluation of clinical skills and professional behaviors of athletic training students for at least one academic year.

All student participants were potential model clinicians. Because this study examined peer assessment, the students scored their classmates on the observed clinical skills and professional behaviors. Every student enrollee participated in one triad for data collection, and was placed in a pool for random selection as a model clinician. No student was the model clinician more than once. Because two students of the same academic level were collecting data and only one student in the model clinician in each triad, only half of the students in the triad were selected to be a model clinician for each group of students (junior-level students, senior-level students.

The enrollees who were model patients during data collection were recruited from introductory courses in the university's undergraduate athletic training education program. These volunteers were college freshman and sophomores and were not currently enrolled in professional coursework as an athletic training major. The model patients did not need to speak or portray any particular signs or symptoms to the model clinicians during data collection. The model patients were required to follow the instructions of the model clinician and had no previous understanding of how to perform the clinical skills that were demonstrated. The instructions to the model patients included body positioning directions and relaxation of the involved body parts that were assessed during the clinical skills demonstration.

Instrument

There were currently measurements available in the field of athletic training education designed and tested for use in this study. The measurement tool used for data collection was adapted, with permission, from an athletic training textbook designed for athletic training clinical skills documentation (Amato et al., 2006). The textbook was developed by its authors as an accumulation of “best practice” steps to complete the clinical skills contained within the textbook. As the textbook from which the data collection instrument was taken included skills expected to be taught to athletic training students within their respective education programs, as opposed to specific instruments designed to measure outcomes, the authors (Amato et al., 2006) did not test the instruments for validity or reliability. The textbook was designed for use as a mechanism to track student progress in learning and mastering the skills, not as a formal measurement for research, hence the need to test the instrument prior to use in this study. Many evaluative tests have modifications that can make their use in a research study problematic, but the items chosen for this study did not have such modifications and the proper performance of each was not ambiguous (Kendall, McCreary, Provance, Rodgers, & Romani, 2005; Prentice, 2011; Starkey, Brown, & Ryan, 2010).

The textbook from which the data collection instrument (Amato et al., 2006) was adapted listed three items for global rating following each clinical skill. The primary investigator evaluated the three items and found them to be too vague for use as the professional behaviors evaluation of the current study. The primary investigator determined the addition of more items with specific areas of professional behavior allowed for more clear interpretation by study participants.

The data collection instrument (Appendix A) contained five clinical skills and seven professional behaviors. The instrument used a simple 2- point nominal scale for assessment of clinical skills and a 5-point Likert scale for global rating of professional behaviors. The 2- point nominal scale (Yes/No) was based on completion of the individual tasks needed for each clinical skill, as described on the data collection instrument. The 5- point Likert scale (5 = Always, a score of 4 = Frequently, a score of 3 = Occasionally, a score of 2 = Rarely, and a score of 1 = Never) was designed to measure the frequency with which the participants observed the clinician's professional behaviors during the performance of the five clinical skills. The professional behaviors were assessed at the conclusion of the performance of the five clinical skills and used as a global rating.

Each of the five clinical skills tests had three subscales: (a) patient position, (b) clinician position, and (c) test performance. The patient position subscale referred to the initial posture the model patient was instructed to assume in order to perform the clinical skill effectively. The clinician position referred the location of the model clinician relative to the patient during the clinical skill performance. Test performance related to the procedures used by the model clinician to perform the clinical skill on the model patient.

Raw data was used for the Cohen's kappa coefficient to determine accuracy and reliability of each student group to score their peers on clinical skills. The nominal dichotomous Yes/No scale used for clinical skills scoring supported the use of a 2x2 contingency table for Cohen's kappa (Howell, 2002; Warner, 2008; von Eye & von Eye, 2008). For each treatment, the triad of one instructor and two students of the same academic level collected data concurrently during the treatment. The individual components within each clinical skill scored the same between raters measured the level of accuracy (for instructor-student) and reliability (for student-student) of the students. The total number of items scored on the Yes/No nominal scale for each clinical skill ranged from six to ten. The three subscales (patient position, clinician position, test performance) were measured for accuracy and reliability across all five clinical skills. The total number of items within the subscales ranged from one to seven. Table 1 demonstrates the breakdown of scored items for each clinical skill and its corresponding subscales.

Table 1

Clinical Skills Breakdown by Subscale Items

Clinical skill test	Breakdown by subscale	Total items
Biceps Femoris Manual Muscle Test	Patient Position: 1 Clinician Position: 2 Test Performance: 7	10
Kleiger Test	Patient Position: 2 Clinician Position: 3 Test Performance: 5	10
Lachman Test	Patient Position: 2 Clinician Position: 2 Test Performance: 4	8
Noble’s Compression Test	Patient Position: 2 Clinician Position: 3 Test Performance: 3	8
Thompson’s Test	Patient Position: 2 Clinician Position: 1 Test Performance: 3	6

The seven professional behaviors scored on the 5- point Likert scale contained no subscales and therefore had a total of seven items scored for each treatment. The individual components of the professional behaviors scored the same between raters measured the level of accuracy (for instructor-student) and reliability (for student-student) of the students.

For each treatment, the triad of one instructor and two students of the same academic level collected data concurrently during the treatment. The participants were in view of each other, but were unable to see the data of the other participants. The model clinician and model patient during each treatment were be able to see the participants as they collected data, but were unable to see the data each participant collected.

Because this study investigated peer assessment among athletic training students, it was appropriate to use items that were taught and learned within an athletic training education program. All the clinical skills selected for inclusion were in the Amato et al. (2006) textbook because they are commonly used skills that athletic training students are expected to be able to perform in preparation of becoming a professional athletic trainer. Additionally, the professional behaviors used were based on the *Foundational Behaviors of Professional Practice* (NATA, 2011). These are considered basic professional behaviors that permeate professional practice. The National Athletic Trainers' Association stated these behaviors be infused into instruction and assessment throughout the athletic training education program.

The known limitation in the number of participants for the full study meant an even smaller number of participants were required in a pilot test to validate the data collection instrument. Because the group sizes for a pilot test needed to be very small (n= 2 for each of the three groups), sufficient data to validate the instrument properly could not be collected prior to use in the full study. A field test using participants who shared the characteristics of the full study's participants was logical method to test the instrument since the selected items were taught universally to athletic training students.

The participants for the field test included the senior-level athletic training students who were alumni at the time of data collection, athletic training educators, and clinical athletic trainers. A total of 25 people were contacted (15 students, seven athletic training educators, three clinical athletic trainers). Thirteen people provided feedback (seven students, three athletic training educators, three clinical athletic trainers).

Lack of validity testing of peer assessment instruments is an ongoing concern in the literature. First noted by Topping (1998), the variability of designs, tools and measurements used in peer assessment research was recently reiterated in a literature review by Speyer et al. (2011). The purpose of the review was to provide an overview of the instruments and questionnaires that have been used for peer assessment in medical and allied health education, and to then present the psychometric properties of these tools, as described in the literature.

Independent literature searches using five databases (Pubmed, ERIC, Embase, PsychINFO, and Web of Science) were performed by two reviewers. The searches yielded 2,899 articles, but only 28 met the inclusion criteria of the authors (Speyer et al., 2011). These criteria were (a) articles that described a peer assessment tool in allied health or medical education settings and (b) original articles that described peer assessment tools and articles that presented information about the validity or reliability of any of the peer assessment tools. Within these final 28 articles, the authors identified 22 different assessment instruments used. Most authors developed their own peer assessment instrument and applied it to their individual educational setting, in accordance with their own criteria and scoring system.

Findings included only two articles that did not occur in a medical education setting; one performed in pharmacy education, and one performed on a combined group of medical and dental students. Also, only three articles described the concept of validity on any level and six articles provided no psychometric data at all. The authors concluded that their attempt at a statistical pool of data was not possible for their designed review. The included articles offered too much heterogeneity in their designs, instrument diversity, and restricted data availability of psychometric characteristics to be able to provide a useful review (Speyer et al., 2011).

The use of a tested data collection instrument was preferred, however, for this particular study, the data collection instrument was field tested with both content experts and subjects who mirrored the characteristics of the participants in the full study. The field test was performed only after an attempt was made by the author to acquire IRB approval for a pilot study was essentially rejected due to the constraints on the number of participants available for the pilot study.

Operational Definition of Variables

For this study's investigation into the accuracy and reliability of undergraduate athletic training students to perform peer assessment of clinical skills, the independent variables for the accuracy and reliability measures of clinical skills were identified as the group membership of the participants: junior-level students, senior-level students, and instructors; the type of clinical test (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble's Compression Test, Thompson Test); and the clinical skills test subscale (patient position, clinician position, test performance). The dependent variables for the measure of accuracy of clinical skills were the clinical performance ratings on the individual components and subscales collected for each of the clinical skills procedures (Table 1). The dependent variables for the accuracy measure of professional behaviors were individual professional behavior ratings. The independent and dependent variables for the reliability measures were the same as those for the accuracy measures. Cohen's kappa coefficient was used to compute inter-rater agreement for the accuracy and reliability measures for the clinical skills variables and a linear weighted Cohen's kappa coefficient was used to compute inter-rater agreement for the accuracy and reliability measures for the professional behaviors variable.

Clinical Skills. The clinical evaluation tests performed during data collection where each skill was assessed by completion of its component parts. Because the total number of component parts differed among the five clinical evaluation tests (Table 1), and each component part aimed at measuring different clinical skills, each test was analyzed on its own. However, the number of subscales (patient position, clinician position, test performance), are consistent across all clinical skills will therefore were analyzed across all skills.

Individual components of each evaluated clinical skill test was scored as either "Yes" or "No" as they pertained to the participant's observation of the model clinician's ability to complete each individual task, as described on the measurement tool (Appendix A). The 2- point nominal scale was coded as 1= "Yes" and 0= "No". The subscales for each clinical skill were coded as 1= "Patient Position", 2= "Clinician Position", and 3= "Test Performance". The clinical skills were coded 1= "Biceps Femoris Manual Muscle

Test”, 2= “Kleiger Test”, 3= “Lachman Test”, 4= “Noble’s Compression Test”, and 5= “Thompson Test”. Table 2 displays a sample of coded data for the Biceps Femoris Manual Muscle Test from one data collection session in order to illustrate how the raw data appeared for data analysis.

Table 2

Coded Data Sample for Analysis

Biceps Femoral Manual Muscle Test	Clinical skill (1-5)	Subscales (1-3)	Instructor score (0, 1)	Student 1 score (0, 1)	Student 2 score (0, 1)
Prone	1	1	1	1	1
Stabilizes thigh firmly against the table	1	2	1	1	1
Places the other hand against the distal lower leg	1	2	1	1	1
Has patient actively flex the knee between 50 to 70 degrees	1	3	0	1	1
Hip is placed in external (lateral) rotation	1	3	0	1	1
Lower leg is placed in external (lateral) rotation	1	3	1	0	1
Instructs model patient to maintain position of hip and lower leg external (lateral) rotation	1	3	0	0	0
Applies resistance to the distal lower leg in the direction of knee extension	1	3	1	1	1
Holds resistance for 5 seconds	1	3	1	1	1
States what indicates a positive test	1	3	0	0	0

Professional Behaviors. The affective qualities displayed during clinical skills demonstration, assessed as a summative performance following completion of all clinical skills.

Each of the seven professional behaviors were scored on a 5-point Likert-type scale where a score of 5 = Always, a score of 4 = Frequently, a score of 3 = Occasionally, a score of 2 = Rarely, and a score of 1 = Never (Appendix A). The professional behaviors were: (a) performed the skills completely and in the appropriate order, (b) showed confidence during the interaction with the model patient, (c) provided clear instructions, without the need for clarification, to the model patient, (d) showed respect towards the model patient, (e) allowed adequate time for model patient's response to instructions and in answering questions, f) portrayed a friendly and approachable manner towards the model patient, and (g) maintained the physical and emotional safety of the model patient throughout their interaction (Appendix A).

Data Collection, Processing, and Analysis

Individual information sessions that included explanation and instruction on the use of the data collection instrument (Appendix A) and the informed consent process was scheduled for each interested participant within one week of initial contact from the investigator. As previously stated, all student subjects completed the coursework in which the clinical skills to be assessed during data collection were taught and evaluated within the athletic training education program, therefore all participants were familiar with the skills demonstrated during data collection. During the information sessions, participants were allowed the opportunity to ask questions regarding all aspects of the study prior to providing informed consent. For consistency among participants, within seven days of a participant’s information session, data collection was scheduled. This allowed for consistency in time between treatments across the groups, decreasing the threat to internal validity due to maturation (Millsap & Maydeu-Olivares, 2009; Rubin & Babbie, 2010). All student participants were enrolled in clinical coursework that provided opportunities to use the clinical skills assessed during this study. The longer the duration between information session and data collection, the more likely it was the student participants improved their abilities to evaluate clinical skills and professional behaviors.

As expected with peer assessment practices, all subjects were taught and had practiced the clinical skills both in and out of the classroom, therefore the student participants acted as clinicians during data collection. The student participants acted as model clinicians in order to maintain the authenticity of the study’s purpose of peer assessment evaluation. Because the number of student participants was larger than the number of treatments, student participants were randomly selected to play the role of the model clinician during data collection. The model patients during data collection were recruited from introductory courses in the university’s undergraduate athletic training education program. These volunteers were college freshman and sophomores who were not enrolled in professional coursework as an athletic training major, and did not have any experience or knowledge of the clinical skills that were performed during data collection. The volunteer patients also participated in an information session similar to the study’s participants prior to providing informed consent.

For each treatment, a triad consisting of one participant from the instructor group and two participants from one of the student groups (junior-level students or senior-level students) were randomly assigned from the pool of available participants based on the elapsed time from their training session. The model clinician was randomly assigned from the available pool for the respective student group (junior-level students or senior-level students).

The room used for data collection was served as a medical examination room used by physicians. The room was set-up to limit any extraneous stimuli during data collection: was well lit, climate controlled, and had no other activity occurring at the same time as data collection. Within the room, a padded treatment table was placed in direct view of three chairs for triad members (Figure 1). Participants had the option of sitting or standing during data collection.

Each data collection session began with the primary investigator (PI) reading introductory information and directions to the model patient, model clinician, instructor, and student (two senior-level or two junior-level students) participants. The model patient was seated on the treatment table, with the model clinician standing next to the model patient (Figure 1).

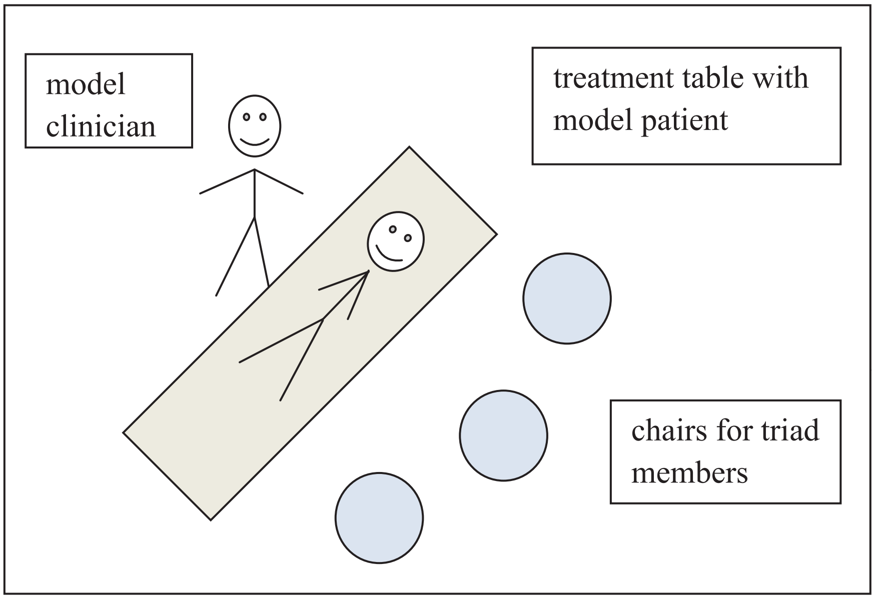

Figure 1. Room setup for data collection. This figure shows how the treatment table and chairs were setup for data collection.

As previously stated, student participants were randomly selected to play the role of the model clinician during data collection. The model clinician was directed to perform the clinical skills in the same manner they would during a real physical examination of a patient, including verbal and non-verbal communication and mannerisms towards the model patient. The model patient was instructed to follow the instructions of the model clinician to the best of their ability. The model patient was directed to ask the model clinician if they needed clarification about what that model clinician asked them to do during the skills demonstration.

Next, the PI announced the first clinical skill performed on the model patient by the model clinician, for example, the Lachman Test. The model clinician then performed the Lachman Test on the model patient, as they would on a real patient during a knee exam.

Once the model clinician was finished performing the LachmanTest, they were instructed to say, "done" or "finished" to signal they completed the skill. The triad then completed the section of the data collection sheet that pertained to the Lachman Test, circling either "Yes" or "No" for each component part to the Lachman Test as the parts pertained to the patient position, clinician position and test performance (Appendix A). This process was repeated for the remaining four clinical skills (e.g. Biceps Femoris Manual Muscle Test, Kleiger Test, Noble's Compression Test, Thompson Test). All data collection sessions included the performance and scoring of the same five clinical skills, with the order of the skills randomly assigned prior to the start of each session by the PI.

Upon conclusion of the fifth clinical skill demonstration, the model patient and clinician exited the room. The triad was instructed to complete the professional behaviors section of the data collection instrument as a summative assessment of the performance of the model clinician during clinical skills demonstration, circling one number on the 5-point scale for each of the seven professional behaviors (Appendix A). Once the participants completed their data collection forms, the forms were collected by the PI and the triad was excused.

Data analyses were chosen based the research questions and the hypotheses. All analyses were conducted after the necessary assumptions were tested and the data was screened for outliers. Data analyses for Cohen's kappa and linear weighted Cohen's kappa were performed through the statistical software AgreeStat 2013.1 (Gwet, 2009-2013).

Demographic data.

Demographic data collected from the participants was analyzed using the frequencies function and frequencies or means for demographic variables were reported, based on the type of variable. For categorical variables, frequencies were used and means were used for continuous demographic variables.

Hypothesis testing.

There were four types of hypotheses in this study. The first (hypotheses one through four) reflected the research questions (RQ1 and RQ2) about the level of accuracy of clinical skills ratings between the instructors and each student group. These hypotheses were examined via Cohen's kappa (κ) inter-rater reliability coefficient. Cohen's kappa is the most widely used measure of inter-rater reliability for dichotomously scored data (Howell, 2002; von Eye & von Eye, 2008; Warner, 2008). As previously stated, all participants witnessed and scored the same procedures for each data collection session.

For example, the first data collection session included triad #1 consisting of one instructor and two junior-level students. The data from each student was compared to the instructor's data in order to determine how similar each student scored the model clinician in comparison to how the instructor scored the model clinician. This process was repeated for every triad that included junior-level students. Data from all triads that included junior-level students was then used to calculate the group (junior-level students) level of inter-rater agreement (κ) with instructors for the final analysis of the accuracy measure of clinical skills and subscales. The same procedure was used for the triad that included senior-level students.

The second type of hypothesis (five through eight) reflected the research questions about the reliability of clinical skills ratings within each of the two student groups in measuring clinical skills of their peers (RQ3 and RQ4). These hypotheses will be examined through a measure of Cohen's kappa (κ) inter-rater reliability coefficient. Using the previous example of triad #1 consisting of two junior-level students, the ratings of each student was compared to each other to determine inter-rater reliability for the data collection session. This process was repeated for every triad that included junior-level students. Data from all triads that included junior-level students was then used to calculate the group (junior-level students) level of inter-rater agreement (κ) with instructors for the final analysis of the reliability measure of clinical skills and subscales. The same procedure was used for the triad that included senior-level students.

The third type of hypothesis reflected the research questions about the level of accuracy of professional behaviors of undergraduate athletic training students between the instructors and each student group. Hypotheses nine and ten addressed the research questions about the accuracy of professional behaviors ratings (RQ5 and RQ6). These hypotheses were examined via a linear weighted Cohen's kappa (κ_ω) inter-rater reliability coefficient. Weighted kappa was used for the ordinal data because it takes into account the differences between raters that are not all equal, as is found in dichotomous variables (Vanbelle & Albert, 2009; Warrens, 2011). The data from the individual triads was collected and analyzed in the same manner for the professional behaviors as it was previously explained for the clinical skills.

The fourth type of hypothesis reflected the research questions about the reliability of each of the two students groups in measuring professional behaviors of their peers; hypotheses 11 and 12 addressed RQ7 and RQ8. These hypotheses were examined via a linear weighted Cohen's kappa (κ_ω) inter-rater reliability coefficient.

Assumptions

The assumptions about the study's design pertained to group assignment and group demographics. Assignment to groups was not random; therefore, the demographic of each group (instructors, senior-level students, junior-level students) was fairly consistent. The non-random assignment to groups allowed the author to use a between groups design to compare the assessment abilities of each student group to the instructors for the research questions that pertained to accuracy. The non-random group assignment also allowed the author to use a within-groups design to determine inter-rater reliability within each student group.

There were three assumptions made in this study regarding the sample. First, the recruitment of all participants from the same athletic training education program allowed the author to know the demographics and experience of the participants. All student participants were taught, had practiced, and been assessed on, the clinical skills used during data collection at least once. Because they had one additional year of clinical experience, the senior-level students had more opportunities than the junior-level students to use, observe, and be assessed on the skills. The results of the data analyses allowed the author to investigate whether academic level was a factor in the students' abilities to accurately and reliably evaluate their peers on clinical skills. Also, all instructors recruited for participation had at least one academic year of experience in evaluating athletic training students on the clinical skills used during data collection.

Second, it was assumed that all participants were physically able to perform the tasks required of them during data collection. All student participants we required to have a current, signed technical standards form on record with the athletic training program in which they were enrolled as students. The technical standards form was signed by a physician allowing the students to participate in the clinical experience portion of the athletic training curriculum, including physical activities that were considered routine for an athletic trainer. Such tasks required more physical exertion than what was asked of all participants during the study. The non-athletic training students recruited as model patients were asked to sign an informed consent form that included confirmation they were in good overall health and were note suffering from an injury to the lower extremity at the time of data collection. In addition, any model patient with a history of injury to the lower extremity was excluded from participation. Instructors recruited as participants also signed an informed consent form stating they were physically able to perform the duties required during data collection (sitting on a stool, writing on a clipboard, observing at a distance of approximately 10-15 feet).

The final assumption made about the participant sample was that none of the student participants had experience with formal peer assessment practices. Formal peer assessment practices were defined for the purpose of this study as any evaluation of a peer where the evaluation was used for formal grade assignment. Peer-assisted learning activities, providing feedback, and group work were not considered formal peer assessment, unless the participant(s) awarded, or were awarded, a grade that counted towards a final grade for an academic course.

Limitations

Sample size was a limitation to the current study. The small sample size and use of participants from the same athletic training program limited the generalizability of the study greatly, thus affecting external validity (Ary, Jacons, Sorensen, & Razavieh, 2010). This was the first peer assessment investigation to use undergraduate athletic training students, collect data concurrently with multiple participants, and the participants were all affiliated with the athletic training program at the institution where data was collected. The results of this study cannot be projected to the larger population of undergraduate athletic training students, but can serve as a starting point for other athletic training education programs to perform their own peer assessment studies, or for multiple data collection sites to be used across athletic training programs.

The small sample size was also the reason for using a field test in place of a pilot test. Lack of truly validated measurement instruments is a known issue in the field of peer assessment, particularly in medical and allied health education (Speyer et al., 2011; Topping, 1998). Although performing a pilot test of the data collection instrument would be best to rigorously test for validity and reliability, the small sample size did not allow the ability to gather enough pilot data to do so. Therefore, a field test of the data collection instrument using athletic training students, faculty, and clinical athletic trainers was performed.

The students and instructors recruited for participation were involved on a day-to-day basis with the athletic training education program, thus providing the opportunity to discuss the study with each other outside of data collection. Although all data was coded for privacy, participant anonymity was not guaranteed, there was no definitive measure that could be taken to completely ensure the participants did not discuss their role in the study with others.

For any study, proper training in the use of the assessment tool is paramount to maintaining the greatest amount of construct validity as possible. Participants did not have an opportunity to practice use of the data collection instrument prior to data collection. Participants did have the opportunity to look over the data collection instrument and ask questions during the information session prior data collection, but inclusion of a practice session may have improved the quality of the data.

Additional threats to construct validity that pertain to this peer assessment study included the amount of exposures of the participants to the treatment. With the study using only one treatment per participant, there could have been an issue of mono-operation bias; meaning there was limited data to be able to make quality inferences from the findings (Trochim & Donnelly, 2008). However, the reverse can also be an issue. The incorporation of only one treatment significantly limited the potential for a maturation effect among the student participants (Millsap & Maydeu-Olivares, 2009; Rubin & Babbie, 2010). The student participants were enrolled in clinical coursework at the time of data collection. Scheduling multiple treatments would have allowed time for the student participants to gain experience using and evaluating the clinical skills that were used in the study. Additionally, the more the students were exposed to the treatment, the more they may have been able to learn from each exposure and carry that experience over to successive treatments.

Another threat to construct validity for this study was one that must be considered anytime students are involved as participants. The anxiety levels of the students may have been a factor (Ary et al., 2010). It was hoped that with the information session and the knowledge that the results of data collection will not be used for any course grading, the anxiety levels of the student participants was greatly reduced. Additionally, participation was completely voluntary, data collection occurred outside of class time, and all contact with the investigator regarding the study will be done outside of the investigator's faculty responsibilities.

Delimitations

The primary delimitation to this investigation into peer assessment among undergraduate athletic training students was the participant population. As mentioned previously, the study's research questions supported the use of a small sample because the purpose of the study was to compare athletic training student and instructor scores during clinical skills demonstration (Trochim & Donnelly, 2008). It was best to compare student scores to the scores assigned by the instructors who taught and evaluated the clinical skills of the students participating in the study. Using participants that fall outside of the desired population characteristics would not provide useful data regarding the purpose of the study.

Another delimitation made by the author was the category of clinical skills selected for use in the study. All of the clinical skills chosen for inclusion were designed for use by clinicians during lower extremity injury evaluations. The curriculum design of the athletic training education program where data collection took place separated courses that instruct evaluation of the upper and lower extremities into two courses. The lower extremity evaluation course was taken by the students during the first academic session they were enrolled in the athletic training education program. The upper extremity evaluation course was taken in the second academic session the students were enrolled in the athletic training education program. By using only lower extremity clinical skills, the investigator had a longer amount of time during the academic year to schedule data collection for a time when both groups of students had been taught, had practiced, and been evaluated on the clinical skills assessed during data collection.

Ethical Assurances

Ethical issues may arise whenever research involving humans is conducted. The responsible conduct of research (RCR) in the United States is supported by the Office of Research Integrity (ORI) in the Department of Health and Human Services (Horner & Minifie, 2011; Steneck, 2007). RCR is an umbrella term that concerns all parts of the research process; including new knowledge discovery and sharing, maintaining scientific integrity, and partaking in responsible science practices (Horner & Minifie, 2011). In order to maintain the scientific integrity of their discipline’s body of knowledge, researchers must demonstrate a thorough understanding of the ethical dimensions of performing research (Horner & Minifie, 2011).

IRB (expedited review) approval was obtained from both Northcentral University (NCU) and the institution where data was collected prior to any data collection conducted. The study posed no more than minimal risk to the participants, and participants were evaluating students on their ability to perform common non-invasive clinical skills used to diagnose orthopedic injuries.

Instructors and senior-level and junior-level athletic training students of the author's home institution were recruited as participants; and data was collected concurrently, with the instructor group and one of the student groups represented during each treatment. Specific issues that the IRB considered with the current study included stress to the participants, possible coercion among participants to enroll, and privacy and confidentiality of the participants (Adu-Gyamfi & Okech, 2010; Brogt, Dokter, & Antonellis, 2007; Brogt, Foster, Dokter, Buxner, & Antonellis, 2009; Moon, 2011). The student-faculty relationship contains an inherent power balance that can produce undue influence on students (Brogt et al., 2007).

The potential for psychological, emotional, or social stress existed in the current study (Brogt et al., 2007; Brogt et al., 2009). The student participants assessed their classmates performing clinical skills. Students and instructors collected data in the same room in order to ensure participants witnessed the skills performed under the same circumstances. These design elements may increase the stress level of anyone, regardless of their familiarity with the study participants or their comfort level with the subject matter assessed. Also, some students may be uncomfortable assessing the performance of a classmate under any circumstances.

Recruiting undergraduate students from the education program from where the investigator was employed created the potential for coercion among students to participate and could have represented a conflict of interest for the author (Brogt et al., 2007). This issue again resulted from the familiarity the investigator and participants all had with each other. Data collection did not occur during scheduled class times, or as part of any coursework, and the clinical skills used for data collection were not part of the coursework taught by the investigator. Despite these steps to convince the students there were no repercussions if they did not participate, the fact that they were students invited to participate in a study performed by one of their faculty members may have had unintentional coercive effects.

The familiarity the participants had with each other created the potential for privacy and confidentiality concerns. Complete privacy was not guaranteed because of the participant population, however, confidentiality regarding the data and results for each participant was managed more easily (Brogt et al., 2007). No identifiable information was used during the course of the study, therefore the participants' identities were held from any outside entities. However, the participants, because of their familiarity with each other, may have discussed the study with each other, which can lead to sharing of data amongst the participants. This can lead to breaches of privacy (Brogt et al., 2007). There were no reported issues from participants regarding breaches of privacy or confidentiality during the course of this study.

All information pertaining to the study, including signed informed consent forms, data collection schedules, data collection sheets, and all analyses and reporting documents were kept off-site from the university where the study was conducted. Hard copies were stored in a locked, fire-safe box. All computerized data was be maintained on the investigator's personal computer that did not leave their residence.

Informed consent was obtained by the author for all participants after the information session, and prior to the start of data collection. Participants were able to ask questions about their role in the study prior to providing informed consent. The informed consent form for the dissertation research of the author included specific items that addressed participants' rights as a research participant and the potential risks to participation, including the ability to stop participation at any time, and for no reason.

Summary

This quantitative, quasi-experimental study used participants from the students and instructors associated with the undergraduate athletic training program at the institution where data was collected. Accuracy of students to assess clinical skills and professional behaviors of their peers was determined through a between-groups design and reliability of students to assess clinical skills and professional behaviors of their peers was determined through a within-groups design. Data analysis included the use of Cohen's Kappa coefficient for the clinical skills and a linear weighted Cohen's kappa coefficient for the professional behaviors. Assumptions, limitation, and delimitations of this study focused primarily on the small, non-randomly assigned groups used for the sample. This study posed no more than minimal risk to participants and IRB approval was obtained prior to data collection.

Chapter 4: Findings

The purpose of this study was to investigate the accuracy and reliability of undergraduate athletic training students' assessment of their peers on clinical skills and professional behaviors. The results of this study provide information that can guide future research into peer assessment in athletic training education. Student and instructor scores on clinical skills and professional behaviors were used to determine inter-rater agreement for accuracy among students. Inter-rater agreement between students was used to determine reliability among students on clinical skills and professional behaviors. This chapter concludes with an evaluation of the findings, including a brief interpretation of the results and the contributions made to the peer assessment literature in athletic training education.

Results

Demographic characteristics. A total of 28 volunteers participated in this study. Seventeen of the participants identified themselves as male and eleven identified themselves as female. The average number of years of clinical athletic training experience was 16.4 for the instructors. The junior-level and senior-level students completed 0.5 years and 1.5 years of clinical education, respectively. The instructor group had an average of 7.7 years of experience evaluating athletic training students, ranging between two and 20 years. The complete breakdown of the demographic characteristics of the study participants is displayed in Table 3.

Research question 1. The first research question of this study was: In relation to instructors, how accurate are junior-level students in scoring the clinical skills

Table 3

Demographic Characteristics of Study Participants

	Instructors	Senior-level Students	Junior-level Students
Number of participants	9	10	9
Gender	6:3	5:5	6:3
Age	36.6 [26.0, 63.0]	24.1 [21.0, 38.0]	21.6 [20.0, 23.0]
Education level completed	Bachelor's Degree: 4 Master's Degree: 5	High school: 8 Associate's Degree: 2	High school: 7 Associate's Degree: 2
Clinical experience	16.4 [5.0, 41.0]	1.5[a]	.5[a]

Note. Gender represented as male:female ratio. Age and clinical experience represented as mean [range].

[a] Senior-level and junior-level students had no variance in years of clinical experience within their respective groups, therefore there is no range given for these values.

performance of undergraduate athletic training students for the a) Biceps Femoris Manual Muscle Test, b) Kleiger Test, c) Lachman Test, d) Noble's Compression Test, and e) Thompson Test? The corresponding null hypotheses were:

> **$H1_0$:** There is no statistically significant agreement between instructor and junior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble's Compression Test, Thompson Test) performance ratings, and
>
> **$H3_0$:** There is no statistically significant agreement between instructor and junior-level student scores among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

To answer Research Question 1, a Cohen's kappa coefficient was calculated to determine agreement between instructors and junior-level students.[1] A separate kappa value was calculated for each clinical skill and for each subscale across all five clinical skills. Additionally, values for p_{pos} and p_{neg} were calculated to identify the consistency between raters when agreement occurred in the positive and negative directions of a 2x2 contingency table. Table 4 displays the results from the kappa analysis for the clinical skills.

Table 4

Results for Junior-level Students' Level of Agreement with Instructors on Clinical Skills

Clinical Skill	Kappa Value (κ)	Amount of Agreement	*p* value	p_{pos}	p_{neg}
Biceps Femoris Manual Muscle Test	.5589	Moderate	.000	.885	.667
Kleiger Test	.2593	Fair	.041	.808	.444
Lachman Test	.1982	Slight	.348	.944	.250
Noble's Compression Test	-.0862	Poor	.005	.913	.000
Thompson Test	.7296	Substantial	.000	.958	.769

Based on the *p* values listed in Table 4, there was no statistically significant agreement between instructor and junior-level students' scores, and therefore the null hypothesis was not rejected for the LachmanTest (p> .05). However, the null hypothesis was rejected for the Biceps Femoris Manual Muscle Test, Kleiger Test, Noble's Compression Test, and Thompson Test, as the *p* values for these tests were less than .05.

[1] All values for Amount of Agreement from "The Measurement of Observer Agreement for Categorical Data," by T.R. Landis and G.G. Koch, 1977, *Biometrics*, *33*, p. 165.

While the *p* level was established at .05, both the Biceps Femoris Manual Muscle Test and Thompson Test were significant at the .001 level. There was statistically significant agreement between instructor and junior-level students' scores on these tests and we accepted the alternate hypothesis.

The separate indices of agreement (p_{pos} and p_{neg}) indicated high levels of positive agreement between instructors and junior-level students for all clinical skills and high levels of negative agreement for the Biceps Femoris Manual Muscle Test and Thompson Test.

Table 5 displays the results of the kappa analysis for the clinical skills subscales across all five clinical skills tests.

Table 5

Results for Junior-level Students' Level of Agreement with Instructors on Subscales

Subscale	Kappa Value (κ)	Amount of Agreement	*p* value	p_{pos}	p_{neg}
Patient position	.7835	Substantial	.000	.983	.800
Clinician position	.1875	Slight	.152	.900	.273
Test performance	.3236	Fair	.000	.855	.462

Based on the *p* values listed in Table 5, there was no statistically significant agreement between instructor and junior-level students' scores for clinician position, and therefore the null hypothesis was not rejected ($p > .05$). However, the null hypothesis was rejected for patient position and test performance as the *p* values were less than .05. While the *p* level was established at .05, patient position and test performance were both significant at the .001 level. There was statistically significant agreement between instructor and junior-level students' scores on these subscales and we accepted the alternate hypothesis.

The separate indices of agreement (p_{pos} and p_{neg}) indicated high levels of positive agreement between instructors and junior-level students for all subscales and high levels of negative agreement for patient position.

Research question 2. The second research question of this study was: In relation to instructors, how accurate are senior-level students in scoring the clinical skills performance of undergraduate athletic training students for the a) Biceps Femoris Manual Muscle Test, b) Kleiger Test, c) Lachman Test, d) Noble's Compression Test, and e) Thompson Test? The corresponding null hypotheses were:

> **$H2_0$:** There is no statistically significant agreement between instructor and senior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble's Compression Test, Thompson Test) performance ratings, and
>
> **$H4_0$:** There is no statistically significant agreement between instructor and senior-level student scores among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

To answer Research Question 2, a Cohen's kappa coefficient was calculated to determine agreement between instructors and senior-level students. A separate kappa value was calculated for each clinical skill and for each subscale across all five clinical skills. Additionally, values for p_{pos} and p_{neg} were calculated to identify the consistency between raters when agreement occurred in the positive and negative directions of a 2x2 contingency table. Table 6 displays the results from the kappa analysis for the clinical skills.

Table 6

Results for Senior-level Students' Level of Agreement with Instructors on Clinical Skills

Clinical Skill	Kappa Value (κ)	Amount of Agreement	*p* value	p_{pos}	p_{neg}
Biceps Femoris Manual Muscle Test	.3802	Fair	.000	.758	.605
Kleiger Test	.5540	Moderate	.000	.886	.667
Lachman Test	.4860	Moderate	.000	.919	.560
Noble's Compression Test	.3735	Fair	.037	.943	.429
Thompson Test	.4068	Fair	.025	.935	.462

Based on the *p* values listed in Table 6, there was statistically significant agreement between instructor and senior-level students' scores for all of the clinical skills ($p< .05$), and therefore the null hypothesis was rejected for all skills. Additionally, the Biceps Femoris Manual Muscle Test, Kleiger Test, and Lachman Test were all significant at the .001 level.

The separate indices of agreement (p_{pos} and p_{neg}) indicated high levels of positive agreement between instructors and senior-level students for all clinical skills, except the Biceps Femoris Manual Muscle Test.

Table 7 displays the results of the kappa analysis for the clinical skills subscales across all five clinical skills tests.

Table 7

Results for Senior-level Students' Level of Agreement with Instructors on Subscales

Subscale	Kappa Value (κ)	Amount of Agreement	*p* value	p_{pos}	p_{neg}
Patient position	.7887	Substantial	.000	.988	.800
Clinician position	.2577	Fair	.019	.882	.353
Test performance	.3062	Fair	.000	.799	.488

Based on the *p* values listed in Table 7, there was statistically significant agreement between instructor and senior-level students' scores for all of the subscales ($p<$ .05), and therefore the null hypothesis was rejected for all subscales. Additionally, patient position and test performance were both significant at the .001 level.

The separate indices of agreement (p_{pos} and p_{neg}) indicated high levels of positive agreement between instructors and senior-level students for all subscales and high levels of negative agreement for patient position.

Research question 3. The third research question of this study was: How reliable are junior-level students to each other in their ability to evaluate the clinical skills performance of undergraduate athletic training students for the a) Biceps Femoris Manual Muscle Test, b) Kleiger Test, c) Lachman Test, d) Noble's Compression Test, and e) Thompson Test? The corresponding null hypotheses were:

> **$H5_0$:** There is no statistically significant agreement between junior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble's Compression Test, Thompson Test) performance ratings, and

H7$_0$: There is no statistically significant agreement between junior-level students among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

To answer Research Question 3, a Cohen's kappa coefficient was calculated to determine agreement between junior-level students. A separate kappa value was calculated for each clinical skill and for each subscale across all five clinical skills. Additionally, values for p_{pos} and p_{neg} were calculated to identify the consistency between raters when agreement occurred in the positive and negative directions of a 2x2 contingency table. Table 8 displays the results from the kappa analysis for the clinical skills.

Table 8

Results for Junior-level Students' Level of Agreement on Clinical Skills

Clinical Skill	Kappa Value (κ)	Amount of Agreement	*p* value	p_{pos}	p_{neg}
Biceps Femoris Manual Muscle Test	.3793	Fair	.030	.852	.526
Kleiger Test	.1342	Slight	.461	.848	.286
Lachman Test	.0000	Poor	n/a	.984	.000
Noble's Compression Test	.5200	Moderate	.039	.947	.571
Thompson Test	1.000	Almost Perfect	n/a	1.00	1.00

Based on the *p* values listed in Table 8, there was no statistically significant agreement between junior-level students' scores for the Kleiger Test ($p > .05$), and therefore the null hypothesis was not rejected. However, the null hypothesis was rejected for the Biceps Femoris Manual Muscle Test and Noble's Compression Test as the *p* value for these tests was less than .05. There was statistically significant agreement between junior-level students' scores on these tests and we accepted the alternate hypothesis.

The data for Thompson Test resulted in perfect agreement when kappa was calculated, but such data does not allow for calculation of a *p* value. Based on the kappa value of 1.000, the null hypothesis was rejected, and we accepted the alternate hypothesis that there was statistically significant agreement between junior-level students' scores on this test. The data for the Lachman Test resulted in total disagreement when kappa was calculated, but such data does not allow for calculation of a *p* value. Based on the kappa value of .0000, there was no statistically significant agreement between junior-level students' scores on this test, and therefore the null hypothesis was not rejected for the Lachman Test.

The separate indices of agreement (p_{pos} and p_{neg}) indicated high levels of positive agreement between junior-level students for all clinical skills and high levels of negative agreement for the Thompson Test.

Table 9 displays the results of the kappa analysis for the clinical skills subscales across all five clinical skills tests. Based on the *p* values listed in Table 9, there was no statistically significant agreement between junior-level students' scores for clinician position (p> .05), and therefore the null hypothesis was not rejected for this subscale. However, the null hypothesis was rejected for patient position and test performance as the *p* values were less than .05. There was statistically significant agreement between junior-level students' scores on these subscales and we accepted the alternate hypothesis.

Table 9

Results for Junior-level Students' Level of Agreement on Subscales

Subscale	Kappa Value (κ)	Amount of Agreement	*p* value	p_{pos}	p_{neg}
Patient position	.6250	Substantial	.016	.958	.667
Clinician position	.3529	Fair	.231	.951	.200
Test performance	.3295	Fair	.023	.868	.923

The separate indices of agreement (p_{pos} and p_{neg}) indicated high levels of positive agreement between junior-level students for all subscales, but low level of negative agreement for clinical position.

Research question 4. The fourth research question of this study was: How reliable are senior-level students to each other in their ability to evaluate the clinical skills performance of undergraduate athletic training students for the a) Biceps Femoris Manual Muscle Test, b) Kleiger Test, c) Lachman Test, d) Noble's Compression Test, and e) Thompson Test? The corresponding null hypotheses were:

$H6_0$: There is no statistically significant agreement between senior-level students in clinical skills (Biceps Femoris Manual Muscle Test, Kleiger Test, Lachman Test, Noble's Compression Test, Thompson Test) performance ratings, and

$H8_0$: There is no statistically significant agreement between senior-level students among the clinical skills subscales (patient position, clinician position, test performance) in clinical skills performance ratings.

To answer Research Question 4, a Cohen's kappa coefficient was calculated to determine agreement between senior-level students. A separate kappa value was calculated for each clinical skill and for each subscale across all five clinical skills. Additionally, values for p_{pos} and p_{neg} were calculated to identify the consistency between raters when agreement occurred in the positive and negative directions in a 2x2 contingency table. Table 10 displays the results from the kappa analysis for the clinical skills.

Table 10

Results for Senior-level Students' Level of Agreement on Clinical Skills

Clinical Skill	Kappa Value (κ)	Amount of Agreement	*p* value	p_{pos}	p_{neg}
Biceps Femoris Manual Muscle Test	.4286	Moderate	.003	.829	.600
Kleiger Test	.3077	Fair	.076	.894	.400
Lachman Test	.3750	Fair	.174	.960	.400
Noble's Compression Test	.1702	Slight	.430	.914	.250
Thompson Test	.6296	Substantial	.014	.963	.667

Based on the *p* values listed in Table 10, there was no statistically significant agreement between senior-level students' scores for the Kleiger Test, Lachman Test, and Noble's Compression Test (p> .05), and therefore the null hypothesis was not rejected. However, the null hypothesis was rejected for the Biceps Femoris Manual Muscle Test and Thompson Test, as the *p* values for these tests were less than .05. There was statistically significant agreement between senior-level students' scores on these tests and we accepted the alternate hypothesis.

The separate indices of agreement (p_{pos} and p_{neg}) indicated high levels of positive agreement between senior-level students for all clinical skills.

Table 11 displays the results of the kappa analysis for the clinical skills subscales across all five clinical skills tests.

Table 11

Results for Senior-level Students' Level of Agreement on Subscales

Subscale	Kappa Value (κ)	Amount of Agreement	*p* value	p_{pos}	p_{neg}
Patient position	1.000	Almost Perfect	n/a	1.00	1.00
Clinician position	.0277	Slight	.615	.898	.167
Test performance	.3407	Fair	.002	.852	.478

Based on the *p* values listed in Table 11, there was no statistically significant agreement between senior-level students' scores, and therefore the null hypothesis was not rejected for clinician position ($p > .05$). However, the null hypothesis was rejected for test performance as the *p* value was less than .05. There was statistically significant agreement between senior-level students' scores on this subscale and we accepted the alternate hypothesis. Patient position resulted in perfect agreement when kappa was calculated, but such data does not allow for calculation of a *p* value. Based on the kappa value, the null hypothesis was rejected, and we accept the alternate hypothesis that there was statistically significant agreement between senior-level students' scores on this subscale.

The separate indices of agreement (p_{pos} and p_{neg}) indicated high levels of positive agreement between senior-level students for all subscales and high levels of negative agreement for patient position.

Research question 5. The fifth research question of this study was: In relation to instructors, how accurate are junior-level students in scoring professional behaviors during clinical skills performance of undergraduate athletic training students?

The corresponding null hypothesis was:

H9$_0$: There is no statistically significant agreement between instructor and junior-level students in professional behaviors ratings

To answer Research Question 5, a linear weighted Cohen's kappa coefficient was calculated to determine agreement between instructors and junior-level students. Additionally, the value for p_o was calculated to identify the consistency between raters when agreement occurred in a 5x5 contingency table. Table 12 displays the results from the kappa analysis for the professional behaviors.

Table 12

Results for Students' Level of Agreement with Instructors on Professional Behaviors

Student Level	Linear Weighted Kappa Value (κ_ω)	Amount of Agreement	*p* value	p_o
Juniors	.2083	Fair	.053	.471
Seniors	.2559	Fair	.008	.486

Based on the *p* values listed in Table 12, the null hypothesis was not rejected for the junior-level students, as the *p* value was greater than .05. There was no statistically significant agreement between junior-level students' and instructor scores on the professional behaviors.

Research question 6. The sixth research question of this study was: In relation to instructors, how accurate are senior-level students in scoring professional behaviors during clinical skills performance of undergraduate athletic training students?
The corresponding null hypothesis was:

$H10_0$: There is no statistically significant agreement between instructor and senior-level students in professional behaviors ratings

To answer Research Question 6, a linear weighted Cohen's kappa coefficient was calculated to determine agreement between instructors and senior-level students. Additionally, the value for p_o was calculated to identify the consistency between raters when agreement occurred in a 5x5 contingency table. Table 12 displays the results from the kappa analysis for the professional behaviors.

Based on the *p* values listed in Table 12, there was statistically significant agreement between senior-level students' and instructor scores (p< .05), and therefore the null hypothesis was rejected.

The *p* values of the senior-level students' and junior-level students' scores differed enough to reject the null hypothesis for the seniors, but not reject the null hypothesis for the juniors. However, the weighted kappa values and p_o values were similar between the two groups of students and the instructors, indicating similar levels of agreement between each student group and instructors on scoring professional behaviors.

Research question 7. The seventh research question of this study was: How reliable are junior-level students to each other in their ability to evaluate professional behaviors during clinical skill performance of undergraduate athletic training students? The corresponding null hypothesis was:

H11$_0$: There is no statistically significant agreement between junior-level students in professional behaviors ratings.

To answer Research Question 7, a linear weighted Cohen's kappa coefficient was calculated to determine agreement between junior-level students. Additionally, the value for p_o was calculated to identify the consistency between raters when agreement occurred in a 5x5 contingency table. Table 13 displays the results from the kappa analysis for the professional behaviors.

Table 13

Results for Students' Level of Agreement on Professional Behaviors

Student Level	Linear Weighted Kappa Value (κ_ω)	Amount of Agreement	*p* value	p_o
Juniors	.0000	Poor	1.00	.643
Seniors	.4094	Fair	.000	.545

Based on the *p* values listed in Table 13, the null hypothesis was not rejected for the junior-level students, as the *p* value was greater than .05. There was no statistically significant agreement between junior-level students' scores on the professional behaviors. Although the linear weighted kappa value was .0000, the p_o value was .643, indicating overall agreement between junior-level students was higher than the calculated weighted kappa value indicated.

Research question 8. The eighth research question of this study was: How reliable are senior-level students to each other in their ability to evaluate professional behaviors during clinical skill performance of undergraduate athletic training students? The corresponding null hypothesis was:

H12$_0$: There is no statistically significant agreement between senior-level students in professional behaviors ratings.

To answer Research Question 8, a linear weighted Cohen's kappa coefficient was calculated to determine agreement between senior-level students. Additionally, the value for p_o was calculated to identify the consistency between raters when agreement occurred in a 5x5 contingency table. Table 13 displays the results from the kappa analysis for the professional behaviors.

Based on the *p* values listed in Table 13, there was statistically significant agreement between senior-level students' scores (p<.05), and therefore the null hypothesis was rejected. Additionally, the senior-level students had significant agreement at the .001 level.

Although the *p* values of the senior-level students' and junior-level students' scores differ enough to reject the null hypothesis for the seniors, but not reject the null hypothesis for the juniors, the weighted kappa values and p_o values are similar between the two groups of students, indicating similar levels of agreement between each group of students on scoring professional behaviors.

Evaluation of Findings

Upon review of the results, junior-level students had a similar number of instances of significant agreement with instructors and each other as did the senior-level students for research questions one through four, which related to the scoring of clinical skills. For Research Questions 1 and 2, which explored each student groups' scores compared to instructor scores, the senior-level students had significant levels of agreement for all clinical skills and subscales, and the junior-level students had significant levels of agreement for all but one clinical skill (Lachman Test) and one subscale (clinician position). For Research Questions 3 and 4, which explored the inter-rater reliability within each student groups' scores, the junior level-students agreed on three (Biceps Femoris Manual Muscle Test, Noble's Compression Test, Thompson Test) of the five clinical skills, and the senior-level students agreed on two (Biceps Femoris Manual Muscle Test, Thompson Test) of the five clinical skills. Neither student group had significant levels of agreement for the Lachman Test. Both the senior-level students and junior-level students had significant levels of agreement for the same two of the three subscales. The subscale both groups did not have significant levels of agreement was the clinician position subscale.

For Research Questions five through eight, the student groups had opposing results. The senior-level students had significant amounts of agreement with instructors and each other in the rating of professional behaviors. The junior-level students, however, did not have significant levels of agreement with either the instructors, or each other, for the rating of professional behaviors.

Given the current literature in athletic training education, the results on this study are not surprising. The only other studies about peer assessment in athletic training education (Marty et al., 2010; Marty et al., 2011) found students to be accurate graders compared to certified athletic trainers. In both studies, Marty et al. used simple percentage agreement to determine level of accuracy between students and certified athletic trainers in their grading of clinical skills recorded on video; and reported very high levels of agreement (> 94%). Although percentage agreement was not reported in data analysis for the current study, and students and instructors recorded scores during live demonstration, the p_{pos} and p_{neg} values represented the patterns of overall agreement recorded for each clinical skill and subscale, similar to the use of a percentage agreement. Although the p_{pos} values in particular were high for all clinical skills and subscales, the use of Cohen's kappa allowed a more discerning statistical value to be assigned for amount of agreement.

Marty et al. (2010) determined athletic training students to be unreliable when scoring their peers on recorded clinical skills for a one-time measurement. The authors used a generalizability study, which differed from the kappa measurement used in this study. For the current study, the student groups had mixed results for reliability in grading their peers during live skills demonstration. The juniors were reliable in scoring the Biceps Femoris Manual Muscle Test, Noble's Compression Test, and Thompson Test, and the seniors were reliable in scoring the Biceps Femoris Manual Muscle Test and Thompson Test.

Marty et al. (2010; 2011) did not investigate professional behaviors among athletic training students, and no studies in medical education or other allied health education programs were found to be comparable to the current study in design. The majority of the authors that investigated professional behaviors used qualitative methods and/or longitudinal designs in order to see the effect(s) peer assessment had on the development of professionalism. For the current study, senior-level students were found to be accurate and reliable, but junior-level students were found to be inaccurate and unreliable when scoring a classmate on professional behaviors during live clinical skills demonstration.

Although the decisions regarding whether or not to reject this study's null hypotheses were based on the *p* values, use of Cohen's kappa for such conclusions comes with the effects of a known marginal dependency of Cohen's kappa. Multiple authors noted the occurrence of a paradox in some instances of calculating Cohen's kappa; wherein there is high raw agreement between raters, but a low resulting kappa value (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990; von Eye & von Eye, 2008; Warrens, 2012). With this in mind, further consideration must be given to the data where kappa results were low, but p_{pos} and/or p_{neg} (for clinical skills) or p_o (for professional behaviors) were high.

There were two cases in the data where very low values of kappa were calculated despite high levels of raw agreement. The Noble's Compression Test (κ= -.0862, p_{pos}= .913, p_{neg}=.000) between junior-level students and instructors and the Lachman Test between junior-level students (κ= .0000, p_{pos}= .984, p_{neg}=.000) both had very high agreement in the positive direction and no agreement in the negative direction. The p_{neg} values of .000 do not reflect total disagreement in the negative direction, but rather that neither rater selected "No" for that test, creating asymmetry in the margins of the 2x2 contingency tables for those clinical skills. Such asymmetry in the respective margins of a 2x2 contingency table used to calculate Cohen's kappa yields a very low kappa value due to a relatively high value for the proportion of agreement expected by chance (Feinstein & Cicchetti, 1990; von Eye & von Eye; 2008). The standard 2x2 contingency table for two observers (A and B) using a dichotomous rating system from which Cohen's kappa is calculated is illustrated in Table 14 (Feinstein & Cicchetti, 1990).

Table 14

2x2 Contingency Table for Computing Cohen's kappa

	Ratings by Observer A		
Ratings by Observer B	YES	NO	Totals
YES	a	b	g_1
NO	c	d	g_2
Totals	f_1	f_2	N

Based on the information in Table 14, the values for p_o (proportion of observed total agreement), p_e (proportion of expected agreement by chance), p_{pos} (observed proportion of positive agreement), p_{neg} (observed proportion of negative agreement) are calculated using Equations 1 through 4, respectively (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990).

$$p_{pos} = \frac{2a}{f_1 + g_1} \tag{1}$$

$$p_{neg} = \frac{2d}{f_2 + g_2} \tag{2}$$

$$p_o = \frac{(a + d)}{N} \tag{3}$$

$$p_e = \frac{(f_1 g_1 + f_2 g_2)}{N^2} \tag{4}$$

Cohen's kappa is calculated using Equation 5 (Cohen, 1960).

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{5}$$

As illustrated in the Cohen's kappa equation, if the value of p_e is high relative to the value of p_o, then κ will be low. The value of p_e is dependent on the products of the sums on each column and row in the 2x2 contingency table. When the margin representing total positive agreement (*a*) is significantly greater than the margin representing total negative agreement (*d*), p_e will be high relative to p_o; thus resulting in a low result for κ. For both the Noble's Compression Test between junior-level students and instructors and the Lachman Test between junior-level students, the very low levels of kappa can be attributed the this asymmetry, resulting in the paradox of high agreement, but low resulting kappa.

This paradox may have affected the decision not to reject the null hypothesis for the Lachman Test between junior-level students. As mentioned in Chapter four, the raw data for this clinical skill resulted on total disagreement (.0000) for a kappa value, thus disallowing a *p* value to be calculated. Without a *p* value, the null hypothesis was not rejected based on the κ value of .0000. However, the very low kappa value was explained through the known paradox that occurs with kappa when marginal asymmetry occurs in the 2x2 contingency table.

There was one case in the data where the kappa value was very low, the *p* value very high, but there was a moderate amount of agreement in the raw data. The weighted kappa value for the junior-level students' level of agreement on professional behaviors was found to be κ= .0000, with a p_o value of .643 and a *p* value of 1.00. The null hypothesis was not rejected for this measure due to the *p* value being >.05. However, the raw data showed 18 of 28 data points fell into total agreement between raters with the ten remaining data points spread throughout a 5x5 contingency table, thus resulting in asymmetrical margin distribution. To put this in perspective, the senior-level students' raw data for the same measure (reliability of professional behaviors) was κ= .4094, with a p_o value of .545 and a *p* value of .0000. The senior-level students had a lower value for raw agreement than the junior-level students, but the null hypothesis was rejected for the seniors due to the *p* value of .0000. This difference between the groups is explained due to the more symmetrical distribution of the raw data for the senior-level students in the 5x5 contingency table.

Summary

The results of this study varied between the level of students and between the clinical skills and professional behaviors analyses. The senior-level students were accurate graders of their peers when compared to instructors for all clinical skills and the majority of subscales during clinical skills assessment, and for grading professional behaviors of their peers. The junior-level students were accurate graders of their peers for the majority of the clinical skills and subscales, but not accurate for professional behaviors. The senior-level students were reliable in the grading of their peers for two of the five clinical skills, two of the three subscales and the professional behaviors. The junior-level students were reliable graders in the grading of their peers for three of the five clinical skills, two of the three subscales, but were not reliable for the professional behaviors.

As there were no previous studies found in athletic training education that used live demonstrations for data collection, and the majority of studies that investigated professional behavior in students were not similar in design to the current study, it is difficult to draw comparisons to earlier research. However, this study does contribute to peer assessment in athletic training education since it is the first to incorporate live action scoring and the grading of professional behaviors.

Chapter 5: Implications, Recommendations, and Conclusions

Athletic training educators need to research the use of peer assessment as an evaluation tool in order to better prepare students to practice as healthcare professionals, or risk placing students at a clinical disadvantage to their counterparts in the medical and other healthcare fields upon graduation. The purpose of this quantitative study was to investigate the accuracy and reliability of undergraduate athletic training students to assess their peers on clinical skills and professional behaviors. This was a quasi-experimental study with three, relatively small ($n \leq 11$) groups; non-randomly assigned as instructors, senior-level students, and junior-level students. The measurement tool used a simple dichotomous nominal Yes/No scale for assessment of clinical skills and a 5-point Likert-type (5= Always, 1= Never) scale for global ratings of professional behaviors. For the clinical skills, Cohen’s kappa coefficient measured inter-rater agreement between instructor and student scores for accuracy and among each student group for reliability. Cohen's kappa is the most widely used measure of inter-rater reliability for dichotomously scored data (Howell, 2002; von Eye & von Eye, 2008; Warner, 2008). For the professional behaviors, a linear weighted Cohen’s kappa coefficient measured inter-rater agreement between instructor and student scores for accuracy and among each student group for reliability. Weighted kappa is used with ordinal data because it takes into account the differences between raters that are not all equal, as is found in dichotomous variables (Vanbelle & Albert, 2009; Warrens, 2011).

Limitations for this study included the low number of participants, lack of data collection instrument testing, and lack of training in peer assessment practices among the participants. In order to preserve internal validity, participants were recruited from only one athletic training education program. This ensured all clinical skills used in the study were taught and assessed in a similar manner across the student participants. Recruiting participants from more athletic training education programs would have improved external validity, but as the first study to use live skills demonstrations for peer assessment in athletic training education, the preservation of internal validity was considered a higher priority.

As previously mentioned, the data collection instrument was field tested prior to use in this study. A pilot study to measure the validity and reliability of the instrument was the preferred choice, but low participant numbers in the pilot study would have lead to insufficient data from which to draw conclusions. Since the participant pool for the larger study was already small (28 total participants) the use of one or two participants from each group for a pilot study would not have provided enough data to draw conclusions about the data collection instrument's value, and would have decreased the number of participants available for the full study.

Multiple authors stressed the importance of training participants prior to implementing a peer assessment program (Falchikov & Goldfinch, 2000; Garner et al., 2010; Marty et al., 2011; van Zundert et al., 2010; Topping, 2010; Vickerman, 2009). While the current study was not a long-term peer assessment program, it may have been beneficial for the participants to have a practice session with the data collection instrument. All participants were able to look over the data collection instrument and ask questions about it during the informed consent process and prior to the start of their data collection session, but they did not use the instrument until data collection.

Use of students and instructors from the same academic program did not guarantee anonymity for the participants. Data was coded for anonymity and privacy, but it was likely that participants were familiar with each other prior to, during, and after data collection. Ethical research practices were maintained through purposeful, clear, and thorough information provided to all participants regarding the purpose and design of the study, and their rights as research participants. Institutional Review Board approval was granted by both Northcentral University and the institution were data was collected prior to initiating the recruitment process. Implications from the results of this study and recommendations for future research based on those implications are presented in the following sections.

Implications

The first research question of this study compared junior-level students' scores to instructor scores for inter-rater agreement on clinical skills and subscales. The junior-level students had significant levels of agreement to the instructors on four (Biceps Femoris Manual Muscle Test, Kleiger Test, Noble's Compression Test, Thompson Test) of the five clinical skills and two (patient position, test performance) of the three subscales. However, the p_{pos} values were high between the groups for all clinical skills and subscales for the first research question.

Despite the rejection of the null hypothesis for some items and the failure to reject the null hypothesis for other items, the high levels of agreement represented by the p_{pos} values for all clinical skills and subscales should not be ignored. The lowest p_{pos} value for items related to Research Question 1 was .808, indicating that the junior-level students and instructors agreed in the positive direction better than 80% of the time.

Chenot et al. (2007) used Cohen's kappa to measure inter-rater agreement between medical students and instructors when scoring objective structured clinical examinations by third-year medical students. The authors reported a κ range of .41- .64 for checklist and global ratings for four stations, resulting in moderate to good agreement levels between groups. The Chenot et al. study had 214 participants, allowing for greater amounts of data and thus, a lower likelihood of marginal asymmetry affecting kappa values than the current study did.

The second research question of this study compared senior-level students' scores to instructor scores for inter-rater agreement on clinical skills and subscales. The senior-level students had significant levels of agreement with instructors for all of the clinical skills or subscales. Not surprisingly, the p_{pos} values were high (≥.799) between the groups for all clinical skills and subscales except one (Biceps Femoris Manual Muscle Test), indicating the senior-level students and instructors agreed in the positive direction better than 79% of the time.

The high number of incidences of significant agreement between the students and instructors on clinical skills was in-line with previous peer assessment studies in athletic training education. Marty et al. (2010; 2011) found students to be accurate graders compared to certified athletic trainers. The authors used simple percentage agreement to determine level of accuracy between students and certified athletic trainers in their grading of clinical skills recorded on video; and reported very high levels of agreement for all skills (> 94%).

Outside of athletic training education, the findings in the current study support the findings of previous studies in medical and other allied health education research. Evans et al. (2007) reported excellent agreement between dental students and instructors on molar removal and Bucknall et al. (2008) had very high levels of agreement between students and instructor grading on CPR skills. Chenot et al. (2007) used Cohen's kappa as the inter-rater agreement measure between instructor and medical student scores during clinical skills assessment and found moderate-to-substantial levels of agreement between groups (κ= .41- .64).

This study found senior-level students had more instances of significant levels of agreement with instructors on clinical skills and subscales than junior-level students, suggesting year in school affected the students' abilities to evaluate their peers. The senior-level students had one more year of didactic and clinical experience than the junior-level students, and therefore should be more likely to score more closely to an instructor than the junior-level students. However, this finding is in contrast to previous studies performed by Marty et al. (2010; 2011). The authors in both studies determined that year in school had no effect on the accuracy of athletic training students to assess their peers or provide accurate feedback to their peers. The current study differs from the Marty studies in two key ways that may explain this difference in findings. First, this study used undergraduate students for participants, whereas the Marty studies used graduate-level students. Second, the participants in the Marty studies all had previous experience and training in the use of peer assessment skills, but the participants in the current study did not have prior experience in peer assessment. Further research is needed in order to determine whether or not academic year, academic level (graduate or undergraduate), and prior experience affect the ability of athletic training students to accurately assess their classmates on clinical skills.

The third research question of this study compared junior-level students' scores for inter-rater agreement on clinical skills and subscales. The junior-level students had significant levels of agreement on three (Biceps Femoris manual Muscle Test, Noble's Compression Test, Thompson Test) of the five clinical skills, and two (patient position, test performance) of the three subscales. However, the p_{pos} values were high between the junior-level students for all clinical skills and subscales for the third research question. The separate indices of agreement (p_{pos} and p_{neg}) indicated high levels of positive agreement between junior-level students for all clinical skills and high levels of negative agreement for the Thompson Test; and high levels of positive agreement for all subscales. Additionally, the Lachman Test kappa analysis demonstrated a very low k value (.0000), but was explained through the known paradox that occurs with marginal asymmetry with Cohen's kappa.

Despite the rejection of the null hypothesis for some items and the failure to reject the null hypothesis for other items, the high levels of agreement represented by the p_{pos} values for all clinical skills and subscales should not be ignored. The lowest p_{pos} value for items related to Research Question 3 was .848, indicating that the junior-level students agreed in the positive direction just under 85% of the time.

There were no studies in athletic training education that used inter-rater reliability measures to determine the reliability of students to assess their peers. Marty et al. (2010) used a generalizability (G) study, followed by a decision (D) study in order to produce a summary coefficient (φ), and determined graduate-level athletic training students were not reliable graders of their peers when the measurement was taken for a one-time occasion. The authors did find the students became more reliable as the number of measurements increased over time.

In medical education, Machado et al. (2008) also found students improved in reliability over time in assessing their peers. The authors compared self-, peer and instructor grades over seven semesters. Self- and peer assessment grades were found to be significantly different than instructor grades, but were significant to each other. Because of the significant difference from instructor grades, the authors (Machado et al., 2008) determined that peer assessment was not a valid summative assessment tool for medical students.

The fourth research question of this study compared senior-level students' scores for inter-rater agreement on clinical skills and subscales. The senior-level students had significant levels of agreement on two (Biceps Femoris Manual Muscle Test, Thompson Test) of the five clinical skills, and two (patient position, test performance) of the three subscales. The separate indices of agreement (p_{pos} and p_{neg}) indicated high levels of positive agreement between senior-level students for all clinical skills ($\geq .829$) and subscales ($\geq .852$), demonstrating the senior-level students agreed in the positive direction better than 82% of the time.

Overall, the student groups had a similar number of clinical skills and subscales where they were found to be significantly reliable. Both groups had significant levels of agreement on the Biceps Femoris Manual Muscle Test, Thompson Test, patient position, and test performance; indicating all students in the study scored their classmates similarly on clinical skills.

The author anticipated the senior-level students would have higher levels of agreement with the instructors than the junior-level students due to the 12 additional months the senior students practiced in a clinical setting after learning the skills used in this study. The same logic allowed the author to conclude the senior-level students would have had a higher amount of agreement for the reliability measure. The junior-level students learned the clinical skills only six months prior to data collection and therefore should not have been as consistent in scoring the skills as the senior-level students.

The fifth and sixth research questions of this study compared each group of students' scores to instructor scores for inter-rater agreement on professional behaviors. The senior-level students had significant levels of agreement to the instructors and the junior-level students did not have significant levels of agreement to the instructors. The senior-level students had one more year of clinical experience than the junior-level students, and therefore should be more likely to be able to score more closely to an instructor that the junior-level students.

The seventh and eighth research questions of this study compared each group of students' scores for inter-rater agreement on professional behaviors. The senior-level students had significant levels of agreement between each other and the junior-level students did not have significant levels of agreement between each other. However, as explained earlier in this chapter, the junior-level students had a p_o value of .643, which was higher than the senior-level students' p_o value of .545, indicating the junior-level students had a higher level of raw agreement that the senior-level students for this measure.

There were no currently published studies in athletic training education about peer assessment of professional behaviors among athletic training students, but medical education produced a number of studies in this area. Hulsman et al. (2013) and Kovach et al. (2009) performed quantitative studies in which student and faculty scores were compared on different aspects of professionalism. Similar to the current study, the authors had conflicting results. Despite a weak but significant correlation, Hulsman et al., (2013) found medical students were significantly more lenient than instructors on grading communication skills. Alternately, Kovach et al. (2009) found medical students to be harsher graders than instructors for professionalism during a clerkship.

Lurie et al. (2006) investigated the longitudinal stability of peer assessment ratings of professionalism traits in second and third year medical students. Chronbach's alpha was used to determine consistency levels within the groups. Two classes were followed over two academic years. The authors found high levels of internal consistency for both groups of students from one year to the next, and the peer assessment scores were predictive from year two to year three for both groups of students.

Nofziger et al. (2010) and Garner et al. (2010) used qualitative research methods to investigate student views on the use of peer assessment for professionalism in medical education. Nofziger et al. (2010) determined that peer assessment was a great tool for the development of professional behaviors, especially interpersonal skills. Both groups of authors asserted that training and practice in peer assessment were keys to the success of the programs with students.

Similar to the other results in this study, it was anticipated that senior-level students would produce higher levels of agreement on professional behaviors than the junior-level students. With an extra year of clinical and classroom experience, the senior-level students proved better able to consistently grade their peers in a similar manner to each other than the junior-level students.

In the current study, assessment of live skills demonstrations was used in order to obtain results that would provide athletic training educators an opportunity to better identify the quality of peer assessment among athletic training students over videotaped presentations. Athletic training students enter a workforce in healthcare that uses peer assessment practices on a daily basis to evaluate the performance of fellow professionals and foster professional growth (Speyer et al., 2011); and therefore, live data collection for peer assessment studies need to mimic as closely as possible, conditions in which athletic training students are employed upon graduation.

The results of this study addressed the problem in various ways. First, the raw data did show that there were high levels of agreement in the positive direction (p_{pos}) for all clinical skills and subscales between students and instructors and among the student groups. The Cohen's kappa results were generally similar between the student groups. The senior-level students were accurate (compared to instructor scores) for all of the clinical skills and subscales, and reliable (compared to other senior scores) for half of the clinical skills and subscales (Biceps Femoris Manual Muscle Test, Thompson Test, patient position, test performance). The junior-level students were accurate (compared to instructor scores) for all but one each of the clinical skills (Lachman Test) and subscales (clinician position), and reliable (compared to other junior scores) for one more clinical skill and subscale than the senior-level students (Biceps Femoris Manual Muscle Test, Thompson Test, Noble's Compression Test, patient position, test performance).

Second, the professional behaviors results used a linear weighted Cohen's kappa on a 5x5 contingency table. The raw data showed moderate levels of total agreement (p_o) for both the accuracy and reliability tests between students and instructors and among the student groups. However, there were mixed results for the linear weighted Cohen's kappa. The senior-level students agreed significantly with instructors and each other, but the juniors did not agree significantly with the instructors or each other for professional behaviors assessment.

As the first study to use live skills demonstration and concurrent scoring among participants, the results were a promising next step for athletic training educators to identify peer assessment as an evaluation tool to prepare athletic training students for professional practice after graduation.

The use of Cohen's kappa to measure inter-rater agreement fit with the study's purpose and design to determine accuracy and reliability of undergraduate athletic training students in evaluation of clinical skills and professional behaviors. The mixed results between the raw data agreement levels (p_{pos}, p_{neg}, p_o) and the Cohen's kappa results showed some discrepancies that can be attributed to the noted marginal dependency issues with kappa that resulted in the paradox of high agreement, low kappa (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990; Warrens, 2012; von Eye & von Eye, 2008).

This was the first study of peer assessment in athletic training education to use undergraduate students as participants, live and concurrent data collection, and evaluation of professional behaviors; all significant to the advancement of peer assessment in athletic training education. The majority of accredited athletic training education programs are undergraduate (CAATE, 2013b), therefore, more research in the field needs to focus on this student population. Live and concurrent data collection allowed for better comparisons between and among groups for analysis. The inclusion of professional behaviors as part of the analysis was vital to initiating this area of research in athletic training education to better prepare athletic training students for the workforce where peer assessment is a common practice; similar to its use in medical education (Finn & Garner, 2011; Garner et al., 2010; Speyer et al., 2011).

The results of this study contributed to the application of peer assessment practices in athletic training education in multiple ways. First, peer assessment of clinical skills and professional behaviors can be used in the classroom setting. During practice skills sessions for formative purposes, or during practical examinations for grading purposes, students can evaluate their peers on a level close to an instructor. Second, peer assessment can be used in the clinical setting where students interact with patients, clinical preceptors (supervisors), physicians and other healthcare providers, coaches, parents, and administrators. Particularly for professional behaviors, use of peer assessment during clinical education is a valuable tool to prepare athletic training students for the workplace after graduation.

Recommendations

This study's promising results demonstrated research in peer assessment in undergraduate athletic training students should continue. Further research in this area needs to include training in peer assessment practices for the participants. The medical education research cited the importance and value of training medical students in peer assessment prior to implementing a peer assessment program (Chenot et al., 2007; Garner et al., 2010; Nofziger et al., 2010). With training, students may score peers more closely to instructors than what was reported in the current study.

The current study used one-time data collection for analysis. It is recommended that future studies use repeated measures designs, including longitudinal research. Multiple data collection sessions for participants will provide more data and allow the authors to draw better conclusions. Additionally, repeated measures allow for tracking a participant's data over time, including over a semester, academic year, or entire matriculation. Marty et al. (2010) found graduate athletic training students became more reliable in scoring their peers on clinical skills over time, but were not reliable for a one-time occurrence.

Instructors (clinical preceptors and academic faculty) need to be included in inter-rater reliability studies in peer assessment. As the two groups currently responsible for the majority of formal grading, the amount of consistency between the groups should be tested for validity and reliability. The current study used the instructor group for comparative purposes only, but the information gained from comparing these individuals to each other, and whether they are as effective as students in scoring, can provide additional information regarding the use of peer assessment over traditional assessment methods in athletic training education. If the consistency levels are similar between instructors as they are between students, than use of peer assessment is further supported.

The final recommendation for future research is to develop consistent use of statistical tests to measure between groups in peer assessment studies. A common issue in peer assessment overall, the various methods and analyses used in this area of research do not allow for a clear and coherent message about the value of peer assessment (Strijbos and Sluijsmans, 2010; van Zundert et al., 2010). Prior studies in athletic training education used simple percentage agreement to determine the accuracy of students to score their peers on clinical skills with excellent results (Marty et al., 2010; 2011). Use of Cohen's kappa theoretically provided greater value to the current study's data due to kappa's inclusion of the amount of agreement that occurred by chance between the participants (Cohen, 1960). However, the established paradox that occurred with marginal dependency in kappa lead to low kappa scores even though the raw data showed high levels of agreement (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990; Warrens, 2012; von Eye & von Eye, 2008). Thus, investigation into the best measure for peer assessment research needs to occur. Other than percentage agreement and Cohen's kappa, Pearson's chi-square, intraclass correlation coefficient, and Scott's Pi, among others, should be considered for use in peer assessment studies.

Conclusions

As the first study in athletic training education to use undergraduate students as participants, live and concurrent data collection, and evaluation of professional behaviors, the findings were promising for future research. Overall, the data for this study showed high levels of observed agreement for most clinical skills, subscales and the professional behaviors, but some had low Cohen's kappa values, most likely due to the marginal dependency of the kappa statistic. The senior-level students had higher levels of agreement with instructors and each other, as compared to the junior-level students. This was not a surprising outcome, as the senior-level students had an additional year of clinical experience in which to use and practice the skills, as opposed to the junior-level students, who only had six months of clinical experience.

Future research into peer assessment in athletic training education needs to include training in peer assessment for participants, use of repeated measures designs, and comparison of instructors' (academic faculty and clinical preceptors) scores in order to better understand the use of peer assessment in this population. Additionally, as indicated by a number of earlier studies, there is a need to establish consistent, quality measures in peer assessment research, including those which are performed in athletic training education.

References

Adu-Gyamfi, K. & Okech, A. (2010). Ethics in research in mathematics education. *Journal of Academic Ethics, 8*(2), 129-135. doi: 10.1007/s10805-010-9111-2

Accreditation Council for Graduate Medical Education. (2013). Common Program Requirements. Retrieved from http://www.acgme-nas.org/assets/pdf/CPR-Categorization-TCC.pdf

Amato, H., Hawkins, C.D., & Cole, S.L. (2006). *Clinical skills documentation guide for athletic training* (2nd ed.). Thorofare, NJ: SLACK Inc.

American Board of Medical Specialists (ABMS) (2006-2012). MOC competencies and criteria. Retrieved from: http://www.abms.org/maintenance_of_certification/ MOC_ competencies.aspx

Ary, D., Jacobs, L.C., Sorensen, C., & Razavieh, A. (2010). *Introduction to research in education* (8th ed.). Belmont, CA: Wadsworth, Cengage Learning.

Association of American Medical Colleges. (1998). *Medical schools objectives project. Report I: Learning objectives for medical school education- Guidelines for medical schools.* Washington, DC: Association of American Medical Colleges.

Blaik-Hourani, R. (2011). Constructivism and revitalizing social studies. *The History Teacher, 44*(2), 227-249.

Board of Certification, Inc. (2013). What is the BOC? Retrieved from http://www. bocatc.org/ index.php?option=com_content&view=article&id=27& Itemid=29

Brogt, E., Dokter, E., & Antonellis, J. (2007). Regulations and ethical considerations for education research. *Astronomy Education Review, 6*(1), 43-49.

Brogt, E., Foster, T., Dokter, E., Buxner, S., & Antonellis, J. (2009). Regulations and ethical considerations for astronomy education research III: A suggested code of ethics. *Astronomy Education Review, 7*(2), 57-65.

Bucknall, V., Sobic, E.M., Wood, H.L., Howlett, S.C., Taylor, R., & Perkins, G.D. (2008). Peer assessment of resuscitation skills. *Resuscitation, 77*(2), 211-215. doi: 10.1016 /j.resuscitation.2007.12.003

Casey, D., Burke, E., Houghton, C., Mee, L., Smith, R., Van Der Putten, D.,…Folan, M. (2011). Use of peer assessment as a student engagement strategy in nurse education. *Nursing and Health Sciences, 13*(4), 514-520. doi: 10.1111/j.1442-2018.2011.00637.x

Chenot, J., Simmenroth-Nayda, A., Koch, A., Fischer, T., Scherer, M., Emmert, B… Himmel, W. (2007). Can student tutors act as examiners in an objective structured clinical examination? *Medical Education, 41*(11), 1032-1038. doi: 10.111/j.1365-2923.2007.02895.x

Cicchetti, D.V. & Feinstein, A.R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology, 43*(6), 551-558.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. doi: 10.1177/001316446002000104

Commission on Accreditation of Athletic Training Education. (2013a). About. Retrieved from http://caate.net/about/

Commission on Accreditation of Athletic Training Education. (2013b). Professional program BOC exam pass rate data. Retrieved from http://caate.net/accredited-programs/pass-rate/

Commission on Accreditation of Athletic Training Education. (2012). *Standards for the Accreditation of Educational Programs for the Professional Preparation of the Athletic Trainer.* Retrieved from http://caate.occutrain.net/wp-content/uploads/2014/01/2012-Professional-Standards.pdf

Dochy, F. & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation, 23*(4), 279-298.

Earl, S. (1986). Staff and peer assessment- measuring an individual's contribution to group performance. *Assessment and Evaluation in Higher Education, 11*(1), 60-69.

Edwards, E. & Sutton, A. (1991). A practical approach to student-centered learning. *British Journal of Educational Technology, 23(1*), 4-20.

Elton, L. & Johnson, B. (2002). Assessment in universities: A critical review of research. Retrieved from http://eprints.soton.ac.uk/59244/1/59244.pdf

English, R., Brookes, S.T., Avery, K., Blazeby, J.M., & Ben-Shlomo, Y. (2006). The effectiveness and reliability of peer-marking in first-year medical students. *Medical Education, 40*(10), 965-972. doi: 10.111/j.1365-2929.2006.02565.x

Evans, R., Elwyn, G., & Edwards, A. (2004). Review of instruments for peer assessment of physicians. *BMJ, 328*(7450), 1240-1244. doi: 10.1136/bmj.238.7450.1240

Evans, A.W., Leeson, R.M.A., & Petrie, A. (2007). Reliability of peer and self-assessment scores compared with trainer's scores following third molar surgery. *Medical Education, 41*(9), 866-872. doi: 10.1111/j.1365.2923.2007.02819.x

Falchikov, N. (2005). *Improving assessment through student involvement: Practical solutions for aiding in higher and further education.* New York: Poutledge Falmer.

Falchikov, N. & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287-322.

Feinstein, A.R. & Cicchetti, D.V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 534-549.

Finn, G.M. & Garner, J. (2011). Twelve tips for implementing a successful peer assessment. *Medical Teacher, 33*(6), 443-446.doi: 10.3109/0142159X. 2010.546909

Friesen, D.D. & Dunning, G.B. (1973). Peer evaluation and practicum supervision. *Counselor Education and Supervision, 12*, 229-235.

Garner, J., KcKendree, J., O'Sullivan, H., & Taylor, D. (2010). Undergraduate medical student attitudes to the peer assessment of professional behaviours in two medical schools. *Education for Primary Care, 21*(1), 32-37.

Gielen, S., Dochy, F., Onghena, P., Struyven, K., & Smeets, S. (2011). Goals of peer assessment and their associated quality concepts. *Studies in Higher Education, 36*(6), 719-735. doi: 10. 1080/03075071003759037

Gwet, K.L. (2009-2013). AgreeStat2013.1 [Computer software]. Gaithersburg, MD: Advanced Analytics, LLC.

Harris, J.R. (2011). Peer assessment in large undergraduate classes: An evaluation of a procedure for marking laboratory reports and a review of related practices. *Advances in Physiology Education, 35*(2), 178-187. doi: 10.1152/advan. 00115.2010

Henning, J.M. & Marty, M.C. (2008). A practical guide to implementing peer assessment in athletic training education. *Athletic Therapy Today, 13*(3), 30-33.

Hodges, H.F. (2011). Preparing new nurses with complexity science and problem-based learning. *Journal of Nursing Education, 50*(1), 7-13. doi: 10.3928/01481834-20102029-01

Horner, J. & Minifie, F.D. (2011). Research ethics I: Responsible conduct of research (RCR)-Historical and contemporary issues pertaining to human and animal experimentation. *Journal of Speech, Language, and Hearing Research, 54*(Suppl.), S303-S329. doi: 10.1044/1092-4388(2010/09-0265)

Howell, D. C. (2002). *Statistical Methods for Psychology* (5th ed.). Pacific Grove, CA: Duxbury Thomson Learning.

Hulsman, R.L., Peters, J.F., & Fabriek, M. (2013). Peer-assessment of medical communication skills: The impact of students' personality, academic and social reputation in behavioural assessment. *Patient Education and Counseling, 92*(2), 346-354.

Iqbal, Z. & Mahmood, N. (2008). Compatibility of peer assessment and teacher assessment in observational situations: An emerging assessment tool in higher education. *Bulletin of Education and Research, 31*(2), 61-77.

Kaufman, J.H. & Schunn, C.D. (2011). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional Science, 39*(3), 387-406. doi: 10.1007/s11251-010-9133-6

Kendall, F.P., McCreary, E.K., Provance, P.G., Rodgers, M.M., & Romani, W.A. (2005). *Muscles: Testing and function with posture and pain* (5th ed.). Baltimore, MD: Lippincott Williams & Wilkins.

Kollar, I. & Fischer, F. (2010). Peer assessment as collaborative learning: A cognitive process. *Learning and Instruction, 20*(4), 344-348. doi: 10.1016/j.learninstruc .2009.08.005

Kovach, R.A., Resch, D.S., & Verhulst, S.J. (2009). Peer assessment of professionalism: A five-year experience in medical clerkship. *Journal of General Internal Medicine, 24*(6), 742-746. doi: 10.1007/s11606-009-0961-5

Kubany, A.J. (1957). Use of sociometric peer nominations in medical education research. *Journal of Medical Education, 46*, 670-673.

Kubiszyn, T. & Borich, G. (2007). *Educational testing and measurement, 8th edition*. Hoboken, NJ: John Wiley & Sons, Inc.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Langendyk, V. (2006). Not knowing what they don't know: Self-assessment accuracy in third-year medical students. *Medical Education, 40*(2), 173-179. doi: 10.1111/ j.1365-2929-2005.023732.x

Liaison Committee on Medical Education. (2013). About the Liaison Committee on Medical Education. Retrieved from http://www.lcme.org/about.htm

Liaison Committee on Medical Education. (2012). *Functions and Structure of a Medical School: Standards for Accreditation of Medical Education Programs Leading to the M.D. Degree*. Retrieved from http://www.lcme.org/functions.pdf

Liu, N., Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education, 11*(3), 279-290.

Lurie, S.J., Nofziger, A.C., Meldrum, S., Mooney, C., & Epstein, R.M. (2006). Temporal and group-related trends in peer assessment amongst medical students. *Medical Education, 40*(9), 840-847. doi: 10.1111/j.1365-2929.2006.02540.x

Luxton-Reilly, A. & Denny, P. (2010). Constructive evaluation: A pedagogy of student-contributed assessment. *Computer Science Education, 20*(2), 145-167. doi: 10.1080/08993408.2010.486275

Machado, J.L.M., Machado, V.M.P., Grec, W., Bollela, V.R., & Vieira, J.E. (2008). Self- and peer assessment may not be an accurate measure of PBL tutorial process. *BMC Medical Education, 8*(55). doi: 10.1186/1472-6920-8-55

Marty, M.C., Henning, J.M., & Willse, J.T. (2010). Accuracy and reliability of peer assessment of athletic training psychomotor laboratory skills. *Journal of Athletic Training, 45*(6), 609-614.

Marty, M., Henning, J.M., & Willse, J.T. (2011). Students provide accurate, but not always detailed, feedback to a peer performing an orthopedic assessment skill. *The Internet Journal of Allied Health Sciences and Practice, 9(*1). Retrieved from http://ijahsp. nova.edu/articles/Vol9Num1/pdf/marty.pdf

Mathews, B.P. (1994). Assessing individual contributions: Experience of peer assessment evaluation in major group projects. *British Journal of Educational Technology, 25*(1),19-28.

Millsap, R.E. & Maydeu-Olivares, A. (Eds.). (2009). *The SAGE handbook of quantitative methods in psychology.* Thousand Oaks, CA: SAGE Publications Ltd.

Moon, T.R. (2011). A primer on research ethics in the field of gifted education. *Gifted Child Quarterly, 55*(3), 223-229. doi: 10.1177/0016986211412163

Moore, D.S. (2004). *The basic practice of statistics, 3rd edition*. New York: W.H. Freeman and Co.

Moreland, R., Miller, J., & Laucka, F. (1981). Academic achievement and self-evaluation of academic performance. *Journal of Educational Psychology, 73*(3), 335-344.

National Athletic Trainers' Association. (2014). About NATA. Retrieved from http://www.nata.org/aboutNATA

National Athletic Trainers' Association. (n.d.). About us. Retrieved from http://athletictrainers.org/about-us/

National Athletic Trainers' Association. (2011). *Athletic Training Educational Competencies* (5th ed.). Dallas, TX: National Athletic Trainers' Association.

NCSS. (2014). PASS Sample Size Software (Version 13) [Computer software]. Kaysville, UT: NCSS, LLC.

Nofziger, A.C., Naumburg, E.H., Davis, B.J., Mooney, C.J., & Epstein, R.M. (2010). Impact of peer assessment on the professional development of medical students: A qualitative study. *Academic Medicine, 85*(1), 140-147. doi: 10.1097/ACM. 0b013e3281c47a5b

Ploegh, K., Tillema, H.H., & Segers, M.S.R. (2009). In search of quality criteria in peer assessment practices. *Studies in Educational Evaluation, 35*(2/3), 102-109. doi: 10.1016/j.stuedu.2009.05.001

Powell, K.C., & Kalina, C.J. (2009). Cognitive and Social Constructivism: Developing tools for an effective classroom. *Education, 130*(2), 241-250.

Prentice, W.E. (2011). *Principles of athletic training: A competency-based approach* (14th ed.). New York, NY: McGraw-Hill.

Roloff, M. (2010). A constructivist model for teaching evidence-based practice. *Nurse Education Perspectives, 31*(5), 290-293.

Rubin, A. & Babbie, E.R. (2010). *Essential research methods for social work* (2nd ed.). Belmont, CA: Brooks/Cole, Cengage Learning.

Rush, S., Firth, T., Burke, L., & Marks-Maran, D. (2012). Implementation and evaluation of peer assessment of clinical skills for first year student nurses. *Nurse Education in Practice, 12*(4), 219-226. doi: 10.1016/j.nepr.2012.01.014

Schönrock-Adema, J., Heijne-Penninga, M., van Duijn, M.A., Geertsma, J., & Cohen-Schotanus, J. (2007). Assessment of professional behaviour in undergraduate medical education: Peer assessment enhances performance. *Medical Education, 41*(9), 836-842. doi: 10.1111/j.1365-2923.2007.02817.x

Schunk, D.H. (2012). *Learning theories: An educational perspective* (6th ed.). Boston: Pearson Education, Inc.

Speyer, R., Walmari, P., Van Der Kruis, J., & Brunning, J.W. (2011). Reliability and validity of student peer assessment in medical education: A systematic review. *Medical Teacher, 33*(11), e-572-e585. doi: 10.13109/0142159X.2011.610835

Starkey, C., Brown, S., & Ryan, J. (2010). *Orthopedic and athletic injury examination handbook* (2nd ed.). Philadelphia, PA: F.A. Davis Company.

Steneck, N.H. (2007). *ORI introduction to the responsible conduct of research*. Washington, DC: U.S. Government Printing Office.

Strijbos, J. & Sluijsmans, D. (2010). Unravelling peer assessment: Methodological, functional, and conceptual developments. *Learning and Instruction, 20*(4), 265-269. doi: 10.1016/ j.learninstruc. 2009.08.002

Tiew, F. (2010). Business students' views of peer assessment on class participation. *International Education Studies, 3*(3), 126-131.

Tillema, H., Leenknecht, M., & Segers, M. (2011). Assessing assessment quality: Criteria for quality assurance in design of (peer) assessment for learning- A review of research studies. *Studies in Educational Evaluation, 37*(1), 25-34. doi: 10.1016/ j.stueduc.2011. 03.004

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249-276.

Topping, K.J. (2009). Peer assessment. *Theory Into Practice, 48*(1), 20-27. doi: 10.1080/ 00405840802577569

Topping, K.J. (2010). Methodological quandaries in studying process and outcomes in peer assessment. *Learning and Instruction, 20*(4), 399-343. doi: 10.1016/ j.learninstruc. 2009.08.003

Trochim, W.M.K. & Donnelly, J.P. (2008). *The research methods knowledge base, 3rd edition.* Mason, OH: Cengage Learning.

Vanbelle, S. & Albert, A. (2009). A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology, 6*(2), 157-163. doi: 10.1016/j.stamet.2008. 06.001

van den Berg, I., Admiraal, W. & Pilot, A. (2006). Design principles and outcomes of peer assessment in higher education. *Studies in Higher Education, 31*(3), 341-356. doi: 10.1080/03075070600680836

van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction, 20*(4), 270-279. doi: 10.1016/j.learninstruc.2009.08.004

Vickerman, P. (2009). Student perceptive on formative peer assessment: An attempt to deepen learning? *Assessment & Evaluation in Higher Education, 34*(2), 221-230. doi: 10.1080/02602930801955986

von Eye, A. & von Eye, M. (2008). On the marginal dependency of Cohen's κ. *European Psychologist, 13*(4), 305-315. doi: 10.1027/1016-9040.13.4.305

Vu, T.T. & Dall'Alba, G. (2007). Students' experience of peer assessment in a professional course. *Assessment & Evaluation in Higher Education, 32*(5), 541-556. doi: 10.1080/02602930601116896Warner, R.M. (2008). Applied statistics: From bivariate through multivariate techniques.Thousand Oaks, CA: SAGE Publications, Inc.

Warner, R.M. (2008). *Applied statistics: From bivariate through multivariate techniques.* Thousand Oaks, CA: SAGE Publications, Inc.

Warrens, M.J. (2010). A formal proof of a paradox associated with Cohen's Kappa. *Journal of Classification, 27*(3), 322-332. doi: 10.1007/s00357-010-9060-x

Warrens, M.J. (2011). Cohen's linearly weighted kappa is a weighted average of 2x2 kappas. *Psychometrika, 76*(3), 471-486. doi: 10.1007/s11336-011-9210-z

Welsh, M. (2007). Engaging in peer assessment in post-registration nurse education. *Nurse Education in Practice, 7*(2), 75-81. doi: 10.1016/j.nepr.2006.04.006

Wright, S. & Grenier, M. (2009). Examining effective teaching via a social constructivist pedagogy "case study". *Education, 130*(2), 255-264.

Young, G. (1999). Using portfolios for assessment in teacher preparation and health sciences. In S. Brown and A. Glaser (Eds), *Assessment matters in higher education: Choosing and using diverse approaches* (pp. 121-131). Buckingham and Philadelphia, PA: The Society for Research in to Higher Education and Open University Press.

Yurdabakan, I. (2011). The investigation of peer assessment in primary school cooperative learning groups with respect to gender. *Education 3-13: International Journal of Primary, Elementary and Early Years Education, 39*(2), 153-169. doi: 10.1080/030042709033 13608

Appendixes

Appendix A

Accuracy and Reliability of Peer Assessment of Clinical Skills
Among Undergraduate Athletic Training Students
Data Collection Instrument

Clinical Skills Evaluation

Instructions to the model patient: The clinician will perform five clinical skills that are commonly used by athletic trainers in the diagnosis of lower extremity injuries. Please follow the clinician's instructions to the best of your ability. If you are uncertain of what is asked of you, please ask the clinician for clarification. If you feel pain or discomfort at any time during this session, please inform the clinician immediately.

Biceps Femoris Manual Muscle Test
Instructions to clinician: This task allows you the opportunity to demonstrate the manual muscle test for the Biceps Femoris muscle on the patient.

Instructions to examiner: Please circle "Yes" or "No" as it pertains to your observation of the clinician's ability to complete each individual task, as described.

Patient Position	Completed as Described	
Prone	YES	NO
Clinician Position		
Stabilizes the thigh firmly against the table	YES	NO
Places the other hand against the distal lower leg	YES	NO
Test Performance		
Has patient actively flex the knee between 50 to 70 degrees	YES	NO
Hip is placed in external (lateral) rotation	YES	NO
Lower leg is placed in external (lateral) rotation	YES	NO
Instructs model patient to maintain the position of hip and lower leg external (lateral) rotation.	YES	NO
Applies steady resistance to the distal lower leg, in the direction of knee extension	YES	NO
Holds resistance for 5 seconds	YES	NO
States what indicates a positive test	YES	NO

Adapted from Amato, H., Hawkins, C.D., & Cole, S.L. (2006). *Clinical skills documentation guide for athletic training* (2nd ed.). Thorofare, NJ: SLACK Inc.

Kleiger's Test for Deltoid Ligament and Syndesmosis Instability
Instructions to the clinician: This task allows you the opportunity to demonstrate the Kleiger Test on the patient.

Instructions to the examiner: Please circle "Yes" or "No" as it pertains to your observation of the clinician's ability to complete each individual task, as described.

Patient Position	Completed as Described	
Seated or kneeling in front of patient	YES	NO
Knees flexed to 90 degrees (legs over the edge of the table)	YES	NO
Clinician Position		
Stabilizes lower leg without compressing the distal tibiofibular area	YES	NO
Grasps the medial aspect of the foot	YES	NO
Foot and ankle placed in neutral position (0 degrees of dorsiflexion)	YES	NO
Test Performance		
Instructs patient to relax leg during test	YES	NO
Externally rotates the foot (calcaneus and talus)	YES	NO
Repeats the test with ankle positioned in dorsiflexion	YES	NO
Maintains lower leg stabilization during external rotation movement	YES	NO
States what indicates a positive test	YES	NO

Adapted from Amato, H., Hawkins, C.D., & Cole, S.L. (2006). *Clinical skills documentation guide for athletic training* (2nd ed.). Thorofare, NJ: SLACK Inc.

Lachman's Test for ACL Laxity
Instructions to the clinician: This task allows you the opportunity to demonstrate the Lachman's Test for ACL Laxity on the patient.

Instructions to the examiner: Please circle "Yes" or "No" as it pertains to your observation of the clinician's ability to complete each individual task, as described.

Patient Position	Completed as Described	
Supine	YES	NO
Knee flexed 10 to 25 degrees	YES	NO
Clinician Position		
Stabilizes posteriorly on proximal calf with one hand	YES	NO
Stabilizes anteriorly on distal femur with other hand	YES	NO
Test Performance		
Instructs patient to relax leg during test	YES	NO
Attempts to anteriorly displace tibia on femur (draws anteriorly)	YES	NO
Maintains adequate stabilization of leg during test		
States what indicates a positive test	YES	NO

Adapted from Amato, H., Hawkins, C.D., & Cole, S.L. (2006). *Clinical skills documentation guide for athletic training* (2nd ed.). Thorofare, NJ: SLACK Inc.

Noble's Compression Test for ITB Friction Syndrome

Instructions to the clinician: This task allows you the opportunity to demonstrate the Noble's Compression Test for ITB Friction Syndrome on the patient.

Instructions to the examiner: Please circle "Yes" or "No" as it pertains to your observation of the clinician's ability to complete each individual task, as described.

Patient Position	Completed as Described	
Supine or seated	YES	NO
Knees flexed to 90 degrees	YES	NO
Clinician Position		
Standing or seated to the side of the patient, on side to be tested	YES	NO
Places thumb over the lateral femoral epicondyle on the side to be tested	YES	NO
Places other hand around the lower leg for support	YES	NO
Test Performance		
Applies pressure over the lateral femoral epicondyle	YES	NO
Instructs the patient to actively extend the knee at a slow pace (may be performed passively by clinician instead)	YES	NO
States what indicates a positive test	YES	NO

Adapted from Amato, H., Hawkins, C.D., & Cole, S.L. (2006). *Clinical skills documentation guide for athletic training* (2nd ed.). Thorofare, NJ: SLACK Inc.

Thompson Test for Achilles Tendon Rupture

Instructions to the clinician: This task allows you the opportunity to demonstrate the Thompson's Test on the patient.

Instructions to the examiner: Please circle "Yes" or "No" as it pertains to your observation of the clinician's ability to complete each individual task, as described.

Patient Position	Completed as Described	
Prone	YES	NO
Feet over the edge of the table	YES	NO
Clinician Position		
Standing or seated to the side of the patient, on side to be tested	YES	NO
Test Performance		
Instructs patient to relax leg during test	YES	NO
Squeezes the belly of the calf muscle group	YES	NO
States what indicates a positive test	YES	NO

Adapted from Amato, H., Hawkins, C.D., & Cole, S.L. (2006). *Clinical skills documentation guide for athletic training* (2nd ed.). Thorofare, NJ: SLACK Inc.

Professional Behaviors Evaluation

Instructions to the examiner: Use the provided scale to rate the following statements based on the clinician's overall performance, as you observed it, during the clinician's interaction with the patient during demonstration of all clinical skills. For clarification: to receive a score of "5" for a particular behavior, the clinician should have completed all aspects of the behavior during all of the clinical skills. If the clinician completed all aspects of the behavior for some of the clinical skills or some of the aspects of the behavior for all of the clinical skills, then the score should range between "2" and "4". If the clinician did not complete any aspect of the behavior for any of the clinical skills, then they should receive a score of "1".

Rating	**Criteria**
5	Always
4	Frequently
3	Occasionally
2	Rarely
1	Never

Each psychomotor skill was performed completely and in the appropriate order (patient position, clinician position, then test performance). 5 4 3 2 1
The clinician showed confidence in their actions during interaction with the patient (was poised, spoke with purpose, and acted with assurance in their abilities). 5 4 3 2 1
The patient was able to follow the instructions of the clinician without needing clarification. 5 4 3 2 1
The clinician showed respect towards the patient by being considerate of their modesty and polite in giving instructions. 5 4 3 2 1
The clinician allowed an adequate amount of time for the patient to respond to instructions and was courteous in answering questions and providing clarification, as needed. 5 4 3 2 1
The clinician portrayed his or herself in a friendly and approachable manner by smiling, making eye contact with the patient, keeping arms uncrossed, etc. 5 4 3 2 1
The clinician performed the skills in a manner that ensured the physical and emotional safety of the patient. 5 4 3 2 1