Accuracy and Reliability of Peer Assessment of Athletic Training Psychomotor Laboratory Skills

By: Melissa C. Marty, Jolene M. Henning, and John T. Willse

Marty, M. C., Henning, J. M. , & Willse, J. (2010). Accuracy and Reliability of Peer Assessment of Athletic Training Psychomotor Laboratory Skills. *Journal of Athletic Training, 45 (6)*.

Abstract:

Peer assessment is defined as students judging the level or quality of a fellow student's understanding. No researchers have yet demonstrated the accuracy or reliability of peer assessment in athletic training education. To determine the accuracy and reliability of peer assessment of athletic training students' psychomotor skills. Cross-sectional study. Entry-level master's athletic training education program. First-year (n = 5) and second-year (n = 8) students. Participants evaluated 10 videos of a peer performing 3 psychomotor skills (middle deltoid manual muscle test, Faber test, and Slocum drawer test) on 2 separate occasions using a valid assessment tool. Accuracy of each peer-assessment score was examined through percentage correct scores. We used a generalizability study to determine how reliable athletic training students were in assessing a peer performing the aforementioned skills. Decision studies using generalizability theory demonstrated how the peer-assessment scores were affected by the number of participants and number of occasions. Participants had a high percentage of correct scores: 96.84% for the middle deltoid manual muscle test, 94.83% for the Faber test, and 97.13% for the Slocum drawer test. They were not able to reliably assess a peer performing any of the psychomotor skills on only 1 occasion. However, the ψ increased (exceeding the 0.70 minimal standard) when 2 participants assessed the skill on 3 occasions (ψ = 0.79) for the Faber test, with 1 participant on 2 occasions (ψ = 0.76) for the Slocum drawer test, and with 3 participants on 2 occasions for the middle deltoid manual muscle test (ψ = 0.72). Although students did not detect all errors, they assessed their peers with an average of 96% accuracy. Having only 1 student assess a peer performing certain psychomotor skills was less reliable than having more than 1 student assess those skills on more than 1 occasion. Peer assessment of psychomotor skills could be an important part of the learning process and a tool to supplement instructor assessment.

Key words: peer-assisted learning, athletic training education, clinical education

Key Points:

* Based on the decision study, acceptable reliabilities would be obtained with 1 participant on 2 occasions (0.76) or 2 participants on 1 occasion "(0.80) for the Slocum drawer test; with 2 participants on 2 occasions (0.75) or 3 participants on 1 occasion (0.74) for the Faber test; or with 3 participants on 2 occasions for the middle deltoid manual muscle test (0.72).
* Athletic training students were highly accurate in their peer assessments for all 3 tests, and scores of the first-year and second-year students did not differ.
* Peer assessment of psychomotor skills may be a valuable contribution to the learning process.

Article:

Peer assessment is a pedagogic tool used in higher education to enhance students' learning. Defined as "the process whereby individuals or groups of students assess the work of their peers,"[1] peer assessment is a student-centered approach[2] that promotes active involvement and deeper thinking.[3] Authors in athletic training education have suggested using peer assessment of psychomotor skills to enhance understanding and skill performance.[4-7]

Athletic training education programs (ATEPs) can benefit from the use of peer assessment of psychomotor skills in multiple ways. Used as part of the learning-overtime process, peer assessment can increase the frequency with which students practice skills, augment instructor feedback, decrease anxiety, increase confidence, and enhance clinical competence.[7] Peer assessment of psychomotor skills may also benefit students by preparing them for future roles as clinical instructors and for peer assessment of future professional colleagues.[8] Educational goals, such as becoming lifelong learners or working effectively in a team, can be accomplished through peer assessment.[3] Other benefits include increased critical thinking,[9] enhanced learning of the material,[10] simultaneous self-assessment,[8,10] improved confidence,[9] and decreased anxiety.[11]

Although athletic training education researchers have yet to specifically examine the accuracy and reliability of peer assessment of psychomotor skills, we can glean some insight into peer interactions based on past studies of the broader topic of peer-assisted learning (PAL). In a national survey,[4] athletic training students (ATSs) described their perceptions of different PAL activities. The result was that ATSs practiced a moderate to large number of clinical skills with their peers and felt more confident practicing skills with other ATSs than with clinical instructors.[4] It is logical to assume that ATSs assessed their peers to some extent while practicing psychomotor skills. However, the helpfulness of the feedback that students receive from their peers in general is unclear.[4]

In a separate experimental study,[6] ATSs who attended a review session led by a peer tutor felt less anxious and were more confident performing psychomotor skills with the peer tutor than with a laboratory instructor. The ATSs also felt the review session was more collaborative than competitive, and some students commented that peers understood the barriers to learning better and could explain things more effectively. However, the ATSs were undecided as to whether feedback from peers was more helpful than that from the laboratory instructor. The authors[6] suggested using peer tutors to assess individual psychomotor skills rather than clinical proficiencies.

Thus, based on the aforementioned research, ATSs are engaging in teaching and learning exchanges with their peers and assessing and providing feedback on psychomotor skills. However, whether ATSs can accurately and reliably assess their peers' performance of psychomotor skills is unknown. Demonstrating the accuracy and reliability of peer assessment of psychomotor skills may encourage students and educators to have more confidence in the peer-assessment process and to use it more frequently. Therefore, our purpose was to determine ATSs' accuracy and reliability when assessing a peer performing 3 athletic training psychomotor skills. In addition, the reliability was examined further to determine the sources of error and how it was affected by increasing the number of participants and the number of times the skills were assessed. Furthermore, we compared the accuracy of peer assessment performed by ATSs currently enrolled in an orthopaedic-evaluation course with that of ATSs who had taken the course previously.

METHODS

Participants

A total of 13 students enrolled in an accredited entrylevel master's ATEP during the fall 2007 semester volunteered for this study: 5 first-year and 8 second-year ATSs. The first-year ATSs were currently enrolled in an orthopaedic-assessment course, and the second-year ATSs had taken the same course the prior academic year. All participants had experience assessing their peers' psychomotor skills in their previous athletic training courses ("Anatomical Basis of Athletic Injury," "Athletic Training Foundations," and "Therapeutic Modalities Laboratory"). The second-year students had experience in the aforementioned courses as well as a course on general medical conditions and a therapeutic exercise laboratory.

Instrumentation

Data collection was completed using skill-assessment sheets adapted with permission from a peer-reviewed athletic training education text.[12] On the assessment sheets, each skill was divided into 9 components addressing patient position, examiner position, and skill performance. Participants were asked whether the peer accurately performed the skill; they responded by circling yes or no next to each specific component. Each psychomotor

skill was taught to all ATSs in the same manner as documented on the assessment sheets. The ATSs were given the assessment sheets 2 weeks before data collection so they could familiarize themselves with the tool.

Procedures

Approval for the study was granted by the university's institutional review board, and informed consent was obtained before data collection began. We used a repeated-measures design to determine the accuracy and reliability of participants' ability to assess an ATS peer performing 3 psychomotor laboratory skills on 2 separate occasions. The middle deltoid manual muscle test, Faber test for hip conditions, and Slocum drawer test with internal rotation were selected because they represent various assessment techniques (ie, manual muscle test, special test, ligamentous stress test) and skills for a variety of regions of the body (ie, shoulder, hip, knee). The first peer assessment was completed the week after the skill was taught to the first-year ATSs in their orthopaedicassessment class and then again the following week. For example, the manual muscle test of the middle deltoid was taught to the first-year ATSs during week 6 of the semester. The first peer assessment for the middle deltoid manual muscle test was completed during week 7 of the semester, and the second assessment was completed during week 8.

During each data-collection session, participants viewed 11 different video segments of a peer performing 1 of the 3 psychomotor skills (eg, 11 video segments of the Faber test). Each video session started with instructions, and the first video segment shown each time demonstrated accurate performance of the skill before peer assessment began. The subsequent 10 video segments each had various intentional errors (eg, incorrect patient positioning, incorrect hand placement). Between video segments, 45 seconds was allowed for participants to complete the peer-assessment form; a 5-second prompt was given before the next video began. Video segments could not be rewound, and participants were instructed to use visual and auditory cues present in the video in order to assess the peer's performance. To counteract a learning effect, participants were shown the same video segments in a different order during the second assessment session.

Data Analysis

Before data collection, the principal investigator assessed each video segment for each psychomotor skill to determine the expert score for comparison purposes. The scores of the principal investigator were reviewed for accuracy by a panel of 5 certified athletic trainers (minimum of 7 years' experience) to ensure correct assessment of the videos. The participants' peer-assessment scores for the 10 erroneous video segments for each skill were compared with this expert-assessment score.

Accuracy of the participants' assessments was examined through percentage correct scores. An independent t test was calculated to compare scores of the first-year and second-year ATSs. The oc level was set at $P \leq .05$, and SPSS software (version 14.0; SPSS Inc, Chicago, IL) was used to analyze t tests and descriptive statistics.

Reliability of the peer-assessment scores was determined through a generalizability study (G study), a technique used by other authors to characterize the reliability of peer assessment.[13] A G study uses generalizability theory (G theory), an extension of classical test theory with a series of analyses of variance to determine the dependability of behavioral measurements.[14] This study included 2 sources of error variance: participants (number of ATSs) and occasion (testing sessions 1 and 2). Each video had intentional errors to assist in characterizing the reliability of the peer assessment; thus, variance attributed to the video was considered to reflect the true differences among individuals and not error.

Although the sample size for this study was not large, we highlight several important points. The videos, which served as our sample (ie, objects of measurement), were created to specifically represent a range of possible performances that the participants might see when actually practicing psychomotor skills with their peers. Because we purposefully controlled the variance through this manipulation of performances, the small sample size should not negatively affect data representativeness. Often G studies are conducted with a small number of participants (eg, 2 or 3 raters). So our data-collection design actually had a large sample (n = 13) compared with many studies.[15-18] For example, in a G study[19] recently published in the Journal of Athletic Training, ankle

laxity was measured by 2 raters. We must remember that G studies are not designed for calculating inferential statistics regarding mean differences. Therefore, many of the usual concerns about sample size from analysis of variance are less relevant here. For instance, no assumption of normality or homogeneity of variance is required in G theory. With fewer such concerns, we are more comfortable operating with smaller sample sizes than those found in current research using inferential statistics.

A G study allows for estimates of many sources of error in a single test, such as the error variance associated with comparison of the participants' assessments of skill components at 2 different times. The primary difference is that classical test theory coefficients only include one source of error at a time (eg, either interrater or stability), whereas G coefficients can include multiple sources or errors (eg, interrater and stability) and describe each source of error's contribution to the overall variance. These G coefficients tend to be lower, but they more accurately reflect the true reliability. For example, interrater reliability examines how 1 participant marks (eg, consistently low or high) compared with another, but it does not take into account how each participant assesses on different days (stability).[20]

Crossed sources of variability can also be measured through a G study. A participant-by-video interaction tells us that the scale was applied differently, depending on which video was being assessed. The residual error, video-by-participant-by-occasion variance, is random error that cannot be explained by any measurement facet (eg, participants or occasions). This residual error is analogous to the concept of random error variance in classical test theory. Overall, the information obtained from G studies can be used to determine the relative amount of error variation associated with the facets of measurement. By examining these relative contributions to variability, a researcher can determine which aspects of the assessment are leading to a lack of scoring consistency (eg, high variation among participants) and which aspects are working well (eg, small occasion variance). Dependability is the G-theory analog to reliability, and all sources of error variance can be included in a G coefficient or f coefficient summarizing the dependability of an assessment.[14] The G coefficient only uses variance from the interaction terms as error (eg, participant by video). The interactions suggest that videos are ordered differently (ie, relative standing). The ψ coefficient also counts the absolute magnitude as error, what we would normally call main effects (eg, participants). The main effects do not indicate people being ordered differently but rather the consistency of difficulty (eg, severity of ratings, with some participants rating more strictly than others).

Results from a G study can be used to conduct a decision study (D study). In the D-study phase of the analysis, a summary coefficient can be produced that is similar to a reliability coefficient in classical test theory.[14] A major feature of D studies is that they can be used to determine how dependability might change if numbers of participants or occasions were different. This aspect of D studies is similar to applying the Spearman-Brown prophecy formula used in classical test theory.

A ψ coefficient was calculated as part of the D study. A f coefficient is similar to an ordinary reliability coefficient except that it includes the absolute value of a score obtained on a scale and not just the rank ordering of assessments.[14] That is, typical reliability coefficients describe the extent to which assessed individuals could be expected to be ranked in the same order on another assessment (or assessment occasion). A ψ coefficient instead indicates the extent to which examinees are expected to receive the same score (not just the same ranking) under different measurement conditions. The f coefficient was chosen because the assessments used in this study were not for determining which videos were the best. Rather, the assessments were intended to indicate the level at which a given skill was being performed on a video and the dependability or reliability of the participants' assessments. Previous authors[13] who used G theory to analyze dependability of peer assessments in higher education deemed ψ values of 0.60 and 0.80 as ideal. We treat 0.70 as the minimal acceptable level for results that are to be used for group-level research. The information from a D study allows us to predict how many participants (ie, raters) and occasions are needed for the peer assessments to reach that level of dependability. Genova software was used for all G studies and D studies (http://www.education.uiowa.edu/ casma/GenovaPrograms.htm).

RESULTS

All 13 participants assessed 10 separate video segments of a peer performing 3 different psychomotor skills on 2 occasions. The G study allowed variance to be partitioned into components as discussed in the "Data Analysis" section (Table 1). The total variance was partitioned into single sources of variance (video, participant, and occasion) and crossed sources of variance (video by participant, video by occasion, and participant by occasion). The remaining variance (video by participant by occasion) was random error. The largest variances for all 3 skills were related to the video: 36.2% for the middle deltoid test, 50.7% for the Faber test, and 66.3% for the Slocum drawer test. This finding was expected because the videos were filmed with different intentional errors in each video. Variation in video scores was, therefore, desirable, and this variance represented the true differences in the videos. The largest source of error was the video-by-participant variance (21.3% for the middle deltoid test, 10.9% for the Faber test, and 9.1% for the Slocum drawer test), which means that the participants were somewhat inconsistent in their assessments of individual videos. This source of error is related to the concept of interrater reliability. Occasion and participant-by-occasion variances are related to intrarater reliability and were minimal for all 3 skills (0.8% and 6.9%, respectively, for the middle deltoid test; 0.6% and 0.9%, respectively, for the Faber test; 0% and 1.6%, respectively, for the Slocum drawer test), indicating that the participants were highly consistent across occasions.

Results from the D study are found in Table 2. A ψ coefficient was calculated to examine the absolute dependability of the assessments. The D study allows for determination of how the f coefficient would change if the number of participants or occasions was changed. None of the 3 skills was projected to have acceptable dependability of 0.70 with 1 participant on 1 occasion; the Slocum drawer test had the largest f value (0.67). Based on the D study, the Slocum drawer test would have acceptable reliabilities with 1 participant on 2 occasions (0.76) or with 2 participants on 1 occasion (0.80). The Faber test would have acceptable ratings with 2 participants on 2 occasions (0.75) or with 3 participants on 1 occasion (0.74), and the middle deltoid test would reach an acceptable level of reliability with 3 participants on 2 occasions (0.72). Each of these results suggests that peer assessments must be based on multiple measurement opportunities (eg, multiple participants, multiple occasions, or both) if the stability of the result is important.

The participants were highly accurate in their peer assessments for all 3 skills (Table 3). Each skill had a total of 90 possible points and no differences were noted between the scores of the first-year and second-year ATSs. The minimal acceptable score for this study was 72 (80%) because 80% is the minimal acceptable level for course work in the program in which the students were enrolled. All skills were rated at an acceptable level. However, high accuracy of assessment is somewhat at odds with the reliability findings. Reliability is affected by the variance of the object of measurement (eg, the videos). If videos were created with more variance in the displayed skill and the participants' accuracy remained high, we would expect the reliability of these assessments to increase. However, variance was not added to avoid inserting errors that students would be unlikely to make, even though our results might have improved.

DISCUSSION

Overall, ATSs were highly accurate when assessing peers. The ATSs rated the peers performing 3 different psychomotor skills at acceptable levels with few errors. We also found that the academic year of the ATSs did not affect their ability to accurately assess peers. Thus, the second-year ATSs were able to retain the information learned the prior year well enough to assess the skill, and the first-year ATSs were able to accurately assess a peer even though they just learned the material. These results are similar to those of a study[21] conducted with fifth-year medical students in a surgery department: The students' assessments were highly correlated with staff peer-assessment scores. The medical students overwhelmingly believed they made fair assessments of their peers' knowledge, clinical ability, attitude and interest, and attendance; the students' confidence in their assessments was supported by high correlations with staff assessments.[21]

In contrast to our findings, studies in medical education indicate that students are not always accurate when assessing the psychomotor skills of their peers. For example, second-year medical students who assessed a videotaped performance of a peer conducting a physical examination had low correlations with expert ratings of

the performance.[22] Similarly, in a separate study,[23] peer and staff assessments of medical students' clinical performance in a surgery course showed low correlations. Researchers in these 3 studies[21-23] compared peer assessments and staff assessments to determine accuracy. We used a peerreviewed assessment to show accuracy of the peer assessments, which may be a more accurate method than correlating with the assessment of 1 staff member.

Our results indicated that students did not reliably assess each other when the evaluations occurred 1 time and by only 1 person. This finding is different from the results of a study[24] in which junior medical students performed peer assessments of knowledge (ie, clinical judgment) and relationships (ie, leadership) during a ward assignment. Test-retest correlations were high, indicating that the students were reliable.[24] Again, test-retest correlations only determine intrarater reliability and would be expected to be higher than the reliability measured by a G study.

The results of the D study indicate that student-assigned scores may be more reliable with each of the 3 skills when the number of participants or occasions is increased. Athletic training students may perform skills in groups and on more than 1 occasion when practicing and preparing for practical examinations. Therefore, it is plausible that students may be assessed more reliably by their peers in this situation. Athletic training educators should purposefully structure peer-assessment opportunities in laboratory courses so the students can assess one another in groups and repeatedly. Other researchers[2] have suggested examining peer assessment in groups to help minimize variance. This practice would be consistent with the implications of our findings.

Even though ATSs could not always reliably assess their peers' performances of psychomotor skills, we believe that their high level of accuracy makes peer assessment of psychomotor skills an important part of the learning process in athletic training education. Some authors[25] have suggested that the reliability of peer assessment has been overemphasized in the literature and that the feedback provided to correct mistakes may be more important than simply recognizing an incorrectly performed skill component. Although we did not specifically examine ATSs' ability to provide corrective feedback, the topic is worthy of examination in future studies. For example, an ATS may recognize through peer assessment that a peer's hand placement during the Faber test was incorrect, and the ATS can then provide feedback on how to correct the hand placement. Using this feedback, the 2 ATSs can use critical-thinking skills to start a dialogue on how to best perform the Faber test and what constitutes a positive test.

Peer assessment of psychomotor skills can be an important component of athletic training clinical education to help ATSs progress from supervised practice to independent, collaborative practice.4 Peer assessment should supplement and not replace clinical-instructor assessment or feedback.[4] In peer-assessment relationships, both ATSs benefit, with the ATS being assessed receiving feedback and the ATS assessing being required to review the material and think critically about skills.[2] Allied health and medical students who participated in peer-assessment activities have recognized benefits including encouragement in skill practice,[22] enhanced self-assessment,[2,8] better understanding of the assessment criteria,[26] and assistance in preparing for peer assessment as a professional.[27,28] Peer assessment also had psychosocial benefits, such as decreased anxiety,[6,8] increased collaboration,[6] and increased confidence.[28] Peer assessment could provide similar benefits to ATSs, leading to better-prepared clinicians.

LIMITATIONS AND SUGGESTIONS FOR FURTHER RESEARCH

Our study had several limitations, which should be considered when determining the generalizability of our results. Although the sample size was small (N = 13), G theory allows meaningful results. The participants were all enrolled in the same entry-level master's ATEP. Perhaps the results would be different for undergraduate ATSs due to potential differences in maturity levels and comfort in critically analyzing a peer's performance. Furthermore, involving more participants from different ATEPs would provide a broader view of peer assessment. Peer assessment could also be beneficial for the evaluation of cognitive competencies and affective skills, such as professionalism.

Examining ATSs' abilities to provide corrective feedback is another area for further research. Our participants did not have an opportunity to describe what the peer did correctly or incorrectly. Perhaps requiring the ATSs to give feedback would improve how closely the ATSs observed the psychomotor skill and improve reliability. In addition, the type of assessment sheet used to guide ATSs' peer assessment could have affected reliability and validity. We had a detailed assessment sheet that included patient position, examiner position, and skill performance. Detailed assessment sheets used in a physical therapy study[29] allowed a wider distribution of scores and were associated with a decreased tendency to rate more highly. However, students who used the assessment sheets that were less detailed gave more feedback. In a psychology study,[30] detailed assessment sheets resulted in more consistent findings. Students who constructed marking criteria had a greater appreciation for the criteria[9,31] and a sense of ownership.[32] It has been suggested[33] that not giving students grading criteria increases critical thinking, but some students may have difficulty without such guidance. Thus, the use of assessment sheets could alter the reliability, accuracy, and outcomes of peer assessment.

In addition, the observational skills of the ATSs could have been a limiting factor in their ability to reliably assess peers performing psychomotor skills. Several researchers[1,2,30,31] in peer assessment have stated that some type of training or guidelines on how to assess psychomotor skills would be beneficial and may enhance reliability. We believe a peerassessment training program based on adult learning theory that emphasizes observational strategies, constructive feedback, role-playing scenarios, and resources could increase reliability and accuracy and warrants further investigation.

CONCLUSIONS

This study provides a starting point for examining and discussing peer assessment of psychomotor skills in athletic training education. The accuracy and reliability of peer assessment are factors because ATSs should receive accurate information; several suggestions for further research on this topic have been provided. It is plausible that ATSs' performance of psychomotor skills could improve as a result of the peer-assessment process. However, more studies are warranted to determine the direct effect of peer assessment on skill performance.

REFERENCES

1. Orsmond P, Merry S, Callaghan A. Implementation of a formative assessment model incorporating peer and self-assessment. Innov Educ Teach Int. 2004;41(3):273-290.
2. Ladyshewsky R, Gotjamanos E. Communication skill development in health professional education: the use of standardised patients in combination with a peer assessment strategy. J Allied Health. 1997; 26(4):177-186.
3. Welsh MM. Engaging with peer assessment in post-registration nurse education. Nurse Educ Pract. 2007;7(2):75-81.
4. Henning JM, Weidner TG, Jones J. Peer-assisted learning in the athletic training clinical setting. / Athl Train. 2006;41(1):102-108.
5. Knight K. Assessing Clinical Proficiencies in Athletic Training: A Modular Approach. 3rd ed. Champaign, IL: Human Kinetics; 2001:41, 54, 209.
6. Weidner TG, Popp JK. Peer assisted learning and orthopaedic evaluation psychomotor skills. J Athl Train. 2007;42(1):113-119.
7. Henning JM, Marty MC. A practical guide to implementing peer assessment in athletic training education. Athl Ther Today. 2008; 13(3):30-33.
8. Erickson GP. Peer evaluation as a teaching-learning strategy in baccalaureate education for community health nursing. J Nurs Educ. 1987;26(5):204-206.
9. Orsmond P, Merry S, Reiling K. The use of student derived marking criteria in peer and self-assessment. Assess Eval High Educ. 2000; 25(l):23-38.
10. O'Moore LM, Baldock TE. Peer assessment learning sessions (PALS): an innovative feedback technique for large engineering classes. Eur J Eng Educ. 2007;32(1):43~55.
11. Yates P, Cunningham J, Moyle W, Wollin J. Peer mentorship in clinical education: outcomes of a pilot programme for first year students. Nurse Educ Today. 1997;17(6):508-514.
12. Amato H, Hawkins C, Cole S. Clinical Skills Documentation Guide for Athletic Training. 2nd ed.

Thorofare, NJ: Slack Ine; 2006.
13. Sluijsmans MA, Moerkerke G, van Merrienboer JT, Dochy F. Peer assessment in problem-based learning. Stud Educ Eval. 2001;27(2): 153-173.
14. Shavelson RJ, Webb NM. Generalizability Theory: A Primer. Newbury Park, CA: Sage; 1991:1-26, 99-114.
15. Roebroeck ME, Harlaar J, Lankhorst GJ. The application of generalizability theory to reliability assessment: an illustration using isometric force measurements. Phys Ther. 1993;73(6):386-395.
16. Hagtvet KA, Hanin YL. Consistency of performance-related emotions in elite athletes: generalizability theory applied to the IZOF model. Psychol Sport Exerc. 2007;8(l):47-72.
17. Lafave MR, Katz L, Donnon T, Butterwick DJ. Initial reliability of the Standardized Orthopedic Assessment Tool (SOAT). J Athl Train. 2008;43(5):483-488.
18. Gross DP, Battie MC. Reliability of safe maximum lifting determinations of a functional capacity evaluation. Phys Ther. 2002;82(4): 364-371.
19. Heitman RJ, Kovaleski JE, Pugh SF. Applicability of generalizability theory in estimating the reliability of ankle-complex laxity measurement. J Athl Train. 2009;44(l):48-52.
20. Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. Med Educ. 2002;36(10):972-978.
21. Burnett W, Cavaye G. Peer assessment by fifth-year students of surgery. Assess Eval High Educ. 1980;5(3):273-278.
22. Calhoun JG, Woolliscroft JO, Ten Haken JD, Wolf FM, Davis WK. Evaluating medical student clinical performance: relationships among self, peer and expert ratings. Eval Health Prof. 1988;11(2):201-212.
23. Morton JB, MacBech WAAG Correlations between staff, peer and self assessments of fourth-year students in surgery. Med Educ. 1977; 11(3): 167-1 70.
24. Linn BS, Arostegui M, Zeppa R. Performance rating scale for peer and self assessment. Br J Med Educ. 1975;9(2):98-101.
25. Liu NF, Carless D. Peer feedback: the learning element of peer assessment. Teach High Educ. 2006;ll(3):279-290.
26. Bloxham S, West A. Understanding the rules of the game: marking peer assessment as a medium for developing students' conceptions of assessment. Assess Eval High Educ. 2004;29(6):721-733.
27. Papinczak T, Young L, Groves M, Haynes M. An analysis of peer, self, and tutor assessment in problem-based learning tutorials. Med Teach. 2007;29(5):el22-el32.
28. Flynn JP, Marcus MT, Schmadl JC. Peer review: a successful teaching strategy in baccalaureate education. J Nurs Educ. 1981;20(4):28-32.
29. Miller PJ. The effect of scoring criteria specificity on peer and selfassessment. Assess Eval High Educ. 2003;28(4):383-394.
30. Smith H, Cooper A, Lancaster L. Improving the quality of undergraduate peer assessment: a case for student and staff development. Innov Educ Teach Intern. 2002;39(1):71-81.
31. Sivan A. The implementation of peer assessment: an action research approach. Assess Educ. 2000;7(2): 193-21 3.
32. Orsmond P, Merry S, Reiling K. The importance of marking criteria in the use of peer assessment, yteesi Eval High Educ. 1996;21(3): 239-250.
33. Heron J. Assessment revisited. In: Boud D, ed. Developing Student Autonomy in Learning. 2nd ed. London, UK: Kogan Page; 1988: 77-91.