Journal of Athletic Training 2014;49(3):373–380
doi: 10.4085/1062-6050-49.1.12

www.natajournals.org

Validity and Reliability of the Standardized Orthopedic Assessment Tool (SOAT): A Variation of the Traditional Objective Structured Clinical Examination

Mark R. Lafave, PhD, CAT(C)*; Larry Katz, PhD†

*Department of Physical Education and Recreation Studies, Mount Royal University, Calgary, AB, Canada; †Faculty of Kinesiology, University of Calgary, AB, Canada

Context: Health care professions have replaced traditional multiple choice tests or essays with structured and practical, performance-based examinations with the hope of eliminating rater bias and measuring clinical competence.

Objective: To establish the validity and reliability of the Standardized Orthopedic Assessment Tool (SOAT) as a measure of clinical competence of orthopaedic injury evaluation.

Design: Descriptive laboratory study.

Setting: University.

Patients or Other Participants: A total of 60 undergraduate students and 11 raters from 3 Canadian universities and 1 standardized patient.

Intervention(s): Students were required to complete a 30-minute musculoskeletal evaluation in 1 of 2 randomly assigned mock scenarios involving the knee (second-degree medial collateral ligament sprain) or the shoulder (third-degree supraspinatus muscle strain).

Main Outcome Measure(s): We measured interreliability with an intraclass correlation coefficient (ICC) (2,k) and stability of the tool with standard error of measurement and confidence intervals. Agreement was measured using Bland-Altman plots. Concurrent validity was measured using a Pearson product moment correlation coefficient whereby the raters' global rating of a student was matched to the cumulative mean grade score.

Results: The ICCs were 0.75 and 0.82 for the shoulder and knee cases, respectively. Bland-Altman plots indicated no systematic bias between raters. In addition, Pearson product moment correlation analysis demonstrated a strong relationship between the overall cumulative mean grade score and the global rating score of the examinees' performances.

Conclusions: This study demonstrated good interrater reliability of the SOAT with a standard error of measurement that indicated very modest stability, strong agreement between raters, and correlation indicative of concurrent validity.

Key Words: clinical competence, psychometrics, health professionals

Key Points

- The Standardized Orthopedic Assessment Tool (SOAT) demonstrated good interrater reliability with very modest stability, strong agreement between raters, and concurrent validity.
- Using the SOAT in final, summative-type examinations may help determine clinical competence of orthopaedic injury evaluation, but it should be put into context within a broader and diverse programmatic evaluation plan and should not be considered the final authority on clinical competence.
- Future researchers will help build the psychometric soundness of the SOAT.

Development and evaluation of clinical competence orthopaedic evaluation skills is important in both the medical and allied health care professions. Orthopaedic injury evaluation and management composes between 15% and 30% of all primary care visits.[1–3] Despite its high prevalence rate, little attention has been paid to teaching and evaluating orthopaedic clinical skills at the undergraduate level in medicine.[4,5] Arguably, orthopaedic evaluation of injuries is a substantial part of the workload of every athletic trainer and therapist.

Experts agree that to develop and evaluate clinical competence, practical performance-based examinations, such as an objective structured clinical examination (OSCE), are necessary.[6] A number of variations of OSCEs may or may not include any or all of the following: standardized patients (SPs), other observer ratings, short written tests, evaluation of history taking, evaluation of physical examination, and evaluation of communication skills.[6] Despite the accolades and ubiquity of OSCEs throughout the medical and allied health care professions, these examinations also have been criticized.

Many research groups have questioned the validity and reliability of OSCEs.[6–15] One reason that OSCEs originally were introduced was to theoretically improve reliability among examinees by removing subjective bias.[11] Objective, dichotomous checklists were introduced to enhance interrater reliability and increase the number of competencies that could be sampled in a brief period.[16] However, many researchers[8,12,13] believe checklists objectify the process so much that they remove the meaningful evaluation of clinical competence. Seemingly, the very thing that was aimed at improving reliability actually decreased validity; the checklists took on greater meaning than the outcomes. Regehr et al stated succinctly: "checklists may reward

thoroughness rather than competence."[11(p994)] Furthermore, thoroughness is generally not a good indicator of expertise or clinical competence.[12,14] Experts tend to take shortcuts in their history and physical examination process, resulting in lower grades or scores on traditional OSCEs that reward thoroughness.[11,17,18] Moreover, candidates in the testing process, and more specifically in clinical practice, rarely follow a consistent protocol from task to task or patient to patient.[17] As practitioners vary their processes to accommodate client needs and characteristics, the intercase reliability of specific checklists is put into question.[17] Therefore, the efficacy of checklists in OSCEs designed to measure clinical competence is uncertain.[14]

When designing the original OSCE, examination creators assumed that removing expert opinion would lead to greater reliability.[7,8] However, some researchers[8,9,18] have demonstrated that a hybrid of traditional OSCE-type checklists along with more subjective global rating scales or continuous scales could maintain the overall reliability. Theoretically, global rating scales with continuous scales should permit expert opinion to be factored into the final grade or performance. Whereas the efficacy of incorporating checklist and global-rating-scale weighting into final scores or grades has not been determined,[10] their combination appears to have promise for more valid and reliable results.[16]

Measuring clinical competence through practical, performance-based examinations has been guided by a long-standing paradigm: Miller's pyramid.[19] This model has guided the development and evolution of clinical-competence evaluations in the medical and allied health care professions for more than 3 decades despite challenges to its merit.[12,15] The flaws in measuring clinical competence in a valid and reliable manner have led to the development of new workplace-based evaluation tools, such as the Mini-Clinical Evaluation Exercise, and in-training assessments whereby faculty members observe and grade medical residents with actual patients.[20–22] The variability in cases, the number of exposures to cases, and judgments of raters are a few of the issues that jeopardize reliability in exchange for the higher validity necessary for high-stakes examinations. Clearly, an ideal evaluation of clinical competence still is needed.

The Standardized Orthopedic Assessment Tool (SOAT) has undergone initial content validation and reliability testing.[23,24] The next logical progression for establishing validity and reliability of a tool is stringent testing in a wider audience. The focus for teaching, learning, and evaluating musculoskeletal or orthopaedic assessment skills seems to be destined for the specialist level of care. Attaining clinical competence in orthopaedic assessment is common for many medical and allied health care specialists, including orthopaedic surgeons,[20] sports medicine physicians,[21] rheumatologists,[25] physiotherapists,[26] and athletic therapists.[24,27] Therefore, the overarching purpose of our study was to establish the validity and reliability of the SOAT to measure clinical competence of orthopaedic assessment in third-year and fourth-year undergraduate athletic therapy students in 3 Canadian universities. To accomplish that purpose, the 4 objectives included the following measures: reliability, stability, agreement, and concurrent validity. Their associated statistical analyses are outlined in the Methods section.

METHODS

Participants

Standardized Patient. The primary author (M.R.L.) acted as the SP for all testing sites (Mount Royal University, University of Winnipeg, and Concordia University). He also trained the raters at every testing site with a standardized and pilot-tested 3-hour training course described elsewhere.[24]

Raters. The 11 raters were chosen from a convenience sample using the following criteria: minimum of 5 years of experience with some testing experience, availability to test during specified periods, and ability to attend the 3-hour training. A total of 5 raters were from Mount Royal University; 4 raters, University of Winnipeg; and 2 raters, Concordia University. All raters met the minimal criteria and also were actively involved in clinical education and testing at their respective universities.

Examinees. Participants consisted of 60 third-year and fourth-year undergraduate athletic therapy students (age = 28.83 ± 4.06 years, grade point average = 3.31 ± 0.39) from a convenience sample of volunteers at Mount Royal University (n = 27), the University of Winnipeg (n = 9), and Concordia University (n = 24). These students (and universities) were selected for participation due to the similarity of their programs and curricular designs. Volunteers were included if they were in the third or fourth year of their undergraduate programs and had completed a minimum of 2 years (or 20 three-credit-hour classes) in their programs and at least 1 undergraduate course in orthopaedic injury-assessment skills. The testing in this study was not a formal part of their educational requirements, but students from 2 universities (Mount Royal University and Concordia University) received bonus marks in a class if they participated. All participants provided written informed consent, and the study was approved by the human research ethics boards of the University of Calgary, Mount Royal University, the University of Winnipeg, and Concordia University.

Rater Training

Rater training involved a review of the scenario diagnosis and an item-by-item review of the correct answers and how the examinee should be graded using either the dichotomous or continuous scales throughout the assessment. Raters discussed appropriate grading based on hypothetical variations of the correct answers because a number of acceptable pathways would accomplish the same diagnosis.

The Standardized Orthopedic Assessment Tool

The SOAT has been described in greater detail elsewhere and has undergone content validation, as well as initial reliability testing.[24,27] It is an assessment instrument or tool that is used during a 30-minute practical, performance-based examination. The examinee is expected to complete a history and physical examination and to generate a diagnosis or conclusion for an SP. The raters grade the examinee using the SOAT, which is a unique combination of task checklists and continuous and global rating scales depending on the subcategory being measured (see Supplemental Appendix, available online at http://dx.doi.

org/10.4085/1062-6050-49.1.12.S1). The examinee enters the room and is instructed to interact only with the SP and to treat the SP as he or she would treat a patient in a clinical setting. He or she is instructed to ignore the 2 raters in the room, and the raters are not permitted to interact with the examinee or SP in any way throughout the examination process.

Procedures

The SOAT comprises 10 subcategories: 2 are dichotomous (history and observation), and 8 are continuously scaled components (clearing joints above and below the injury site, scanning examination, active range of motion, passive range of motion, strength testing, special testing, palpation, and diagnosis or conclusion). The history subcategory consists of a checklist or essentially 39 dichotomous items that the examinee must complete. The observation subcategory also has dichotomous items that the examinee must complete, but the remaining subcategories are continuously scaled items that raters assess on a 6-point scale, with 0 indicating *the examinee did not complete the task* and 5 indicating *the examinee did an outstanding job with the task*. The raters complete a global rating scale at the end of each subcategory and a final overall-performance global rating scale after all other measures are finished. Examinees are expected to complete each task listed in the history and observation components, but completion of the tasks for the remaining 8 components is based on personal judgment of the examinee. He or she can complete more or fewer tasks listed under the 8 continuously scaled components of the SOAT to ultimately obtain a more accurate diagnosis or conclusion. Concomitantly, the raters use their personal expertise to judge whether the examinee follows a correct and complete pathway to obtain the diagnosis or conclusion.

Raters graded examinees based on the quality of the task completed and the appropriateness of the task completion. If an examinee skipped a task, the raters could grade that item as *not applicable* or could award a grade of zero if they believed the task should have been completed. After each of the 10 components of the SOAT, the SP cued the examinee by asking: "What do you think is wrong with me?" The examinee had been oriented and instructed to provide at least 3 indices of suspicion to the SP after each component. The indices of suspicion can change, and the raters also factor these responses into their overall expert evaluations. The cue from the SP is intended to give the raters a glimpse into the thought processes and rationale the examinees used to choose whether to complete certain tasks. Given the unique nature of the SOAT, each examinee could be graded with a different denominator, so grades are presented as a percentage to permit comparison among other examinees.

Examinees were randomly assigned a knee ($n = 30$) or a shoulder ($n = 30$) scenario or case using a random-number table. They were permitted 30 minutes to complete all 10 components of the SOAT. The knee diagnosis was a second-degree medial collateral sprain and the shoulder diagnosis was a complete rupture of the supraspinatus muscle; both have been content validated.[27] Examinees were stopped at 30 minutes regardless of whether they were finished and were instructed to leave the room. After the examinees left the room, the SP was available to the raters to clarify tasks they could not see or experience themselves, such as touch, pressure, hand position, and general disposition. However, the raters were blinded to one another's grading, and the SP was not permitted to talk to the raters unless they needed clarification about what happened. The blind rating of the examinees from raters removed bias or influence from the SP or another rater. The final SOAT grade was converted to a percentage score for each of the 10 subcategories, which included the global rating scale at the end of each section. The final SOAT score was a mean of the 10 subcategory percentage scores. The SP scoring was slightly different from the complete SOAT with every item on it. Rather, the SP only completed the global rating scale at the end of each of the 10 subcategories and that score was converted to a percentage score for comparative purposes (see Supplemental Appendix, available online at http://dx.doi.org/10.4085/1062-6050-49.1.12.S1). In addition, both the SP and raters completed a final, overall global rating scale of the examinee's performance for the entire 30-minute assessment. This global rating scale was not factored into the final examination score but rather was used only for the Pearson product moment correlation coefficient or concurrent validation in this study.

Statistical Analysis

Interrater reliability was measured using an intraclass correlation coefficient (ICC) (2,k)[28–30] based on the experimental design.[31–33] The ICC (2,k) was performed both with and without the SP global rating to determine the effect of another rater on the overall reliability. In addition, SP scores from the global rating scale at the end of each of the 10 subcategories were used to evaluate potential bias of the SP in the overall reliability coefficient. We used the following formula to calculate the standard error of measurement (SEM) and confidence intervals (CIs), which indicate stability of the tool to measure future musculoskeletal evaluation competence[31]:

Step 1: SEM = standard deviation (SD) $\sqrt{1 - \text{ICC}}$
Step 2: Mean score $\pm$ 1.96 $\times$ SEM

Bland-Altman plots were calculated for the shoulder and knee using the mean scores of the 2 raters and the mean difference between the raters. Limits of agreement (95%) were calculated by multiplying the sum (upper limit) or difference (lower limit) of the mean and SD score by 2 SDs (1.96) from the mean score. If 95% of the differences lie within 2 SDs of the mean of the differences, then the 2 raters are thought to have good agreement.[32] A Pearson product moment correlation was used to evaluate the relationship between 2 seemingly similar yet different measures of student performance: the overall cumulative mean grade from the 10 subcategories and the global rating scale for the entire 30-minute assessment, both of which had been converted to percentage scores. We used the concurrent validity measure not only to establish the relationship between the detailed checklist and the SOAT but also the relationship between raters for overall impression of the examinee's performance. The α level was set at .001. We used SPSS (version 14.0; SPSS Inc, Chicago, IL) to analyze the data.

Table 1. Descriptive Statistics and Reliability Coefficients for Raters and the Standardized Patient

	Shoulder Case (n = 30)			Knee Case (n = 30)		
		Intraclass Correlation Coefficient (2,k)			Intraclass Correlation Coefficient (2,k)	
Rater	Grade, % (Mean ± SD)	Without Standardized Patient	With Standardized Patient	Grade, % (Mean ± SD)	Without Standardized Patient	With Standardized Patient
Rater 1	63.26 ± 14.58	0.75	0.75	68.21 ± 16.90	0.82	0.83
Rater 2	64.21 ± 16.64			68.93 ± 15.32		
Standardized patient	62.37 ± 16.65			68.28 ± 15.16		

RESULTS

The mean grades and ICC (2,k) for the shoulder and knee cases are listed in Table 1. The mean grades of both raters were 68.57% (95% CI = 49.45%, 78.01%) and 63.73% (95% CI = 55.41%, 81.33%) for the knee and shoulder, respectively.

The mean score differences between raters were –0.71 ± 9.65 and –0.95 ± 11.01 for the knee and shoulder, respectively. Upper and lower limits of agreement (95%) indicated no systematic bias. The upper and lower limits of agreement for the knee were 17.52% and –20.31%, respectively. The upper and lower limits of agreement for the shoulder were 19.71% and –23.44%, respectively. A graphical representation of the limits of agreement for the knee and shoulder is presented in the Figure.

The Pearson *r* correlation, which was used to evaluate the relationship between the overall cumulative mean grade

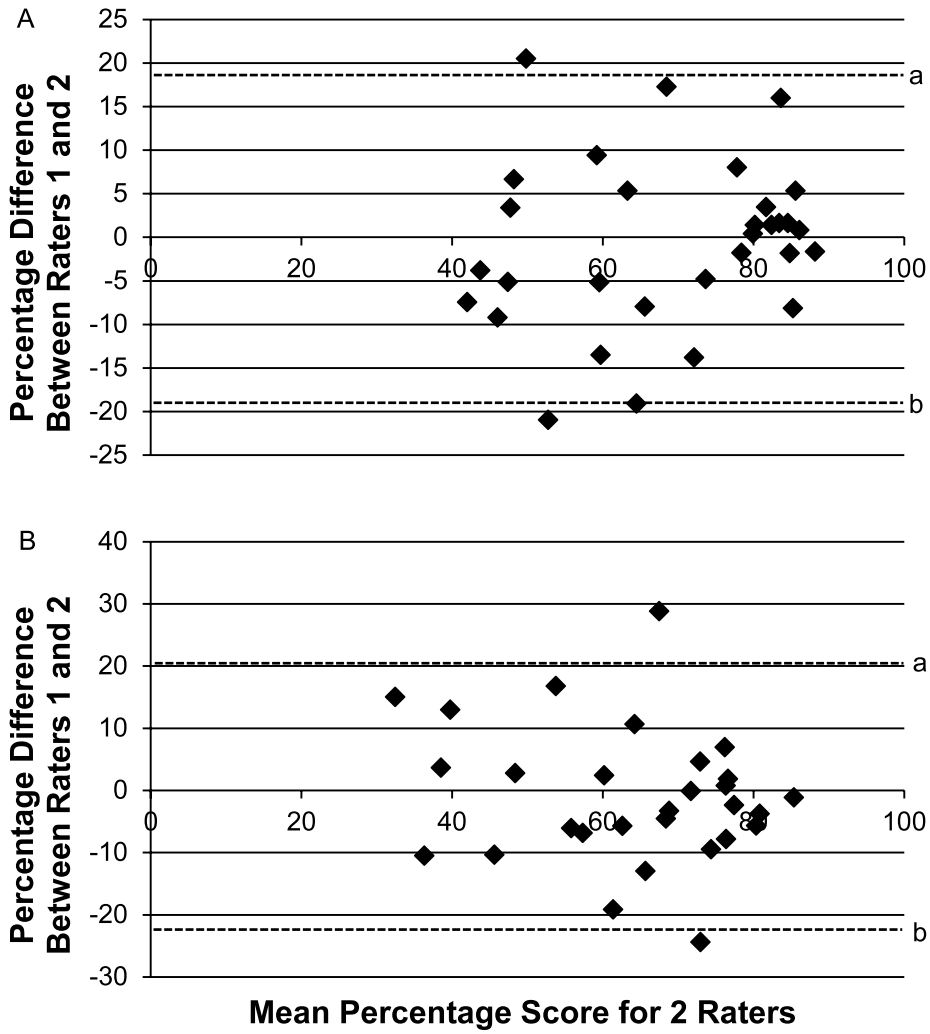

Figure. Bland-Altman plots measured in percentage scores for the knee and shoulder examinations depicting differences between raters 1 and 2 plotted against the mean score of those 2 raters. A, In the knee scenario, the upper (18.20%) and lower (–19.62%) limits of agreement demonstrate that 95% of the raters fell within this range. B, In the shoulder scenario, the upper (20.63%) and lower (–22.53%) limits of agreement demonstrate that 95% of the raters fell within this range. [a] Indicates the upper limit of agreement (95%). [b] Indicates the lower limit of agreement (95%).

Table 2. Pearson *r* Correlation Coefficient Determining the Relationship Between the Overall Cumulative Mean Grade and the Global Rating Scale

Body Region	Overall Cumulative Mean Grade, %	Global Rating Scale Mean Grade, %	Pearson *r* Correlation Coefficient
Shoulder	63.73	60.00	0.826[a]
Knee	67.85	66.13	0.617[a]

[a] Indicates difference ($P < .001$).

from the 10 subcategories and the global rating scale for the entire 30-minute assessment, revealed correlations between the measures for the shoulder ($r = .826$, $P < .001$) and knee ($r = .617$, $P < .001$) (Table 2).

DISCUSSION

The SOAT was created as a practical, performance-based examination intended to measure clinical competence in orthopaedic injury evaluation. It was designed to address some of the shortcomings of the traditional OSCE. The tool includes a combination of required and optional tasks for examinees to complete based on their clinical judgment. Theoretically, objectivity should be heightened with dichotomous-scale tasks that each rater and examinee is required to complete. However, the removal of the expert judgment from both the examinee and raters actually reduces the validity.[12] The SOAT is a unique blend of dichotomous-scale and continuous-scale items, thus introducing the concept of expert judgment for each rater. Furthermore, it does not require examinees to follow a linear progression through the 8 continuously scaled items but rather permits them to use their clinical judgment about the most appropriate tests to adequately and competently evaluate the condition. An assessment tool that affords flexibility to the examinee and the rater to use their judgment is not necessarily congruent with the original intent of the OSCE, whereby objectivity is maximized and the overall reliability and stability of the tool are left in question.

The SOAT was designed so thoroughness could be rewarded but efficiency would not be penalized. In fact, thoroughness requires efficiency to complete all tasks. If the examinees are not efficient, they might not complete the entire test and thus might be penalized. The test was designed to be completed in 30 minutes or less to avoid rewarding examinees who are thorough but not competent. The flexibility built into the SOAT design also may address another shortcoming: OSCEs do not capture varying levels of expertise.[14] This study was not designed to distinguish among levels of expertise. However, the SOAT was designed to permit varying levels to be captured and, more important, not to penalize higher levels of expertise. The ability of the SOAT to distinguish among varying levels of expertise, such as fourth-year students compared with those who have been practicing clinically for 5 years, may be an area of future study, thus capturing construct validity of the SOAT.

The rationale for the SOAT design, with a combination of both dichotomous- and continuous-scaling responses, was based on the nature of the tasks being measured. Streiner and Norman[33] maintained that inappropriate scaling responses lead to error in the overall measurement and consequently an inefficient measurement instrument. The tasks that are measured during the history and observation period are organized in a manner that requires the rater to determine whether the question was relevant at the time. A similar argument is made for the continuous scales regarding special testing, for example. If an examinee decided to complete the empty-can test for the shoulder, an expert rater would consider many variables to judge the quality of the performance. Therefore, simply dichotomizing this variable actually may lead to greater error in measurement and lower reliability.[8,12] The magnitude of the continuous-scaling response also appears to have been ignored in the literature concerning practical, performance-based examinations.[33] Designers of the SOAT attempted to reduce error by applying a scaling response for each task that was long enough to maximize the discriminability between raters and examinees.[33,34] Streiner and Norman[33] suggested a continuous scale of 7 points, which can be traced back to a study by Symonds.[35] The ability of human cognition and processing seems to be optimal when the scale has 7 points of discrimination.[35] However, McKelvie[36] concluded that a scale greater than 5 or 6 in length had no statistical advantage. The SOAT was designed with a 6-point scale for the pragmatic reason of fitting it onto 1 page. We believe the scaling-response design in the SOAT is one factor that has led to strong reliability.

Traditional OSCEs typically separate the various subcategories of orthopaedic injury evaluation into its various components and test them separately from one another in different, shorter stations. Separation of these subcategories would result in a low-fidelity, artificial environment that is further removed from how an examinee might act in a real-life environment.[37,38] The SOAT requires examinees to complete the entire continuum in 1 station from history to diagnosis. Raters who observe the entire continuum have a greater understanding of the overall performance of the examinees and, therefore, should be better judges of their clinical competence. Perhaps the SOAT is closer to measuring what Miller[19] intended practical, performance-based examinations to measure: clinical competence. The SOAT has been designed to address this shortcoming of the OSCE and thus may possess stronger validity.[19] However, a constant tension exists between validity and reliability with assessment tools.[39] An assessment tool can be valid but not reliable, yet it can be reliable without being valid.[33,39] Researchers[33,39,40] must establish validity and reliability for an assessment tool to be psychometrically sound.

The SOAT is thought to have good validity due to the previous content validation,[27] good initial reliability, or internal consistency,[24] but the questions of interrater reliability, stability, agreement, and criterion (concurrent) validity have remained. The results of our study demonstrate that the SOAT has good interrater reliability coefficients for the knee and shoulder body regions, both with and without the SP grading included. Several experts have indicated that ICCs greater than 0.70 and less than 0.90 are required for good reliability, and the SOAT achieves that requirement for the knee and shoulder both with and without the SP rating the performance.[33,39] The SOAT falls within those acceptable limits of reliability.

Weir[31] differentiated between types of reliability measures as either (1) absolute consistency or reliability as measured by the SEM or CIs or (2) relative consistency or reliability as measured by ICCs. Again, the ICC values would indicate good reliability if greater than 0.70,[33,39] but the large CIs (approximately 19% for the knee and approximately 22% for the shoulder), as calculated with the SEM, would indicate these scores may not be as precise as they should be in future research. In other words, the study results indicate the SOAT has good relative consistency as indicated by high ICCs but moderate absolute consistency or stability based on the rather large CIs. A question of whether it is acceptable to permit rater scores that range from 49.45% to 78.01% and 55.81% to 81.33% for the shoulder and knee, respectively, arises. Those CIs are likely where the minimal passing score may lie. Thus, with the lack of stability in the score, it is challenging to know where the error may lie with the SOAT: the tool or the rater? Generalizability analysis (theory) is the only statistical technique that provides greater detail of where the error may lie and, thus, should be the focus of future study.[41]

In contrast to the CI data, agreement between raters appeared to lack systematic biases, as evidenced from the Bland-Altman plots. The plots show that 95% of the raters' scores fall within 2 SDs of the mean differences from the mean rater scores and, thus, demonstrate good agreement between raters without systematic bias (Figure).

The final objective of this study was to establish concurrent validity through a concomitant comparison of the mean overall score of the SOAT and a global rating scale completed by the rater after all other items in the SOAT were completed. The global rating scale theoretically represents the expert opinions of whether the raters believed the examinee was competent. A Pearson *r* correlation coefficient was used to measure the relationship between these measures. The strong correlations indicated that the finite details of each item in the SOAT, when summed and a percentage score is provided, correlate well with what an expert rater believes the examinee's score should be. These findings differ somewhat from those of Ringsted et al.[42] They found poor agreement between raters who used a checklist and those who used a global rating scale.[42] However, their methods differed from ours, whereby raters in our study benefitted from using the checklist and the global rating scale.[42] In contrast to the findings of Ringsted et al,[42] some researchers[11,18,43,44] have demonstrated that global rating scales are superior to checklists. One possible explanation for the conflicting results is related to the level of expertise the OSCE is intended to measure.[43,44] Individuals with lower levels of expertise tend to process clinical cases in a stepwise, tasklike fashion, whereas individuals with higher levels of expertise tend to skip some tasks to focus more time on what they deem relevant to each case.[39] However, the SOAT was designed to capitalize on global rating scales, checklists, and dichotomous and continuous measures all in 1 tool, perhaps providing a hybrid among the tools that measured competence with only 1 of those systems.

Our study had several limitations that may frame some of the results. The SP was also the primary author, which could lead to the potential for bias. However, we took steps to prevent bias, such as blinding the SP from the rater evaluations and limiting discussion about student performance to only answering questions for clarification. Attempts to limit the information given to the raters and not to bias their evaluation of the student were made throughout. Having the same SP for all cases also had some advantages. Many sources of error are present in testing students, and by keeping the same SP for all students, greater consistency likely resulted and, thus, reduced the overall error. This has been demonstrated in other examinations in which multiple SPs participated at multiple testing sites and greater SP error resulted.[45] Researchers should have multiple SPs to determine the amount of error that they contribute to the overall error with reliability testing of students using the SOAT.

A convenience sample of both raters and students participated in our study. Raters were chosen based on the minimal criteria described, but all were associated with the examinees from the respective university where the testing took place. Students were chosen based on the minimal criteria described, but they were the students who had attended the university in the testing year only. The results of our study possibly cannot be generalized to a population beyond this cohort, and only further testing on another group will address that question. The scenarios that we tested were specific to the knee and shoulder regions. The question of whether the results could be generalized to other joints or other diagnoses has not been answered.

CONCLUSIONS

The 4 objectives designed to address the purpose of our study have each contributed somewhat mixed results. Specifically, the question of whether the SOAT is a valid and reliable tool to assess clinical competence of orthopaedic injury evaluation has been addressed, but more questions remain. Whereas we found strong interrater reliability, the error associated with the measurements leaves doubt about the stability of the SOAT in future research. The only way to parcel out the error and study it more closely is to conduct a generalizability theory study. On a positive note, no systematic bias was apparent based on the Bland-Altman plots, which is an indication of the agreement between the raters. Finally, the SOAT demonstrated concurrent validity due to the strong correlation between the global rating scale scores of the raters and the summary of their ratings or scores of examinees throughout the SOAT. The use of the SOAT in final, summative-type examinations may be helpful in determining clinical competence of orthopaedic injury evaluation. However, it should not be considered the final authority on clinical competence but rather be put into context within a broader and diverse programmatic evaluation plan.[22] Validity is iterative, so future research on the SOAT will help build its psychometric soundness.

ACKNOWLEDGMENTS

We thank the program leaders at Concordia University (Dr Richard DeMont) and the University of Winnipeg (Dr Glen Bergeron) for their assistance in this study.

REFERENCES

1. MacKay C, Canizares M, Davis AM, Badley EM. Health care utilization for musculoskeletal disorders. *Arthritis Care Res (Hoboken)*. 2010;62(2):161–169.
2. Decker SL, Schappert SM, Sisk JE. Use of medical care for chronic conditions. *Health Aff (Millwood)*. 2009;28(1):26–35.
3. Jordan K, Clarke AM, Symmons DP, et al. Measuring disease prevalence: a comparison of musculoskeletal disease using four general practice consultation databases. *Br J Gen Pract*. 2007; 57(534):7–14.
4. Oswald AE, Bell MJ, Snell L, Wiseman J. The current state of musculoskeletal clinical skills teaching for preclerkship medical students. *J Rheumatol*. 2008;35(12):2419–2426.
5. Abou-Raya A, Abou-Raya S. The inadequacies of musculoskeletal education. *Clin Rheumatol*. 2010;29(10):1121–1126.
6. Turner JL, Dankoski ME. Objective structured clinical exams: a critical review. *Fam Med*. 2008;40(8):574–578.
7. Mavis BE, Henry RC, Ogle KS, Hoppe RB. The emperor's new clothes: the OSCE reassessed. *Acad Med*. 1996;71(5):447–453.
8. Cunnington JPW, Neville AJ, Norman GR. The risk of thoroughness: reliability and validity of global ratings and checklists in an OSCE. *Adv Health Sci Educ Theory Pract*. 1996;1(3):227–233.
9. Solomon DJ, Szauter K, Rosebraugh CJ, Callaway MR. Global ratings of student performance in a standardized patient examination: is the whole more than the sum of the parts? *Adv Health Sci Educ Theory Pract*. 2000;5(2):131–140.
10. McIlroy JH, Hodges B, McNaughton N, Regehr G. The effect of candidates' perceptions of the evaluation method on reliability of checklist and global rating scores in an objective structured clinical examination. *Acad Med*. 2002;77(7):725–728.
11. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklist and global rating scales for assessing performance on an OSCE-format examination. *Acad Med*. 1998;73(9):993–997.
12. Norman G. Editorial: checklists vs. ratings, the illusion of objectivity, the demise of skills and the debasement of evidence. *Adv Health Sci Educ Theory Pract*. 2005;10(1):1–3.
13. Norman GR, van der Vleuten CP, De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Med Educ*. 1991;25(2):119–126.
14. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Acad Med*. 1999;74(10):1129–1134.
15. Rethans JJ, Norcini JJ, Baron-Maldonado M, et al. The relationship between competence and performance: implications for assessing practice performance. *Med Educ*. 2002;36(10):901–909.
16. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. *Med Educ*. 2004;38(2):199–203.
17. Wass V, van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet*. 2001;357(9260):945–949.
18. Reznick R, Regehr G, Macrae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg*. 1997;173(3):226–230.
19. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med*. 1990;65(suppl 9):S63–S67.
20. Griesser MJ, Beran MC, Flanigan DC, Quackenbush M, Van Hoff C, Bishop JY. Implementation of an objective structured clinical exam (OSCE) into orthopedic surgery residency training. *J Surg Educ*. 2012;69(2):180–189.
21. Mohtadi NG, Harasym PH, Pipe AL, Strother RT, Mah AF. Using an objective structured clinical exam to evaluate competency in sport medicine. *Clin J Sport Med*. 1995;5(2):82–85.
22. Schuwirth LW, van der Vleuten CP. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach*. 2011; 33(6):478–485.
23. Lafave MR, Butterwick DJ, Donnon T, Mohtadi N. The feasibility and practicality of employing generalizability theory in performance-based, practical examinations: the SOAT experience [abstract]. *Clin J Sport Med*. 2008;18(2):198.
24. 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.
25. Vivekananda-Schmidt P, Lewis M, Hassell AB; ARC Virtual Rheumatology CAL Research Group. Cluster randomized controlled trial of the impact of a computer assisted learning package on the learning of musculoskeletal examination skills by undergraduate medical students. *Arthritis Rheum*. 2005;53(5):764–771.
26. Nayer M. An overview of the objective structured clinical exam. *Physiother Can*. 1993;45(3):171–178.
27. Lafave M, Katz L, Butterwick D. Development of a content-valid standardized orthopedic assessment tool (SOAT). *Adv Health Sci Educ Theory Pract*. 2008;13(4):397–406.
28. Shrout P, Fleiss J. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420–428.
29. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Method*. 1996;1(1):30–46.
30. Armstrong GD. The intraclass correlation coefficient as a measure of inter-rater reliability of subjective judgments. *Nurs Res*. 1981;30(5): 314–320.
31. Weir JP. Quantifying test-rest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res*. 2005; 19(1):231–240.
32. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986; 1(8476):307–310.
33. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 3rd ed. New York, NY: Oxford University Press; 2003.
34. Neufeld VR, Norman GR. *Assessing Clinical Competence*. New York, NY: Springer; 1985.
35. Symonds PM. On the loss of reliability in ratings due to coarseness of the scale. *J Exp Psychol*. 1924;7(6):456–461.
36. McKelvie SJ. Graphical rating scales: how many categories? *Br J Psychol*. 1978;69(2):185–202.
37. Norcini JJ, Blank LL, Duffy FD, Forna GS. The mini-CEX: a method for assessing clinical skills. *Ann Intern Med*. 2003;138(6): 476–481.
38. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med*. 1995;123(10):795–799.
39. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 3rd ed. Upper Saddle River, NJ: Pearson/ Prentice Hall; 2009.
40. Nunnally JC, Bernstein IH. *Psychometric Theory*. 3rd ed. New York, NY: McGraw-Hill; 1994.
41. Brennan R. Performance assessment from the perspective of generalizability theory. *Appl Psychol Meas*. 2000;24(4):339–353.
42. Ringsted C, Ostergaard D, Ravn L, Pedersen JA, Berlac PA, van der Vleuten CP. A feasibility study comparing checklists and global rating forms to assess resident performance in clinical skills. *Med Teach*. 2003;25(6):654–658.
43. Reznick RK, Regehr G, Yee G, Rothman A, Blackmore D, Dauphinee D. High-stakes examinations: what do we know about measurement? Process-rating forms versus task-specific checklists in an OSCE for medical licensure. *Acad Med*. 1998;73(10):S97–S99.
44. Hodges B. Validity and the OSCE. *Med Teach*. 2003;25(3):250–254.
45. Tamblyn RM, Klass DJ, Schnabl GK, Kopelow ML. The accuracy of standardized patient presentation. *Med Educ*. 1991;25(2):100–109.

SUPPLEMENTAL MATERIAL

Supplemental Appendix. The Standardized Orthopedic Assessment Tool (SOAT) for the knee**.** Abbreviations: AROM, active range of motion; ASIS, anterior-superior iliac spine; IR, internal rotation; IT, iliotibial; Jt, joint; LCL, lateral collateral ligament; MCL, medial collateral ligament; MOI, mechanism of injury; n/a, not applicable; PROM, passive range of motion; PSIS, posterior-superior iliac spine; Psycho, psychological; Rehab, rehabilitation; ROM, range of motion; exam, examination; SHARP, swelling, heat, altered function, redness, and pain; Tib-Fib, tibia-fibula

Found at DOI: http://dx.doi.org/10.4085/1062-6050-49.1.12.S1

Address correspondence to Mark R. Lafave, PhD, CAT(C), Department of Physical Education and Recreation Studies, Mount Royal University, 4825 Mount Royal Gate SW, Calgary, AB, Canada T3E 6K6. Address e-mail to mlafave@mtroyal.ca.