The Intra-rater Reliability of Nine Content-Validated Technical Skill Assessment Instruments (TSAI) for Athletic Taping Skills.

Niko G. Lagumen MSc, CAT(C), ATC; Dale J. Butterwick, MSc, CAT(C); David M. Paskevich, PhD; Tak S. Fung, PhD; Tyrone L. Donnon, PhD
University of Calgary, Calgary, Alberta, Canada

**Objective:** To establish the intra-rater reliability of nine content-validated Technical Skill Assessment Instruments (TSAI) for the skills of athletic taping.

**Setting:** University of Calgary.

**Subjects:** Canadian Certified Athletic Therapists, CAT(C), with a mean ± SD of 9.6 ± 10.8 years as a CAT(C), 7.8 ± 10.9 years as a Supervisory Athletic Therapist, 8.5 ± 12.0 years teaching athletic taping skills, and 9.2 ± 11.5 years evaluating athletic taping skills.

**Design:** Six Certified Athletic Therapists from Canada completed the repetitive evaluations of nine different athletic taping scenarios. Each rater evaluated the performance of a student therapist taping a standardized patient while using the appropriate TSAI designed for each athletic taping scenario. Evaluations occurred once per month for five successive months. Raters viewed the performances on a portable DVD player at a central testing site.

**Measurements:** The percent scores of 270 completed TSAIs were used for analysis. ICC (3, k) was used to quantify the intra-rater reliability. We used a One-way ANOVA with repeated measures to determine if mean differences across testing months existed within raters. Significance was achieved with $\alpha = 0.05$.

**Results:** ICC values for the nine TSAIs ranged from 0.65 to 0.95 with Ankle 3 and Thumb 2 achieving the lowest and highest ICC values respectively. One-way ANOVA with repeated measures did not provide significant mean differences between testing months within each rater.

**Conclusion:** The nine TSAIs possess substantial to almost perfect reliability with seven TSAIs appropriate for summative evaluations and two appropriate for formative evaluations.

**Key Words:** Intra-class correlation (ICC), repeatability.

The skill of athletic taping is fundamental in the scope of practice of Athletic Therapists and Athletic Trainers (AT). Thus, athletic taping and bandaging courses within the accredited programs of AT are requisites for graduation and certification in Canada. Athletic taping skills in academic institutions are both taught and evaluated by educators and practicing Certified Athletic Therapists [CAT(C)]. Evaluation has generally been conducted using instruments that are created by instructors and applicable to various taping scenarios and class requirements. These instruments may incorporate a Likert-type scale and may include evaluations of the history, components of the taping performance, materials, design, and functional capacity. Importantly, there is little evidence in the published literature regarding the validity or reliability of evaluation instruments for the technical skill of athletic taping.

The technical skills of athletic taping are complex. A therapist must take into account various factors that will affect the techniques needed to facilitate a safe return to activity. Evaluations of technical skills differ from the evaluation of knowledge. Technical skills require the interpretation of a case, recall of applicable knowledge, and the subsequent physical performance of a technique appropriate for that case. As Winckel states, “technical ability is elaborate...drawing on issues of motor skills, previous experience, cognitive elements, and adaptability to new and changing situations.” In addition, the element of observation or witness by the rater or examiner requires a level of objectiveness. The complexity of evaluating technical skills lies in the need to accurately translate the examinee’s performance into a representative mark or score. Thus there is a need for instruments to be robust enough to accommodate for the intricacies of a technical skill yet also contain...
a level of objectiveness to facilitate reliability.²

Reliability can simply be defined as consistency.³ In regard to educational measurement reliability speaks to the consistency of an instrument to repetitively produce comparable assessments.⁴⁶ The significance of reliability in educational measurement is the need to consistently reproduce accurate measurements. Thus, in terms of Classical Test Theory, reliability also speaks to the amount of measurement error contained within an observed score.³ Reliability in educational measurement is categorized as Test-retest, Internal Consistency, Rater, or Equivalence models of reliability.³⁷ Rater reliability examines the consistency of raters and can be further classified into inter-rater and intra-rater reliability. Inter-rater and intra-rater reliability are described as the outcome of consistency between and within raters respectively.¹² Eight In education there is a need for educators to evaluate students with both valid and reliable instruments. It is important that educators use instruments that can consistently and accurately evaluate the performance of students, particularly the performance of student’s technical skills.

There is paucity in the academic literature regarding the validity and/or reliability of the instruments used to evaluate the proficiency of athletic taping skills. Only one content-valid checklist specific to the technical skill of athletic taping is found in the literature.⁹

Study Objective and Research Hypothesis

The objective of our study was to establish the intra-rater reliability of nine content validated Technical Skill Assessment Instruments (TSAI). Our primary hypothesis was that nine content validated⁹ TSAIs would produce consistent and repeatable evaluations across five different testing months. Our secondary hypothesis was that no differences in rater scoring between testing months would occur. Relative to our study design, intra-rater reliability is further defined as the consistency of successive measurements within an individual rater.²⁸ The significance of our study is the establishment of reliability to nine content-validated⁹ TSAIs. Additionally, the majority of intra-rater reliability studies in the literature adopt a test-retest model in which the re-evaluation by raters occurs only one additional time after the initial evaluation.¹⁰¹⁵ In contrast, our current study will involve five successive evaluations.

Technical Skill Assessment Instruments (TSAI)

We created nine TSAIs for three anatomical joints (ankle, elbow, and thumb) and content-validated the TSAIs by simple and committee validation.⁹ For each anatomical joint we designed three specific athletic taping techniques that increase in technical difficulty. The technical difficulty of each athletic taping technique increases in relation to the nature and severity of each injury scenario thus requiring additional taping techniques to accomplish the objective. Therefore, Ankle 3 is more difficult than Ankle 1 etc. The ankle TSAIs represent injuries to medial and/or lateral ligaments, with combined ligament injuries representing the athletic taping technique with the highest technical difficulty. The elbow and thumb contained components of hyperextension and/or valgus (abduction) injuries with combinations of injuries again representing the tape job with the highest technical difficulty.

The TSAIs were composed of items that were organized in the logical progression of a taping performance. Evaluations begin with judgments on how the subject prepares for a taping encounter such as ‘Selection of Materials’, ‘Starting Position of the Joint’, and ‘Skin Protection’. Evaluations then end with judgments on ‘Circulation Assessment’, ‘Design’, and ‘Functional Testing’.

To use the TSAIs, raters evaluated a range of 58 to 75 items. Items are dichotomous and the evaluations are conducted by deducting items that are considered inappropriate or insufficient by the rater. There is evidence in the literature that the use of dichotomous scales (yes/no) in structured checklists contribute to higher reliability.¹,¹⁶

Items in each TSAI are not weighted evenly as the validation committee determined some items have more significance than other items. Ebel weights are representative of the importance and difficulty of each item as assigned by the members of the validation committee.⁹ Thus, we assigned Ebel weightings to all items of each of the nine TSAI’s during the validation process.⁹ Overall scores are calculated by subtracting the sum of deducted Ebel points from the numerator and the sum of N/A items from the denominator. Overall scores represent a proportion of these two values.

Methods

Ethical Approval

This project received ethical approval from the Conjoint Health Research Ethics Board of the faculties of Medicine, Nursing and Kinesiology, at the University of Calgary, Grant #17237.

Raters

Six Certified Athletic Therapists [CAT(C)] provided consent to participate in the evaluation of nine video-recorded student therapist (ST)-standardized patient (SP) interactions. Our raters represent a convenience sample of certified ATs who volunteered for the study based on a recruitment announcement sent via email to a database of all certified ATs in the province. The raters had a mean ± SD of 9.6 ± 10.8 years of experience as a CAT(C) and 7.8 ±10.9 years as a Supervisory Athletic Therapist (SAT). Academically, the raters had a mean ± SD of 8.5 ± 12.0 and 9.2 ± 11.5 years teaching and evaluating athletic taping skills, respectively. Rater demographics are located in Table 1.

Rater Training

We conducted a mandatory standardized training program for each rater. The session began by providing an introduction to the TSAIs including their purpose, design, development, and implementation. Training then focused on the use of each TSAI such as: how to score, use of the non-applicable (N/A) designation, and scoring of the ‘Special Techniques’ section.
Table 1. Rater Demographics

<table>
<thead>
<tr>
<th></th>
<th>Years as CAT(C)</th>
<th>Years as Supervisory Athletic Therapist (SAT)</th>
<th>Years teaching athletic taping skills</th>
<th>Years evaluating athletic taping skills</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>27</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

We demonstrated to raters that TSAIs are scored by deducting items that are considered inappropriate or insufficient. In addition, each item on the TSAI may be scored ‘not applicable’ by a rater. For example, if a rater believes the use of under-wrap is not necessary and also not incorrect, the rater may mark that particular item ‘N/A’ and the item is deducted from the denominator value for the total score. Scoring the items within the section of special techniques was then emphasized to correctly guide the rater’s evaluation of different taping techniques that could be used by the ST. Because a student could choose varying techniques to accomplish the objective of each tape job, a scenario dependent list of taping techniques was included in each TSAI. Scoring of this section depended upon the examiner’s functional impression of the technique(s) used and therefore influenced the scoring of the remaining techniques available.

Next, time was then allotted for the raters to familiarize themselves with the associated injuries and their appropriate TSAI. Finally, in addition to the clarification of taping nomenclature in the TSAIs, we reviewed and demonstrated taping techniques for the raters. Training sessions for raters lasted two hours.

We trained the SP separately and we familiarized them with the injury characteristics of each scenario and the standardized responses required. The SP was first introduced to each injury for all of the nine TSAIs. We concluded SP training once each SP gained understanding of the injury and the standard responses required. Training of the SP lasted for two hours.

Technology

To evaluate the same athletic taping performance repetitively, we created a DVD of all nine athletic taping scenarios. Video-recorded interactions provided the ability to control variations in SP performance and variations as a result of time. The videos displayed the SP being taped by the ST. Prior to the filming of each scenario we presented the ST with a written briefing which contained details of the nature and severity of the injury, functional capacity, rehabilitation, physician clearance to return to activity, and objectives of the scenario. Upon reading the briefing we instructed the ST to accomplish the athletic taping objective without structure or guidelines from the investigators.

The videos displayed a standard lateral view of the athletic taping performance in progress. During the mid-way point and at the completion of each athletic taping performance, we included still close-up views (lateral, medial, and superior/frontal) to emphasize details of the athletic taping techniques used. In addition to these three views, we included close-up views of the desired restricted range at the end of each scenario. For example, if the athletic taping objective required restriction of thumb extension, a photograph was taken when the SP was asked to bilaterally fully extend the thumb. This allowed the rater to visually inspect functional restriction in comparison to the contralateral joint.

Data Collection

We collected data over a period of five months at one month intervals. We required raters to evaluate all nine taping scenarios at a central testing site. We used a random ordering of scenarios at each data collection session. Raters viewed the videos on a portable DVD player. Prior to the evaluation of each scenario, we gave raters time to re-familiarize themselves with the appropriate TSAI, sport, injury, and considerations for marking. We provided flowcharts to raters during each data collection session to guide the appropriate use of the N/A column for special techniques, under-wrap, and tape adherent. We allowed raters to view and evaluate each of the nine scenarios once per data collection session. At each testing session we did not allow raters to rewind or repeat the current scenario being evaluated. The principle investigator supervised all testing sessions for each rater. Each TSAI was scored and the percent score calculated and used for analysis once raters completed their evaluations.

Statistical Analysis

We conducted statistical evaluations on SPSS v. 14.0 with a statistical significance level of α = 0.05. Each rater contributed five evaluations for each of the nine separate TSAIs. Thus, the percent score of 270 completed TSAIs were used for analysis. We calculated Intraclass Correlation Coefficients (ICC) to establish intra-rater reliability. The ICC model provides both a degree of correlation and agreement of scores and can be applied to more than two raters. We chose ICC (3, k) based on the purpose and design of the study as well as the type of measurement taken. To reiterate, our purpose of the study was to establish the intra-rater reliability. This requires raters to make multiple evaluations of a single performance. Although the fixed-effect model of ICC (3, k) limits generalizability to only the raters of this study, it is recognized as the appropriate ICC for studies of multiple measurements. An ICC (3, k) is equivalent to Cronbach’s coefficient alpha. Additionally, we used the mean of several evaluations by raters as the unit of analysis. Classifications of reliability coefficients developed by Landis and Koch were used to classify the reliability results. We performed a one-way ANOVA with repeated measures to establish if mean differences across
testing months existed within each rater.\textsuperscript{1,15,22} Results of the ANOVA may indicate variance in raters’ monthly evaluations. Based on evidence in the literature to achieve 80\% power for reliability values of 0.7 to 0.9 with five repeated measurements, we required six subjects.\textsuperscript{23}

### Results

All subjects completed all evaluations thus the analysis contains no drop-outs or missing data. Table 2 outlines the mean, standard deviation, range of scores, and a brief description of each TSAI. The mean score (SD) of each TSAI ranged from 83.10\% to 89.57\% (5.28 to 9.79). Elbow 2 achieved the lowest mean score and Thumb 3 achieved the highest mean score. Of the three anatomical locations, Ankle and Elbow achieved the narrowest and widest range of scores respectively. Of all TSAIs, Ankle 1 and Elbow 2 displayed the narrowest and widest range of scores respectively.

Figures 1-9 provide graphic representation of each rater’s scoring for each individual TSAI. No discernable trends in evaluation are apparent within each TSAI. However, Ankle 3 and Thumb 3 appear to have the most erratic marking trend between testing months indicated by the sharp changes in evaluations between months. This is seen most in Examiners 3 & 4 in Ankle 3 (Figure 3) and Examiners 1, 4 & 6 in Thumb 3 (Figure 9). The

### Table 2. Summary Statistics of Each TSAI

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Athletic Taping Objective</th>
<th>Mean ± SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ankle 1</td>
<td>Limit inversion</td>
<td>89.37 ± 5.30</td>
<td>82.95 to 99.25</td>
</tr>
<tr>
<td>Ankle 2</td>
<td>Limit inversion</td>
<td>88.39 ± 7.08</td>
<td>74.44 to 98.33</td>
</tr>
<tr>
<td>Ankle 3</td>
<td>Limit inversion &amp; eversion</td>
<td>86.54 ± 5.28</td>
<td>79.52 to 97.90</td>
</tr>
<tr>
<td>Elbow 1</td>
<td>Limit hyperextension</td>
<td>87.62 ± 8.34</td>
<td>66.00 to 100.00</td>
</tr>
<tr>
<td>Elbow 2</td>
<td>Limit valgus</td>
<td>83.10 ± 9.79</td>
<td>56.07 to 97.66</td>
</tr>
<tr>
<td>Elbow 3</td>
<td>Limit hyperextension &amp; valgus</td>
<td>84.95 ± 9.29</td>
<td>65.26 to 100.00</td>
</tr>
<tr>
<td>Thumb 1</td>
<td>Limit MCP hyperextension</td>
<td>86.69 ± 8.45</td>
<td>68.52 to 97.53</td>
</tr>
<tr>
<td>Thumb 2</td>
<td>Limit MCP valgus</td>
<td>85.49 ± 8.81</td>
<td>67.76 to 100.00</td>
</tr>
<tr>
<td>Thumb 3</td>
<td>Limit MCP hyperextension &amp; valgus</td>
<td>89.57 ± 7.82</td>
<td>70.23 to 98.86</td>
</tr>
</tbody>
</table>

Table 3. ICC (3, k) and 95\% Confidence Intervals

<table>
<thead>
<tr>
<th>Scenario</th>
<th>ICC (3, k)</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ankle 1</td>
<td>0.89</td>
<td>0.65 to 0.98</td>
</tr>
<tr>
<td>Ankle 2</td>
<td>0.93</td>
<td>0.76 to 0.99</td>
</tr>
<tr>
<td>Ankle 3</td>
<td>0.65</td>
<td>-0.15 to 0.095</td>
</tr>
<tr>
<td>Elbow 1</td>
<td>0.71</td>
<td>0.04 to 0.95</td>
</tr>
<tr>
<td>Elbow 2</td>
<td>0.92</td>
<td>0.74 to 0.99</td>
</tr>
<tr>
<td>Elbow 3</td>
<td>0.92</td>
<td>0.73 to 0.99</td>
</tr>
<tr>
<td>Thumb 1</td>
<td>0.94</td>
<td>0.80 to 0.99</td>
</tr>
<tr>
<td>Thumb 2</td>
<td>0.95</td>
<td>0.84 to 0.92</td>
</tr>
<tr>
<td>Thumb 3</td>
<td>0.85</td>
<td>0.51 to 0.98</td>
</tr>
</tbody>
</table>

Table 4. One-way ANOVA with Repeated Measures

<table>
<thead>
<tr>
<th>Scenario</th>
<th>One-way ANOVA with Repeated Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ankle 1</td>
<td>$F = 0.453$ (df = 4, 20) $p = 0.769$</td>
</tr>
<tr>
<td>Ankle 2</td>
<td>$F = 0.231$ (df = 4, 20) $p = 0.918$</td>
</tr>
<tr>
<td>Ankle 3</td>
<td>$F = 0.722$ (df = 4, 20) $p = 0.587$</td>
</tr>
<tr>
<td>Elbow 1</td>
<td>$F = 0.762$ (df = 4, 20) $p = 0.562$</td>
</tr>
<tr>
<td>Elbow 2</td>
<td>$F = 0.911$ (df = 4, 20) $p = 0.477$</td>
</tr>
<tr>
<td>Elbow 3</td>
<td>$F = 0.460$ (df = 4, 20) $p = 0.764$</td>
</tr>
<tr>
<td>Thumb 1</td>
<td>$F = 0.681$ (df = 4, 20) $p = 0.613$</td>
</tr>
<tr>
<td>Thumb 2</td>
<td>$F = 0.234$ (df = 4, 20) $p = 0.916$</td>
</tr>
<tr>
<td>Thumb 3</td>
<td>$F = 0.271$ (df = 4, 20) $p = 0.893$</td>
</tr>
</tbody>
</table>

graphs reveal that Examiner 6 is the most stringent examiner and Examiner 3 the most lenient.

ICC values ranged from 0.65 to 0.95 for all nine TSAI with Ankle 3 and Thumb 2 achieving the lowest and highest ICC values respectively. Table 3 provides the ICC and corresponding 95\% confidence interval for each TSAI. One-way ANOVA with repeated measures did not provide evidence of mean differences between testing months. Table 4 provides the one-way ANOVA results. Ankle 3 produced the lowest ICC of all nine TSAIs. Additionally, it is worthy to note that it also produced the largest 95\% confidence interval and contained the smallest range of scores.
Figure 1. Results of ankle TSAI 1 rater scoring

Figure 2. Results of ankle TSAI 2 rater scoring
Figure 3. Results of ankle TSAI 3 rater scoring

Figure 4. Results of elbow TSAI 1 rater scoring
Figure 5. Results of elbow TSAI 2 rater scoring

Figure 6. Results of elbow TSAI 3 rater scoring
Figure 7. Results of thumb TSAI 1 rater scoring

Figure 8. Results of thumb TSAI 2 rater scoring
Discussion

Within medical education there has long been a trend of objective and structured assessments of competence in knowledge and physical skills. Subsequently, general, specialty, and subspecialty surgeons are urging the need for objective and structured assessment of surgical skills. The significance of objective and structured assessment of surgical skill is rooted in the consequences of poor or inadequate surgical competency. Proficiency of surgical skill is directly related to the positive outcomes of cure, decreased complication rates, and return to daily activities. Consequently the objective measurement of technical skill in surgery can lead to identification of substandard skill and influence the reform of training programs. Although athletic taping is far removed from the technical skill of surgery, the importance of proficiency is also directly related to the outcomes of dealing with patient populations. Application of athletic taping skills requires sound decision making and technical proficiency to ensure an athlete will return and remain safe throughout activity.

The measurement of technical ability in athletic taping has yet to be studied vigorously. As an outcome of a structured process of development and testing, results of our nine different TSAIs have shown to possess both content validity and recently established intra-rater reliability. According to classifications of reliability coefficients developed by Landis and Koch, two TSAIs possess substantial reliability (coefficients of 0.61 to 0.80) and the remaining seven TSAIs possess almost perfect reliability (coefficients of 0.81 to 1.00). Various studies have indicated minimum standards of reliability dependent upon the implication of the testing results. Some studies have suggested evaluation instruments producing reliability coefficients lower than 0.8 or 0.9 should not be used for high-stakes examinations or pass/fail decisions. However, for acceptance examinations, in-training evaluations, or less important situations it has been suggested that reliability coefficients of 0.6 to 0.8 may be acceptable. In light of this, all nine of our TSAIs possess reliability suitable for formative evaluation and seven of our nine TSAIs are also suitable for use in high-stakes evaluations.

Our ANOVA results indicate no mean differences occurred across testing months within raters. The relevance of this finding is applicable to the training of raters. With no significant variations in evaluations over one-month intervals, training of raters can be done once every five months with no re-training required if the TSAIs were used in either summative evaluations once per semester or at shorter formative evaluation intervals (ie. monthly).

We established the inter-rater reliability of our nine TSAIs in a previous study. A total of 114 different students were evaluated by 14 different raters from five of the six accredited AT programs across Canada. We randomly allocated each student to one of the nine scenarios. We required three raters to simultaneously evaluate the performance of the ST on a SP. Cronbach’s coefficient alpha was used to quantify inter-rater reliability. Our Cronbach’s coefficient for all nine TSAIs ranged from 0.604 to 0.945 with Thumb 3 and Elbow 1 achieving the lowest and highest Cronbach’s values respectively (unpublished data).
In light of our inter-rater reliability study our intra-rater reliability results parallel the relationship of concurrent reliability studies in the literature. The literature indicates that intra-rater reliability studies tend to produce higher reliability coefficients than inter-rater reliability studies.\textsuperscript{12,13,15} Six of our 9 TSAIs produced higher intra-rater reliability coefficients than their comparative inter-rater reliability coefficients. Lower intra-rater versus inter-rater reliability values were produced for Ankle 1 and 3 and Thumb 1.

In regards to Ankle 3, our low ICC value attained may be attributable to the homogeneity of the scoring. Evidence of homogeneity may be seen in the small range of scoring produced by the raters (Table 2). Within the literature it is noted that large variations in scoring produce greater reliability values.\textsuperscript{3} However, Ankle 1 produced a smaller range of scores than Ankle 3. This smaller range of scores for Ankle 1 subsequently produced a higher ICC value and smaller confidence interval, thus indicating the attribute of homogeneity as a less likely cause of the low ICC value in Ankle 3. By examining Figure 3, it may be more likely that the cause of the lower ICC value of Ankle 3 may be better attributed to the erratic marking by examiners. Ankle 3 required the student to perform a neutral inversion/eversion ankle tape job. Conceptually, this required the student to apply techniques such as Stirrups and Heel Locks in opposing manners to maintain a neutral positioning of the subtalar joint. In regard to heel locks, if a rater is not accustomed to identifying the appropriate direction of applying the heel lock they may be inclined to arbitrarily mark the technique as either correct or incorrect regardless of the actual performance. Thus it is speculated that our low ICC and large 95% confidence interval may be a result of this factor.

All nine of our TSAIs contain items that evaluate history taking, performance of the taping skill, and examiner judgments on functional capacity. The video-taped intra-rater performances used still photographs comparing the restricted range to the contralateral joint in lieu of an examiner's hands-on evaluation. It is common in the literature for studies to separately examine the skills of history taking, physical examination, technical skills, or underlying traits such as competency or communication.\textsuperscript{11,26,27} It may be suspected that our inclusion and aggregate analysis of these different items may affect the reliability. Thus, further research should include item analysis to describe and understand the relationship of individual items despite the established intra-rater reliability of the current TSAIs.

Additionally, ICC (3,k) which involves multiple evaluations by the same rater does not afford generalizability to a larger population of raters.\textsuperscript{3} Therefore, it is suggested that the intra-rater reliability should be continually re-assessed when used with different sets of raters.

Although our reliability evidence is presented, evidence of generalizability or item analysis will add to the internal structure validity of the TSAIs. Due to the erratic marking evident in the data (Figure 3 and 9), further generalizability and item analysis studies are suggested to determine the precise sources of measurement error. It is speculated that the use of mean scores due to the nature of the statistic used, ICC (3, k), may have masked the effects of these erratic marking trends. As checklists provide higher objectivity,\textsuperscript{1,16} it is worthy to explore the large variation of scoring within some examiners and TSAIs. Generalizability studies may highlight sources of error unaccounted for in our present study design. Item analysis may provide valuable insight into the performance of items within the TSAIs. In light of these limitations, it is suggested that Ankle 3 and Thumb 3 may require further evaluation and revision to improve their reliability.

Conclusion

The results of our nine TSAIs have shown substantial to almost perfect intra-rater reliability according to classifications outlined by Landis and Koch. According to these classifications, seven of the nine TSAIs possess reliability suitable for high-stakes examinations and two possess reliability suitable for in-training or formative evaluations. However, two of the nine TSAIs (Ankle 3 and Thumb 3) require further investigation and revision to improve their reliability to the standard of the other seven TSAIs. Generalizability studies and/or item analysis will provide comprehensive assessments of reliability to the nine content-validated TSAIs. Specifically, these further investigations may provide insight into the lower reliability of TSAIs 3 and 9, which required the highest technical proficiency for the respective joint. It is believed that a Generalizability study will provide exact sources of measurement error and item analysis will indicate the performance of each item in the TSAIs in order to improve the reliability of TSAI Ankle 3 and Thumb 3.

References


**Acknowledgments**

The authors would like to thank the Social Sciences and Humanities Research Council (SSHRC) of Canada for funding this project.