2018

Inter-Rater Agreement and Validity of a Tackling Performance Assessment Scale in Youth American Football

Eric Schussler
Old Dominion University, eschussl@odu.edu

Richard J. Jagacinski

Susan E. White

Ajit M. Chaudhari

John A. Buford

See next page for additional authors

Original Publication Citation

Schussler, E., Jagacinski, R. J., White, S. E., Chaudhari, A. M., Buford, J. A., & Onate, J. A. (2018). Inter-rater agreement and validity of a tackling performance assessment scale in youth American football. I*nternational Journal of Sports Physical Therapy, 13*(2), 238-246.

IJSPT

ORIGINAL RESEARCH

INTER-RATER AGREEMENT AND VALIDITY OF A TACKLING PERFORMANCE ASSESSMENT SCALE IN YOUTH AMERICAN FOOTBALL

Eric Schussler, PhD, PT, ATC[1]
Richard J. Jagacinski, PhD[2]
Susan E. White, PhD[2]
Ajit M. Chaudhari, PhD[2]
John A. Buford, PhD, PT[2]
James A. Onate, PhD, ATC[2]

ABSTRACT

Background: Long term neurologic injury and concussion have been identified as risks from participation in American football. Altering tackling form has been recommended to reduce the risk of neurologic injury caused by head accelerations when tackling. The purpose of this research is to determine the inter-rater agreement and validity of the Qualitative Youth Tackling System (QYTS), a six-item feedback scale to correct tackling form, when utilized by novice and expert raters.

Hypothesis: Experienced raters will have higher levels of agreement with each other and with motion capture when compared to novice raters.

Methods: Both novice and experienced raters viewed video of youth athletes (ages 9-13) tackling a dummy in a laboratory setting along. The raters identified successful performance according to a binary rating scale for each component. Analysis of both the raters' agreement with each other and with an objective motion capture measure were completed.

Results: Fliess' Kappa measures between all raters were found to be moderate for head placement (k = .48), fair for cervical extension (k = .38), trunk inclination (k = .37), shoulder extension (k = .27) and step length (k = .29), and there was no agreement for pelvic height (k = .-16). When compared to the dichotomized validation measures of each of the five components provided by the motion capture system the average Cohen's Kappa agreement was substantial for pelvic height (k = .63), fair for step length (k = .34), cervical extension (k = .40), trunk inclination (k = .35), and slight for shoulder extension (k = .16). The experienced raters outperformed the novice raters in all categories.

Conclusion: The results of this study indicate that skilled raters are better able to identify the movement patterns included in the QYTS when compared to a validation measure as well have higher rates of inter-rater agreement than novice raters.

Level of Evidence: 3b

Keywords: Concussion, Feedback, Football, Motor Learning

CORRESPONDING AUTHOR
Eric Schussler
School of Physical Therapy and Athletic Training
Norfolk, VA 23529
Phone: (757) 683-4492,
Fax: (757) 683-4410
E-mail: eschussl@odu.edu

[1] Old Dominion University, Norfolk, VA, USA
[2] The Ohio State University, Columbus, OH, USA

The Ohio State University Biomedical Sciences Institutional Review Board Protocol Number: 2015H0010. The authors report no conflicts of interest.

INTRODUCTION

A 2015 position statement by the American Academy of Pediatrics recommended, "officials and coaches must enforce the rules of proper tackling, including zero tolerance for illegal, head-first hits."[1] Concussions in high school football occur at a rate of 6.71 injuries per 10,000 athlete exposures, this number jumps to 30.07 injuries per 10,000 athlete exposures in competition.[2] Poor form, creating head contact during blocking and tackling is the most prevalent mechanism of injury or activity associated with concussion in American football.[2] An appropriate instruction and feedback methodology to improve tackling form has yet to be determined. Verbal feedback is the standard mechanism utilized to improve movement technique in athletes of all ages and sports. The ability to provide consistent and valid feedback is crucial to the success of any coaching intervention, yet often high rates of variability exist in the provision of feedback.

Coaches and medical professionals often visually estimate activity in order to provide feedback. Caution should be exercised when providing feedback developed solely from visual estimation, as this technique can create highly variable feedback. While visual estimation of movement patterns is standard practice in coaching[3], the use of additional measurement techniques such as video applications has increased.[4–7] Visual estimation of joint motion has been reported to be highly variable with limitations in its accuracy.[8–11] Despite these concerns, visual estimation of movement requires no equipment and can be performed immediately without data processing. Due to this simplicity, visual estimation of movement is commonly utilized in movement instruction and training. Rater training and utilization of standardized procedures has been shown to improve rater agreement in assessment of dynamic movements.[12–14]

Providing consistent feedback to learners is important to develop the skill being learned. When developing motor strategies, learners are better able to attain a higher level of performance when the model or feedback they receive is consistent.[15–19] Combined feedback from visual estimation and other sources are common in feedback mechanisms and with training can be reliable. The purpose of this study was to identify the inter-rater agreement and validity of a six-criteria tackling scale utilizing video review. Identification of the rater's ability to provide both consistent and accurate feedback is important in developing training tools to improve tackling form. The development of a standardized tackling feedback tool will give sport and movement coaches the ability to provide appropriate feedback both in a verbal only mechanism as well as in combination with other modalities. Therefore, the purpose of this research is to determine the inter-rater agreement and validity of the Qualitative Youth Tackling System (QYTS), a six-item feedback scale to correct tackling form, when utilized by novice and expert raters.

METHODS

The Qualitative Youth Tackling Scale (QYTS) is a head up, vertical style tackle developed to limit athlete head contact while completing an effective tackle. The QYTS (Figure 1) is a visually observed, objective based scale created to instruct a vertical, head up tackling form that mimics the Heads Up Tackle© form previously recommended by USA Football.[20] This scale is designed to provide feedback on the components of the technique believed to be most related to safety while maintaining performance. This system applies quantifiable, objective actions during the tackle to a subjective feedback mechanism that aligns with the overall form requirements of the Heads Up Tackle©. To determine an overall score, participants are subjectively assigned a point for successful completion of the specified movement measure.

Inter-rater agreement was examined utilizing two experienced clinicians with six (ATC) and ten years (PT, ATC) of post-certification experience, respectively, and two novice raters with no formal training in movement evaluation. All participants were informed of the benefits and risks of participating in this study and signed an IRB approved consent form. Participants were provided with an interactive text and video training module on the components of the Qualitative Youth Tackling Scale (QYTS). The rater training included an explanation of the correct tackling form, examples of expert tackling, and an immediate feedback pre-test utilizing video examples of youth athletes performing both correct and

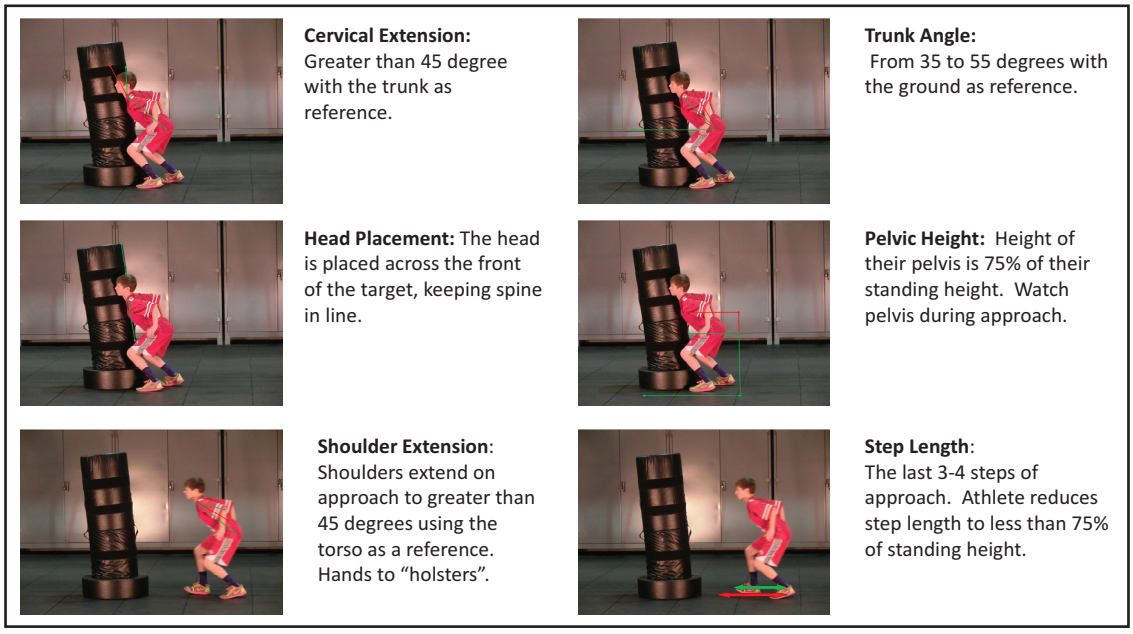

Figure 1. *Six item criteria evaluated during tackling training.*

incorrect tackling. Each rater reported their evaluation of the performance as correct or incorrect as it pertained to the guidelines for each movement item. Participants were required to achieve 80% accuracy on the pre-test prior to rating experimental trials. The total time spent on the training prior to rating the experimental videos was recorded to determine training exposure. Participants were then given 20 video trial examples to rate independently. The raters were able to review the video as many times as needed and were given full control over the playback of each video. The total time to complete the rating was recorded. Overall rater agreement was calculated utilizing a Fleiss' Kappa score. Rater agreements between two experienced, two novices and between experienced and novices were calculated utilizing Cohen's Kappa scores and positive (PA) and negative agreement (NA).

In order to understand the relationship between the raters' evaluation of the performance and the movement being performed, agreement between the raters' scores and a validation standard were performed utilizing a dichotomous split of the motion capture data, within or outside of the desired range of motion of the movement goal, to calculate averaged Cohen's Kappa scores, PA and NA. Because accurate visual estimation is inherently difficult, the validation measure was dichotomized in increasing bands of five percent accuracy from 100% to 80% using a Banded Cohen's Kappa. This expanding band is utilized to determine the potential accuracy of raters. An increasing rate of agreement indicates the raters could be more accurate if they are allowed increased latitude with their response. A decreasing trend indicates increased latitude does not positively affect the agreement outcome and the raters were already at their highest level of agreement. A level line indicates no change in agreement with increased latitude and that the measure is stable. For example, the dichotomized acceptable shoulder movement was adjusted in increments of 5% of 45°: 95% = 42.75°, 90% = 40.5°, 85% = 38.25, 80% = 36°. Averaged Cohen's Kappa scores, PA and NA were then calculated for each point to determine if an expanded definition of accuracy increased rater agreement.

RESULTS

Fliess' Kappa measures between all raters were found to be moderate for head placement ($k = .48$), fair for cervical extension ($k = .38$), trunk inclination ($k = .37$), shoulder extension ($k = .27$) and step length ($k = .29$), and there was no agreement for pelvic height ($k = .-16$) (Table 1). Cohen's Kappa measures

between experienced found substantial agreement between ratings of cervical extension ($k = .69$), head placement ($k = .61$), pelvic height ($k = .73$) and shoulder extension ($k = .70$). Step length results indicate moderate agreement ($k = .49$) and trunk inclination results indicate fair agreement ($k = .24$) (Table 2). Cohen's Kappa measures between the two novice raters found moderate agreement for head placement ($k = .41$). Step length ($k = .34$), trunk inclination ($k = .40$), and shoulder extension ($k = .34$) were found to have fair agreement. Slight agreement was found for cervical extension ($k = .15$) and pelvic height ($k = .11$) (Table 3).

When compared to the dichotomized validation measures of each of the six components provided by the motion capture system the Experienced rater's average Cohen's Kappa agreement was substantial for pelvic height ($k = .68$), moderate for step length ($k = .44$) and cervical extension ($k = .55$) and fair for trunk inclination ($k = .31$) and shoulder extension ($k = .27$) (Table 4). The novice raters had lower levels of agreement, moderate for pelvic height ($k = .57$), fair for cervical extension ($k = .25$), trunk inclination ($k = .39$), and step length ($k = .24$) and slight for shoulder extension ($k = .05$) (Table 5).

Banded Cohen's Kappa comparisons utilizing averaged measures from raters and the values derived from motion capture found increasing agreement in measures of trunk inclination ($k = .35$ to $.50$) and shoulder extension ($k = .16$ to $.55$) with decreasing required accuracy while the agreement between raters and motion capture in pelvic height ($k = .62$

Table 1. *Fleiss Kappa Measures between all raters.*

	Cervical extension	Trunk Inclination	Head placement	Pelvic height	Shoulder extension	Step length
Fleiss' Kappa	0.38	0.37	0.48	-0.16	0.27	0.29
Lower Bound	0.20	0.19	0.30	-0.34	0.09	0.11
Upper Bound	0.55	0.54	0.66	0.02	0.45	0.47

Fleiss Kappa Measures: Almost Perfect: 0.81-1, Substantial: 0.61-0.80, Moderate:0.41-0.60, Fair: 0.21-0.40, Slight: 0-0.20. Negative scores indicate no agreement between scoring.

Table 2. *Cohen's Kappa, Positive and Negative Agreement percentage between AT raters.*

	Cervical extension	Trunk Inclination	Head placement	Pelvic height	Shoulder extension	Step length
Cohen's Kappa	0.69	0.24	0.61	0.73	0.70	0.49
Positive Agreement	87%	40%	94%	80%	84%	60%
Negative Agreement	82%	80%	67%	93%	86%	87%

Cohens Kappa Measures: Almost Perfect: 1-0.81, Substantial: 0.80-0.61, Moderate: 0.60-0.41-, Fair: 0.40-0.21, Slight: 0.20-0.

Table 3. *Cohen's Kappa Positive and Negative Agreement percentage between Novice Raters.*

	Cervical extension	Trunk Inclination	Head placement	Pelvic height	Shoulder extension	Step length
Cohen's Kappa	0.15	0.40	0.41	0.11	0.34	0.34
Positive Agreement	64%	57%	88%	57%	52%	77%
Negative Agreement	40%	77%	50%	53%	35%	57%

Cohens Kappa Measures: Almost Perfect: 1-0.81, Substantial: 0.80-0.61, Moderate: 0.60-0.41-, Fair: 0.40-0.21, Slight: 0.20-0.

Table 4. *Cohen's Kappa Positive and Negative Agreement percentage between rater and validation measure for AT raters.*

		Cervical extension	Trunk Inclination	Pelvic height	Shoulder extension	Step length
Cohen's Kappa	Rater 1	0.50	0.15	0.74	0.30	0.39
	Rater 2	0.60	0.48	0.63	0.24	0.48
	Average	0.55	0.31	0.68	0.27	0.44
Positive Agreement	Rater 1	76%	36%	80%	46%	55%
	Rater 2	82%	57%	73%	40%	57%
	Average	79%	47%	76%	43%	56%
Negative Agreement	Rater 1	74%	76%	93%	74%	83%
	Rater 2	78%	91%	90%	80%	91%
	Average	76%	83%	91%	77%	87%

Cohens Kappa Measures: Almost Perfect: 1-0.81, Substantial: 0.80-0.61, Moderate: 0.60-0.41-, Fair: 0.40-0.21, Slight: 0.20-0.

Table 5. *Cohen's Kappa Positive and Negative Agreement percentage between rater and validation measure for Novice raters.*

		Cervical extension	Trunk Inclination	Pelvic height	Shoulder extension	Step length
Cohen's Kappa	Rater 1	0.30	0.38	0.44	0.01	0.24
	Rater 2	0.20	0.40	0.69	0.10	0.24
	Average	0.25	0.39	0.57	0.05	0.24
Positive Agreement	Rater 1	63%	50%	67%	25%	47%
	Rater 2	69%	57%	80%	31%	47%
	Average	66%	54%	73%	28%	47%
Negative Agreement	Rater 1	67%	88%	73%	50%	61%
	Rater 2	43%	77%	88%	67%	61%
	Average	55%	82%	80%	58%	61%

Cohens Kappa Measures: Almost Perfect: 1-0.81, Substantial: 0.80-0.61, Moderate: 0.60-0.41-, Fair: 0.40-0.21, Slight: 0.20-0.

to .00) comparisons decreased with decreasing required accuracy (Figure 2). Banded positive agreement increased between 100% and 90% accuracy for step length (51% to 57%) and trunk inclination (50% to 65%), while shoulder extension continued to improve (35% to 78%) through 80% of the validity measure (Figure 3). Banded negative agreement remained stable for all measures with the exception of pelvic height which decreased from 86% agreement at 100% of the validity measure to 75% at 90% of the validity measure then sharply to 0% at 80% of the validity measure (Figure 4). Average time to complete the training was 34 ± 8 minutes. Average time to complete the rating of the 20 videos was 20.5 ± 3 minutes.

DISCUSSION

Raters of the QYTS obtained substantial to slight agreement (dependent on the specific movement) when identifying the movements performed during a tackle when compared between themselves, experienced to novice, and themselves to motion capture (a validation standard). A higher degree of accuracy and agreement was found between raters with experience evaluating human movement. Raters with movement evaluation training (Physical Therapists and Athletic Trainers) had higher levels of agreement than non-certified novices through most movements as well as a higher level of agreement with the validation measurements when compared to the novice raters. Banded Kappa analysis indicated the agreement between raters improved when accepting a lower percentage of accuracy compared to the motion capture system for measures of shoulder extension and trunk inclination, decreased agreement at lower measures of accuracy for pelvic height and no change for step length and cervical angle.

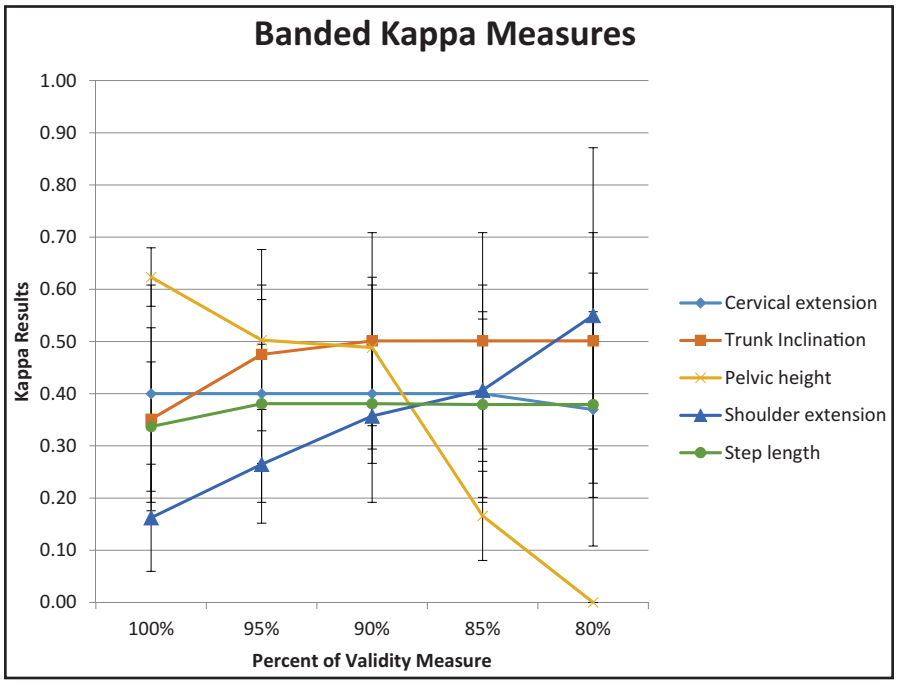

Figure 2. *Banded Average Cohen's Kappa Measures between all raters and motion capture system, banded from 100% of desired movement to 80% of desired movement.*

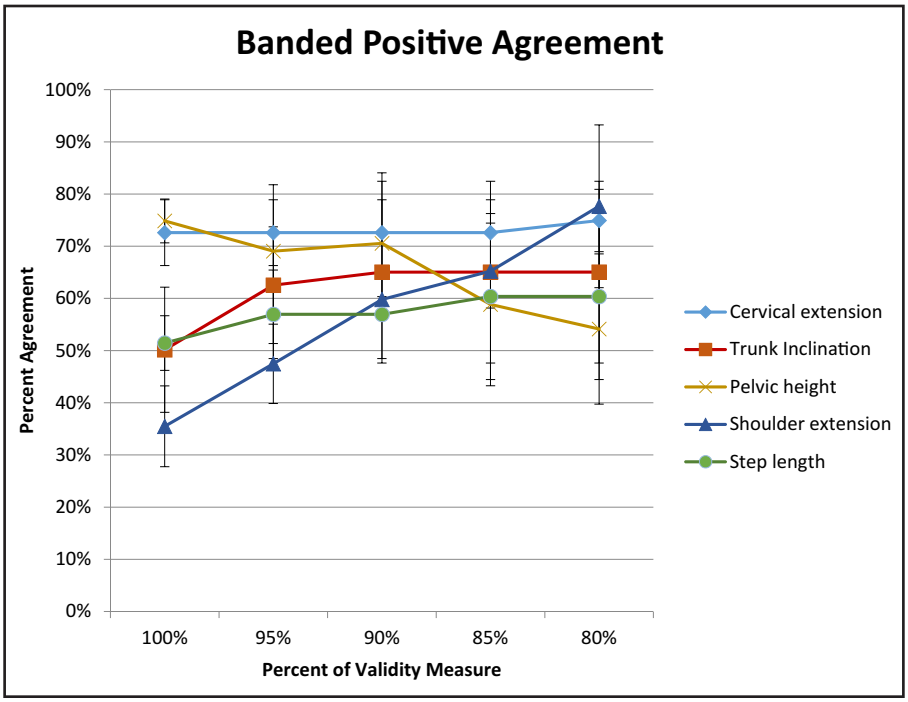

Figure 3. *Banded Average Positive Agreement Measures between all raters and motion capture system.*

Overall agreement measures between raters found fair agreement between all raters. When rating cervical extension, trunk inclination, head placement, shoulder extension and step length the raters were able to achieve Fleiss' Kappa ratings within the fair (k = .21 to .41) range. While agreement may be low in these results, the outcomes are comparable to other studies of visual estimation[9,11,12,21] and better than

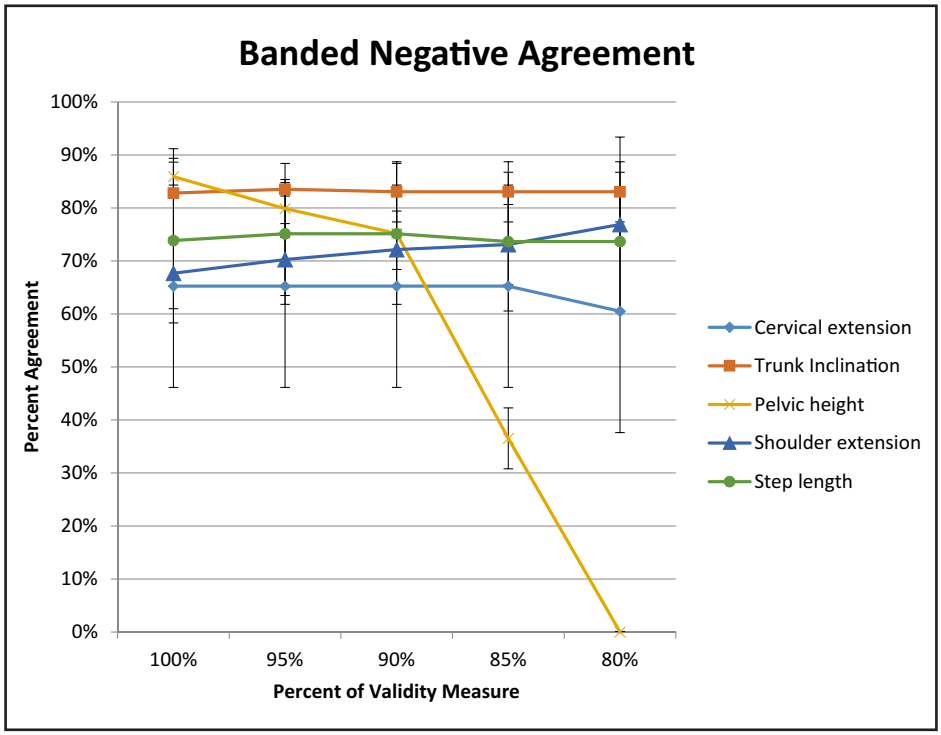

Figure 4. *Banded Average Negative Measures between all raters and motion capture system.*

others.[22] Visual estimation of movement is often hampered by difficulty judging the movements produced. This result is seen across many areas of study, such as knee motion during running [8] and cervical spine motion.[11] In the case of the QYTS, the raters were able to utilize video playback to improve their evaluation of the movement, though the results of this study may have been affected by the number of variables evaluated[7] and the use of one camera angle.

Physical Therapists, Athletic Trainers and other human movement practitioners are experienced in evaluating human movement. This training may explain why experienced raters were able to achieve both higher agreement between raters and between raters and the validity measures.[12] In their training, human movement practitioners would have been exposed to many cases of evaluating movement visually. This may have allowed the raters to gain a perspective or evaluation technique to improve their accuracy and reliability when viewing human movement. Human movement practitioners also have a better understanding of the visual appearance of the range motion referenced in the training, having had experience measuring and evaluating movement. They are likely better able to understand the reference to 45 degrees of shoulder extension during QYTS training, having measured such movements themselves as part of their training. The increased agreement seen in both the inter-rater comparison and between experienced raters and validity measures may be a function of the additional training of the experienced raters.

In the banded Kappa analysis, when the percentage of accuracy required by the validity rating is reduced, the validity agreement for trunk inclination and shoulder extension improved. This result indicates raters were capable of higher agreement with less stringent requirements. They were less able to identify the movement exactly, but a small allowance in the accuracy requirement increased their agreement. As the accuracy required was reduced, shoulder extension agreement improved across all bands. While it appears that raters had difficulty identifying movements over 45 degrees, they were able to separate those who extended the shoulder to at least 80% of the desired movement profile. Further investigation reveals raters consistently responded affirmatively down to 50% of the desired rating, or 25 degrees shoulder extension indicating

their estimation of shoulder range of motion to be a rather course measure. They could not identify a movement difference between 45 degrees and 25 degrees, but could identify that they had passed into a measure beyond 25 degrees and considered that successful shoulder extension. Rating of trunk inclination improved with a shift to 95% accuracy at which time the improvement in agreement stabilized. Raters reached their highest consistency in agreement when the movement was considered correct between 43 and 57 degrees trunk angle. This would indicate raters either just missed correctly identifying the motion correctly or were not close in their estimation. These expanded movement parameters may be satisfactory for proper execution of the tackle, though this answer is beyond the scope of this project.

Agreement on pelvic height between the raters and motion capture decreased with an increased acceptable range. Raters achieved their highest agreement with 100% accuracy to the pelvic height requirement, indicating they were achieving their best possible accuracy at the desired goal. Accepting measures beyond 100% accuracy caused measures that were correctly identified as outside of the goal motion to be included in the desired range, creating less agreement. Cervical extension and step length measures stayed stable with an expanded range. This indicates raters did not benefit from a relaxation of the standard. This most likely is caused by large errors in the estimated range of motion for those who incorrectly identified the motion.

The pre-assessment training and assessment for the raters may not have been sufficient to ensure a thorough understanding of the method of movement evaluation.[23] The training program for raters should be evaluated, though additional training maybe ineffective due to the inherent limitations of visual estimation of movement.[8] All of these variables may have played a part in the less than perfect agreement seen in the comparison between all raters.

Limitations to this research include a small sample of raters with limited training on the QYTS. Future studies should include a larger cohort of raters, both experienced and inexperienced, who have participated in a more in-depth training program. This research also only identifies trends in video evaluation of the movement, real-time rating of movement presents a different set of requirements and should be considered separately. Coaches, trainers and health care professionals often provide verbal feedback to players without the aid of video; thusly additional research should examine the ability of the QYTS to be utilized in real time. Additional research should also examine the intra-rater reliability of the QYTS scale over time.

CONCLUSION

The results of this study indicate that the inter-rater agreement and validity measures for the QYTS show a range of agreement from substantial to slight across the five rated movement components. With refinement this system may function as a mechanism to provide feedback during video review of tackling practice in American football. More experienced and movement trained raters showed a higher level of agreement both with each other and with a validation standard. It is important when providing feedback during motor learning that the learner to be provided with consistent and correct information regarding their performance. This study indicates those with more experience analyzing human movement are able to provide more accurate and reliable feedback to the learner.

REFERENCES

1. Council on Sports Medicine and Fitness. Tackling in youth football. *Pediatrics*. 2015; 136(5):e1419-e1430.
2. Marar M, McIlvain NM, Fields SK, Comstock RD. Epidemiology of concussions among united states high school athletes in 20 sports. *Am J Sports Med*. 2012; 40(4):747-755.
3. Markland R, Martinek, TJ. Descriptive analysis of coach augmented feedback given to high school varsity female volleyball players. *J Teach Phys Educ*. 1988; 7(4):289-301.
4. Boden BP, Torg JS, Knowles SB, Hewett TE. Video analysis of anterior cruciate ligament injury: abnormalities in hip and ankle kinematics. *Am J Sports Med*. 2009; 37(2):252-259.
5. Boyer E, Miltenberger RG, Batsche C, Fogel V. Video modeling by experts with video feedback to enhance gymnastics skills. LeBlanc L, ed. *J Appl Behav Anal*. 2009; 42(4):855-860.
6. Hewett TE, Torg JS, Boden BP. Video analysis of trunk and knee motion during non-contact anterior

cruciate ligament injury in female athletes: lateral trunk and knee abduction motion are combined components of the injury mechanism. *Br J Sports Med*. 2009; 43(6):417-422.

7. Olsen O-E. Injury mechanisms for anterior cruciate ligament injuries in team handball: a systematic video analysis. *Am J Sports Med*. 2004; 32(4):1002-1012.
8. Krosshaug T, Nakamae A, Boden B, et al. Estimating 3d joint kinematics from video sequences of running and cutting maneuvers—assessing the accuracy of simple visual inspection. *Gait Posture*. 2007; 26(3):378-385.
9. Rose V. Visual estimation of finger angles: do we need goniometers? *J Hand Surg J Br Soc Surg Hand*. 2002; 27(4):382-384.
10. Terwee CB, de Winter AF, Scholten RJ, et al. Inter-observer reproducibility of the visual estimation of range of motion of the shoulder. *Arch Phys Med Rehabil*. 2005; 86(7):1356-1361.
11. Whitcroft KL, Massouh L, Amirfeyz R, Bannister G. Comparison of methods of measuring active cervical range of motion. *Spine*. 2010; 35(19):E976-E980.
12. Onate J, Cortes N, Welch C, Van Lunen BL. Expert versus novice interrater reliability and criterion validity of the landing error scoring system. *J Sport Rehabil*. 2010; 19(1):41-56.
13. Padua DA, Marshall SW, Boling MC, Thigpen CA, Garrett WE, Beutler AI. The landing error scoring system (less) is a valid and reliable clinical assessment tool of jump-landing biomechanics: The JUMP-ACL study. *Am J Sports Med*. 2009; 37(10):1996-2002.
14. Padua DA, Boling MC, Distefano LJ, Onate JA, Beutler AI, Marshall SW. Reliability of the landing error scoring system-real time, a clinical assessment tool of jump-landing biomechanics. *J Sport Rehabil*. 2011; 20(2):145-156.
15. Buchanan JJ, Dean N. Consistently modeling the same movement strategy is more important than model skill level in observational learning contexts. *Acta Psychol (Amst)*. 2014; 146:19-27.
16. Gerson RF. Effect of observer and objective feedback on performance of a novel motor task. *Percept Mot Skills*. 1978; 46(2):624.
17. Ryan LJ, Fritz MS. Erroneous knowledge of results affects decision and memory processes on timing tasks. *J Exp Psychol Hum Percept Perform*. 2007; 33(6):1468-1482.
18. Ryan LJ, Robey TB. Learning and performance effects of accurate and erroneous knowledge of results on time perception. *Acta Psychol (Amst)*. 2002; 111(1):83-100.
19. Buekers MJA, Magill RA, Hall KG. The effect of erroneous knowledge of results on skill acquisition when augmented information is redundant. *Q J Exp Psychol Sect A*. 1992; 44(1):105-117.
20. USAFootball. Heads Up Tackling. http://usafootball.com/headsup. Published 2012.
21. Hayes K, Walton JR, Szomor ZR, Murrell GA. Reliability of five methods for assessing shoulder range of motion. *Aust J Physiother*. 2001; 47(4):289-294.
22. Fedorak C, Ashworth N, Marshall J, Paull H. Reliability of the visual assessment of cervical and lumbar lordosis: how good are we?: *Spine*. 2003; 28(16):1857-1859.
23. Whatman C, Hume P, Hing W. The reliability and validity of visual rating of dynamic alignment during lower extremity functional screening tests: a review of the literature. *Phys Ther Rev*. 2015; 20(3):210-224.