

The reliability of tests for sport-specific skill amongst elite youth rugby league players

Item Type	Article
Authors	Waldron, Mark;Worsfold, Paul R.;Twist, Craig;Lamb, Kevin L.
Citation	Waldron, M., Worsfold, P., Twist, C. & Lamb, K. (2014). The reliability of tests for sport-specific skill amongst elite youth rugby league players. European Journal of Sports Sciences, 14, S471-S477. http://dx.doi.org/10.1080/17461391.2012.714405
DOI	10.1080/17461391.2012.714405
Publisher	Taylor & Francis
Journal	European Journal of Sport Science
Download date	2026-05-21 17:05:15
Item License	http://creativecommons.org/licenses/by/4.0/
Link to Item	http://hdl.handle.net/10034/620702

The reliability of tests for sport-specific skill amongst elite youth rugby league players

Abstract

In rugby league, tests of sport-specific technical skill often involve subjective assessments of performance by observers differing in their levels of coaching qualification. However, the reliability of such subjective assessments has yet to be investigated via appropriate statistical techniques. Therefore, the aims of the current study were to investigate: (i) the intra-observer reliability of a non-qualified observer ('novice'), and (ii) the inter-observer reliability of the three observers (two qualified 'experts' and one novice observer) in the assessment of catching, passing and tackling (stages 1 and 2) ability in elite adolescent rugby league players (age: 14.7 ± 0.5 years). Players performed each skill element within a simulated practice drill and were assessed in 'real time' by the observers according to pre-defined skill criteria. The presence of an overall bias ($P < 0.05$) was revealed between the observers in stage 1 of catching and stage 1 of passing, the differences specifically being higher for the novice compared to both expert coaches for each stage of catching and the first stage of passing, and between expert 2 and the novice for stage 2 of tackling. No comparisons met the pre-determined analytical goal of 'perfect agreement', for any of the skill components. Similarly, comparisons between the expert observers did not reach perfect agreement, with the lowest values occurring for both tackling skill stages (60% to 65%). Therefore, none of the tests employed were sufficiently reliable to permit their application for discerning between players of differing ability, which may mean up to approximately 56% of players' technical ability being misinterpreted. Accordingly, the credibility of such assessments of sport-specific skill has to be questioned and alternative tests considered.

Key Words: Rugby League; Skill testing; Reliability

Introduction

In rugby league, tests of sport-specific skill have been used to differentiate between higher and lower playing standards in both adult (Gabbett, Jenkins & Abernethy, 2011; Gabbett, Kelly & Pezet, 2007) and junior players (Gabbett, Jenkins & Abernethy, 2010). Such tests are typically technique (process) rather than outcome based, involving subjective assessments relating to the quality of the performed skill, performed within simulated playing scenarios. Such tests meet the 'open' nature of skill within the context of team sport, requiring players to execute the correct technique in a realistic playing environment (Ali, 2011). For example, studies have employed the use of highly qualified coaching staff (Australian Rugby League Level 3 coaching accreditation) to devise standard criteria for the assessment of fundamental game-related skills such as tackling, catching and receiving the ball during sport-specific conditioning practices or rugby league matches (Gabbett, 2008; Gabbett, et al., 2007). The proficiency of players has been subsequently based upon a Likert scale rating provided by an observer (Gabbett et al., 2010; Gabbett, 2008; Gabbett, et al., 2007). Such tests have grown in popularity since the assessment of technical skills performed within an open environment may offer a more realistic playing scenario in comparison to the closed skill testing often utilised in skill test batteries (Ali, 2011).

While process-driven measures ostensibly reveal a deeper dimension of technical ability (Williams & Reilly, 2000), such tests remain scientifically questionable owing to the subjective nature of their scoring or assessment, even amongst experienced and appropriately qualified coaches (i.e. Level 3 Rugby League Coaching Qualification). Although Gabbett et al. (2007) demonstrated a seemingly reliable testing procedure for the assessment of tackling an opponent and passing or receiving of the ball (ICC = 0.85 to 0.98 and CV = 5.1 to 5.3%),

the degree of attainable reliability for observers without qualification or experience in the sport has been reported less favourably (CV = 4.7 to 8.7%; Gabbett, 2008). Indeed, assessments of this type lack procedural consistency between studies and often overlook the potential for perceptual (systematic) differences between experienced or non-experienced observers. The higher degree of experience and level of coaching qualification is broadly considered to support the credibility of the coach to discern between correctly or incorrectly executed sport-specific skills (Ste-Marie, 1999). Therefore, the recognition of, and differentiation between, systematic bias and random error are important for research of this type (see Atkinson & Nevill, 1998). However, assessment of *relative* reliability (instead of *absolute* reliability) or statistics that fail to quantify systematic bias and random error have been commonly applied to skill tests (Gabbett, 2008). Furthermore, the use of certain statistical procedures, such as the CV or, indeed, traditional parametric analyses such as 95% limits of agreement (LoA), to test for agreement between ordinal data sets (Likert scales), is also questionable (Cooper, Hughes, O'Donoghue & Nevill, 2007).

Parametric statistical tests are carried out on the assumption that the dependent variables follow a normal distribution (Atkinson & Nevill, 1998). However, ordinal data often follow a non-normal distribution and, accordingly, should be treated with non-parametric analyses (Cooper et al., 2007; Bland & Altman, 1999). Further considerations, such as the tolerable degree of error when using a 1 to 5 Likert scale (Gabbett et al., 2007), should be made in the context of previous findings. For example, previous research using Likert scales to discern between lower and higher ability players in the skills of catching, passing and tackling in rugby league players, has demonstrated 'significant' differences equating to 0.5 and 0.6 on the Likert scale in adults (Gabbett et al., 2007) or 0.75 (mean difference over various skills) in junior players (Gabbett et al., 2010). Using a non-parametric reliability analysis on such

ordinal data sets will quantify the repeatability of assessments ranging from zero to five (without decimals). Therefore, notwithstanding the erroneous presentation of unattainable mean scores in previous studies, it is clear that to recognize such minor differences below the score of 1, the observer must achieve a ‘perfect’ agreement (i.e. less than 1). As a result, any error in subjective assessment would be intolerable. In this context, the *a-priori* analytical goal of any researcher attempting to administer such tests in order to discern between playing standards in rugby league players, should be to achieve a high proportion of perfect agreement. For this reason, and those highlighted above, the credibility of subjectively scored tests in rugby league motor skill tests remains to be established, thus limiting the application of such tests for talent identification purposes. Accordingly, the aims of the current study were to investigate: (i) the intra-observer reliability of a non-qualified observer (‘novice’), and (ii) the inter-observer reliability of the three observers (two qualified ‘experts’ and one novice observer) in the assessment of catching, passing and tackling (stages 1 and 2) ability in elite adolescent rugby league players.

Methods

Participants

Twenty elite youth male rugby league players (8 forwards, 6 backs & 6 adjustables; King, Hume & Clark, 2010) contracted to a professional club in the North West of England volunteered to participate in the study (age: 14.7 ± 0.5 years; body mass: 72.8 ± 10.7 kg; stature: 176.5 ± 6.5 cm). All participants were asked not to exercise on the day of testing and to follow their normal dietary guidelines. Each player and the coaches were familiar with the testing protocols (see Procedures section) via their usual training practices and had 7.2 ± 1.2

years of formal playing experience, defined as a minimum of one training session and weekend match with a rugby league club. Consent was obtained from the players and their parents/guardians and approval for the study was granted by the Faculty of Applied Sciences Ethics Committee.

Procedure

Skill simulation

All testing procedures took place outdoors on a grass training pitch under dry, mild weather conditions, over a period of one-to-two hours on the same day. Using examples from previous research (Gabbett et al., 2010; Gabbett, 2008), a simulated sport-specific match scenario was devised and implemented (as shown in Figure 1). The skills of passing, tackling and catching were selected since they represent fundamental game skills in rugby league that are performed by all players (see Sirotic, Coutts, Knowles & Catterick, 2009). The players performed a 'warm-up' (in groups of three) led by the club coach, consisting of moderate intensity running and upper and lower body dynamic stretching exercises, immediately prior to the skills tests..

The players were randomly selected to complete the test as either one of two attacking players who retained possession of the ball or a defensive player. Set within a 10 x 10 m grid, attacking players (ball carriers) were required to advance from one side of the grid to the other and complete one pass each before being tackled by the defending player. After one cycle of this protocol, the players were instructed to wait for a brief recovery period (remaining on their feet) at the opposite end of the grid before repeating the drill in the

opposite direction. The test was designed to obligate catching, passing and tackling from both the player's left and right hand sides. If an action was performed that was deemed to be outside of the skills being assessed, such as an incorrect sequence of passing, the players were allowed to re-start the trial. To avoid such issues, demonstrations from qualified coaches (Level 3 Rugby League Coaching Qualification, UK) were performed prior to the testing procedures in order to enhance players' understanding of the test and to provide them with a reference for the required match-like intensity. The practice was continued until the coaches notified the researcher that they had completed their assessment (see criteria below), which lasted between four and six repetitions for each trial (~ 2 min). Once the observer had provided a score out of five for each of the three skill components, the players were required to exchange roles, with one player per drill under assessment. Once the first set of players had completed their rotation as the tackler, the next group of three commenced an identical testing procedure. A camera (Canon MV 700i, 50 Hz, Japan) was set up approximately at eye level of the coaches in a static position at a distance of 15 m from one end of the grid and used to film all proceedings (Figure 1). This was later used by the novice observer for technical skill assessments (see following sections).

*****Figure 1 here*****

Skill assessment

The skills of the players were assessed by two expert coaches with 10 and 15 years of coaching, respectively, using set criteria (Table 1) previously established by Gabbett et al. (2007). The aim for the observer was to rate the players (in real-time) on their overall proficiency in each skill using a Likert scale ranging from one to five, with five representing an optimal score and one representing the lowest score possible. The expert and novice observers were provided with the assessment criteria one week prior to the testing, and subsequently given explicit instruction to refer to the criteria during the testing procedures. For consistency, the expert observers were positioned equi-distant either side of the camera, enabling a similar perspective of the players. Each observer was not made aware of the other's scores. The inclusion of a novice observer (having watched rugby league for the previous two seasons but no coaching qualification) enabled a comparison with the expert assessors. To be consistent with the analyses of the experts, the novice observer was required to analyze, continuously, the video footage (without slowing or re-watching the footage) of the players' performances using the set criteria, albeit post-event. In order to evaluate the consistency of his subjective assessments (intra-observer reliability), he was required to repeat this task a week later. Following the recommendation of Gabbett et al. (2007), two stages of assessment (approach, Stage 1; execution, Stage 2) were included for each skill yielding two scores per skill performed.

*****Table 1 here*****

Statistical analyses

The distributions of the six skill elements (approach and execution of catching, passing and tackling) were initially checked for normality using the Shapiro-Wilk test and where violations were observed ($P < 0.05$), non-parametric Kruskal-Wallis tests were applied to test

for differences between observers (expert 1, expert 2 and the novice). Post-hoc Mann Whitney-U tests were used for pairwise comparisons between each of the observers. The presence of bias between the test and re-test trials of the novice observer was checked via a median sign test. Owing to the multiple (six) comparisons made between each different observer (novice, expert 1 and expert 2), the Benjamini Hochberg False Discovery Rate (FDR) technique was applied to control for the potential increase in the type I error rate. The technique involves, firstly, ranking the P -values ($p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(k)}$) obtained from a series of multiple comparison tests performed under a shared hypothesis, from smallest to largest (six comparisons between each observing pair in the current case). The formula $k\alpha/n$ is used to derive the FDR where; k = rank, α = alpha level (0.05), n = number of tests. Beginning with the largest (step-up), each original P -value is compared to the FDR (i.e. compare $p_{(k)}$ to $k\alpha/n$). At the point at which $p_{(k)} \leq k\alpha/n$, the null hypothesis was rejected and every value thereafter (Benjamini & Hochberg, 1995). The degree of random variation between or within observers was evaluated using the non-parametric technique advocated by Cooper et al. (2007). This technique involved calculating the percentage of agreement and associated 95% confidence intervals (CIs) between or within observers inside a ‘practically important’ reference value (Nevill, Lane, Kilgour, Bowes & Whyte, 2001). As established above, a reference value of perfect agreement (zero difference between observations) was deemed as ‘practically important’ for each type of skill assessed. A secondary reference value of ± 1 (a difference of one in either direction) was also set in order to demonstrate the portion of agreement between observers in the presence of the smallest possible error that can be made on the 1-5 Likert scale. Additionally, the coefficient of variation (CV) was calculated to enable comparisons with the findings of previous research.

Results

*****Table 2 here*****

Based upon the data presented in Table 2, the Kruskal-Wallis tests identified significant observer effects on stage 1 ($X^2_{(2)} = 10.5, P = 0.005$) and 2 ($X^2_{(2)} = 9.7, P = 0.008$) of catching performance, and stage one of passing ($X^2_{(2)} = 5.8, P = 0.046$). Post-hoc comparisons revealed higher scores ($P < 0.05$) recorded by both experts than the novice observer for each stage of catching, and the first stage of passing (Table 3). Likewise, there was a significant observer effect on stage two of tackling ($X^2_{(2)} = 5.76, P = 0.049$) which was attributable to the score of expert 2 being higher than that of the novice ($P < 0.05$). Based upon the analytical goal of a ‘perfect agreement’, further analysis showed that the degree of variation between the expert coaches and the novice was as low as 30%, and no better than 65%. The CV statistics for the same comparisons ranged between 7.9% and 14.3%, and included only four values below 10% (Table 3).

Systematic bias was not present between expert observers and whilst there were no instances of 100% perfect agreement between them, it ranged from 75% to 90% in all passing and catching skills. However, for the tackling skills the agreement was notably lowered (60% to 65%). Nonetheless, all the CVs for the three skills were below 10% (1.6% to 8.1%). For the novice, intra-observer analysis revealed no overall difference ($P > 0.05$) between any scores, and the levels of agreement in the range 70% to 85%. CVs were below the 10% threshold for all scores, ranging from 3.4% to 6.0%.

Based upon the less stringent analytical goal of plus or minus '1' on the Likert scale, better agreement was achieved for all comparisons. For example, Table 3 shows that between expert observers and the intra-reliability of the novice observer, agreement was 100% in all but one comparison (tackling stage two for expert 1 to expert 2). Expert versus novice agreement remained sub-optimal, though was as high as 95% for most of the scores.

*****Table 3 here*****

Discussion

It was the analytical goal of the present study to obtain 'perfect agreement' between expert observers in order to meet the requirements outlined in previous research (i.e. a difference of less than '1' on the Likert scale). As it emerged, in no case was 100% perfect agreement obtained between the expert observers and, given the width of the 95% confidence intervals (approximately 44% to 100% for catching and passing skills), it is likely that some talented players could be incorrectly appraised using such tests, which may contribute to the coaches' misinterpretation of their playing ability. That is, in the skills of passing and catching, the 'population' agreement between experts could be as high as 100% or low as 44%, rendering the potential for disagreement and performance misinterpretation to be as high as 56%. Importantly, it is noteworthy that, for the same data, the CV ranged from 2.8% to 8.1% which is less than the magnitude often deemed as 'reliable' (< 10%; Atkinson & Nevill, 1998) and, similar to previous research in rugby league (Gabbett et al., 2007). Given that, in the context

of talent identification, it is typically expert coaches who are responsible for discerning between players showing signs of higher or lower ability, and that previous reports in rugby league have failed to establish the inter-observer reliability between expert observers via the correct statistical approach, the general application of subjective rating systems across different expert users has to be questioned.

The current results should be interpreted on behalf of the broader rugby league community in accordance with the tolerable degree of error. That is, those charged with identifying talented players based, in part, upon the construct of sport-specific skill measured in such a way are required to consider what degree of error is acceptable. For example, if a tolerance of plus or minus one on a scale of one to five is deemed satisfactory, then the current data would indicate a much better level of agreement between expert observers than if zero difference reference value was adopted. However, in the context of talent identification, this parity between observers does not support the worthiness of the test for correctly interpreting skilled performance in higher ability players. Rather, the probability of misinterpreting (falling within plus or minus one) the quality of sport-specific skill is reinforced.

The limited agreement between experts was also exacerbated within both stages of tackling. The reliability of the assessment of tackling was the poorest between experts, with a perfect agreement as low as 50%. Such poor agreement may relate to the open nature of the skill in which a simulated collision between two participants induces a less predictable environment in which to base judgements of technical performance. Indeed, previous analyses have assessed such skills within the open match environment (Gabbett et al., 2007), in which a stability of the set criteria, such as the upper and lower body position, cannot be expected. Moreover, it could be argued that the set criteria will vary according to the context in which

the tackle is performed, such as side-on and chasing tackles. In addition, research has shown that only 17% of tackles in rugby league are performed in a one-on-one scenario, with players often tackling in conjunction with other team-mates (King et al., 2010). Such findings support our previous assertion regarding the situational inconsistencies during match time, adding further complication to the assessment of tackling technique. Whilst these suggestions detract from the potential reliability of tackling analysis, the intention of previous researchers to enhance the ecological validity of skill testing should be recognized. Given the current findings and the general disparity between both experts, it remains unclear exactly what criteria expert observers are basing their judgements on. Indeed, it would be useful to evaluate the intra-observer reliability of expert observers' ratings, with and without the use of the set criteria.

The ratings of the experts were found to be systematically higher ($P < 0.05$) than the novice observer in the skills of catching (all stages) and passing stage 1. Such results fundamentally question the validity of the rugby league tests for motor skill ability in the hands of an inexperienced observer and suggest that it would be inappropriate to use the assessments of novice or expert observers interchangeably. Indeed, Gabbett (2008) has discussed the results of previous studies that have used either a novice or an expert observer without consideration of the potential differences in interpretation. In relation to the analytical goals of perfect agreement, the degree of random variation between the scores of the expert coaches and the novice was as low as 30%, with associated CIs ranging from 19.5% to 46.8%. Furthermore, the largest perfect agreement was 65% and in no case did the comparisons between the novice and expert observers indicate the potential (via CIs) for 100% agreement. If it is the intention of future research to compare findings between different studies, than an *a priori*

evaluation similar in nature to the current study should be undertaken in order to establish the reliability of the observer.

The differences found in the present study between novice and expert observers may be owing to the inconsistent use of the set criteria for skill assessment. It has been suggested that inexperienced observers over-rely on operational definitions whilst assessing technical actions during match play (O'Donoghue, 2007). In contrast, an expert observer may choose to underpin interpretations of performance with previously acquired tacit coaching knowledge, using definitions as a vague guide rather than to strictly inform assessment, even when instructed otherwise (O'Donoghue, 2007). Although such reasoning may partly explain the disparity between expert and novice coaches, it is reasonable to question the necessity of 'set criteria', particularly for the expert coaches, if it fails to inform the resultant assessment. However, in the present study the novice observer demonstrated no systematic bias and perfect agreement ranging from 70% to 85% between repeated trials, which may support the utility of set criteria since this alone guided the interpretation of skill in the absence of sport-specific knowledge. It is therefore apparent that the set criteria may be used differently depending upon the user's prior experience of the sport. Consequently, it can only be assumed that the exact construct of skill being assessed will vary between users with more or less experience.

Conclusion

The current analysis has raised a general concern over the use of subjective ratings of rugby league skill in their current form and highlighted potential issues with the application of set

skill criteria in relation to the 1 to 5 Likert scale ratings. Collectively, the inter-observer trials have shown that the application of a Likert scale cannot be used reliably to obtain a perfect agreement, most likely reflecting the subjectivity of the observers. This finding was supported by the novice's higher level of reliability demonstrated over the two repeated trials. Furthermore, it is clear that some skills, such as tackling, are inherently more difficult to assess reliably than others, perhaps owing to the open nature of the assessment method. If sport-specific skill is an underlying facet of talented performance, capable of discerning between the elite or sub-elite players, then a test based upon an objective outcome may provide a more suitable measure. However, whilst such tests offer greater control over the performed skill, a sacrifice in ecological validity is inevitable.

References

- Ali, A., Measuring soccer skill performance: a review. *Scandinavian Journal of Science and Medicine in Sports*, 21, 170-183.
- Atkinson, G. & Nevill, A. M. (1998). Statistical methods for addressing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine*, 26, 217-38.
- Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57, 289-300.
- Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8, 135-160.
- Cooper, S. M., Hughes, M., O'Donoghue, P., & Nevill, A. M. (2007). A simple statistical method for assessing the reliability of data entered into sport performance analysis systems. *International Journal of Performance Analysis in Sport*, 7, 87-109.
- Gabbett, T., Kelly, J. & Pezet, T. (2007). Relationship between physical fitness and playing ability in rugby league players. *Journal of Strength and Conditioning Research*, 21, 1126-1133.
- Gabbett, T. J. (2008). Influence of fatigue on tackling technique in rugby league players. *Journal of Strength and Conditioning Research*, 22, 625-632.
- Gabbett, T. J., Jenkins, D. G., & Abernethy, B. (2010). Physiological and anthropometric correlates of tackling ability in junior elite and subelite rugby league players. *Journal of Strength and Conditioning Research*, 24, 2989–2995.
- Gabbett, T. J., Jenkins, D. G., & Abernethy, B. (2011). Relationships between physiological, anthropometric, and skill qualities and playing performance in professional rugby league players. *Journal of Sports Sciences*, 29, 1655-1664.

- King, D., Hume, P. A., Clark, T. Video analysis of tackles in rugby league matches by player position, tackle height and tackle location. *International Journal of Performance Analysis in Sport*, 10, 241-254.
- Nevill, A. M., Lane, A. M., Kilgour, L. J., Bowes, N. and Whyte, G.P. (2001). Stability of psychometric questionnaires. *Journal of Sports Sciences*, 19, 273 – 278.
- O'Donoghue, P. (2007). Reliability issues in performance analysis. *International Journal of Performance Analysis*, 7, 35-48.
- Sirotic, A. C., Coutts, A. J., Knowles, H, & Catterick, C. (2009). A comparison of match demands between elite and semi-elite rugby league competition. *Journal of Sports Sciences*, 27, 203-211.
- Ste-Marie, D. M. (1999). Expert-novice differences in gymnastic judging: an information-processing perspective. *Applied Cognitive Psychology*, 13, 269-281.
- Williams, A. M. & Reilly, T. (2000). Talent identification and development. *Journal of Sports Sciences*, 18, 657-667.

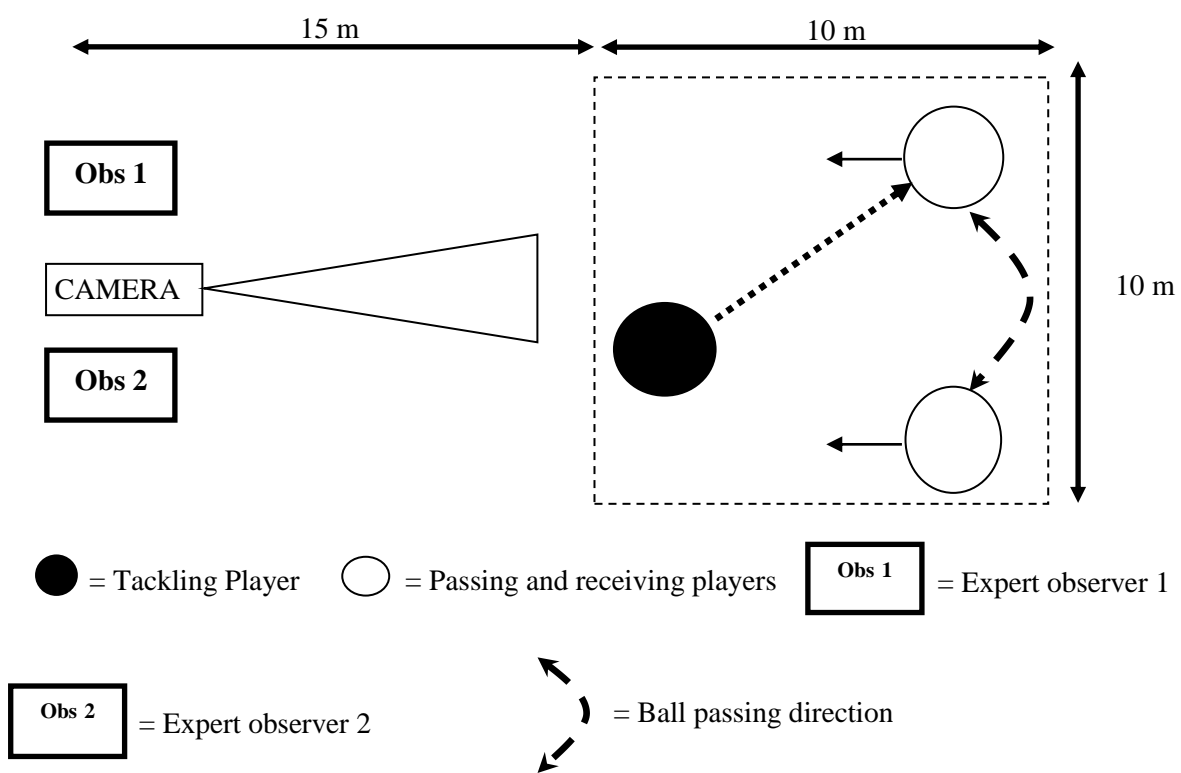


Figure 1. Rugby league tackling, passing and catching protocol (based on Gabbett et al., 2010). *Note: the above diagram shows the protocol in one direction. Players performed the test in both directions. The novice observer was filming the training practice.*

Table 1. Standard criteria for the assessment of tackling, passing and catching techniques.

Skill	Criteria	
Catching	Stage 1	Stage 2
	Hands up	Catch the ball
	Fingers up	Holding the 'body' of the ball
	Palms out	Prepared to carry the ball or execute a pass
	Call for the pass	Minimal breaking of the natural stride
	Take pass early	
Passing	Stage 1	Stage 2
	Pendulum action	Receiver able to catch the ball
	Look where passing	Receiver able to maintain stride
	Single movement	
	Flat and behind	
	Ahead of receiver	
Appropriate ball speed		
Tackling	Stage 1	Stage 2
	Low body position	'Turtle' player
	Arms ready	Hold defensive shape
	Head behind/to one side	Point to remaining player being marked
	Contact with shoulder	
	Wrap arms around waist	
	Drive with legs	
	Pull with arms	
	Maintain grip until on ground	

Table 2. Median and inter-quartile range for the subjective scoring assessments of expert and novice observers.

		Median	25th Percentile	75th Percentile
Catching Stage 1	Expert 1	5	4	5
	Expert 2	5	4	5
	Novice	5	4	5
Catching Stage 2	Expert 1	5	4	5
	Expert 2	5	4	5
	Novice	4	3	4
Passing Stage 1	Expert 1	4	4	5
	Expert 2	5	4	5
	Novice	4	3	4
Passing Stage 2	Expert 1	5	4	5
	Expert 2	5	4	5
	Novice	4	4	5
Tackling Stage 1	Expert 1	4	4	4
	Expert 2	4	4	5
	Novice	4	3	5
Tackling Stage 2	Expert 1	4	3	5
	Expert 2	4	4	5
	Novice	4	3	4

Table 3. Comparisons of the inter- and intra-observer reliability of expert and novice rugby league practitioners.

Inter-observer	P-value	Perfect Agreement		Plus or Minus 1		CV (%)
		PA (%)	95% CIs (%)	PA ± 1 (%)	95% CIs (%)	
Expert 1 to Expert 2						
Catching Stage 1	0.323	80	44.6 to 100	100	83.8 to 100	3.6
Catching Stage 2	0.560	85	47 to 100	100	83.8 to 100	2.8
Passing Stage 1	0.520	75	44.2 to 99.6	100	83.8 to 100	4.2
Passing Stage 2	1.000	90	49.4 to 100	100	83.8 to 100	1.6
Tackling Stage 1	0.324	60	34.9 to 81.8	100	83.8 to 100	7.8
Tackling Stage 2	0.188	65	37.3 to 87.8	95	76.3 to 99.1	8.1
Expert 1 to Novice						
Catching Stage 1	0.002*	45	27.4 to 64.2	95	76.3 to 99.1	11.0
Catching Stage 2	0.002*	45	27.4 to 64.2	95	76.3 to 99.1	11.0
Passing Stage 1	0.005*	50	29.9 to 70	90	69.8 to 97.2	11.5
Passing Stage 2	0.185	65	37.3 to 87.8	95	76.3 to 99.1	7.9
Tackling Stage 1	0.786	50	29.9 to 70	95	76.3 to 99.1	11.6
Tackling Stage 2	0.518	65	37.3 to 87.8	95	76.3 to 99.1	8.12
Expert 2 to Novice						
Catching Stage 1	0.003*	30	19.5 to 46.8	90	69.8 to 97.2	14.3
Catching Stage 2	0.004*	35	22.2 to 52.5	90	69.8 to 97.2	13.7
Passing Stage 1	0.023*	50	29.9 to 70	85	63.9 to 94.7	12.5
Passing Stage 2	0.185	55	32.4 to 75.9	95	76.3 to 99.1	9.5
Tackling Stage 1	0.230	60	34.9 to 81.8	95	76.3 to 99.1	9.0
Tackling Stage 2	0.033*	50	29.9 to 70	90	69.8 to 97.2	11.5
Intra-observer						
Novice trial 1 to 2						
Catching Stage 1	1.000	80	44.6 to 100	100	83.8 to 100	4.6
Catching Stage 2	1.000	85	47 to 100	100	83.8 to 100	3.8
Passing Stage 1	0.219	70	39.8 to 93.7	100	83.8 to 100	6.0
Passing Stage 2	1.000	85	47 to 100	100	83.8 to 100	3.8
Tackling Stage 1	0.625	80	44.6 to 100	100	83.8 to 100	4.4
Tackling Stage 2	0.250	85	47 to 100	100	83.8 to 100	3.4

Note: * = significantly larger for the expert observer based on pairwise comparisons ($n = 20$). Benjamini Hochberg adjusted alpha levels.