Psychology of Sport & Exercise 43 (2019) 261–270

Contents lists available at ScienceDirect

Psychology of Sport & Exercise

journal homepage: www.elsevier.com/locate/psychsport

Video-based testing in sporting officials: A systematic review

Aden Kittel[a,*], Paul Larkin[a,b], Nathan Elsworthy[c], Michael Spittle[a]

[a] Institute for Health and Sport, Victoria University, Australia
[b] Maribyrnong Sports Academy, Melbourne, Australia
[c] School of Health, Medical and Applied Sciences, Central Queensland University, Australia

ARTICLE INFO

Keywords:
Video-based testing
Sports officials
Perceptual-cognitive expertise

ABSTRACT

Objectives: Decision-making is the most important skill for sporting officials, consequently, assessment of this skill is becoming increasingly popular in the literature. There is considerable interest in the use of video-based methods to assess decision-making of officials in controlled, off-field environments.
Design: Systematic review of the literature examining video-based testing in sporting officials.
Methods: Using the keywords "umpire", "referee", "sport officials", "decision making" and "judgement", a comprehensive search was conducted in February 2018 on electronic databases (SPORTDiscus, Medline, PsycInfo, Google Scholar). Inclusion criteria included full text articles from January 2000 to January 2018 published in peer-reviewed journals. Only 'central' or 'field' officials were included in this review (i.e., assistant referees, touch judges were excluded).
Results: The search yielded 27 studies. The majority of articles were specific to soccer officials. Overall, video-based testing appears to be a valid measure of decision-making differentiating between performance levels. This review highlighted a high degree of variability among the methods applied, with varied participation groups, clip type used, and influences on decision-making. The reporting of reliability and implementation of transfer tests was rarely incorporated in the research.
Conclusions: Video-based testing appears to be a valid measure of decision-making of officials in an off-field, controlled environment. This research area would be advanced through further investigation into sports other than soccer, examination of transfer to match performance testing, reporting the reliability of the test, reporting decisional accuracy rather than solely number of decisions, and investigation of additional video modes.

1. Introduction

Perceptual-cognitive skills are an integral aspect of sporting performance for all individuals. Perceptual-cognitive skills are defined as the ability to identify crucial information within the environment, and integrate this information with existing knowledge of motor capabilities to select and execute an appropriate response (Marteniuk, 1976). Decision-making is the foremost perceptual-cognitive skill involved in sport (Williams, Ward, Smeeton, & Allen, 2004), and can be defined as the ability to perceive information, correctly interpret, then select an appropriate response (Baker, Côte, & Abernethy, 2003). The literature has focused on athletes, commonly investigating the development of decision-making (Baker, Cote, & Abernethy, 2003), expert-novice differences (Williams & Ericsson, 2005), and decision-making processes (Araújo, Davids, & Hristovski, 2006). Importantly, perceptual-cognitive skills are an effective means to differentiate between less skilled and higher skilled performers in sport (Berry, Abernethy, & Côté, 2008; Williams & Ericsson, 2005).

Sporting officials, however, are a vital component of the sporting domain along with players. As they are required to perceive sporting actions and react to whether an infringement (i.e., free kick, penalty) has occurred, decision-making is the foremost perceptual-cognitive skill of sporting officials. The accuracy of such decisions can greatly impact the outcome of a match, leading to criticism and impacting club revenue (Larkin, Berry, Dawson, & Lay, 2011). As such, decision-making is commonly cited as the most important overall skill for effective officiating (Helsen & Bultynck, 2004; Morris & O'Connor, 2016). Within the literature, officials are classified as interactors (e.g., soccer referee, Australian football umpire), monitors (e.g., volleyball referee; gymnastic judge), and reactors (e.g., tennis line judge) based upon their decision-making processes, and their interaction or movement within the environment (MacMahon et al., 2014). This review will focus on interactor officials who have high perceptual demands (i.e., cues and players to monitor), and interaction with their environment (high physical demands) (MacMahon et al., 2014).

From a sporting officials perspective, researchers have investigated both on-field and off-field aspects of performance. On-field research has typically examined the movement (Elsworthy & Dascombe, 2011; Emmonds et al., 2015; Krustrup & Bangsbo, 2001) and decision-making (Burnett, Bishop, Ashford, Williams, & Kinrade, 2017; Helsen & Bultynck, 2004; Larkin, Mesagno, Berry, & Spittle, 2016) match demands of interactor officials in a range of sports. In relation to these areas of research, there are inconsistent findings between sports, with research suggesting there is no significant impact of exertion on match decision-making in Australian football umpires (Elsworthy, Burke, Scott, Stevens, &

* Corresponding author. Institute of Health, Exercise and Sport, Victoria University, Footscray, VIC, 3011, Australia.
E-mail address: Aden.Kittel@live.vu.edu.au (A. Kittel).

https://doi.org/10.1016/j.psychsport.2019.03.013
Received 10 September 2018; Received in revised form 22 March 2019; Accepted 29 March 2019
Available online 01 April 2019
1469-0292/

Dascombe, 2014) and rugby league referees (Emmonds et al., 2015), yet does impact the decision-making demands of soccer referees (Mallo, Frutos, Juárez, & Navarro, 2012). Further, the influence of communication (Cunningham, Simmons, Mascarenhas, & Redhead, 2014; Neville, Salmon, & Read, 2018), psychological factors (Johansen & Haugen, 2013; Page & Page, 2010) and physical fitness on match performance (Castagna, Abt, & D'ottavio, 2002) have also been investigated. These studies highlight that although there are similarities across different sporting officials, not all findings are transferrable across sports and therefore researchers should further investigate methods of assessing the decision-making performance of officials from a wide range of sports.

Off-field research in sporting officials has grown in recent years, with an emphasis on examining the perceptual-cognitive aspects of performance (i.e., decision-making) in an isolated manner. This is due to the paramount importance of decision-making to overall performance (Helsen & Bultynck, 2004). Assessing the decision-making of officials in games may be the optimum measure of decision-making performance, however, there is a high degree of variability from game to game. This limits the performance comparisons between officials across different games, and within individual officials from game to game. Off-field video-based testing overcomes this limitation, with the ability to test decision-making in a controlled environment to present consistent scenarios across multiple officials. We define these methods as presenting sport-specific decision-making in a video format to simulate on-field decision-making in an off-field setting. The high degree of variability in the methods applied for video-based research will be outlined as part of this review. Therefore, the aim of this systematic review is to summarise video-based decision-making assessment literature in the domain of interactor officials, and analyse the various methods utilised to simulate match-like decision-making.

2. Method

The method for this systematic review was informed by the PRISMA guidelines (Moher, Liberati, Tetzlaff, Altman, & Group, 2009), summarised in Fig. 1.

2.1. Search strategies

Electronic databases (SPORTDiscus, Medline, PsycInfo, Google Scholar) were searched for articles published between January 2000 until February 2018. Keyword combinations included “decision making” in conjunction with “umpire” and “referee”. The search was restricted to English peer-reviewed articles. 207 articles were identified in this initial search. Following this, keywords of the articles were analysed for further search combinations. Two new search terms were identified and combined with the existing terms including “judgement” and “sport officials”, and subsequently searched within the databases. A further 161 articles were identified with the new search terms in the four databases, resulting in a total of 368.

2.2. Inclusion and exclusion of studies

Studies included in this systematic review adhered to the following criteria: (i) participant groups included interactor officials; (ii) used video-based methods for off-field decision-making assessment; (iii) assessed sport-specific decision-making involving infringement/penalty scenarios (i.e., general perceptual-motor skill assessments such as pattern recognition and reaction time were excluded); and (iv) participants were central referees/field umpires (i.e., assistant referees were excluded when the aim of the study investigated offside decision-making performance). These populations were excluded as this review focused on infringement-based decision-making, rather than onside/offside processing, which is subject to different psychological factors (e.g., flash-lag effect). Studies were included if infringement-based decision-making of central referees was compared to other populations, such as fans or assistant referees.

2.3. Screening articles

Each article was screened by examining the title, abstract and keywords based on the inclusion criteria. If there was any uncertainty over the appropriateness of an article, this was debated by the first and second authors. In the rare circumstance where uncertainty remained, the third author was included in the discussion. Final classification and acceptance of all studies was agreed upon by all authors.

2.4. Quality assessment

The quality of articles was analysed using a scale adapted from previous research (Larkin, Mesagno, Spittle, & Berry, 2015). This assessment scale (Fig. 2) assesses the quality of each study based on three sub-scales (assessment of test measures, groups examined, decision reporting). The quality of the test measures, including different levels of validity (face, construct and concurrent) and reliability of the test were assessed in the first sub-scale. Face validity was demonstrated if the test presented a sport-specific decision-making scenario (Larkin, Mesagno, Berry, & Spittle, 2014). In psychological tests, construct validity refers to the degree to which a test measures a concept or construct that it intends to measure (DeVellis, 2016; Haynes, Richard, & Kubany, 1995). In comparison, for performance assessments, construct validity is also obtained through evaluation of performance between known skill level differences (Gadotti, Vieira, & Magee, 2006; Thomas, Silverman, & Nelson, 2015). The known-group difference method (Thomas et al., 2015) is commonly used in performance-based tests, such as video-based tests, to determine construct validity of the measure (Larkin et al., 2015). For the purpose of this review, this is how construct validity will be defined. Concurrent validity refers to the relationship between the measure (i.e., decision-making performance) and a criterion such as performance ranking, or on-field performance (Gadotti et al., 2006). The second sub-scale examined the demographics of the participants, in relation to the level they typically officiate. This is an example of construct validity, and is important to determine whether the video-based testing tool distinguishes between performance levels. The third sub-scale refers to the reporting given to the decisions provided by the participants. Specifically, did the study report accuracy of the decisions? This is important, as although studies report the number of free kicks or penalties made, differences must be put into perspective by providing the accuracy of the decisions. The quality assessment of each study is presented in Table 2.

3. Results

Full-text review of 95 articles was conducted after being identified as potentially relevant from scrutinising titles, keywords and abstracts. To determine the appropriateness, the full text articles were reviewed and assessed against the aforementioned inclusion criteria. Reference lists of each article were also examined to include articles not located within the above search criteria, with two studies included. Studies were predominantly excluded because they examined on-field decision-making (n = 56), rather than off-field video-based assessment per the criteria. Overall, 27 studies were included in the final analysis.

3.1. Sports investigated

Key results from each study are presented in Table 1. In the studies identified, the most prevalent sport investigated was soccer (n = 16), followed by Australian football umpires (n = 3) and rugby union officials (n = 3), handball referees (n = 2), basketball referees (n = 2), and ice hockey referees (n = 1).

3.2. Reliability & validity of tests

Reliability of the decision-making test was reported in four of the 27 studies. One study conducted a reliability assessment, with intra-class correlation coefficients demonstrating high test-retest reliability (0.76–0.82) (Spitz, Put, Wagemans, Williams, & Helsen, 2017). Three studies reported reliability from previous studies (Larkin, O'Brien, et al., 2014; Paradis, Larkin, & O'Connor, 2016; Spitz, Put, Wagemans, Williams, & Helsen, 2018).

The three types of validity examined were face/content, construct, and concurrent validity. As the inclusion criteria stipulates video clips must include sport-specific decision-making for officials (i.e., presenting an infringement, penalty, or free kick), all the studies had adequate face validity. Although this was not explicitly reported by the researchers, face validity is evident due to the sport-specific decision-making nature of the task. Similarly, construct validity is not explicitly stated by the studies, but can be assumed with a comparison of different performance levels. Construct validity was assessed by differentiating between known skill levels in 13 studies (Catteeuw, Helsen, Gilis, & Wagemans, 2009; Larkin et al., 2011; Larkin, O'Brien, et al., 2014; MacMahon, Helsen, Starkes, & Weston, 2007; MacMahon & Ste-Marie, 2002; Mascarenhas, Collins, & Mortimer, 2005a; Paradis et al., 2016; Plessner & Betsch, 2001; Renden, Kerstens, Oudejans, & Cañal-Bruland, 2014; Spitz, Put, Wagemans, Williams, & Helsen, 2016; Spitz et al., 2017; Spitz et al., 2018; Wilson & Mock, 2013). Of the 13 studies which assessed construct validity, no significant differences were found in three of the studies (MacMahon & Ste-Marie, 2002; Plessner & Betsch, 2001; Wilson & Mock, 2013). This infers the decision-making task was not able to differentiate decision-making skill between known performance levels. Concurrent validity, which examines the correlation between test score and on-field performance or ranking, was assessed in one study (Nazarudin et al., 2015).

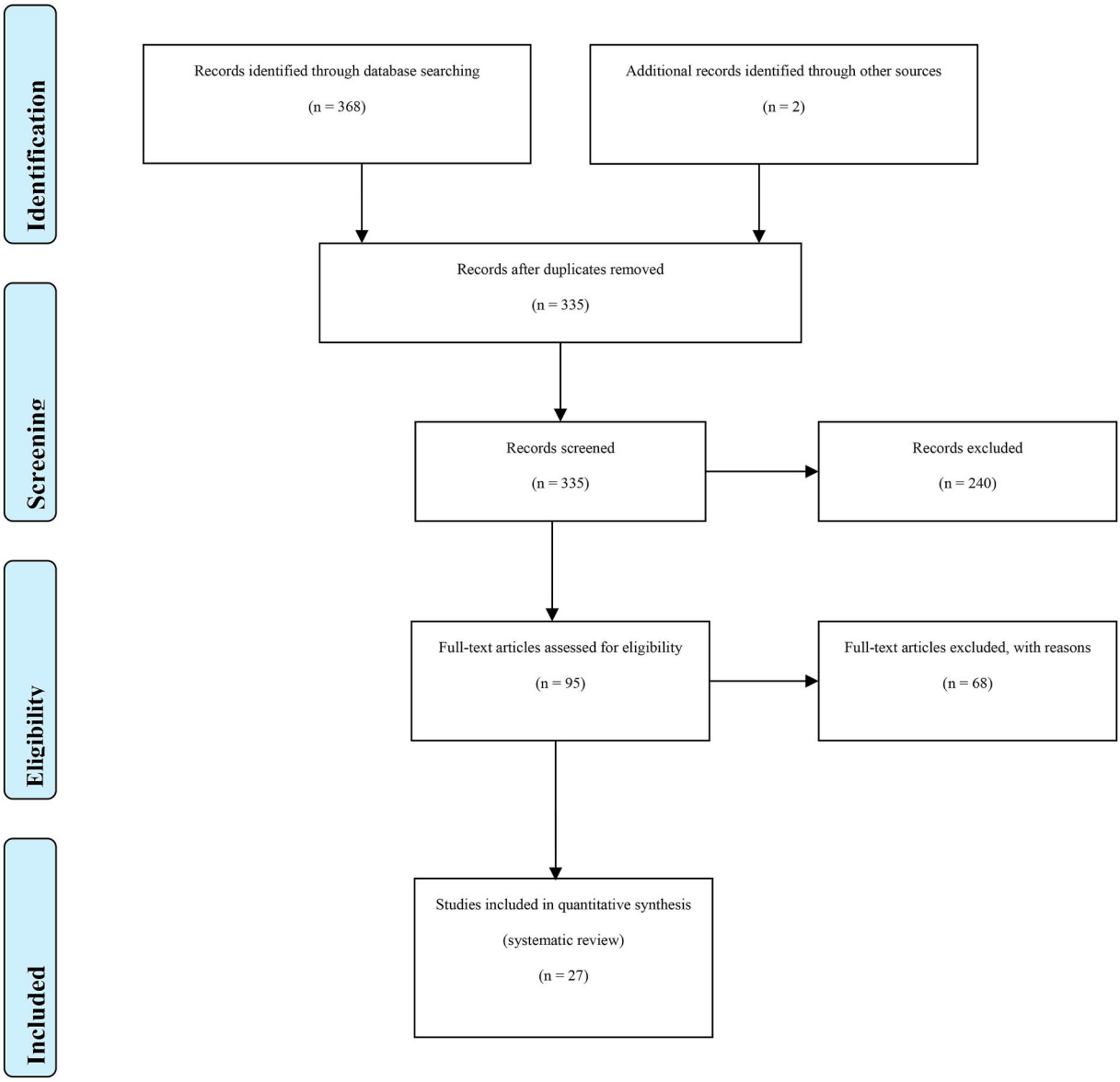

Fig. 1. PRISMA flow diagram.

Strength of evidence	Assessment of Test Measure	Officiating Level Used	Clip Decision
Most Robust	Reliability and validity assessed	Three participating groups (elite, sub-elite, amateur)	Decision accuracy reported
	Reliability only		
	Concurrent validity only		
	Construct validity only	Two participating groups only	
		One participating group only	Only number of decisions reported
	Face validity only		
Least Robust	No reliability or validity assessed	Participating group not reported	No reporting of decision accuracy or number

Fig. 2. Classification scale of the video-based testing tool strength of evidence based on three independent variables (adapted from Larkin et al. (2015).

Table 1
Summary of studies examining video-based decision-making in sporting officials.

	Author (date)	Sport officials' skill level	Other populations	Test overview	Influences on DM in task	Validity and Reliability Assessment	Results overview	Key findings
Soccer	Plessner and Betsch (2001)	Skill level NR (n = 58)	Players (n = 57)	20 match clips.	Impact of previous decisions.	NR (not reported or evident)	No significant differences between referees' and players' decision-making scores. Referees did not want to award a penalty in the second scene if they had awarded a penalty in the first.	Both referees and players were biased by their own earlier decisions.
	Jones et al. (2002)	Amateur (n = 38)	Nil.	50 match clips (20 'certain' where participants called a foul 90% of the time, 20 'uncertain' where a foul was called 45%, 10 'innocuous' where foul committed 11%).	Aggressive reputation. Participants told those wearing blue uniforms had a history of aggressive behaviour (experimental group), or not told (control).	NR	No significant difference in number of decisions awarded against the blue team in either conditions ($p > 0.01$). This was evident for both 'certain' and 'uncertain' clips. Significant difference ($p < 0.01$) by number of cards by the two groups.	Referees who are informed of a team's aggressive reputation respond differently than those who do not receive this information, by awarding more red and yellow cards.
	Nevill et al. (2002)	Amateur (n = 40)	Nil.	47 match clips (22 referees watched the clips with noise, 18 in silence).	Crowd noise.	Concurrent validity only.	Referees in the noise group awarded significantly less fouls against the home team than the silent group ($p < 0.05$). More experienced referees awarded significantly less fouls ($p < 0.05$).	Crowd noise influences decision-making, by reducing the number of fouls against the home team rather than increasing fouls against away.
	Balmer et al. (2007)	Individuals with coaching, playing and/or refereeing experience (n = 26)	Nil.	One game including 47 incidents. Participants randomly assigned to a silent, and a noise group (with commentary). Tested on opposite condition one week later.	Crowd influence, anxiety levels.	NR	In noise condition, referees awarded fewer fouls against the home side ($p < 0.001$), and more no foul decisions ($p = 0.003$). Significant relationship between noise condition, and cognitive anxiety ($r = 0.55$, $p < 0.01$), and mental effort ($r = 0.54$, $p < 0.05$).	Supports the home crowd advantage theory. Biased decision-making was linked to increase in cognitive anxiety and mental effort.
	MacMahon, Helsen, et al. (2007)	Elite (n = 7)	Youth academy players (n = 41)	20 clips.	Nil.	Construct validity only.	Referees scored significantly higher (80.6%) than players (55.1%) ($p < 0.001$). No significant effect of playing experience on decision-making accuracy ($p > 0.05$). Hours per week of practice is a moderate predictor of accuracy ($p < 0.01$).	Soccer referees significantly outperform players in an infringement identification task. Practice activities predict performance.
	Catteeuw et al. (2009)	Elite (n = 27)	Elite assistant referees (n = 27)	24 clips presented twice. First as normal speed. Second set as once normal speed, followed by the same clip twice in slow motion.	Nil.	Construct validity only.	Referees scored significantly higher (72.4%) than assistant referees (66.9%) ($p < 0.01$).	Referees outperform assistant referees in an infringement identification task.
	Ghasemi et al. (2011)	Elite (n = 41). Split into two groups- top 10 ranked, bottom 10 ranked.	Nil.	76 match clips.	Nil.	Concurrent validity only.	Top referees scored significantly higher than bottom ranked.	This video test was able to distinguish between competition levels for decision-making skill.
	Poolton et al. (2011)	Amateur (n = 28). Group split in half according to whether they were high or low ruminators.	Nil.	45 match clips.	Effect of previous decisions.	NR	The high decision rumination group awarded significantly more fouls against the away team ($p = 0.001$).	Ruminating over previous decisions can have an influence on awarding penalties to the home team.
	Wagner-Egger et al. (2012)	Amateur (n = 17)	Amateur players (n = 43) Fans (n = 22)	64 clips of the video game FIFA 2005.	Racism.	NR	Generally, referees were more likely to evaluate challenges as fouls than players and fans ($p < 0.05$). Referees were more likely to evaluate challenges by white players as fouls, than black players ($p < 0.05$). This was more evident in the referee group than players and fans.	There is some evidence of discrimination, but not all in black players.
	Krenn (2014)	Sub-elite (n = 42)	University students with high football law knowledge (n = 81) University students with low football law knowledge (n = 82)	52 match clips (24 chromatic, 28 achromatic).	Uniform colour.	NR	Referees judged tackles less harshly than university students with a high and low knowledge of football ($p < 0.01$). Colour did not have a significant impact on tackle judgements ($p > 0.05$).	Uniform colour does not affect judgement of tackles.
	Renden et al. (2014)	Sub-elite (n = 31)	Sub-elite players (n = 17) Wheelchair-bound fans (n = 12) Novices (n = 18)	54 clips.	Previous motor experience.	Construct validity only.	Players and referees were both significantly more accurate than fans ($p < 0.001$). No significant differences between players and referees ($p < 0.998$), nor fans and novices ($p < 0.799$).	Motor experience playing the sport (i.e., players), is beneficial for decision-making accuracy compared to no motor experience (wheelchair bound fans). Players and referees have similar decision-making skill.

(*continued on next page*)

Table 1 (*continued*)

	Author (date)	Sport officials' skill level	Other populations	Test overview	Influences on DM in task	Validity and Reliability Assessment	Results overview	Key findings
	Lex et al. (2015)	Amateur (n = 50)	Nil.	52 clips. Participants assigned to two groups based on age and refereeing experience. One group first watched a silent, and then a noise condition (with match noise). Tested on opposite condition one week later.	Crowd influence.	NR	Does not measure decision accuracy. No influence of sound on foul decisions ($p = 0.806$). Significant effect of sound on penalty (i.e., yellow or red card) ($p < 0.01$). Referees were significantly more likely to produce a yellow card when players produced audible vocalisations (30.8%) than no vocalisations (23.6%).	When the foul has already been made, players' vocalisations influence whether a yellow card will be produced.
	Spitz et al. (2016)	Elite (n = 20) Sub-elite (n = 19)	Nil.	20 simulation clips from perspective of assistant referee.	Nil.	Construct validity only.	Participants significantly more accurate in corner kick than open play situations ($p < 0.001$). Elite referees significantly more accurate in certain situations.	This video test was able to distinguish between competition levels for decision-making skill.
	Nevill et al. (2017)	Sub-elite (n = 6)	Nil.	One game. 2 referees watched game with no supporters present, 2 watched with one team's supporters present, other pair watched with other team's supporters.	Crowd influence.	NR	The referees with NO supporters present were significantly more likely to disagree with match referee's decision ($p = 0.004$).	Systematic tendency of crowd to influence referees to make less decisions. Evidence of home bias present with referees tending to favour team of supporters present.
	Spitz et al. (2017)	Elite (n = 19) Sub-elite (n = 18)	Elite assistant referees (n = 24)	40 clips presented twice (once in normal speed, once in slow motion. Order randomised).	Slow motion speed.	Construct validity. Reliability assessed by intra-class correlation coefficients over two viewings, high reliability (range 0.76–0.82).	Elite scored higher (66.4%) than sub-elite (60.1%), and assistant referees (58.0%). Both the sub-elite and assistant referees scored significantly lower than elite ($p < 0.001$). All referees were significantly more accurate in slow motion ($p < 0.001$).	Referees are more accurate in slow motion than real time. This video test was able to distinguish between population groups.
	Spitz et al. (2018)	Elite (n = 22) Sub-elite (n = 21)	Nil.	20 simulation clips from perspective of assistant referee.	Nil.	Construct validity. Test-retest reliability from previous Spitz et al. (2017).	Elite referees were significantly more accurate than sub-elite ($p = 0.012$).	This video test was able to distinguish between competition levels for decision-making skill.
Australian football	Larkin et al. (2011)	Elite (n = 15) Sub-elite (n = 23)	Nil.	25 match clips.	Nil	Construct validity only.	Elite umpires scored significantly higher than sub-elite umpires ($p < 0.05$).	Elite umpires have more advanced decision-making skill than their sub-elite counterparts.
	Larkin, O'Brien, et al. (2014)	Amateur (n = 15)	Nil.	32 match clips (8 following the completion of each game quarter).	Game physical exertion levels.	Construct validity. Reliability from Larkin, Mesagno, et al. (2014).	No significant correlations between physical exertion and decision-making in a particular quarter. Significant improvement in quarter 4 ($p = 0.001$) compared to quarters 2 & 3.	No relationship between decision-making and in-game physical exertion. Higher decision-making at the end of the game could be due to the high importance of this period.
	Paradis et al. (2016)	Sub-elite (n = 10) Junior (n = 8)	Nil.	50 clips (10 blocks of 5: 2 × "easy", 1 × "medium", 2 × "hard").	Physical exertion: 10 × 300 m run test.	Construct validity. Reliability from Larkin, Mesagno, et al. (2014).	Sub-elite outperformed junior in decision-making ($p = 0.016$, $r = 0.5$). Only scored significantly higher in "easy" clips ($p = 0.043$, $r = 0.46$). No significant correlation between physical exertion and decision-making overall, and in any difficulty.	Physical exertion does not influence decision-making in an off-field decision-making test. Sub-elite umpires have better decision-making than junior.
Rugby union	MacMahon and Ste-Marie (2002)	Study 1: high experience (n = 12), low experience (n = 12). Study 2: high experience (n = 12)	Study 2: players (n = 12)	Clip n NR.	Nil.	NR	High and low experience had the same accuracy (53.7%) in Study 1. Players (62%) had slightly higher accuracy than referees (59%) in Study 2. Statistical significance NR.	No significant differences in accuracy for either study.
	Mascarenhas et al. (2005a)	Elite (n = 45)	Referee assessors (n = 27) Referee coaches (n = 13) Touch judges (n = 47)	10 match clips. Referees split into three groups based on ranking.	Nil.	Construct validity only.	Top group of referees scored highest. Bottom ranked referees scored higher than the middle group.	Low level of accuracy and agreement of decisions.
	Nazarudin et al. (2015)	Rugby referees (n = 132. Experience: 1–5yrs n = 33, 6–10yrs n = 34, 11–15yrs n = 34, 16 + yrs n = 31)	Nil.	18 match clips (filmed from referees' perspective)	Nil.	NR	Highly experienced referees scored higher in all decision-making than less experienced ($p < 0.05$). Match performance measured; significant relationship between match performance and decision-making performance ($r = 0.61$, $p < 0.05$)	This first person video test was able to predict match performance, and differentiate between experience levels.
Basketball	Brand et al. (2006)	Elite (n = 113)	Nil.	18 match clips. One condition in game/block order. One condition in random order.	Impact of previous decisions.	NR	Referees in the random condition more likely to make more rigorous decisions than original sequence decision.	Referees are less rigorous in their decision-making when presented in game order than random.
	MacMahon, Starkes, et al. (2007)	Amateur (n = 44)	Nil.	44 clips (2 sets of 22). Knowledge primed 1 (KP1): watched first set of infractions (ID1), then rules/signal test (T), then second set of	Priming before decision-making.	NR	Participants were more accurate in the set that had fewer off the ball infractions. Knowledge and infraction priming had little to no effect on decision-making.	More off the ball incidents represent a higher difficulty of decision-making task.

(*continued on next page*)

Table 1 (*continued*)

	Author (date)	Sport officials' skill level	Other populations	Test overview	Influences on DM in task	Validity and Reliability Assessment	Results overview	Key findings
				infractions (ID2). IP1: watched ID2, then 1D1, then did T. KP2: watched ID2, then did T, then watched ID1. IP2: watched ID1, then ID2, then did T.				
Handball	Souchon et al. (2004)	Amateur (n = 30)	Nil.	Watched video clips (n = unknown) of male and female handball incidents. Equal number of penalties for each.	Decision-making differences of male and female sport.	NR	Significantly more penalties for women than men ($p < 0.05$). Women received significantly more disciplinary penalties ($p < 0.001$).	Referees perceive fouls and penalties differently between men & women.
	Souchon et al. (2013)	Elite (n = 47) Sub-elite (n = 48) Amateur (n = 50)	Nil.	122 match clips (60 male, 62 female games).	Player gender.	NR	All participants more likely to apply sanctions to, and intervene in situations with female players than male players ($p < 0.001$), this was more evident in amateur/junior referees ($p < 0.001$). Non-significant difference between sub-elite and elite referees.	More penalties against female players than male players. Does not provide decision accuracy.
Ice hockey	Wilson and Mock (2013)	Ice hockey (high certification n = 15, low certification n = 15). High certification was level 3 or higher.	Nil.	10 match clips. Seven (high accuracy was 5–7 correct calls) were a penalty, three no penalty (high accuracy was 3 correct calls.)	Nil.	NR	Neither certification nor assertiveness levels were significantly associated with making correct calls, whether that be penalty or no penalty ($p > 0.05$). Highly certified referees that had higher assertiveness were more likely to make a correct call, referees with high certification and low assertiveness had lowest decision-making ($p = 0.03$)	Assertiveness levels are more strongly associated with decision-making in ice hockey referees rather than certification level.

3.3. Skill levels

There was a cross-sectional analysis of one skill group investigated in 8 studies. These studies covered a range of skill levels, including amateur officials (n = 6) (Jones, Paull, & Erskine, 2002; Larkin, O'Brien, et al., 2014; Lex, Pizzera, Kurtes, & Schack, 2015; MacMahon, Starkes, & Deakin, 2007; Nevill, Balmer, & Williams, 2002; Souchon, Coulomb-Cabagno, Traclet, & Rascle, 2004). Sub-elite officials alone were investigated in one study (Nevill, Hemingway, Greaves, Dallaway, & Devonport, 2017), and one study with only elite officials (Brand, Schmidt, & Schneeloch, 2006). There were ten studies identified which compared multiple skill levels of sporting officials. Specifically, two studies split one skill level into two participant groups based on the following criteria; high or low ruminators (i.e., considering previous decisions) (Poolton, Siu, & Masters, 2011), and top compared to bottom ranked (Ghasemi, Momeni, Jafarzadehpur, Rezaee, & Taheri, 2011). Two studies investigated the effect of experience (MacMahon & Ste-Marie, 2002; Nazarudin et al., 2015), and one investigated certification level (Wilson & Mock, 2013). In terms of officiating level; amateur, sub-elite and elite were compared in one study (Souchon, Livingstone, & Maio, 2013), elite vs sub-elite in three studies (Larkin et al., 2011; Spitz et al., 2016; Spitz et al., 2018), and sub-elite to junior in one study (Paradis et al., 2016). The differences between each of these groups are presented in Table 2.

The remaining studies (n = 9) investigated decision-making differences of interactor officials to individuals who do not complete 'central' officiating tasks. Comparisons were made against assistant referees (Catteeuw et al., 2009; Spitz et al., 2017), soccer players (MacMahon, Helsen, et al., 2007; Plessner & Betsch, 2001), wheelchair-bound fans, players and novices (Renden et al., 2014), players and fans (Wagner-Egger, Gygax, & Ribordy, 2012), official's assessors, official's coaches, touch judges (Mascarenhas et al., 2005a), university students with high football knowledge and low football knowledge (Krenn, 2014), and finally individuals with soccer refereeing, coaching and playing experience but did not distinguish between groups (Balmer, Nevill, Lane, & Ward, 2007).

3.4. Clip type used

There was a range of decision-making footage in the testing protocols. Most commonly, individual video clips of match play from a broadcast perspective (i.e., third person) were presented in 21 studies (Brand et al., 2006; Catteeuw et al., 2009; Ghasemi et al., 2011; Jones et al., 2002; Krenn, 2014; Larkin et al., 2011; Larkin, O'Brien, et al., 2014; Lex et al., 2015; MacMahon, Helsen, et al., 2007; MacMahon, Starkes, et al., 2007; MacMahon & Ste-Marie, 2002; Mascarenhas et al., 2005a; Nevill et al., 2002; Paradis et al., 2016; Plessner & Betsch, 2001; Poolton et al., 2011; Renden et al., 2014; Souchon et al., 2004; Souchon et al., 2013; Spitz et al., 2017; Wilson & Mock, 2013). Two studies investigated the decision-making of referees while watching one soccer game from a broadcast (i.e., third person) perspective in a single sitting (i.e., watching a full football game at once) (Balmer et al., 2007; Nevill et al., 2017). One study investigated individual clips filmed from the referees' perspective (i.e., first person) using head-mounted glasses (Nazarudin et al., 2015). Two studies filmed simulation clips of players from the assistant referees' perspective (mix of first and third person) attempting to present a stronger simulation of the referees' perspective (Spitz et al., 2016; Spitz et al., 2018). Although this is from the assistant referees' perspective, the decision-making was infringement-based as per the inclusion criteria. Clips from the video game FIFA 2005 were investigated in one study (Wagner-Egger et al., 2012).

3.5. Additional influences on decision-making

Two studies examined the effect of physical exertion on decision-making performance, with one implementing decision-making following 300 m efforts (Paradis et al., 2016), and another in the quarter breaks of an Australian football match (Larkin, O'Brien, et al., 2014). Four studies examined the influence of crowd noise on decision-making performance. Of these, one study watched a game with a team's supporters present in the room (Nevill et al., 2017), and three studies implemented crowd noise into the video clips compared to a silent condition (Balmer et al., 2007; Lex et al., 2015; Nevill et al., 2002). The effect of previous decisions was investigated in three studies (Brand et al., 2006; Plessner & Betsch, 2001; Poolton et al., 2011). One study explored the effect of priming prior to a decision-making test (MacMahon, Starkes, et al., 2007). Two examples of priming were used, including; knowledge priming (i.e., completing a rules and signals test prior to the decision-making task), and infraction priming (i.e., being instructed to focus on specific infractions prior to the decision-making task). The influence of slow motion on decision-making accuracy was assessed in one study (Spitz et al., 2017). Biases were examined as an influence on decision-making in five studies encompassing player gender (Souchon et al., 2004; Souchon et al., 2013), player skin colour

Table 2
Quality assessment of included studies.

Study	Validity and reliability			Skill level of officials				Decision reporting		
	Not reported or evident	Construct or concurrent validity evident	Reliability reported	Not reported	Amateur or junior	Sub-elite	Elite	Not reported	Number of decisions only	Accuracy of decisions reported
Spitz et al. (2018)		✓	✓			✓	✓			✓
Spitz et al. (2017)		✓	✓			✓	✓			✓
Paradis et al. (2016)		✓	✓		✓	✓				✓
Larkin, O'Brien, et al. (2014)		✓	✓		✓					✓
Spitz et al. (2016)		✓				✓	✓			✓
Larkin et al. (2011)		✓				✓	✓			✓
Catteeuw et al. (2009)		✓					✓			✓
MacMahon, Helsen, et al. (2007)		✓					✓			✓
Nazarudin et al. (2015)		✓				✓				✓
Mascarenhas et al. (2005a)		✓					✓			✓
Renden et al. (2014)		✓				✓				✓
Wilson and Mock (2013)	✓				✓	✓				✓
Ghasemi et al. (2011)	✓						✓			✓
Plessner and Betsch (2001)	✓					✓				✓
MacMahon, Starkes, et al. (2007)	✓				✓					✓
MacMahon and Ste-Marie (2002)	✓			✓						✓
Souchon et al. (2013)	✓				✓	✓	✓		✓	
Brand et al. (2006)	✓						✓		✓	
Nevill et al. (2017)	✓					✓			✓	
Krenn (2014)	✓					✓			✓	
Lex et al. (2015)	✓				✓				✓	
Wagner-Egger et al. (2012)	✓				✓				✓	
Poolton et al. (2011)	✓				✓				✓	
Souchon et al. (2004)	✓				✓				✓	
Nevill et al. (2002)	✓				✓				✓	
Jones et al. (2002)	✓				✓				✓	
Balmer et al. (2007)	✓			✓					✓	

(Wagner-Egger et al., 2012), aggressive team reputation (Jones et al., 2002), and uniform colour (Krenn, 2014). There were 11 studies which did not investigate any additional influences on decision-making (Catteeuw et al., 2009; Ghasemi et al., 2011; Larkin et al., 2011; MacMahon, Helsen, et al., 2007; MacMahon & Ste-Marie, 2002; Mascarenhas et al., 2005a; Nazarudin et al., 2015; Renden et al., 2014; Spitz et al., 2016; Spitz et al., 2018; Wilson & Mock, 2013).

3.6. Transfer of skills to match performance

One study examined the transfer of video-based performance to match performance (Nazarudin et al., 2015).

4. Discussion

The primary aim of this review was to provide a summary of the off-field video-based decision-making assessment literature of interactor officials, and analyse the various methods utilised to simulate sport-specific decision-making. The results highlight several key findings, including: i) soccer (football) is emphatically the most researched sport in this domain; ii) off-field video-based methods appear to have high construct validity (i.e., ability to differentiate between skill levels); iii) a high degree of variability in the methods applied to each study, leading to mixed interpretations of the results; iv) most influences on decision-making applied in the tests effectively highlight potential biases that may be present to an official in match play. Importantly, tests that effectively replicate potential biases such as crowd noise present in match play may be able to delineate which officials are affected by certain biases, leading to individualised training programs to promote consistent decision-making.

This review focuses on the off-field decision-making assessment of interactor officials from a range of sports. Soccer officials are the most commonly researched group, followed by Australian football umpires and rugby union referees with three studies focusing on each of these groups. A review of research on sporting officials also indicated the majority focuses on soccer officials (Hancock, Rix-Lièvre, & Côté, 2015). While certain findings can be transferred from research involving soccer referees, there are a number of inherent differences between soccer and other sports. As noted by MacMahon et al. (2014), sporting officials in different sports vary in the number of players and cues they need to monitor as part of their role, which can consequently impact the decision-making requirement of officials across different sports. Furthermore, it has been suggested that a wider range of research is required to determine which officiating approach is most effective given a specific sport or situation (MacMahon et al., 2014). Different approaches to decision-making include officiating in a black and white manner where each decision is taken in isolation, or using contextual judgement to apply the laws in consideration of the environment. Specifically, a key coaching instruction of soccer referees is for each decision to be evaluated in isolation, irrespective of previous decisions (Plessner & Betsch, 2001). Despite these coaching instructions, there are certain number of "unwritten rules" in officiating which can influence decisions (Plessner & Betsch, 2001). Contextual judgement is regarded as an important factor by elite officials themselves (Mascarenhas, Collins, & Mortimer, 2005b; Morris & O'Connor, 2016), whereby decision-making is not black and white, rather officials are somewhat flexible in their final decision based upon certain contextual factors. For example, elite rugby union and rugby league referees consider the impact of a number of factors such as time, score line, momentum and field position when making decisions (Mascarenhas et al., 2005b; Morris & O'Connor, 2016). This knowledge will contribute to the understanding of how officials approach decision-making, by further exploring the effect of contextual judgements and "unwritten rules" in specific sports.

A limited number of studies reported the validity of the video-based test, and fewer reported reliability. It is imperative measures of validity and reliability are provided prior to using a video-based assessment to ensure accurate results (Larkin et al., 2015), and is a necessity to ensure a robust measurement tool for accurate results. Many studies do not explicitly state the validity of their testing measures. Construct validity is, in fact, measured in 13 of the 27 studies included, by comparing decision-making performance across multiple skill groups. Construct validity was evident in 10 of these studies, which were able to differentiate between skill levels using the video-based task. In terms of reliability, this was reported in only four of the 27 studies. One study included a reliability assessment as part of the study, demonstrating high reliability as assessed per intra-class correlation coefficient (0.76–0.82) (Spitz et al., 2017). In addition, three studies reported reliability from a previous paper (Larkin, O'Brien, et al., 2014; Paradis et al., 2016; Spitz et al., 2017). Only one study examined concurrent validity of the decision-making task (Nazarudin et al., 2015). A video-based test with a strong relationship to on-field performance or ranking has the potential to be used in conjunction with a battery of additional performance measures for talent identification. These results highlight reliability is a key consideration to be included in future studies, yet is not in the majority of research on this topic.

A key measure of validity is construct validity, which in this review refers to how the test differentiates between known skill levels (Gadotti et al., 2006). It is evident there is a high degree of inconsistency within this research area, when examining a range of skill levels. Eight studies provided a cross-sectional analysis of

just one skill level. Some of these studies investigated the effect of an external influence on decision-making performance such as crowd (Nevill et al., 2017) or player vocalisation (Lex et al., 2015), yet is it important to consider whether different levels of sporting officials are affected by these external stimuli similarly or differently. Ten studies investigated the decision-making performance across skill levels, of which one study examined three separate performance levels; elite, sub-elite, and novice (Souchon et al., 2013). It is imperative for studies to draw comparisons across multiple skill levels to determine construct validity. Establishing the validity of a video-based tool assists in monitoring individual decision-making progression over time using a consistent measurement. When determining what constitutes the performance level of an official, there is a lack of uniformity among the studies, similar to the general sporting population (Swann, Moran, & Piggott, 2015). The majority of studies consider 'elite' officials to be officiating the top national level for their respective sports (Brand et al., 2006; Larkin et al., 2011; Souchon et al., 2013; Spitz et al., 2016; Spitz et al., 2018). This definition is accurate for Australian football umpires as Australia is the only country to play this sport at a national high performance level. However, this can become problematic when considering sports such as soccer and basketball. National German basketball referees are labelled 'elite' by Brand et al. (2006), yet the highest performance level for basketball would be the Basketball World Championships or the NBA competition in the United States. To assist with the creation of a more uniform system for comparison between studies, we propose a categorisation system based from Swann et al. (2015). This system provides examples from the three most researched sporting official groups in this paper (soccer, Australian football, rugby union) (see Fig. 3). As noted by Swann et al. (2015), classification of officials is dependent on not only performance level, but also the standard of competition in the country the sport is officiated. For multi-national sports such as soccer and rugby union, national-level officials are classified differently in large or small sporting nations. It is anticipated the development of this continuum will assist the uniformity of performance levels examined within, and between sports in this research area.

A number of studies compare decision-making performance of referees to other population groups. For example, foul infringement decision-making was assessed between central soccer referees, and their assistant referee counterparts (Catteeuw et al., 2009; Spitz et al., 2017). These studies were included as they assessed match-specific decision-making of the central referee, compared to the assistant referees completing the same task (i.e., no off-side decision-making was included). In both studies, the central referees scored higher than assistant referees. When comparing soccer referees to players, however, referees outperformed soccer players in decision-making performance in one study (MacMahon, Helsen, et al., 2007), but not another (Plessner & Betsch, 2001). Similarly, Renden et al. (2014) determined there were no significant differences between soccer players and referees, yet higher decision-making performance by these groups than wheelchair-bound fans and novices (with no soccer experience). They concluded sport-specific motor experience may play an important role in decision-making skill. Researchers are encouraged to assess accuracy of decisions, as determined by a panel of experts such as officials' coaches (Larkin, Mesagno, et al., 2014; Spitz et al., 2018). Rugby union referees have been compared to touch-judges, assessors, and referee coaches (Mascarenhas et al., 2005a). The construct validity of this study was quite high, with the highest ranked referees scoring the best on the test. Interestingly though, referee coaches who are considered subject matter experts scored the lowest. Referees, coaches and players have been found to award more fouls when there is crowd noise (Balmer et al., 2007). These results, however, were presented with no differentiation between groups or accuracy of the decisions. There are a multitude of studies (Krenn, 2014; Renden et al., 2014; Spitz et al., 2017) comparing officials' decision-making to other populations based on a number of factors (such as uniform colour (Krenn, 2014), motor experience of playing the sport (Renden et al., 2014), and slow motion footage (Spitz et al., 2017)). These studies only further demonstrate skill-based differences associated with the expertise effect, and video-based activities could be used to promote recruitment of these individuals into the ranks of officials. Upon analysis across this research area, there is a wide discrepancy of methods applied, especially in relation to decision-making accuracy. For results to have significance, reporting of decision-making accuracy within the video-based task is a necessity.

In terms of clip type used, the majority of studies (n = 23) have utilised broadcast footage as the presentation method. This is typically filmed from a fixed location in the grandstand and does not replicate the perspective of an official in a game (Craig, 2013), and as such lacks fidelity. Fidelity refers to the extent a situation replicates reality, and is a key element of transfer of on-field performance to off-field such as in video-based tasks (Alessi, 1988; Farrow, 2013; Lorains, Ball, & MacMahon, 2013). Wagner-Egger et al. (2012) presented decision-making scenarios from the video game FIFA 2005, with the rationale that this allows controlled scenarios from multiple viewpoints. As this task presents animations, however, it potentially lacks fidelity as it does not provide real world footage. To increase fidelity through heightened representativeness of the task, one study filmed match clips from the referees' perspective using head-mounted glasses (Nazarudin et al., 2015). Similarly, research has filmed match simulations from the perspective of assistant referees (Spitz et al., 2016; Spitz et al., 2018). These three studies have provided an advancement of the literature, by implementing first-person/egocentric viewpoints to increase representativeness (Petit & Ripoll, 2008). A limitation which remains in this method is video footage does not automatically update with visual changes from head movements, which limited perception-action coupling (Craig, 2013). Virtual reality presentation of 360° videos is a possible technology to overcome this barrier to increase representativeness of the task, as it allows an individual's head movements to be the visual changes perceived. It is imperative for the video-based task to be representative of the match environment to determine differences between expertise levels (Williams & Ericsson, 2005). For this research area to advance, different technologies such as virtual reality should be compared to traditional methods such as broadcast footage.

Several studies have attempted to increase the representativeness of video-based tasks by introducing potential stressors or examining certain biases that may influence an official in a game. As Australian football umpires have a physically demanding task (Elsworthy et al., 2014), studies examined the effect of physical demands in relation to decision-making performance. These studies reported no relationship between physical exertion and video-based decision-making performance (Larkin, O'Brien, et al., 2014; Paradis et al., 2016), similar to match data (Elsworthy et al., 2014). As sport officials, especially elite, contend with significant crowd noise, this has been examined in the literature. One study involved soccer referees watching a game with fans or no fans present, with results demonstrating participants are more likely to disagree with the decision when there are no supporters present (Nevill et al., 2017). Introduction of match noise (i.e., players, commentary, and crowd) in the task leads to a home team bias in decision-making in two studies (Balmer et al., 2007; Nevill et al., 2002), but not necessarily another study (Lex et al., 2015). Despite the lack of influence match noise had on awarding fouls, Lex et al. (2015) reported that it led to a bias in awarding penalties (i.e., red or yellow card) against the away side. This supports the literature stating crowd noise does influence on-field decision-making (Downward & Jones, 2007; Goumas, 2014). Unlike crowd noise, there appears to be no decision-making bias against a team with an aggressive reputation (Jones et al., 2002). The influence of previous decisions has been examined, with results inferring this does indeed impact subsequent decisions in soccer (Plessner & Betsch, 2001), and basketball officials (Brand et al., 2006). Extending on this research, Poolton et al. (2011) reported referees who ruminate over previous decisions are significantly more likely to award more fouls against the away team, supporting the home advantage theory also. The decision-making of referees in comparison to players and fans has been examined in relation to the skin colour of the soccer player. Wagner-Egger et al. (2012) highlighted there were no decision-making differences between these groups, but challenges made by black players were more likely to be considered fouls, whereas fouls made by white players were considered to be more severe. Uniform colour has also been examined as a possible decision-making bias as per skin colour. Results suggested soccer referees judged tackles less harshly than university students, but there was no overall impact of colour on tackle judgement (Krenn, 2014). Unfortunately though, the results of these two do not present the accuracy of the decisions. Although these findings indicate players of a specific shirt colour for example may commit "harsher" fouls, it is imperative to investigate whether these decisions are more or less accurate. The results suggest decision-making in off-field tasks can be influenced in the same way as on-field performance is, and certain biases can influence decision-making.

To accurately assess the performance of an individual on a video-based test, it is important to measure the transfer to match performance. In this review, however, only one study included the assessment of transfer, by calculating the correlation between video-based test performance and match performance (Nazarudin et al., 2015). Research has suggested including this integral component in video-based tasks for sports officials (Paradis et al., 2016), however this is rarely examined. In athletes, transfer tasks have been used to assess the effectiveness of perceptual-cognitive interventions (Lorains et al., 2013). Results suggest the introduction of a perceptual-cognitive training stimulus may not directly benefit on-field performance, despite performance improvement in an off-field video-based test (Gorman & Farrow, 2009; Lorains et al., 2013), suggesting a low correlation between on-field performance, and off-field video-based tests. Transfer tests have typically not been used in the research due to the inherent difficulties of natural in-game variation, and the lack of control researchers have over this environment. As a result, researchers typically do not incorporate this component as it leads to validity and reliability issues (Larkin et al., 2015). Despite this, a novel transfer test is better than none as sporting officials, and athletes in general, are ultimately measured by on-field performance. Only one study

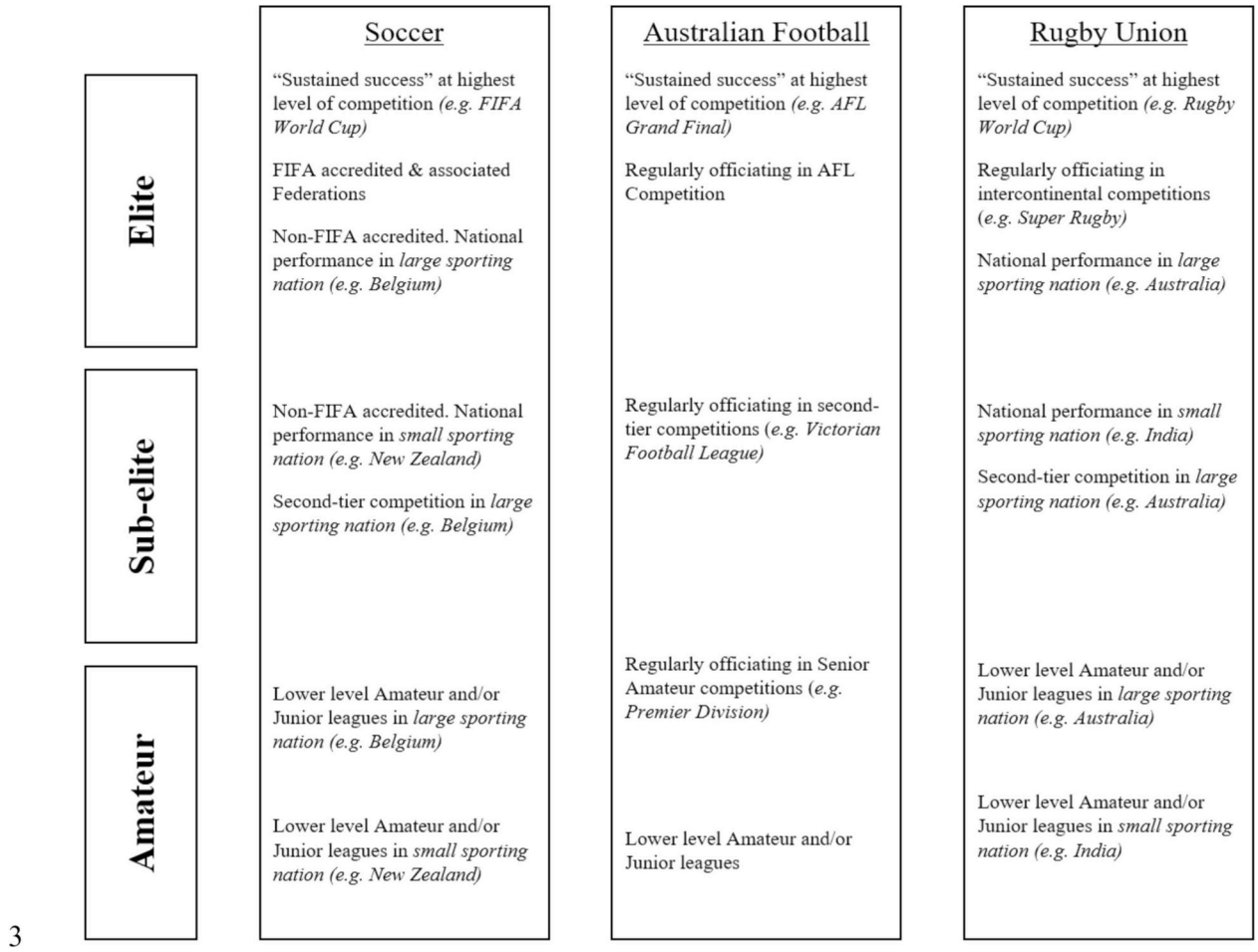

Fig. 3. Classification of the different levels of sporting officials.

included in this review examined concurrent validity (Nazarudin et al., 2015). Concurrent validity can be assessed via skill transfer by correlating the video-based test decision-making accuracy to on-field decision-making accuracy. Alternatively, performance ranking of officials could be assessed, as decision-making is the most important skill for officiating, and a key cornerstone of success (Helsen & Bultynck, 2004; Mascarenhas et al., 2005b). To advance this area of the literature, comparisons to on-field performance are vital.

5. Conclusions

In summary, this review has highlighted several key findings among the literature assessing decision-making skill in interactor officials. Firstly, soccer referees are the most predominantly investigated group of sporting officials. There were 16 studies which researched this group, compared to the next highest being Australian football and rugby union (three studies each). Video-based methods appear to have high construct validity when contrasting decision-making skill of multiple skill levels. Therefore, video-based tests can be used as a consistent and accurate measure of individual decision-making progression over a period of time. There is, however, a high amount of variability (i.e., clip type, number of clips, participating groups) in the methods across the studies identified which may lead to a certain degree of incompatibility among the findings. Finally, this review highlighted the influences which can bias officials' decision-making. Of these influences, crowd noise appears to be a prominent influence on decision-making accuracy.

5.1. Consideration for future research

Based on the key findings outlined above, the authors have several considerations for the direction of research in this area. As discussed, there is a plethora of research investigating decision-making of soccer referees. While it is recognised that is considered the world's most popular sport (Giulianotti, 2012), there is a necessity for research in other interactor officials to promote effective practice in other sports. In addition, validity and, especially, reliability measures need to be implemented in studies of this nature to ensure rigor of the video-based assessment tool. Similar to consistency in the applied methodology of each study, this will ensure compatibility of results across different sports. The most common modality of video presentation was match broadcast footage of sporting games. With the advent of technology, other modalities such as virtual or augmented reality could be considered to be more representative of the in-game decision-making of a sporting official in an off-field controlled environment. For this research area to develop, the assessment of transfer is imperative. Although there are inherent limitations in examining transfer, the exploration of novel transfer assessments will further advance this area of the literature, hence reinforcing the practical implications of this method.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Alessi, S. M. (1988). Fidelity in the design of instructional simulations. *Journal of Computer-Based Instruction, 15*, 40–47.
Araújo, D., Davids, K., & Hristovski, R. (2006). The ecological dynamics of decision making in sport. *Psychology of Sport and Exercise, 7*(6), 653–676.
Baker, J., Cote, J., & Abernethy, B. (2003a). Sport-specific practice and the development of expert decision-making in team ball sports. *Journal of Applied Sport Psychology, 15*(1), 12–25.
Baker, J., Côté, J., & Abernethy, B. (2003b). Learning from the experts: Practice activities of expert decision makers in sport. *Research Quarterly for Exercise & Sport, 74*(3), 342–347.
Balmer, N. J., Nevill, A. M., Lane, A. M., & Ward, P. (2007). Influence of crowd noise on soccer refereeing consistency in soccer. *Journal of Sport Behavior, 30*(2), 130.
Berry, J., Abernethy, B., & Côté, J. (2008). The contribution of structured activity and deliberate play to the development of expert perceptual and decision-making skill.

Journal of Sport & Exercise Psychology, 30(6), 685–708.
Brand, R., Schmidt, G., & Schneeloch, Y. (2006). Sequential effects in elite basketball referees' foul decisions: An experimental study on the concept of game management. *Journal of Sport & Exercise Psychology, 28*(1), 93–99.
Burnett, A. M., Bishop, D. T., Ashford, K. J., Williams, A. M., & Kinrade, N. P. (2017). Decision-making of English netball superleague umpires: Contextual and dispositional influences. *Psychology of Sport and Exercise, 31*, 52–60.
Castagna, C., Abt, G., & D'ottavio, S. (2002). Relation between fitness tests and match performance in elite Italian soccer referees. *The Journal of Strength & Conditioning Research, 16*(2), 231–235.
Catteeuw, P., Helsen, W., Gilis, B., & Wagemans, J. (2009). Decision-making skills, role specificity, and deliberate practice in association football refereeing. *Journal of Sports Sciences, 27*(11), 1125–1136.
Craig, C. (2013). Understanding perception and action in sport: How can virtual reality technology help? *Sports Technology, 6*(4), 161–169.
Cunningham, I., Simmons, P., Mascarenhas, D., & Redhead, S. (2014). Skilled interaction: Concepts of communication and player management in the development of sport officials. *International Journal of Sport Communication, 7*(2), 166–187.
DeVellis, R. F. (2016). *Scale development: Theory and applications*, *26*. Sage publications.
Downward, P., & Jones, M. (2007). Effects of crowd size on referee decisions: Analysis of the FA Cup. *Journal of Sports Sciences, 25*(14), 1541–1545.
Elsworthy, N., Burke, D., Scott, B. R., Stevens, C. J., & Dascombe, B. J. (2014). Physical and decision-making demands of Australian football umpies during competitive matches. *The Journal of Strength & Conditioning Research, 28*(12), 3502–3507.
Elsworthy, N., & Dascombe, B. J. (2011). The match demands of Australian rules football umpires in a state-based competition. *International Journal of Sports Physiology and Performance, 6*(4), 559–571.
Emmonds, S., O'Hara, J., Till, K., Jones, B., Brightmore, A., & Cooke, C. (2015). Physiological and movement demands of rugby league referees: Influence on penalty accuracy. *The Journal of Strength & Conditioning Research, 29*(12), 3367–3374.
Farrow, D. (2013). Practice-enhancing technology: A review of perceptual training applications in sport. *Sports Technology, 6*(4), 170–176.
Gadotti, I., Vieira, E., & Magee, D. (2006). Importance and clarification of measurement properties in rehabilitation. *Brazilian Journal of Physical Therapy, 10*(2), 137–146.
Ghasemi, A., Momeni, M., Jafarzadehpur, E., Rezaee, M., & Taheri, H. (2011). Visual skills involved in decision making by expert referees. *Perceptual & Motor Skills, 112*(1), 161–171.
Giulianotti, R. (2012). *Football*. Wiley Online Library.
Gorman, A. D., & Farrow, D. (2009). Perceptual training using explicit and implicit instructional techniques: Does it benefit skilled performers? *International Journal of Sports Science & Coaching, 4*(2), 193–208.
Goumas, C. (2014). Home advantage and referee bias in European football. *European Journal of Sport Science, 14*(Suppl. 1), S243–S249.
Hancock, D. J., Rix-Lièvre, G., & Côté, J. (2015). Citation network analysis of research on sport officials: A lack of interconnectivity. *International Review of Sport and Exercise Psychology, 8*(1), 95–105.
Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238.
Helsen, W., & Bultynck, J.-B. (2004). Physical and perceptual-cognitive demands of top-class refereeing in association football. *Journal of Sports Sciences, 22*(2), 179–189.
Johansen, B. T., & Haugen, T. (2013). Anxiety level and decision-making among Norwegian top-class soccer referees. *International Journal of Sport and Exercise Psychology, 11*(2), 215–226.
Jones, M. V., Paull, G. C., & Erskine, J. (2002). The impact of a team's aggressive reputation on the decisions of association football referees. *Journal of Sports Sciences, 20*(12), 991–1000.
Krenn, B. (2014). The impact of uniform color on judging tackles in association football. *Psychology of Sport and Exercise, 15*(2), 222–225.
Krustrup, P., & Bangsbo, J. (2001). Physiological demands of top-class soccer refereeing in relation to physical capacity: Effect of intense intermittent exercise training. *Journal of Sports Sciences, 19*(11), 881–891.
Larkin, P., Berry, J., Dawson, B., & Lay, B. (2011). Perceptual and decision-making skills of Australian football umpires. *International Journal of Performance Analysis in Sport, 11*(3), 427–437.
Larkin, P., Mesagno, C., Berry, J., & Spittle, M. (2014a). Development of a valid and reliable video-based decision-making test for Australian football umpires. *Journal of Science and Medicine in Sport, 17*(5), 552–555.
Larkin, P., Mesagno, C., Berry, J., & Spittle, M. (2016). Exploration of the perceptual-cognitive processes that contribute to in-game decision-making of Australian football umpires. *International Journal of Sport and Exercise Psychology*, 1–13.
Larkin, P., Mesagno, C., Spittle, M., & Berry, J. (2015). An evaluation of video-based training programs for perceptual-cognitive skill development. A systematic review of current sport-based knowledge. *International Journal of Sport Psychology, 46*(6), 555–586.
Larkin, P., O'Brien, B., Mesagno, C., Berry, J., Harvey, J., & Spittle, M. (2014b). Assessment of decision-making performance and in-game physical exertion of Australian football umpires. *Journal of Sports Sciences, 32*(15), 1446–1453.
Lex, H., Pizzera, A., Kurtes, M., & Schack, T. (2015). Influence of players' vocalisations on soccer referees' decisions. *European Journal of Sport Science, 15*(5), 424–428.
Lorains, M., Ball, K., & MacMahon, C. (2013). An above real time training intervention for sport decision making. *Psychology of Sport and Exercise, 14*(5), 670–674.
MacMahon, C., Helsen, W., Starkes, J., & Weston, M. (2007a). Decision-making skills and deliberate practice in elite association football referees. *Journal of Sports Sciences, 25*(1), 65–78.
MacMahon, C., Mascarenhas, D., Plessner, H., Pizzera, A., Oudejans, R., & Raab, M. (2014). *Sports officials and officiating: Science and practice*. Oxon: Routledge.
MacMahon, C., Starkes, J., & Deakin, J. (2007b). Referee decision making in a video-based infraction detection task: Application and training considerations. *International Journal of Sports Science & Coaching, 2*(3), 257–265.
MacMahon, C., & Ste-Marie, D. M. (2002). Decision-making by experienced rugby referees: Use of perceptual information and episodic memory. *Perceptual & Motor Skills, 95*(2), 570–572.
Mallo, J., Frutos, P. G., Juárez, D., & Navarro, E. (2012). Effect of positioning on the accuracy of decision making of association football top-class referees and assistant referees during competitive matches. *Journal of Sports Sciences, 30*(13), 1437–1445.
Marteniuk, R. G. (1976). *Information processing in motor skills*. New York: Holt, Rinehart and Winston.
Mascarenhas, D., Collins, D., & Mortimer, P. (2005a). The accuracy, agreement and coherence of decision-making in rugby union officials. *Journal of Sport Behavior, 28*(3), 253–271.
Mascarenhas, D., Collins, D., & Mortimer, P. (2005b). Elite refereeing performance: Developing a model for sport science support. *The Sport Psychologist, 19*, 364–379.
Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), e1000097.
Morris, G., & O'Connor, D. (2016). Key attributes of expert NRL referees. *Journal of Sports Sciences, 35*(9), 852–857.
Nazarudin, M. N., Abdullah, M. R., Suppiah, P. K., Fauzee, M. S. O., Parnabas, V., & Abdullah, N. M. (2015). Decision making and performance of Malaysian rugby sevens referees. *Movement, Health & Exercise, 4*(1).
Nevill, A. M., Balmer, N. J., & Williams, A. M. (2002). The influence of crowd noise and experience upon refereeing decisions in football. *Psychology of Sport and Exercise, 3*(4), 261–272.
Neville, T. J., Salmon, P. M., & Read, G. J. (2018). Radio gaga? Intra-team communication of Australian rules football umpires–effect of radio communication on content, structure and frequency. *Ergonomics, 61*(2), 313–328.
Nevill, A. M., Hemingway, A., Greaves, R., Dallaway, A., & Devonport, T. J. (2017). Inconsistency of decision-making, the Achilles heel of referees. *Journal of Sports Sciences, 35*(22), 2257–2261.
Page, K., & Page, L. (2010). Alone against the crowd: Individual differences in referees' ability to cope under pressure. *Journal of Economic Psychology, 31*(2), 192–199.
Paradis, K., Larkin, P., & O'Connor, D. (2016). The effects of physical exertion on decision-making performance of Australian football umpires. *Journal of Sports Sciences, 34*(16), 1–7.
Petit, J.-P., & Ripoll, H. (2008). Scene perception and decision making in sport simulation: A masked priming investigation. *International Journal of Sport Psychology, 39*(1), 1–19.
Plessner, H., & Betsch, T. (2001). Sequential effects in important referee decisions: The case of penalties in soccer. *Journal of Sport & Exercise Psychology, 23*(3), 254–259.
Poolton, J., Siu, C. M., & Masters, R. (2011). The home team advantage gives football referees something to ruminate about. *International Journal of Sports Science & Coaching, 6*(4), 545–552.
Renden, P. G., Kerstens, S., Oudejans, R. R., & Cañal-Bruland, R. (2014). Foul or dive? Motor contributions to judging ambiguous foul situations in football. *European Journal of Sport Science, 14*(Suppl. 1), S221–S227.
Souchon, N., Coulomb-Cabagno, G., Traclet, A., & Rascle, O. (2004). Referees' decision making in handball and transgressive behaviors: Influence of stereotypes about gender of players? *Sex Roles, 51*(7–8), 445–453.
Souchon, N., Livingstone, A. G., & Maio, G. R. (2013). The influence of referees' expertise, gender, motivation, and time constraints on decisional bias against women. *Journal of Sport & Exercise Psychology, 35*(6), 585–599.
Spitz, J., Put, K., Wagemans, J., Williams, A. M., & Helsen, W. F. (2016). Visual search behaviors of association football referees during assessment of foul play situations. *Cognitive Research: Principles and Implications, 1*(1), 12.
Spitz, J., Put, K., Wagemans, J., Williams, A. M., & Helsen, W. F. (2017). Does slow motion impact on the perception of foul play in football? *European Journal of Sport Science, 17*(6), 748–756.
Spitz, J., Put, K., Wagemans, J., Williams, A. M., & Helsen, W. F. (2018). The role of domain-generic and domain-specific perceptual-cognitive skills in association football referees. *Psychology of Sport and Exercise, 34*(Suppl. C), 47–56.
Swann, C., Moran, A., & Piggott, D. (2015). Defining elite athletes: Issues in the study of expert performance in sport psychology. *Psychology of Sport and Exercise, 16*(1), 3–14.
Thomas, J. R., Silverman, S., & Nelson, J. (2015). *Research methods in physical activity*, *7E*. Human kinetics.
Wagner-Egger, P., Gygax, P., & Ribordy, F. (2012). Racism in soccer? Perception of challenges of black and white players by white referees, soccer players, and fans. *Perceptual & Motor Skills, 114*(1), 275–289.
Williams, M., & Ericsson, A. (2005). Perceptual-cognitive expertise in sport: Some considerations when applying the expert performance approach. *Human Movement Science, 24*(3), 283–307.
Williams, M., Ward, P., Smeeton, N., & Allen, D. (2004). Developing anticipation skills in tennis using on-court instruction: Perception versus perception and action. *Journal of Applied Sport Psychology, 16*(4), 350–360.
Wilson, A. W., & Mock, S. E. (2013). Association of certification level and assertiveness with accuracy of calls among ice hockey referees. *International Journal of Sports Science & Coaching, 8*(3), 505–512.