Ger J Exerc Sport Res
https://doi.org/10.1007/s12662-021-00794-8
Received: 22 May 2021
Accepted: 23 December 2021

Nicolas Bloß[1] · Jörg Schorer[1] · Florian Loffing[1,2] · Dirk Büsch[1]
[1] Institute of Sport Science, Carl von Ossietzky University Oldenburg, Oldenburg, Germany
[2] Institute of Psychology, German Sport University Cologne, Cologne, Germany

Decisions and reasonings of top-class handball referees under physical load

Introduction

A central challenge for referees in sports games is the decision-making in dynamic and complex situations under physical load (Helsen & Bultynck, 2004). In the recent decade, game dynamics of sports games like handball increased, leading to higher physical load referees must cope with (Bilge, 2012; Michalsik, 2018). Specific to handball, a well-trained endurance performance is essential. Handball requires aerobic and anaerobic endurance to perform well on the highest level of competition, as an insufficient physical capacity could impair *referees' decision-making* (RDM; Belcic, Ruzic, & Marošević, 2020; Morillo, Reigal, Hernández-Mendo, Montaña, & Morales-Sánchez, 2017). Also considering that referees are especially assessed through correct RDM, research on the relationship between physical load and RDM is of particular relevance (MacMahon et al., 2015).

RDM, in terms of knowledge and application of the rules of the game is considered a central cornerstone of referees' performance (Mascarenhas, Collins, & Mortimer, 2005b). RDM is understood as a primarily perceptual–cognitive process (Gaoua, de Oliveira, & Hunter, 2017). That is, to ensure high-quality decision-making on dynamically evolving game situations, referees must be capable of picking up, processing and integrating several environmental cues (Helsen, MacMahon, & Spitz, 2019; MacMahon et al., 2015; Plessner & Haar, 2006). Research suggests that perceptual–cognitive functions can be facilitated or impaired through physical load, depending on the task and the exercise intensity (Chang, Labban, Gapin, & Etnier, 2012). Hence, RDM might also be enhanced or impaired by physical load (Bloß et al., 2020; Schmidt et al., 2019). Currently, there are only a few studies investigating the relationship between physical load and RDM in sports, but none in handball.

Current research on the relationship between physical load and RDM can be differentiated into the consideration of external (e.g. running distance) and internal load (e.g. heart rate; Impellizzeri, Marcora, & Coutts, 2019). With regard to studies using *external load* as a parameter for physical load (Bloß et al., 2020; Impellizzeri et al., 2019), findings neither indicate a relationship between running time (Paradis, Larkin, & O'Connor, 2015) nor distance covered and RDM (Emmonds et al., 2015; Mascarenhas et al., 2009). While findings by Oudejans et al. (2005) indicate that more RDM errors occur at higher velocities (> 8 km/h), the study of Gomez-Carmona and Pino-Ortega (2016) point out less accurate RDM at slow velocities (< 8 km/h). Furthermore, findings of Catteeuw, Gilis, Wagemans, and Helsen (2010), Mascarenhas et al. (2009) and Emmonds et al. (2015) suggest that RDM does not deteriorate through progressive match periods, while Ahmed, Davison, and Dixon (2017) and Mallo, Frutos, Juarez, and Navarro (2012) found an impairment of RDM in the second compared to the first half of a match. Also, RDM was found to decrease in the last 10 or 15 minutes of a match respectively under high physical load (rugby; Emmonds et al., 2015; soccer; Mallo et al., 2012). In accordance with previous findings, Samuel, Galily, Guy, Sharoni, and Tenenbaum (2019) showed in a laboratory setting that RDM of soccer referees decreased from quarter two to three as well as from quarter three to four. In contrast, Emmonds et al. (2015) found an improved RDM in rugby referees from the minutes 40–50 to 50–60, as well as findings from Larkin et al. (2014) indicate an improvement with increasing match period (quarters; Australian football). Descriptively, RDM in soccer referees is less accurate in the first 15 minutes of a match and became more accurate after 15 minutes (Mascarenhas et al., 2009). Regarding studies investigating the effects of *internal load* on RDM, one study did not reveal a relationship between RDM and physical exertion (blood lactate; Larkin et al., 2014). Similarly, Emmonds et al. (2015) and Mascarenhas et al. (2009) did not find a relationship between heart rate and RDM, whereas Gomez-Carmona and Pino-Ortega (2016) revealed that RDM errors occur especially above 95% of their maximum heart rate.

Overall, findings on the relationship between physical load and RDM are heterogeneous. A central issue is that previous studies mostly conducted ex-post video analyses and thus systematically controlled neither external nor internal load (Bloß et al., 2020; MacMahon et al., 2015). Also, confounding variables such as psychological load (e.g. crowd noise; Balmer et al., 2007; e.g. rumination; Poolton, Siu, & Masters, 2011) or environmental conditions (e.g. temperature; Watkins et al., 2014) were not

Published online: 21 January 2022

Table 1 Characteristics of the top-class referees who participated in study 1 and study 2

Study	Referees	Age (M ± SD)	Officiating experience (M ± SD)
Study 1	*N* = 87	31.6 ± 6.5 years	11.8 ± 3.9 years
	n = 13 females	30.3 ± 4.5 years	12.7 ± 2.6 years
	n = 74 males	31.9 ± 6.6 years	11.6 ± 4.2 years
Study 2	*N* = 83	32.2 ± 6.3 years	15.7 ± 5.6 years
	n = 14 females	30.5 ± 4.3 years	14.2 ± 2.7 years
	n = 69 males	32.5 ± 6.6 years	16.0 ± 6.0 years
Study 1 and 2	*N* = 59 *n* = 3 females *n* = 56 males	31.6 ± 6.3 years 29.3 ± 4.0 years 31.7 ± 6.4 years	14.9 ± 5.4 years 13.7 ± 3.5 years 15.0 ± 5.5 years

SD standard deviation, *M* mean

controlled in previous research. Consequently, subsequent studies should conduct laboratory studies in which potential confounding variables can be controlled. This approach, in contrast to much of the research to date, would provide the added value of being able to control the internal load as the individual response to an external load (Bloß et al., 2020; Impellizzeri et al., 2019). It is important to note that these studies must be as externally valid as possible, meaning that the task and exercise are representative to examine the effects of physical load on RDM (Bloß et al., 2020; Hancock, Bennett, Roaten, Chapman, & Stanley, 2021).

Furthermore, in the past, RDM has been investigated within theoretical frameworks such as the social information processing model (e.g. see Plessner & Haar, 2006). However, recent research suggests that RDM should be examined under more realistic or representative conditions, respectively, following the perspective of naturalistic decision-making (NDM) and in consequence under physical load for instance (Kittel, Cunningham, Larkin, Hawkey, & Rix-Lièvre, 2021; Mascarenhas et al., 2009; Mascarenhas, Collins, Mortimer, & Morris, 2005; Mascarenhas et al., 2005). NDM is a framework aiming to investigate decision-making in real-world settings likewise in experiments that approximate real world conditions as close as possible. Characteristics of NDM are time pressure, high risks, multiple players, uncertain and dynamic environments, ill-structured problems, shifting or competing goals and action/feedback loops (Orasanu & Connoly, 1993; Zsambok, 1997); however, not all characteristics need to be included for a study being labelled 'naturalistic' (Mascarenhas et al., 2005). If conducting a study oriented towards naturalistic criteria, studies are encouraged to use representative tasks meaning that studies researching RDM in sports games must as best as possible simulate decision-making situations (Mascarenhas et al., 2009). Researchers must therefore not only generate representative tasks for the decision-making, but, for instance, also investigate RDM under physical load. From the NDM perspective, however, studies must also consider how the decision-making of the interest decision-maker—in the case of this study a handball referee—is composed (Zsambok, 1997). Hence, in accordance with the international rule book (International Handball Federation, 2016), a handball referee has to decide if she/he has to whistle or not so as to call a foul or not. Next, it is important to understand the reasoning of the decision, i.e. if a referee calls a foul, there must be a reason for this decision ('why whistling'; Mascarenhas, Collins, & Mortimer, 2005a). The knowledge about the reasoning supports the referees' situation understanding and may enables to anticipate a critical situation before it occurs (Mascarenhas et al., 2005). In handball, the underlying reason for a foul decision is a specific rule violation, which can be differentiated in several type of fouls (e.g. clinging or grab in throwing arm) and which are shown by the referee via hand signals during a match. In addition, in handball, the punishment (e.g. yellow card with progressive 2-minutes, 2-minutes) is reasoned by the type of foul (International Handball Federation, 2016). Hence, collectively, RDM in handball referees is to be differentiated in deciding about calling a foul or not likewise to whistle or not (decision) as well as into correctly determining the type of foul (reasoning 1), as the punishment (reasoning 2) is reasoned on the type of foul. The necessity to correctly determining both the decision and the reasonings fits with the suggestion that referees must be able to avoid categorisation errors to not 'over-punish' or 'under-punish' players (MacMahon et al., 2015). Hence, to systematically rework evidence on the relationship between physical load and handball referees' decision-making with naturalistic criteria, research needs to consider both the referees' decisions and both reasonings.

Here, in two studies we aimed to examine the effects of physical load on RDM, i.e. referees' decisions (calling a foul or not) and reasonings (type of foul and punishment), in top-class handball referees administering external valid tasks for physical load and RDM in a NDM criteria-oriented approach (Bloß et al., 2020; Hancock et al., 2021). Considering recent findings, we hypothesised that physical load affects RDM. Due to the ambiguous evidence about the effects of physical load on referees' decision-making and the lack of research on referees' reasonings, however, we could not formulate well-grounded separate hypotheses regarding the effects of physical load on referees' decisions and reasonings, respectively, and we also refrained from making any directional predictions.

Methods

Participants

In both studies, top-class referees from the German Handball Federation (Deutscher Handballbund e. V. [DHB]) participated (see ◘ **Table 1** for group characteristics). A total of *n* = 59 referees participated in both study 1 and 2.

All participants were informed about the purpose of the studies in advance and in accordance with the Declaration of Helsinki (2013). Participants gave written informed consent prior to testing and were allowed to withdraw from testing at any time. Ethical approval for both studies was obtained from the local commission for research impact assessment and ethics.

Apparatus and stimuli

Video-based decision-making test

To measure the dependent variables of referees' correct decisions and reasonings, a video-based decision-making test (video test) was programmed with the software SR Research Experiment Builder (version 2.1.512, Ottawa, ON, Canada). Research indicates that video-based tasks can be representative in the context of NDM (Mascarenhas et al., 2005). Video tests were run on tablets (Lenovo IdeaPad MIIX 310, 10.1", Lenovo Group Limited, Hongkong, China). As test stimuli, we used video sequences from matches of the Liqui Moly Handball-Bundesliga (first division in Germany) viewed from the television camera perspective (i.e. side view). Participants were highly familiar with this perspective as it is used in the regular video-rule test of the DHB for several years. We prepared 353 video sequences, which were assessed by three independent experts of the referee board of the DHB prior to inclusion in the video test. Video sequences were played to the experts in the same way as they were played to the participants: the three experts had to make an instantaneous decision and reasonings on the sequence without receiving further information. We only integrated video sequences in the video test on which these experts agreed with regard to the following four criteria:

- Foul or no foul
- Type of foul (pushing, clinging, grab in throwing arm, defence inside goal area and offensive foul)
- Punishment (no personal punishment, yellow card, 2-minute penalty and red card)
- Game progress (throw-off, throw-in, goal clearance, free throw and 7 m)

Following this procedure, a total of 45 videos met the above criteria and these clips were integrated into the video test in a random order that was kept constant for all participants. Videos showed either a foul or no foul, clips were prepared using Adobe Premiere Pro CS6 (version 6, Adobe Inc., San José, CA, USA) and displayed at a resolution of 1280 × 720 pixel. Apparatus and stimuli preparation were identical in study 1 and study 2. However, the video sequences shown in study 1 were different to those shown in study 2.

Physical load — The Yo-Yo Intermittent Recovery Test

To expose participants to physical load (as the independent variable), we used the Yo-Yo Intermittent Recovery Test as a reliable and valid exhaustion test (YYT; level 1; Bangsbo, Iaia, & Krustrup, 2008; Krustrup et al., 2003). The test characteristics are close to real physical match demands in handball due to the shift between physical load and recovering periods from physical load.

The YYT began with an audio signal and participants were asked to run 20 m forth and back (40 m in total). Each 20 m run had to be completed within a specific time, which set the minimum required speed. During the entire YYT, participants were guided with audio signals that pointed to the start of a run and to the time of direction reversal at the turning line. After one run (i.e. 2 × 20 m), participants regenerated by walking around a pylon which was located 5 m behind the starting line. The regeneration phase lasted for 10s. Participants had to wait at the starting line and started the next run upon another audio signal. The required running speed increased between the runs as was indicated by the timing of the audio signal. The YYT consists of different stages with a specific number of runs. The test began with one run at 10 km/h in the first stage and increased to one run at 12 km/h in the second stage. In the third stage, the speed increased up to 13 km/h and participants had to complete two runs. In the fourth stage, the speed increased to 13.5 km/h and participants had to finish three runs. From the fifth stage on, the speed increased by another 0.5 km/h after every four runs.

Overall, this procedure ensured the assessment of the individual physiological response (i.e. internal load) to increasing external load induced by the YYT (Impellizzeri et al., 2019). It also ensured

Abstract

Ger J Exerc Sport Res
https://doi.org/10.1007/s12662-021-00794-8
© The Author(s) 2022

N. Bloß · J. Schorer · F. Loffing · D. Büsch

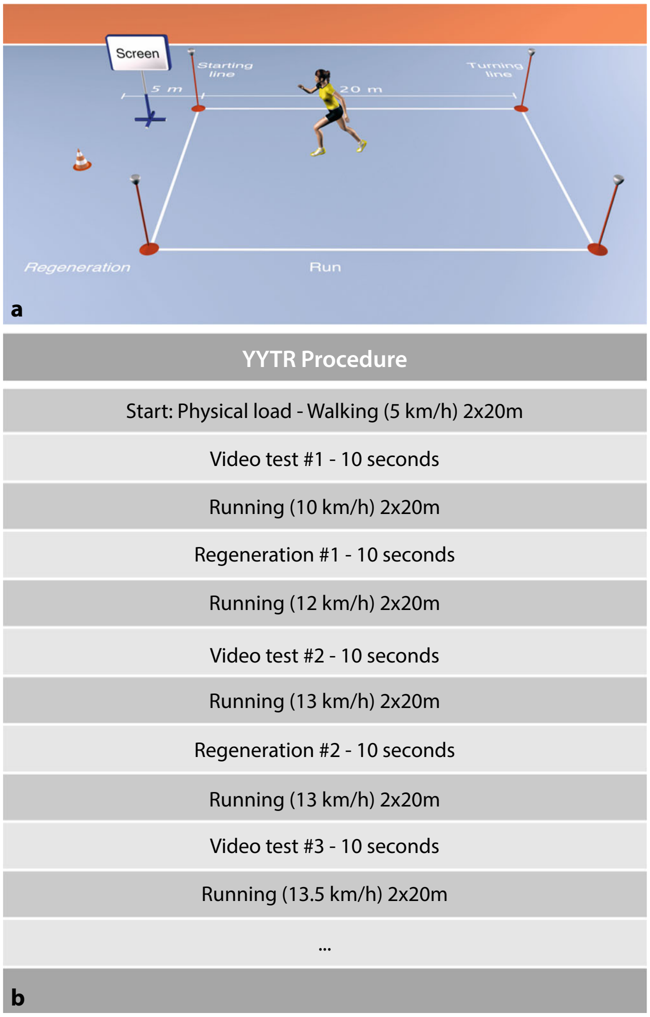
Decisions and reasonings of top-class handball referees under physical load

Abstract

Correct decision-making under physical load is a central challenge for referees in sports games. Handball referees are assumed to make both a decision (calling a foul or no foul) and to call its reasonings (type of foul, punishments). However, the impact of physical load on these two aspects has not been differentiated so far. Here, in two studies, we aimed to investigate the hypothesised impact of physical load on both referees' decisions and reasonings. To this end, $N = 66$ (study 1) and $N = 73$ (study 2) top-class handball referees performed the Yo-Yo Intermittent Recovery Test combined with a video-based decision-making test. Referees' decisions improved from initial to medium physical load and they deteriorated under maximal physical load in study 1, whereas in study 2 the quality of the decisions was constant across physical load conditions. The percent of correct reasonings decreased from initial to medium physical load in study 1, whereas the opposite pattern was found in study 2. In both studies, reasoning performance did not change from medium to maximal physical load. Moreover, referees demonstrated better endurance performance in study 2 than in study 1. Despite some methodological limitations (e.g. familiarisation with the experimental setup in referees who participated in both studies), the present findings tentatively indicate that a well-trained endurance capacity may support referees' decision-making, i.e. to make correct decisions and reasonings. Specifically, enhanced endurance capacity may lead to lower subjectively perceived fatigue, resulting in larger cognitive capacities that may facilitate referees' decision-making.

Keywords

Umpire · Sports officials · Officiating · Sports games · Performance



YYTR Procedure

Start: Physical load - Walking (5 km/h) 2x20m

Video test #1 - 10 seconds

Running (10 km/h) 2x20m

Regeneration #1 - 10 seconds

Running (12 km/h) 2x20m

Video test #2 - 10 seconds

Running (13 km/h) 2x20m

Regeneration #2 - 10 seconds

Running (13 km/h) 2x20m

Video test #3 - 10 seconds

Running (13.5 km/h) 2x20m

...

Fig. 1 ◄ **a** Illustration of the Yo-Yo Test for Referees (YYTR) setup (see main text for details). **b** Procedure of alternating video tests and regeneration phases in the YYTR

that individual participants became physically exhausted, while we continuously checked whether they approached a potentially harmful region of overuse to be able to stop the experiment in time for these participants. Referees' heart rate was monitored and measured through live feed (Polar Team Pro System) on an Apple iPad Pro (12.9", Apple Inc., Cupertino, CA, USA).

Procedure — The Yo-Yo Test for Referees

To investigate the effects of physical load on referees' decisions and reasonings, we combined the YYT with the video test (Yo-Yo Test for Referees [YYTR]). We added an initial stage of 5 km/h, so participants walked the 20 m distance back and forth, as we wanted our participants to become accustomed to the adaption of the regeneration phase: instead of walking around a pylon, participants individually conducted one trial of the video test after every second run (80 m). Overall, in the course of the YYTR, participants were asked to run up until their individual maximal physical capacity. The experimental setup is illustrated in **Fig. 1**.

Participants were divided into groups of eight for data collection. The groups conducted the YYTR in sequence in a gym during the referee preparation and half-season training courses of the DHB. To ensure that the referees could not discuss video sequences with the subsequent groups, later tested groups were not present in the gym during the testing of former groups. In addition, videos were played without sound and the referees shown in the videos were pixelated to avoid any information retrieval from these potential cues. Prior to starting the YTTR, referees were instructed verbally and via a written task description. After the instruction, referees were allowed to conduct a trial version of the video test (nine trials) to become accustomed to the video test procedure before the YYTR started (i.e. not in the gym), but without receiving feedback on RDM correctness.

Each trial of the video test began with a fingertip on a tablet's display to start the presentation of a 4 s video. Upon the end of a video, referees had to make decision and reasoning calls within about 5 s to be ready for the continuation of running afterward. The decision-making procedure differed slightly between study 1 and study 2. In study 1, upon video end, a large decision and reasoning matrix was displayed (**Fig. 2a**) in which referees had to indicate, with one single fingertip on an appropriate cell, whether a foul occurred or not (referees' decision) as well the type of foul and the punishment (reasonings 1 and 2). In a second step, participants rated their confidence in the decision and reasonings on a six-point scale from very certain (1) to very uncertain (6). Ratings were included to examine whether participants become less confident with increased physical load, for example, due to higher perceived stress (Enoka & Duchateau, 2016).

In study 2, the decision-making procedure, i.e. the decision and reasoning matrix, in the video test was modified by separating the foul (decision) and type of foul (reasoning 1) from the punishment (reasoning 2; **Fig. 2b**). That modification was done for three reasons. First, it leads to a better differentiation of referees' decisions and reasonings in the video test. Second, the DHB referees who participated in study 1 reported that the decision-making felt unusual as in real-court refereeing there would be a clearer differentiation between the referees' decisions (calling a foul or not) and the reasonings (type of foul, punishment). Third, confidence ratings were excluded as in study 1 they turned out to be of no added value (see Appendix, **Fig. 5**). We thus rather focused on improving the video test and its representativeness via the mentioned further development of the decision-making procedure.

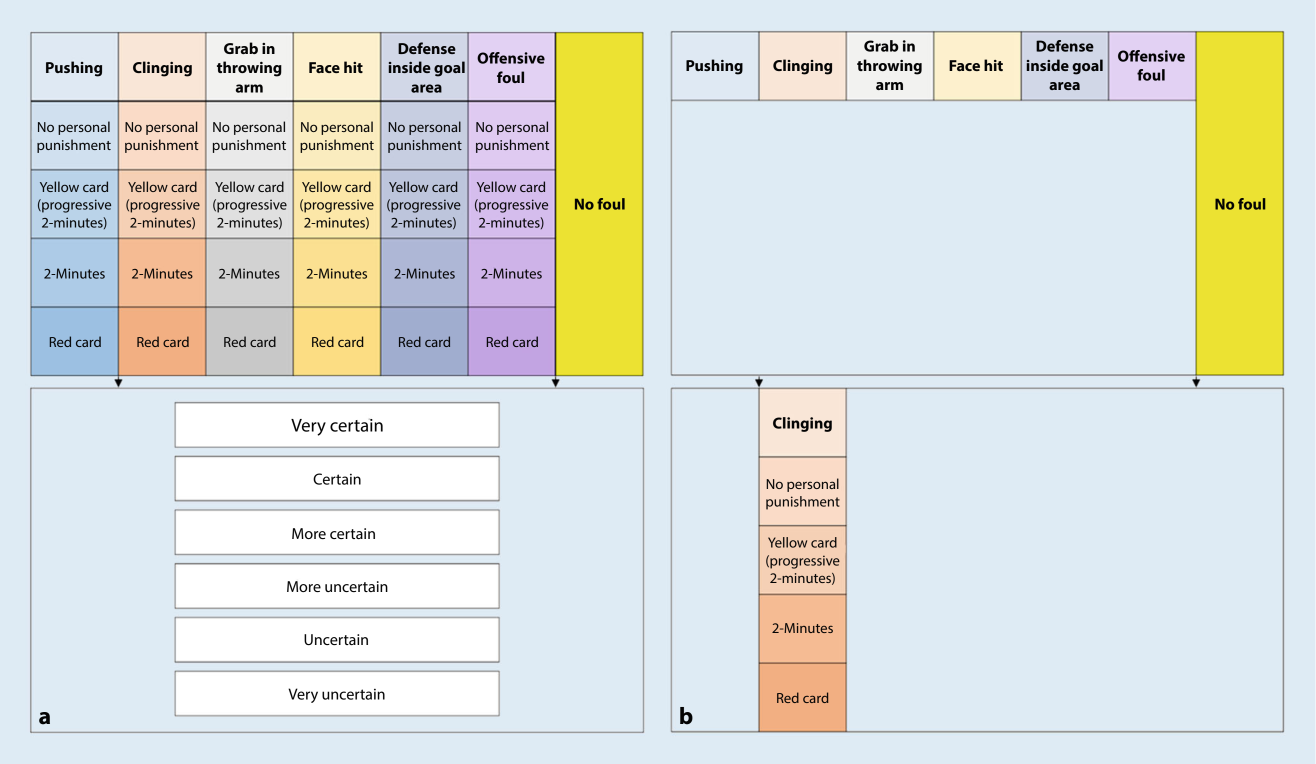


Fig. 2 ▲ Decision and reasoning matrix in the video test of study 1: Video test procedure (**a**) and study 2: Video test procedure (**b**)

The construct validity of the video test realised in study 2 was checked by testing $n = 86$ top-class (first to third division) and $n = 48$ advanced referees (fourth division) under rest. Analysis of the number of correct decisions revealed that the test is able to distinguish between expertise groups with top-class referees outperforming advanced referees ($t[132] = 2.4$, $p = 0.02$, $d = 0.43$, 95% confidence interval [0.10, 0.80]).

Statistical analyses

Heart rate

As a control of the internal load, we measured the participants' heart rate during the entire YYTR. We then calculated the mean individual heart rate based on the raw data for each trial a participant conducted the video test. A customised MatLab script was used to determine the time intervals in which participants conducted a video test trial. As a manipulation check and statistical requirement for the analyses of the decision and reasoning performance, we tested whether the heart rates of study 1 and study 2 were on a comparable level. Also, we tested whether the relative heart rate increased under different physical load conditions (independent variable). To this end, we performed a composite variance analysis with repeated measures (RM-ANOVA) on the factors *study* (levels: study 1 and study 2) and *physical load*. The latter factor had three levels: (1) initial external load, (2) medium external load and (3) maximal external load, which were associated with moderate, submaximal and maximal internal load, respectively.

We composed three blocks corresponding to these levels, each containing four trials of the video test. The first block comprised the first four trials, which counts for every participant. The trials belonging to the second and third block, however, were determined individually. The second block contained an individual's middle four trials, whereas the third block comprised an individual's last four trials. For example, for a participant who completed 12 trials, block 1 (initial external load) included video trials 1–4, block 2 (medium external load) included trials 5–8 and block 3 (maximal external load) was made up of trials 9–12. In another example, a participant completed 19 trials. Since there was an odd number of trials, the middle block—we set this as a general rule in our analyses—was shifted in the direction of the starting block. This means that, for the particular participant who completed 19 trials, block 1 again consisted of trials 1–4, block 2 was made up of trials 8–11 and block 3 comprised trials 16–19. By using such individualised assignments to the three levels of physical load, we aimed to ensure the comparability between participants who differ in their individual physical capacity.

The alpha level was set to $\alpha = 0.05$. However, as we used the same data set for two analyses (decision and reasoning analyses), we used the Bonferroni correction and thus the alpha level was set to $\alpha' = 0.025$. If the assumption of sphericity was not met, we report either the Greenhouse–Geisser epsilon (ε_{GG}) if $\varepsilon < 0.75$ or the Huynh-Feld epsilon (ε_{HF}) if $\varepsilon > 0.75$ (Atkinson, 2001). Effect sizes for the RM-ANOVA are reported as partial eta squared (η_p^2) along with 90% confidence

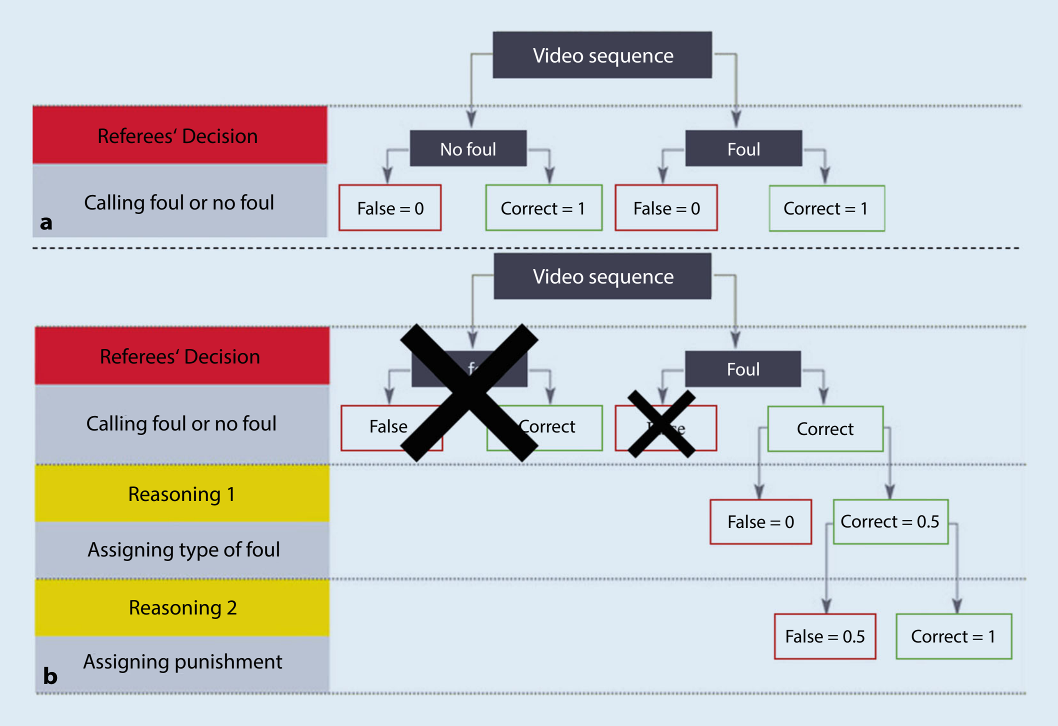


Fig. 3 ◀ Composition and coding of referees' (**a**) decisions and (**b**) reasonings

intervals (CI). Confidence intervals for η_p^2 were calculated using the SPSS syntax by Smithson (2001), downloaded from the website of Wuensch (2017). We used SPSS (version 27, IBM, Armonk, NY, USA) for the statistical analysis, MatLab (version 2019a) for the determination of the heart rate for the single video test trials. We calculated sensitivity analyses via G*Power (version 3.1.9.7, Faul, Erdfelder, Lang, & Buchner et al., 2007) for both studies which will be reported and discussed in the discussion. By doing this, we determine the minimum effect size to which the studies are sensitive by given $\alpha' = 0.025$, $1\text{-}\beta = 0.90$ and $N = 66$ (study 1) likewise $N = 73$ (study 2).

Referees' decisions

For the statistical analysis of correct referees' decisions, the latter were either coded as correct (1) or false (0) for each trial of the video test conducted (◘ **Fig. 3a**). For each of the three levels of the factor *physical load*, a block was formed by the completed trials. We calculated the mean value of correct decisions for each block based on four trials. The formation of the blocks was meant to increase the reliability of the analysis, as it reduced potential interference effects, e.g. of the video sequences to be assessed. The composition of the blocks was equal to the procedure described for the *heart rate analysis* (i.e., first block comprised trials 1–4 for each participant, individually determined second and third block). Hence, only participants with 12 or more completed trials could be considered for analysis. This reduced the sample size in study 1 to $n = 66$ participants (age: $M_{age} = 31.3$ years, $SD = 3.2$ years; officiating experience: $M_{exp} = 11.8$ years, $SD = 3.9$ years) and in study 2 to $n = 73$ participants ($M_{age} = 32.1$ years, $SD = 6.4$ years; $M_{exp} = 15.6$ years, $SD = 5.6$ years). The proportion of correct decisions was subjected to a one-factorial RM-ANOVA with the factor *physical load*. Alpha-level and calculation of effect sizes and the 90% CIs were the same as for the *heart rate analysis*.

Referees' reasonings

To be suited for inclusion in the reasoning analysis, participants must have made correct foul decisions in trials in which a foul was shown. Hence, all trials showing no foul situations as well as false foul decisions were excluded from the analysis (◘ **Fig. 3b**). As a consequence, the composition of the blocks differs from the decision analysis: the first block contained the individual first four trials in which a participant correctly decided that there was a foul. The second block contained the individual middle four trials and the third block the last four trials in which a participant correctly awarded a foul. For the analysis of the referees' reasonings, we used a point system: if the referees correctly awarded a foul in a video test trial, we checked whether they assigned the foul to the correct type of foul. If so, referees received 0.5 points. If the foul was not correctly assigned into the type of foul, they were awarded 0 points (◘ **Fig. 3b**). If referees chose the correct punishment, they received further 0.5 points, so they could receive a maximum of 1 point for the reasonings. The codification of the reasoning analysis is shown in ◘ **Fig. 3b**. Due to the coding procedure, only participants with 12 correct foul decisions or more could be included in the analysis. As a consequence, the sample size reduced to $n = 13$ participants ($M_{age} = 30.0$ years, $SD = 5.1$ years; $M_{exp} = 10.0$ years, $SD = 3.1$ years) in study 1 and to $n = 20$ participants ($M_{age} = 31.7$ years, $SD = 6.4$ years; $M_{exp} = 15.2$ years, $SD = 5.7$ years) in study 2, respectively.

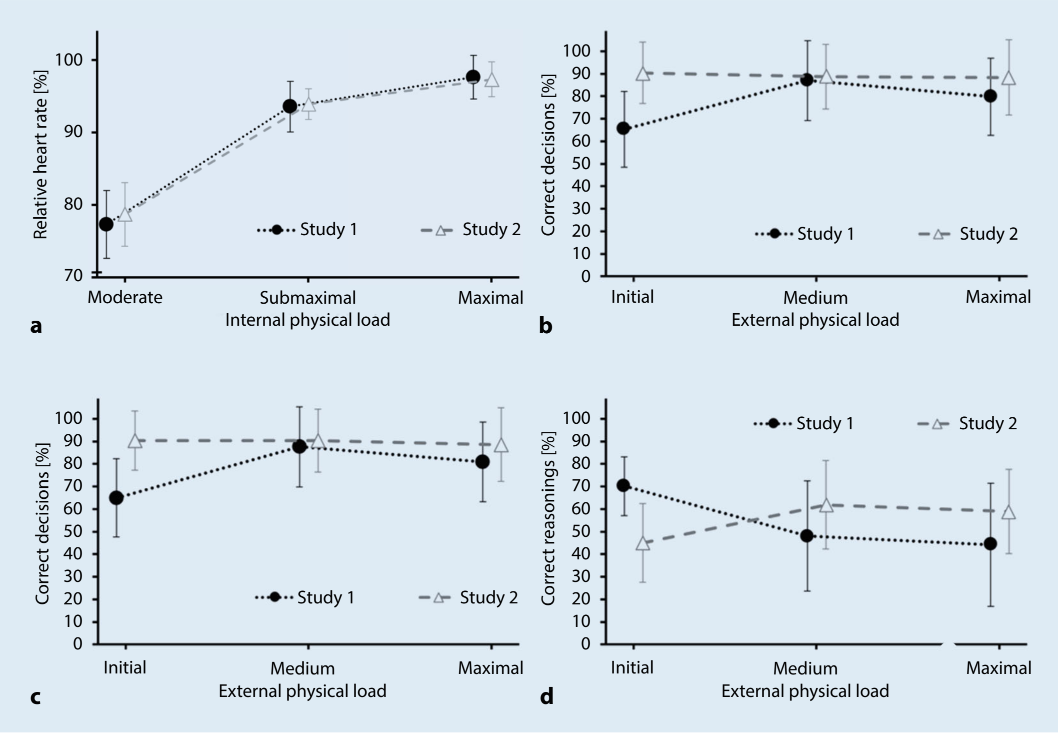


Fig. 4 ◀ Mean relative heart rates (**a**, % maximum). Mean correctness of referees' decisions (**b**, % correctness). **c** The mean correctness of decisions from referees who participated in both studies. **d** The mean correctness of referees' reasoning. In all panels, error bars indicate standard deviation

We conducted a RM-ANOVA with the factor *physical* load to analyse the relationship between physical load and reasoning correctness. The stages as well as the statistical assumptions were identical to those reported in the section about the analysis of *referees' decisions*. For a clearer understanding of the results, we report the values of both decision and reasoning correctness as percentages.

Results

Internal load manipulation check—referees' heart rate

A composite RM-ANOVA did not reveal an interaction between the factors *study* and *physical load*, $F(2, 116) = 3.01$, $\varepsilon_{GG} = 0.747$, $p = 0.05$, $\eta_p^2 = 0.05$, 90% CI [0.12]; thus the change of the relative heart rates may be considered comparable between studies. There was a significant main effect for the factor *physical load*, $F(2, 116) = 1538.93$, $\varepsilon_{GG} = 0.66$, $p < 0.001$, $\eta_p^2 = 0.96$, 90% CI [0.95, 0.97]. In both studies, the referees' heart rate increased from moderate to submaximal internal load, $F(1, 58) = 2089.79$, $p < 0.001$, $\eta_p^2 = 0.97$, 90% CI [0.96,0.98], as well as from submaximal to maximal internal load, $F(1, 58) = 211.61$, $p < 0.001$, $\eta_p^2 = 0.80$, 90% CI [0.70, 0.83] (▣ **Fig. 4a**). Relative heart rates are shown in ▣ **Table 2**.

Referees' decisions—study 1

A RM-ANOVA revealed a significant main effect for the factor *physical load* $F(2, 130) = 25.28$, $p < 0.001$, $\eta_p^2 = 0.28$, 90% CI [0.17, 0.37] (▣ **Fig. 4b**). That main effect was followed up by a posteriori two-sided paired t-tests between the stages with Bonferroni-corrected alpha level $\alpha' = 0.017$. For these comparisons, effect sizes are reported as Cohens' *d* along with 95% confidence intervals (Cohen, 1988). Accordingly, referees made more correct decisions (decision correctness: $M_{Decision} = 87.0\%$, $SD = 18.0\%$) under medium physical load ($M_{HR} = 93.6\%$, $SD = 3.5\%$) than under initial physical load ($M_{Decision} = 65.0\%$, $SD = 17.0\%$; $M_{HR} = 77.3\%$, $SD = 4.6\%$), $t(65) = 6.67$, $p < 0.001$, $d = -0.82$, 95% CI [−1.10, −0.54]. Referees made fewer correct decisions under maximal ($M_{Decision} = 80.0\%$, $SD = 17.0\%$; $M_{HR} = 97.6\%$, $SD = 3.0\%$) than under medium physical load, $t(65) = 2.60$, $p = 0.01$, $d = 0.32$, 95% CI [0.07, 0.57]. Referees made more correct decisions under maximal than under initial physical load, $t(65) = 4.4$, $p < 0.001$, $d = -0.54$, 95% CI [−0.80, −0.28].

Based on the above results, we updated the initial undirected hypothesis for study 2. Specifically, we expected that referees would make more correct decisions under medium than under initial physical load and that they would make more correct decisions under medium than under maximal physical load. Hence, in the following analysis of data from study 2, we did not perform adjusted t-tests but used planned contrast analyses obtained from the RM-ANOVA.

Referees' decisions—study 2

A RM-ANOVA revealed no significant main effect for the factor *physical load*, $F(2, 144) = 0.38$, $p = 0.69$, $\eta_p^2 = 0.01$, 90% CI [0, 0.03]. As is shown in ▣ **Fig. 4b**, referees did not make more correct decisions under medium ($M_{Decision} = 89.0\%$, $SD = 14.0\%$; $M_{HR} = 93.9\%$, $SD = 2.1\%$) than under initial physical load ($M_{Decision} = 90.0\%$, $SD = 14.0\%$; $M_{HR} = 78.7\%$, $SD = 4.4\%$), $F(1, 72) = 0.47$, $p = 0.50$, $\eta_p^2 = 0.01$, 90% CI [0, 0.07]. Also, referees did not make fewer correct decisions under

Table 2 Relative heart rate (%) during referees' decisions and reasonings

Study	Moderate internal load		Submaximal internal load		Maximal internal load	
	M	SD	M	SD	M	SD
Study 1	77.3	4.7	93.6	3.5	97.6	3.0
Study 2	78.1	4.0	93.6	2.3	97.1	2.7

SD standard deviation, *M* mean

maximal ($M_{Decision} = 88.0\%$, $SD = 17.0\%$; $M_{HR} = 97.3\%$, $SD = 2.4\%$) than under medium physical load, $F(1, 72) = 0.02$, $p = 0.90$, $\eta_p^2 < 0.001$, 90% CI [0, 0.02].

Referees' decision performance—differences between studies

In addition to the study-wise analyses of correct decisions, given that 59 referees took part in both studies, we additionally run a composite RM-ANOVA with *study* as another within-subject factor (levels: study 1, study 2). We used the above-described blocks of the decision analysis. Statistical assumptions correspond to those described for the *decision analysis*. This analysis was intended to compare performance differences between the studies.

As illustrated in **Fig. 4c**, there was a significant interaction between the factors *physical load* and *study*, $F(2, 116) = 13.20$, $p < 0.001$, $\eta_p^2 = 0.19$, 90% CI [0.08, 0.28]. There was an interaction between the factors when restricted to initial and medium physical load, $F(1, 58) = 19.45$, $p < 0.001$, $\eta_p^2 = 0.25$, 90% CI [0.10, 0.39], meaning that referees performed similar under both conditions in study 2, whereas their performance increased from initial to medium load in the study 1 (**Fig. 4c**). There was no two-way interaction when restricted to medium and maximal physical load, $F(1, 58) = 1.27$, $p = 0.27$, $\eta_p^2 = 0.02$, 90% CI [0, 0.11].

Referees' reasonings—study 1

A RM-ANOVA showed a significant main effect for the factor *physical load*, $F(2, 24) = 5.35$, $p = 0.01$, $\eta_p^2 = 0.31$, 90% CI [0.05, 0.47]. As illustrated in **Fig. 4d**, referees made fewer correct reasonings under medium (reasoning correctness: $M_{Reasoning} = 48.0\%$, $SD = 24.0\%$; $M_{HR} = 95.5\%$, $SD = 2.0\%$) than under initial physical load ($M_{Reasoning} = 70.0\%$, $SD = 13.0\%$; $M_{HR} = 82.0\%$, $SD = 3.6\%$), $t(12) = 2.64$, $p = 0.02$, $d = 0.73$, 95% CI [0.10, 1.34]. The proportion of correct reasoning only differed descriptively under maximal ($M_{Reasoning} = 44.0\%$, $SD = 27.0\%$; $M_{HR} = 98.2\%$, $SD = 1.7\%$) and medium physical load, $t(12) = 0.37$, $p = 0.72$, $d = 0.1$, 95% CI [−0.45, 0.64]. Referees made fewer correct reasonings under maximal than under initial physical load, $t(12) = 4.16$, $p = 0.001$, $d = 1.15$, 95% CI [0.43, 1.85].

Here as well, we updated the initial undirected hypothesis for study 2 as the correctness of reasoning decreased about 22% from initial to medium physical load. Hence, we expected that the quality of referees' reasonings would drop from initial to medium physical load in study 2. Also, we expected that referees' reasoning would not change, on average, from medium to maximal physical load. Given the now directed hypothesis, we did not perform adjusted t-tests but used planned contrast analyses of the RM-ANOVA in study 2.

Referees' reasonings—study 2

A RM-ANOVA showed a significant main effect for the factor *physical load*, $F(2, 38) = 5.64$, $p = 0.01$, $\eta_p^2 = 0.23$, 90% CI [0.04, 0.38] (**Fig. 4d**). Referees made more correct reasonings under medium ($M_{Reasoning} = 61.9\%$, $SD = 19.7\%$; $M_{HR} = 94.4\%$, $SD = 1.8\%$) than under initial physical load ($M_{Reasoning} = 45.0\%$, $SD = 17.4\%$; $M_{HR} = 80.1\%$, $SD = 3.7\%$), $F(1, 19) = 9.05$, $p = 0.01$, $\eta_p^2 = 0.32$, 90% CI [0.06, 0.52]. Further, referees made correct reasonings on a comparable level under medium and maximal physical load ($M_{Reasoning} = 59.0\%$, $SD = 18.7\%$; $M_{HR} = 96.9\%$, $SD = 3.2\%$), $F(1, 19) = 0.22$, $p = 0.65$, $\eta_p^2 = 0.01$, 90% CI [0, 0.17].

Due to only four referees who took part in both studies and who were eligible for inclusion in the reasoning analysis, we did not calculate a composite RM-ANOVA to compare the reasoning performance differences between the studies.

Discussion

Our studies aimed to examine the hypothesized effects of physical load on top-class handball referees' decisions and reasonings. Although the results of study 1 indicate that physical load affects referees' decisions, we did not find evidence for an effect in study 2. Furthermore, results of both studies point out an effect of physical load on referees' reasonings.

In study 1, referees' decision correctness improved from initial to medium physical load, which corroborates to previous studies indicating that the decision correctness improves under physical load (Emmonds et al., 2015; Larkin et al., 2014; Mascarenhas et al., 2009). Moreover, the decision correctness decreased under maximal physical load. This complies with previous research indicating that the correctness decreases under maximal physical load, respectively at the end of a match (Ahmed et al., 2017; Elsworthy et al., 2014; Emmonds et al., 2015; Gomez-Carmona & Pino-Ortega, 2016; Mallo et al., 2012; Oudejans et al., 2005; Samuel et al., 2019). The decrease also fits with recent studies pointing out an impairment of cognitive processes under maximal physical load (Schmidt et al., 2019). Furthermore, in study 1, the decision correctness was lowest at the beginning of the test, which corresponds to previous research showing that referees were less accurate at the beginning of a match and that they made more correct decisions during the match (Emmonds et al., 2015; Larkin et al., 2014; Mascarenhas et al., 2009). The latter may be in line with the calibration effect: in early stages of a match, referees observe foul and no foul situations to develop an internal rating scale for a match before making strict decisions (e.g. awarding a yellow card to a foul; Unkelbach & Memmert, 2008). Hence, in the video test referees may have adopted a similar 'cautious' approach in the beginning. Even though decision correctness decreased under maximal physical load, it was still on a higher level than under initial physical load. In study 2, however, decision correctness remained constant across physical load levels. Besides methodological limitations which will be discussed in detail below, one potential explanation for the diverging results could be the improved endurance performance of the participants. A comparison of the mean running distances of participants who took part in both studies revealed that endurance performance improved from $M = 1.290$ meters ($SD = 299$ meters; running time: $M = 11{:}10$ min, $SD = 2{:}23$ min) in study 1 to $M = 1.633$ meters ($SD = 340$ meters; running time: $M = 14{:}00$ min, $SD = 02{:}40$ min) in study 2. The improvement in running distance reflects an enhanced endurance performance, which could, in turn, have led to a differently subjectively perceived fatigue (Enoka & Duchateau, 2016). As previous research indicates that a higher endurance capacity is associated with higher attention (de Sousa et al., 2019) likewise that a higher physical fitness level is associated with better cognitive performance (Luque-Casado, Zabala, Morales, Mateo-March, & Sanabria, 2013), referees may have been more aware and concentrated in study 2 (Morillo et al., 2017; Schmidt et al., 2019), which, ultimately, might have resulted in more correct decisions under different physical load conditions in relation with a familiarisation with the YYTR. Furthermore, sensitivity analyses revealed a smallest effect size to which both studies are sensitive of $\eta_p^2 = 0.10$ (calculations revealed similar values for both studies). Hence, from a statistical point of view, effects of the decision analyses are reliably detected (Perugini, Gallucci, & Costantini, 2018).

With regard to the effects of physical load on the referees' reasonings, the correctness of reasonings significantly decreased in study 1 from initial to medium physical load. Results are in line with previous research indicating that specific processes could be negatively affected by physical load (e.g. information processing efficacy; Schmidt et al., 2019; Tomporowski, 2003). In contrast, reasoning correctness increased particularly under medium physical load in study 2. At this point, while not conclusive, a possible explanation as well could be that because participants had a better endurance capacity in study 2, they might have been more effective in the specific processing under medium to maximal physical load conditions (Helsen et al., 2019; Morillo et al., 2017). However, results of our studies concerning the reasonings should be treated with caution due to the small number of participants and thus results may therefore only indicate initial tendencies that have to be further investigated.

Moreover, referees made decisions on a higher level compared to reasonings. On the one hand, participating referees were accustomed to make decisions and reasonings on video sequences shown from the television camera perspective through the official video-rule test of the DHB. On the other hand, potentially relevant information may not be visible via that perspective, but from an on-field position/perspective instead. The camera perspective could be a limiting factor especially for making correct reasonings, since relevant detailed information (e.g. body parts) may not have been visible, which would have reduced the representativeness of the video test (Kittel et al., 2021). Future studies might therefore consider using video sequences recorded from an on-field perspective (even though referees were tested under rest, e.g. see Spitz, Put, Wagemans, Williams, & Helsen, 2018).

Furthermore, the cornerstone model from Mascarenhas et al. (2005b) provides other factors, like psychological components, that are relevant for a referees' performance. Thus, the cornerstone model can serve as a basis for examining the relationships between the constituent performance characteristics. In this context, the presented YYTR has the potential to integrate further loading factors, next to physical load, such as psychological factors (e.g. see Gillué, Laloux, Alvarez, & Feliu, 2018; Hill, Matthews, & Senior, 2016; Poolton et al., 2011), to more closely approximate real match demands and to enhance external validity, which is in line with the NDM framework. Moreover, reflecting that RDM correctness in our studies corresponds with previous studies conducting video analyses (e.g. Mallo et al., 2012) or controlled laboratory studies (e.g. Samuel et al., 2019), the YYTR with the applied video test seems to be a valid tool to measure RDM under physical load and is thus a further step forward toward controlled analyses of RDM following naturalistic criteria. In addition, the video test used in the YYTR might at least currently also be a useful tool to investigate expertise differences in handball referees' decision-making as well as turn out beneficial to train RDM under more natural conditions (e.g. see Kittel et al., 2021; Mascarenhas et al., 2009).

However, our research is not without limitations. Discussions with the involved referees and officials from the referee board of the DHB lead to the presumption that referees would make decisions and reasonings in real matches more differentiated than illustrated in the video test, i.e. the video test would need a better distinction between a referees' decision and the two reasonings. Furthermore, as the YYT(R) is an exhaustion test, referees rapidly became physically exhausted (see heart rate in **Fig. 4a**). However, in real matches, handball referees officiate on a mean heart rate of about 80% (range 72–87%; da Silva et al., 2010; Pabst et al., 2012). Thus, the YYTR needs adaptations to generate more trials in the heart rate range of real handball matches. Both representing the differentiation between referees' decisions and reasonings as well as the adaption of the YYTR protocol may help further improve the external validity and the representativeness of the YYTR.

Concerning the interpretation of our results, first, even though we used different video sequences in study 2, we presumably simplified the video test procedure due to splitting the decision and reasoning matrix (**Fig. 2**). As we tested a large amount of the same referees in both studies, the simplification of the decision and reasoning matrix might have helped the referees to make more correct decisions right from the beginning. Second, even though referees were familiar with the decision-making about video sequences from the television camera perspective through the official video rule test of the DHB, referees were unaccustomed to the experimental setting (YYTR). To let referees become more accustomed to the YYTR, subsequent studies might integrate trials at the beginning of the YYTR, e.g. by letting referees conduct video test trials under rest and in more walking phases. The improved RDM in study 2 compared to study 1 might partly depend on a familiarisation effect. Third, study 2 took place at the half-season training courses and thus 6 months after study 1, which took place at the season preparation courses. Hence, referees had a long-lasting recovery phase before study 1 in which referees presumably did not officiate and tried to physically recover as well. As a result, referees might have benefitted from the deliberate officiating during the season at the half-season training courses (i.e., study 2) compared to study 1.

Furthermore, by individually assigning the video test trials to blocks of physical load levels (i.e. initial, medium, maximal) we intended to analyse participants at a comparable physical load level and took their individual endurance capacity into account. However, by doing so we might have generated a potential artefact as different trials (i.e. videos of different fouls, types of fouls, punishments and potentially different item difficulty) were considered in the respective blocks. It is, therefore, necessary to not only treat the present results with caution, but to also consider the above potential limitations and artefacts in future research. Here, the importance of running replication studies is highlighted, as both the increase in correct decisions from initial to medium physical load and the decrease of correct decisions under maximal physical load found in study 1 were not confirmed (but different videos were used in study 2) and should thus not be overestimated.

Practical implications

Concerning the practical relevance of the referees' decision results in study 1, referees made about one wrong decision out of four decisions under maximal physical load, which indicates 25% wrong decisions. In study 2, referees only made 0.5 wrong decisions out of four decisions under maximal physical load, indicating 12.5% wrong decisions.

Since the level of correct decisions is higher compared to reasonings, referees should specifically include cognitive tasks, ideally reasoning tasks, in their individual training programs to improve their ability of making correct reasonings. To further support referees in practice, federations have to ensure that referees have the opportunity to achieve good physical fitness, e.g. due to specific training programs. Monitoring tools like sports watches could support referees in their individual training (Climstein et al., 2020). Based on the study by da Silva et al. (2010) in which an estimated maximal oxygen consumption of 49.5 l/min was determined for referees during real handball matches, a first recommendation of a minimum requirement for a handball referees' endurance capacity appears to be the YYT-level 17.4 (Bangsbo et al., 2008). That level approximately corresponds to the value provided by da Silva et al. (2010), which should already be reached in the preparation training courses. This would ensure that referees are capable to withstand the physical demands of a match. In addition, a combination of established endurance tests such as the YYT with video-based decision-making tests may be a useful approach to test RDM under varying load conditions. The Yo-Yo Test for Referees presented here combines both an established endurance test with a video-based decision-making test and is thus an example of a diagnostic tool to analyse top-class handball referees' decision-making performance under physical load. The Yo-Yo Test for Referees might also be recommended for others sports in which referees are required to make decisions and reasonings on a top-level under physical load (as another example for soccer referees see Samuel et al., 2019).

Corresponding address

Nicolas Bloß
Institute of Sport Science,
Carl von Ossietzky University
Oldenburg
Ammerländer Heerstr.
114–118, 26129 Oldenburg,
Germany
nicolas.bloss1@uni-oldenburg.de

Acknowledgements. All authors would like to thank the German Handball Federation for supporting our research. We especially want to thank all participating referees as well as Wolfgang Jamelle, Nils Szuka and Kay Holm. We further would like to thank Sarah Schulte, Kim Huesmann, Jelto Witt and Hannes Wünsch for the support in data collection.

Funding. The work was funded by the Federal Institute of Sport Science based on a resolution of the German Bundestag (official file number: ZMVI4-071001/19/20).

Funding. Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest. N. Bloß, J. Schorer, F. Loffing and D. Büsch declare that they have no competing interests.

All studies performed were in accordance with the ethical standards indicated in each case. Ethical approval for both studies was obtained from the local commission for research impact assessment and ethics (EK/2019/093).

Appendix

Results of the decision and reasoning confidence query in study 1

As reported in the section "Procedure—The Yo-Yo Test for Referees" in

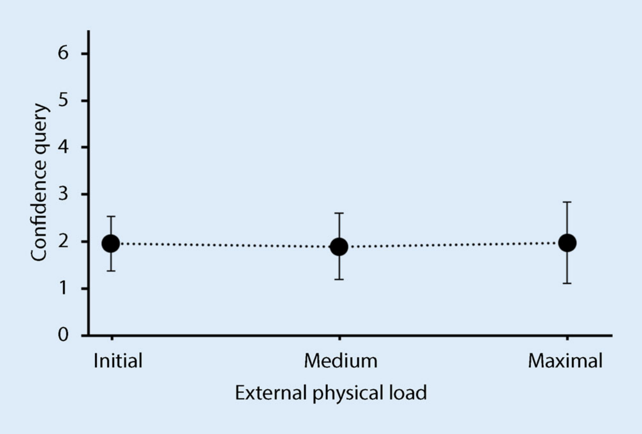


Fig. 5 ◀ Results of the decision and reasoning confidence query (with standard deviation) in study 1. Scaling: (1) very certain, (2) certain, (3) rather certain, (4) rather uncertain, (5) uncertain and (6) very uncertain

the main text, confidence queries did not turn out of added value. As illustrated in **Fig. 5**, referees nearly always decided that they were *certain* about their decisions and reasonings across physical load conditions. A potential explanation could be that referees were focused to make correct decisions and reasonings about the video test and they thus spend more time on the decision and reasoning leading to time pressure when rating the confidence. A second reason could be that top-class referees have such strong self-confidence, as they have to manage the game adequately, even if the decision or reasoning is wrong.

The confidence query was not included in study 2, since we aimed to increase the representativeness of the video test. However, we still recommend to consider confidence queries in replication studies to investigate its potential added value to understanding the processes underlying referees' decision-making.

References

Ahmed, H., Davison, G., & Dixon, D. (2017). Analysis of activity patterns, physiological demands and decision-making performance of elite futsal referees during matches. *International Journal of Performance Analysis in Sport*, *17*(5), 737–751. https://doi.org/10.1080/24748668.2017.1399321.

Atkinson, G. (2001). Analysis of repeated measurements in physical therapy research. *Physical Therapy in Sport*, *2*(4), 194–208. https://doi.org/10.1054/ptsp.2001.0071.

Balmer, N. J., Nevill, A. M., Lane, A. M., Ward, P., Williams, A. M., & Fairclough, S. H. (2007). Influence of crowd noise on soccer refereeing consistency in soccer. *Journal of Sport Behavior*, *30*(2), 130–145.

Bangsbo, J., Iaia, F. M., & Krustrup, P. (2008). The Yo-Yo intermittent recovery test: a useful tool for evaluation of physical performance in intermittent sports. *Sports Medicine*, *38*(1), 37–51. https://doi.org/10.2165/00007256-200838010-00004.

Belcic, I., Ruzic, L., & Marošević, A. (2020). Influence of functional abilities on the quality of refereeing in handball. *Baltic Journal of Health and Physical Activity*, *12*(3), 23–34. https://doi.org/10.29359/BJHPA.12.3.03.

Bilge, M. (2012). Game analysis of olympic, world and european championships in men's handball. *Journal of Human Kinetics*, *35*, 109–118. https://doi.org/10.2478/v10078-012-0084-7.

Bloß, N., Schorer, J., Loffing, F., & Büsch, D. (2020). Physical load and referees' decision-making in sports games: a scoping review. *Journal of Sport Science & Medicine*, *19*(1), 149–157.

Catteeuw, P., Gilis, B., Wagemans, J., & Helsen, W. (2010). Offside decision making of assistant referees in the English Premier League: impact of physical and perceptual-cognitive factors on match performance. *Journal of Sports Sciences*, *28*(5), 471–481. https://doi.org/10.1080/02640410903518184.

Chang, Y. K., Labban, J. D., Gapin, J. I., & Etnier, J. L. (2012). The effects of acute exercise on cognitive performance: a meta-analysis. *Brain Research*, *1453*, 87–101. https://doi.org/10.1016/j.brainres.2012.02.068.

Climstein, M., Alder, J., Brooker, A., Cartwright, E., Kemp-Smith, K., Simas, V., & Furness, J. (2020). Reliability of the Polar Vantage M sports watch when measuring heart rate at different treadmill exercise intensities. *Science Periodical on Research and Technology in Sport*, *8*(9), 1–13. https://doi.org/10.3390/sports8090117.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edn.). Routledge. https://doi.org/10.4324/9780203771587.

Elsworthy, N., Burke, D., Scott, B. R., Stevens, C. J., & Dascombe, B. J. (2014). Physical and decision-making demands of Australian football umpires during competitive matches. *Journal of Strength and Conditioning Research*, *28*(12), 3502–3507. https://doi.org/10.1519/jsc.0000000000000567.

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. *Journal of Strength and Conditioning Research*, *29*(12), 3367–3374. https://doi.org/10.1519/JSC.0000000000001002.

Enoka, R. M., & Duchateau, J. (2016). Translating fatigue to human performance. *Medicine and Science in Sports and Exercise*, *48*(11), 2228–2238. https://doi.org/10.1249/MSS.0000000000000929.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146.

Gaoua, N., de Oliveira, R., & Hunter, S. (2017). Perception, action, and cognition of football referees in extreme temperatures: impact on decision performance. *Frontiers in Psychology*, *8*, 1479. https://doi.org/10.3389/fpsyg.2017.01479.

Gillué, G. S., Laloux, Y. R., Alvarez, M. T., & Feliu, J. C. I. (2018). Sources of stress inside and outside the match in football referees. *Apunts Educacion Fisica y Deportes*, *132*, 22–31. https://doi.org/10.5672/apunts.2014-0983.es.(2018/2).132.02.

Gomez-Carmona, C. D., & Pino-Ortega, J. (2016). Kinematic and physiological analysis of the performance of the football referee and its relationship with decision making. *Journal of Human Sport and Exercise*, *11*(4), 397–414. https://doi.org/10.14198/jhse.2016.114.01.

Hancock, D., Bennett, S., Roaten, H., Chapman, K., & Stanley, C. (2021). An analysis of literature on sport officiating research. *Research Quarterly for Exercise and Sport*, *92*(4), 607–617. https://doi.org/10.1080/02701367.2020.1756198.

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. https://doi.org/10.1080/02640410310001641502.

Helsen, W., MacMahon, C., & Spitz, J. (2019). Decision making in match officials and judges. In A. M. Williams & R. C. Jackson (Eds.), *Anticipation and decision making in sport* (Vol. 1, pp. 250–266). Routledge. https://doi.org/10.4324/9781315146270.

Hill, D. M., Matthews, N., & Senior, R. (2016). The psychological characteristics of performance under pressure in professional rugby union referees. *The Sport Psychologist*, *30*(4), 376–387. https://doi.org/10.1123/tsp.2015-0109.

Impellizzeri, F. M., Marcora, S. M., & Coutts, A. J. (2019). Internal and external training load: 15 years on. *International Journal of Sports Physiology and Performance*, *14*(2), 270–273. https://doi.org/10.1123/ijspp.2018-0935.

International Handball Federation (2016). *IX. Rules of the game.*

Kittel, A., Cunningham, I., Larkin, P., Hawkey, M., & Rix-Lièvre, G. (2021). Decision-making training in sporting officials: past, present and future. *Psychology of Sport and Exercise*, *56*, 102003. https://doi.org/10.1016/j.psychsport.2021.102003.

Krustrup, P., Mohr, M., Amstrup, T., Rysgaard, T., Johansen, J., Steensberg, A., Pedersen, P. K., & Bangsbo, J. (2003). The Yo-Yo intermittent recovery test: physiological response, reliability, and validity. *Medicine & Science in Sports & Exercise*, *35*(4), 697–705. https://doi.org/10.1249/01.Mss.0000058441.94520.32.

Larkin, P., O'Brien, B., Mesagno, C., Berry, J., Harvey, J., & Spittle, M. (2014). Assessment of decision-making performance and in-game physical exertion of Australian football umpires. *Journal of Sports Sciences*, *32*(15), 1446–1453. https://doi.org/10.1080/02640414.2014.896998.

Luque-Casado, A., Zabala, M., Morales, E., Mateo-March, M., & Sanabria, D. (2013). Cognitive performance and heart rate variability: the influence of fitness level. *PloS One*, *8*(2), e56935. https://doi.org/10.1371/journal.pone.0056935.

MacMahon, C., Mascarenhas, D., Plessner, H., Pizzera, A., Oudejans, R., & Raab, M. (2015). *Sports officials and officiating: science and practice*. Routledge. https://doi.org/10.4324/9780203700525.

Mallo, J., Frutos, P. G., Juarez, 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. https://doi.org/10.1080/02640414.2012.711485.

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*(4), 364–379. https://doi.org/10.1123/tsp.19.4.364.

Mascarenhas, D., Collins, D., Mortimer, P., & Morris, R. L. (2005c). Training accurate and coherent decision making in rugby union referees. *The Sport Psychologist*, *19*(2), 131–147. https://doi.org/10.1123/tsp.19.2.131.

Mascarenhas, D., Button, C., O'Hare, D. O., & Dicks, M. (2009). Physical performance and decision making in association football referees: a naturalistic study. *The Open Sports Sciences Journal*, *2*(9), 1–9. https://doi.org/10.2174/1875399X00902010001.

Michalsik, L. B. (2018). On-court physical demands and physiological aspects in elite team handball. In L. Laver, P. Landreau, R. Seil & N. Popovic (Eds.), *Handball sports medicine* (pp. 15–33). Springer Medicine. https://doi.org/10.1007/978-3-662-55892-8_2.

Morillo, J. P., Reigal, R. E., Hernández-Mendo, A., Montaña, A., & Morales-Sánchez, V. (2017). Decision-making by handball referees: design of an ad hoc observation instrument and polar coordinate analysis. *Frontiers in Psychology*, *8*, 1842. https://doi.org/10.3389/fpsyg.2017.01842.

Orasanu, J., & Connoly, T. (1993). The reinvention of decision making. In G. Klein, J. Orasanu, R. Calderwood & C. E. Zsambok (Eds.), *Decision making in action: models and methods* (pp. 158–171). Ablex.

Oudejans, R. R. D., Bakker, F. C., Verheijen, R., Gerrits, J. C., Steinbrückner, M., & Beek, P. J. (2005). How position and motion of expert assistant referees in soccer relate to the quality or their offside judgements during actual match play. *International Journal of Sport Psychology*, *36*(1), 3–21.

Pabst, J., Büsch, D., Pöhler, C., Rauchfuss, P., Pfänder, J., & Sichelschmidt, P. (2012). Mehr als nur ein Pfiff [More than just a whistle]. *Handballschiedsrichter*, *5*(1), 8–11.

Paradis, K., Larkin, P., & O'Connor, D. (2015). The effects of physical exertion on decision-making performance of Australian football umpires. *Journal of Sports Sciences*, *34*(16), 1535–1541. https://doi.org/10.1080/02640414.2015.1122205.

Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology*, *31*(1), 1–23. https://doi.org/10.5334/irsp.181.

Plessner, H., & Haar, T. (2006). Sports performance judgments from a social cognitive perspective. *Psychology of Sport and Exercise*, *7*(6), 555–575. https://doi.org/10.1016/j.psychsport.2006.03.007.

Poolton, J., Siu, C., & Masters, R. (2011). The home team advantage gives football referees something to ruminate about. *International Journal of Sports Science & Coaching*, *6*(4), 545–552. https://doi.org/10.1260/1747-9541.6.4.545.

Samuel, R. D., Galily, Y., Guy, O., Sharoni, E., & Tenenbaum, G. (2019). A decision-making simulator for soccer referees. *International Journal of Sports Science & Coaching*, *14*(4), 480–489. https://doi.org/10.1177/1747954119858696.

Schmidt, S. L., Schmidt, G. J., Padilla, C. S., Simoes, E. N., Tolentino, J. C., Barroso, P. R., Narciso, J. H., Godoy, E. S., & Costa Filho, R. L. (2019). Decrease in attentional performance after repeated bouts of high intensity exercise in association-football referees and assistant referees. *Frontiers in Psychology*, *10*, 2014. https://doi.org/10.3389/fpsyg.2019.02014.

da Silva, J. F., Castagna, C., Carminatti, L. J., Foza, V., Guglielmo, L. G., & de Oliveira, F. R. (2010). Physiological demands of team-handball referees during games. *Journal of Strength and Conditioning Research*, *24*(7), 1960–1962. https://doi.org/10.1519/JSC.0b013e3181ddb019.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: the importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*(4), 605–632. https://doi.org/10.1177/00131640121971392.

de Sousa, A. F. M., Medeiros, A. R., Del Rosso, S., Stults-Kolehmainen, M., & Boullosa, D. A. (2019). The influence of exercise and physical fitness status on attention: a systematic review. *International Review of Sport and Exercise Psychology*, *12*(1), 202–234. https://doi.org/10.1080/1750984X.2018.1455889.

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*, 47–56. https://doi.org/10.1016/j.psychsport.2017.09.010.

Tomporowski, P. D. (2003). Effects of acute bouts of exercise on cognition. *Acta Psychologica*, *112*(3), 297–324. https://doi.org/10.1016/s0001-6918(02)00134-8.

Unkelbach, C., & Memmert, D. (2008). Game management, context effects, and calibration: the case of yellow cards in soccer. *Journal of Sport & Exercise Psychology*, *30*(1), 95–109. https://doi.org/10.1123/jsep.30.1.95.

Watkins, S., Castle, P., Mauger, A. R., Sculthorpe, N., Fitch, N., Aldous, J., Brewer, J., Midgley, A. W., & Taylor, L. (2014). The effect of different environmental conditions on the decision-making performance of soccer goal line officials. *Research in Sports Medicine*, *22*(4), 425–437. https://doi.org/10.1080/15438627.2014.948624.

Wuensch, K. L. (2017). SPSS programs (syntax). http://core.ecu.edu/psyc/wuenschk/SPSS/SPSS-Programs.htm. Accessed 29 June 2021.

Zsambok, C. E. (1997). Naturalistic decision making: Where are we now? In C. E. Zsambok & G. Klein (Eds.), *Naturalistic decision making* (pp. 3–16). Lawrence Erlbaum.