Intrinsic judgment error in men's championship world surf league: WSL 2021

Error de juicio intrínseco en el campeonato mundial masculino de surf: WSL 2021

Authors

Tony Meireles Santos [1,2]
Lucas Eduardo Rodrigues Santos [2]
Ítalo Vinicius [3]
Cayque Brietzke [4]
Lucas Camilo Pereira [5]
Paulo Henrique Melo [1]
Thaiene Camila Beltrão Moura [1]
Taddeo De Negri [6]
Hassan Mohamed Elsangedy [5]
Flávio Oliveira Pires [4]

[1] Federal University of Pernambuco, Pernambuco, Brazil.
[2] University of Pernambuco, Pernambuco, Brazil.
[3] University of São Paulo, São Paulo, Brazil.
[4], Federal University of São Paulo, São Paulo, Brazil.
[5] Federal University of Rio Grande do Norte, Natal, Rio Grande do Norte, Brazil.
[6] Digisport University Research School, University of Rennes 2, France.

Corresponding author:
Tony Meireles Santos
tony.meireles@ufpe.br

How to cite in APA

Santos, T. M., Rodrigues Santos, L. E., Vinicius, Ítalo, Brietzke, C., Pereira, L. C., Melo, P. H., Moura, T. C. B., De Negri, T., Elsangedy, H. M., & Pires, F. O. (2025). Error de juicio intrínseco en el campeonato mundial masculino de surf: WSL 2021. *Retos*, *64*, 311–321. https://doi.org/10.47197/retos.v64.106821

Abstract

Introduction: Surfers' performances are subjectively ranked by 5 judges. Low reliability and validity in judgment may lead to preventable errors and unfair scores.
Objective: The aims of the present study is to describe the judgment error for each wave surfed in world surf league (WSL) championship tour of 2021, establish between-judge reliability, and establish judgment validity comparing each judge score with the final control score of the wave.
Methodology: To describe the judgment error, we analyzed the available WSL data related to the 2021 Men's Championship Tour (4,095 waves; 20,475 scores).
Results: We found an inverted 'U'-shaped pattern for the judgment error score vs. control score, explained by a quadratic regression model (R = 0.52; SEE = 0.10). The reliability produced an excellent intraclass correlation coefficient ($CI_{95\%}$ = 0.97, 1.00), with a between judge (typical) error of 0.15. Validity analyses indicated a minimal real difference of 0.49 in the sum of two waves between the surfers for having 95% certainty for the heat winner.
Conclusion: We recommend WSL incorporate intrinsic judgment errors to increase fairness and trust in the WSL championship tour.

Keywords

Fair game; judgment; reliability; surfing; validity.

Resumen

Introducción: El desempeño de los surfistas es calificado subjetivamente por 5 jueces. La baja confiabilidad y validez en el juicio puede llevar a errores evitables y puntajes injustos.
Objetivo: Los objetivos del presente estudio son describir el error de juicio para cada ola surfeada en el tour del campeonato de la liga mundial de surf (WSL) de 2021, establecer la confiabilidad entre jueces y establecer la validez del juicio comparando el puntaje de cada juez con el puntaje de control final de la ola.
Metodología: Para describir el error de juicio, analizamos los datos disponibles de la WSL relacionados con el Tour del Campeonato Masculino de 2021 (4095 olas; 20475 puntajes).
Resultados: Encontramos un patrón en forma de "U" invertida para el puntaje de error de juicio vs. puntaje de control, explicado por un modelo de regresión cuadrática (R = 0,52; SEE = 0,10). La confiabilidad produjo un excelente coeficiente de correlación intraclase (IC95% = 0,97, 1,00), con un error entre jueces (típico) de 0,15. Los análisis de validez indicaron una diferencia real mínima de 0,49 en la suma de dos olas entre los surfistas para tener un 95 % de certeza sobre el ganador de la serie.
Conclusión: Recomendamos que WSL incorpore errores de juicio intrínsecos para aumentar la imparcialidad y la confianza en el campeonato de WSL.

Palabras clave

Fiabilidad; juego limpio; juicio; surf; validez.

Introduction

Surfing has shown important growth over the last few years, moving from a historically marginalized sport modality to an Olympic event (Rice, 2021). Despite the lack of current statistics, the number of practitioners today is much greater than those estimated in 2012 (Rice, 2021; Román et al., 2022), with have a low probability of abandoning the practice due to social and motivational issues (Santos González, 2024). As a reflection of this popularity, there has been proportional growth in economic activity (> $50 million) in locations with high quality waves (McGregor & Wills, 2016). One of the pillars of the sport's growth is the world championship promoted by the World Surf League (WSL), with the global reach of its broadcasts extending to more than 320 million views only on YouTube.

WSL's judges are responsible for ranking each athletic performance (waves surfed) based on certain criteria, namely, commitment, difficulty, performance of innovative, progressive, and combined maneuvers, speed changes, and power and flow (World Surf League, 2022). In 2021, a new competition model was introduced, which included the five highest-scoring surfers throughout the year for determining the champion. This change in the system may have increased the relevance of each score for each specific event, making it necessary to reduce the subjectivity of the judgments.

The events are contested by 24 athletes, who are divided into eight heats in the first phase (opening round), with three athletes in the water simultaneously. The surfer with the highest score per heat advances, whereas the others compete in a repechage (elimination round). In the repechage, the remaining 16 athletes compete head to head in the heats, in which only the highest score advances to the knockout phase. Until the end of the event, all heats are conducted in a head to head format. Each wave is judged on a scale of 0 to 10 by five different judges for each heat. The highest and lowest scores are excluded, and the average of the remaining three scores is used. The athlete's scores are based on the sum of the two best waves surfed in the 30-minute interval of each heat.

Since athletes can surf as many waves as possible, this creates a need for quick and dynamic evaluation by judges, increasing the possibility of errors in assessment. Even with predefined criteria and rigorous training, the judgment process presents natural subjectivity, leading to inevitable variability, which can be considered as inter-individual error. In sports such as diving, ski jumping and gymnastics, which also require subjective scores, similar phenomena have been observed, in which potential errors by judges may have interfered with the results (Emerson et al., 2009; Lyngstad et al., 2020; Ste-Marie, 1996).

Even considering the existing rotation, judges can be exposed to situations that ultimately increase errors that arise from possible mental fatigue influenced by factors such as resting deprivation, poor nutritional conditions and long-term sustained attention (Muraven & Baumeister, 2000). These errors present characteristics of great inter-individual variation, with each judge establishing their evaluation criteria differently. The inter-individual error during WSL competition may be increased by competitive daily routines imposed on judges during long working hours (< 10 h). Additionally, waves often have considerable variability, direction changes, shape and size, sea conditions, winds and weather, which influence maneuvers, and a permanent sense of urgency in the definition of their multiple judgments in a short period of time can affect the assessment.

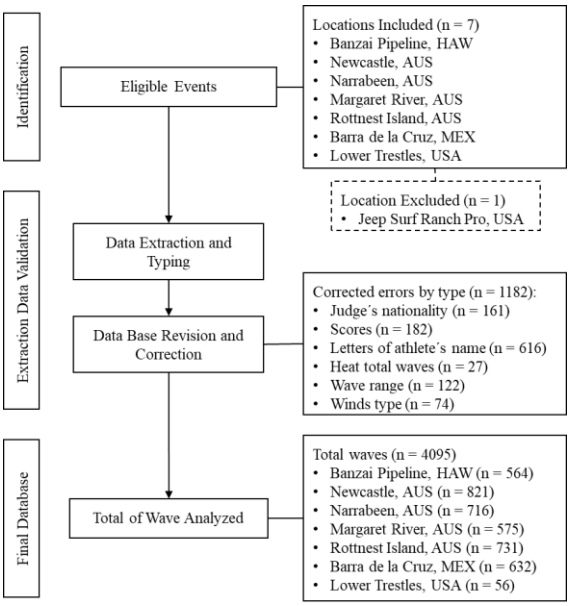
Notably, error is innate to any type of assessment (Flessas et al., 2015). It is not clear how the WSL (and other organizations) addresses the error of judgment, considering the enormous impact of judgments on athletes' careers, including direct (award) or indirect (sponsorship, campaigns, etc.) financial implications, staying in the world surfing elite, and mental health problems, among others. A better understanding of the error of judgment in professional surfing is important to increase the quality, transparency and fairness of the tournament and the overall public image of the organizers. The main goals of the present work were as follows: 1) to describe the judgment error for each wave surfed based on scores; 2) to establish between-judge reliability; and 3) to establish judgment validity comparing each judge score with the final control score (CS) of the wave.

Method

Database

This is a methodological study based on database analysis to determine the reliability (relative and absolute) and validity of the scoring assessment of WSL judges. Available data from all waves surfed (n = 4095) by male participants (n = 55) in 7 events in the 2021 tour were analyzed (except the data from Jeep Surf Ranch Pro1). This study was conducted by researchers who had no relationships with athletes, judges, or the WSL. It is important to note that the WSL's public data do not contain any personal identification of the judges, other than their nationality. Therefore, it is not possible to highlight the level of preparation of each judge as well as their evaluation criteria. An overview of the selection and data treatment steps is presented in Figure 1.

Figure 1. Flow chart of data extraction and analysis.



Procedure

Data extraction was performed manually by the study authors directly from the WSL website (public access - https://www.worldsurfleague.com) between September and December 2021. Data were extracted by pairs of randomly assigned independent researchers, and each researcher was responsible for extracting two events at most. The scores of each judge for waves surfed were recorded in a specific form and later included in a spreadsheet. After merging the two worksheets from each event, typos and miscellaneous inconsistencies were registered for later confirmation (Figure 1). The validation of the extractions was performed by a different pair of researchers. The combined spreadsheets were then returned to the pair of researchers originally responsible for extracting the data to work together to fix the identified inconsistencies. After a new verification and resolution of all problems, the data from a given moment were then integrated into the general database with all events, which was later used for analysis in the present study.

Data analysis

For an overview of the judgment error, the intrinsic judging error variability (IJEV) was adopted as proposed by Heiniger and Mercier (2021) (Equation 1). The exploratory characteristics of this variable

[1] Data from the Jeep Surf Ranch Pro event was not used because it takes place in a mechanical wave pool and has a different competition format than other events of the season.

enable the visualization of the error behavior of each wave judgment relative to its respective CS. After an exploratory approach, the best model for regression analysis was presented between IJEV (dependent variable) and CS (predictor variable). As a result, the determination coefficient (R), p value, standard estimation error (SEE) and prediction equation were presented.

$$IJEV = SD\ [J_{Diff\_1}, J_{Diff\_2}, J_{Diff\_3}, J_{Diff\_4}, J_{Diff\_5}] \quad \text{Eq. 1}$$

Where:
IJEV - Intrinsic judging error variability
SD - standard deviation
J_{Diff_1} - Difference between the control score and the judge score for judge 1 (J_{Diff_2} for judge 2 and so on)

The intraclass correlation coefficient was used considering the model of one-way random effects, the mean of k raters (n = 5) and absolute agreement (ICC(1,k)), as recommended by Koo and Li (2016). The confidence interval for 95% (CI95%) and the level of significance to explore the relative reliability between judges (inter-judge reliability) were also calculated. The ICC was classified as follows: < 0.5 (Poor); between 0.5 and 0.75 (Moderate); between 0.75 and 0.9 (Good); and > 0.9 (Excellent), as suggested by the same authors. For absolute reliability, the standard error between judges (SEBJ) was used to estimate the error between a and his peers (Equation 2). In addition, the minimal real difference of judges (MRDJ) was calculated to represent the threshold of a real error (Equation 3) between judges.

$$SEB_J = SD \times (\sqrt{1} - ICC) \quad \text{Eq. 2}$$

Where:
SEB_J - Standard error between judges
SD - standard deviation
ICC - Intraclass correlation coefficient

$$MRD_J = SEb_J \times 1{,}96 \times \sqrt{2} \quad \text{Eq. 3}$$

Where:
MRD_J - Minimal real difference between judges
SEB_J - Standard error between judges

Considering that reliability promotes between-judge analysis, we also explored the impact of the error magnitude on the difference between each judgment and the CS, previously described as a 'special case of validity' (Leandro et al., 2017). Due to the absence of a 'gold standard' method for wave assessment, the available option is the utilization of a central tendency parameter of the panel of judges. For this purpose, the median of five judges was used as the CS based on the following arguments presented by Heiniger and Mercier (2019): a. to be the best proxy of an 'actual wave score'; and b. to be more robust against misjudgments and biased judges than the trimmed average (approach utilized by WSL). Moreover, other arguments to use the median could be considered: c. the low number of wave scores (n = 5) to generate a normal distribution enabling an average calculation; and d. your robustness to prevent interference from an erratic score (discrepant or outlier). The magnitude of the error for validity for one wave was established by the typical error of judgment in reference to CS (TEJCS_1W, Equation 4a) and its superior confidence interval for 95%, defined as the minimal real difference for CS (MRDCS_1W, Equation 5a), a new variable mixing a traditional approach for error measurement in sports science (the typical error of measurement equal to the standard deviation of the differences divided by the square root) (Hopkins, 2000). Because the WSL utilizes the sum of two waves compared between athletes in the heats, TEJCS_1W and MRDCS_1W were also presented for the sum of two waves by multiplication by 2 (TEJCS_2W and MRDCS_2W, respectively).

$$(a)\ TEJ_{CS\_1W} = SD_{Overall}\ [J_{Diff\_1}, J_{Diff\_2}, J_{Diff\_3}, J_{Diff\_4}, J_{Diff\_5}] \div \sqrt{2} \quad \text{Eq. 4}$$
$$(b)\ TEJ_{CS\_2W} = TEJ_{CS\_1W} \times 2$$

Where:
TEJ_{CS_1W} - Typical error of judgement for control score for one wave
TEJ_{CS_2W} - Typical error of judgement for control score for two waves
$SD_{Overall}$ - Standard deviation for the data matrix
J_{Diff_1} - Difference between the control score and the judge score for judge 1 (J_{Diff_2} for judge 2 and so on)

$$MRD_{CS\_1W} = \sqrt{(DF \times TEJ_{CS\_1W}^2)} \div \chi^2_{97.5\%} \quad \text{Eq. 5}$$

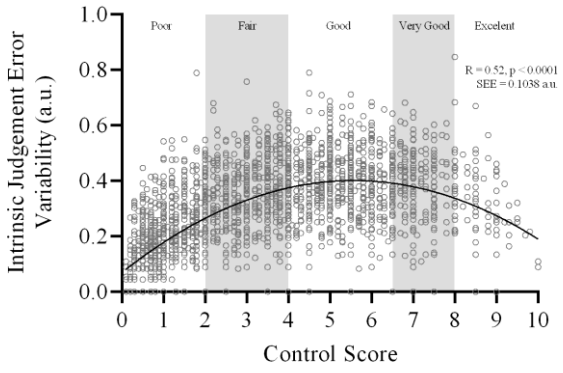
$MRD_{CS\_2W} = MRD_{CS\_1W} \times 2$
Where:
MRD_{CS_1W} - Minimal real difference for control score for one wave
MRD_{CS_2W} - Minimal real difference for control score for two waves
DF - Degrees of freedom
TEJ_{CS_1W} - Typical error of judgement for control score
$\chi 2$ - Chi square

Analyses of reliability and validity were performed for the overall database and for the subgroups of interest: location, round in regular competition, round in finals, wave level, number of athletes in the heat, wave size and wind conditions. The dataset was organized in Microsoft Excel (Microsoft Corporation, Redmond, WA, USA) and later analyzed in RStudio (R Core Team, 2021). Intraclass correlations were calculated using the Psych package (Revelle, 2021), whereas IJEV, SEBJ, MRDJ, TEJCS_1W, and MRDCS_1W were calculated by using RStudio native functions. The level of significance was adjusted for $p < 0.05$.

Results

The representation of the IJEV for each surfed wave in the WSL 2021 championship as a function of the CS is presented in Figure 2. The inverted 'U' pattern described by the second-order (quadratic) polynomial model produced $R = 0.52$, $p < 0.001$ and $SEE = 0.1038$. The smallest magnitudes of IJEV were observed for waves classified as Poor (≅ 0.18) and Excellent (≅ 0.28), and higher magnitudes were observed for waves classified as Good (≅ 0.40).

Figure 2. Intrinsic Judging Error Variability by Control Score for all 2021 season. R - coefficient of determination; SEE - standard estimation error; a.u. - arbitrary unit; the vertical tracks represent 2021 World Surf League criteria for classification the wave quality; The prediction equation for IJEV based on second order polynomial quadratic model is IJEV = 0,06966 + 0,1192 x CS - 0,01070 x CS2.



The global reliability of judgments performed in 2021 (overall) or segmented by the conditions of interest (location, rounds, wave level, number of athletes in the water, wave size and wind intensity) are reported in Table 1. The relative reliability, expressed by the ICC, confidence interval and p value, suggests that the WSL judges showed 'near perfect' performance, with ICC indices consistently close to 1 (CI95% = 0.970, 0.996; p < 0,001) and classified most times (92%) as Excellent. A classification of Good (8%) was observed only when the data were segmented at the wave level. The absolute reliability presented by between judgment error was stable for the average of the whole dataset results, with SEBJ ≅ 0.15 (5.3%) and MRDJ ≅ 0.41 (14.6%).

Table 1. Overall and categorized relative and absolute reliability for surf judgments during 2021 WSL surf season for males

Variables of Interest	n	Relative				Absolute			
		$ICC_{(1,k)}$	Class.	$CI_{95\%}$	p value	SEB_J		MRD_J	
						Raw	%	Raw	%
Overall	4095	0.996	E	0.996, 0.996	< 0.001	0.14	4.8	0.40	13.3
By Location									
Banzai Pipeline, HAW	564	0.997	E	0.996, 0.997	< 0.001	0.14	11.5	0.38	31.8
Newcastle, AUS	821	0.996	E	0.995, 0.996	< 0.001	0.15	5.4	0.40	14.9
Narrabeen, AUS	716	0.996	E	0.995, 0.996	< 0.001	0.14	5.5	0.38	15.2
Margaret River, AUS	575	0.996	E	0.996, 0.997	< 0.001	0.15	3.8	0.42	10.5
Rottnest Island, AUS	731	0.997	E	0.996, 0.997	< 0.001	0.14	4.7	0.39	13.1
Barra de la Cruz, MEX	632	0.996	E	0.996, 0.997	< 0.001	0.15	3.7	0.41	10.3
Lower Trestles, EUA	56	0.998	E	0.998, 0.999	< 0.001	0.13	3.1	0.37	8.6
By Round in Regular Competition									
Seeding Round	1207	0.996	E	0.996, 0.996	< 0.001	0.15	5.2	0.41	14.5
Elimination Round	350	0.996	E	0.995, 0.996	< 0.001	0.15	4.6	0.41	12.7
Round 32	1271	0.996	E	0.996, 0.997	< 0.001	0.15	4.8	0.40	13.4
Round 16	678	0.997	E	0.996, 0.997	< 0.001	0.14	5.5	0.38	15.4
Quarter Final	290	0.997	E	0.997, 0.998	< 0.001	0.14	5.1	0.39	14.3
Semi Final	154	0.997	E	0.997, 0.998	< 0.001	0.14	4.9	0.38	13.6
Finals	89	0.998	E	0.997, 0.999	< 0.001	0.13	4.9	0.37	13.5
By Round in Finals									
Match 1	10	0.995	E	0.990, 0.998	< 0.001	0.16	5.5	0.44	15.2
Match 2	7	0.998	E	0.994, 0.999	< 0.001	0.14	2.3	0.38	6.4
Match 3	8	0.999	E	0.997, 1.000	< 0.001	0.11	1.6	0.30	4.3
Match 4	31	0.999	E	0.998, 0.999	< 0.001	0.12	4.8	0.33	13.2
By Wave Level									
Poor (< 2.0)	1605	0.968	E	0.966, 0.970	< 0.001	0.09	12.2	0.24	33.8
Fair (2.0 - 3.9)	1297	0.963	E	0.960, 0.966	< 0.001	0.18	5.1	0.49	14.1
Good (4.0 - 6.4)	659	0.828	G	0.810, 0.845	< 0.001	0.23	4.3	0.65	11.8
Very Good (6.5 - 7.9)	398	0.827	G	0.803, 0.849	< 0.001	0.23	3.3	0.63	9.0
Excellent (≥ 8)	136	0.898	G	0.873, 0.919	< 0.001	0.19	2.2	0.53	6.2
By Number of athletes in the heat									
Two	2538	0.996	E	0.996, 0.997	< 0.001	0.14	4.8	0.40	13.3
Three	1557	0.995	E	0.994, 0.995	< 0.001	0.15	5.1	0.42	14.1
By Wave Size (ft)									
1 to 4	2629	0.996	E	0.996, 0.996	< 0.001	0.14	4.8	0.40	13.3
4 to 6	766	0.997	E	0.997, 0.998	< 0.001	0.14	6.0	0.39	16.7
6 to 8	298	0.996	E	0.996, 0.997	< 0.001	0.15	4.6	0.42	12.6
8 to 10	121	0.997	E	0.996, 0.997	< 0.001	0.15	3.9	0.41	10.8
Not Reported	281	0.997	E	0.996, 0.997	< 0.001	0.15	4.2	0.41	11.7
By Wind Conditions									
Calm	1768	0.996	E	0.996, 0.997	< 0.001	0.14	4.8	0.40	13.3
Cross	329	0.995	E	0.994, 0.995	< 0.001	0.15	5.1	0.42	14.1
Light	536	0.997	E	0.996, 0.997	< 0.001	0.14	9.4	0.39	26.0
Offshore	1167	0.997	E	0.996, 0.997	< 0.001	0.14	4.7	0.39	13.0
Onshore	14	0.990	E	0.982, 0.996	< 0.001	0.14	14.0	0.39	38.8
Not Reported	281	0.997	E	0.996, 0.997	< 0.001	0.15	4.2	0.41	11.7

Legend: n - Number of waves; $ICC_{(1,K)}$ - Intraclass correlation model one way random (1,5); Class. - ICC classification; $CI_{95\%}$ - ICC confidence interval for 95%; p value - ICC p value; SEB_J - Standard error between judges; MRD_J - Minimal real difference between Judges.

The judgment errors in determining the final score (e.g., CS) considering different variables of interest are presented in Table 2. For all waves surfed (Overall), results of 0.22 and 0.25 were found for TEJCS_1W and MRDCS_1W, respectively. As the WSL uses the sum of the two best waves as a criterion to compare the performance between surfers, the results of TEJCS_2W (0.44) and MRDCS_2W (0.49) were available. Most of the investigated conditions with potential for judgment disruption resulted in a low mean variability for TEJCS_1w (≅ 0.22) of the whole dataset. Considering the segmented conditions, a low mean variation in TEJCS_1w was also found for different locations that hosted the events (CI95%

= 0.21, 0.23), different rounds of regular competition (CI95% = 0.21, 0.22), different wave sizes (CI95% = 0.21, 0.23) and different wind conditions (CI95% = 0.21, 0.23). On the other hand, a slightly greater mean variation was found for the analyses per round in the finals (CI95% = 0.15, 0.29), by wave level (CI95% = 0.16, 0.31) and by the number of surfers disputing the heat (CI95% = 0.17, 0.26).

Table 2. Overall and categorized validity for surf judgments during 2021 WSL surf season for males

Variables of Interest	One Wave		Two Waves	
	TEJ_{CS_1W}	MRD_{CS_1W}	TEJ_{CS_2W}	MRD_{CS_2W}
Overall	0.22	0.25	0.44	0.49
By Location				
Banzai Pipeline, HAW	0.21	0.24	0.42	0.48
Newcastle, AUS	0.22	0.24	0.44	0.48
Narrabeen, AUS	0.21	0.24	0.42	0.48
Margaret River, AUS	0.23	0.24	0.47	0.48
Rottnest Island, AUS	0.22	0.24	0.43	0.48
Barra de la Cruz, MEX	0.23	0.24	0.46	0.48
Lower Trestles, EUA	0.21	0.22	0.42	0.43
By Round in Regular Competition				
Seeding Round	0.22	0.24	0.45	0.48
Elimination Round	0.23	0.24	0.45	0.47
Round 32	0.22	0.24	0.44	0.48
Round 16	0.21	0.24	0.42	0.48
Quarter Final	0.21	0.23	0.42	0.47
Semin Final	0.21	0.23	0.42	0.46
Final	0.21	0.22	0.42	0.43
By Round in Finals				
Match 1	0.28	0.18	0.55	0.36
Match 2	0.23	0.17	0.45	0.35
Match 3	0.19	0.18	0.38	0.35
Match 4	0.18	0.21	0.36	0.41
By Number of Athletes in the Heat				
Two	0.22	0.24	0.43	0.49
Three	0.22	0.24	0.45	0.49
By Wave Level				
Poor (< 2.0)	0.12	0.24	0.25	0.49
Fair (2.0 - 3.9)	0.25	0.24	0.51	0.48
Good (4.0 - 6.4)	0.28	0.24	0.55	0.48
Very Good (6.5 - 7.9)	0.27	0.24	0.54	0.47
Excellent (≥ 8)	0.25	0.23	0.49	0.45
By Wave Size (ft)				
1 to 4	0.22	0.24	0.44	0.49
4 to 6	0.21	0.24	0.43	0.48
6 to 8	0.23	0.23	0.46	0.47
8 to 10	0.23	0.23	0.45	0.45
Not Reported	0.23	0.23	0.45	0.47
By Wind Conditions				
Calm	0.22	0.24	0.44	0.49
Cross	0.23	0.23	0.47	0.47
Light	0.21	0.24	0.43	0.48
Offshore	0.22	0.24	0.43	0.48
Onshore	0.22	0.19	0.44	0.38
Not Reported	0.23	0.23	0.45	0.47

Legend: TEJ_{CS_1W} - Typical error of judgement for control score for one wave; TEJ_{CS_2W} - for two waves; MRD_{CS_1W} - Minimal real difference for judgement for control score for one wave; MRD_{CS_2W} - for two waves.

Discussion

The present study is the first to explore intrinsic judgment error in male professional surf championships organized by the WSL, analyzing the IJEV results, reliability and validity. The judgment error was described for global scores and different conditions of interest (e.g., location, round, wave size, number of athletes involved in the heat, wave level and wind). In addition, the use of TEJCS and MRDCS was proposed to compare the performance of surfers similarly to what has been practiced in the interpretation of statistical tests in clinical areas of health and sports performance, incorporating measurement error for its prognostic relevance (Copay et al., 2007). As far as could be observed, there are no studies with other sports dedicated to establishing the magnitude of the judgment error and,

mainly, its incorporation in the interpretation of the competitive results. This new approach could impact the way scores are assigned in certain sports, as well as provide a more reliable process from a scientific point of view.

Intrinsic Judgment Error Variability

The behavior of the IJEV as a function of the CS depends on the modality investigated. Heiniger and Mercier (2019) described three main possible kinetics: a) Descending (snowboard halfpipe, acrobatic gymnastics, aerobic gymnastics, artistic gymnastics, rhythmic gymnastics, and artistic swimming); b) 'U' pattern (standard and artistic presentation on dressage) and, as in the present study; c) inverted 'U' pattern (diving, figure skating, ski jumping, snowboard slopestyle, trampoline and aerials from skiing). According to the authors, the different kinetic patterns are influenced by the high number of items to be evaluated in each modality, as well as the number of errors to be deducted from a given execution. In surfing, aspects such as height, speed, wave size, difficulty of the maneuver and body posture are considered.

In addition, the shape of the parabola seems to depend on judgments with results close to zero (Heiniger & Mercier, 2019), which occurs more often in surfing (40.07% of waves surfed in the 2021 season were classified as Poor, 32.0% as Fair, 16.1% as Good, 8.9% as Very Good, and 2.9% as Excellent). It is possible that the low IJEV observed in waves classified as Poor (wave score < 2) is determined by the lower complexity of judgment, considering the low number of elements to be observed, since these scores are usually attributed to a wave with an incomplete maneuver or surfer mistake. At the opposite end of the scale in waves with higher scores, the smallest error observed may be related to a ceiling effect of the judgment process provided by the proximity of perfection of the evaluated performances, which may facilitate the process. Jointly, these results partially confirm the arguments of Heiniger and Mercier (2019), who suggest that "judges are more accurate when evaluating outstanding or atrocious performances than when evaluating mediocre ones". Based on the present results, the WSL judges achieve better consistency when evaluating poor waves.

The magnitude of predictive power observed on weighted least-squares exponential regression models developed for other sports modalities resulted in superior values of R (0.75 ± 0.19, CI95% = 0.65, 0.84) and inferior values for the root mean square standard deviation (0.14 ± 0.28, CI95% = 0.00, 0.29) compared with those in the present study (0.52 and 0.10, respectively). However, those differences appear to be determined by a different process to generate regression models than what was used in the present study. While in the present study, the full database was utilized to generate our regression, Heiniger and Mercier (2019) used the mean value for each CS as a predictable variable, resulting in a substantial reduction in residuals with a direct impact on the root mean standard deviation (≅ -48%) and inflating the R (≅ +37%). Considering this and the purpose of the present study related to error scaling, this direct comparison between studies could not be possible.

Relative and Absolute Reliability

Reliability, especially relative reliability, is one of the most commonly used metrics in the investigation of the psychometric quality of judgment in sports. The relative reliability of the present study for the overall database (ICC(1,k) ≅ 0.99, Table 1), with the exception of segmentation by wave level (ICC(1,k) ≅ 0.90), was much greater than the results of Premelč et al. (2019) for different categories of dance sport (≅ 0.62), Pajek et al. (2013) for artistic gymnastics in different competitive phases and apparatus (≅ 0.83) and Leandro et al. (2017) in rhythmic gymnastics for athletes in different ranking positions (≅ 0.66).

The absolute dimensioning of the inter-judge error was produced only for sports dance (Premelč et al., 2019), a modality with an evaluation scale similar to that of surf (0 to 10) but with a competitive dynamic that results in higher mean scores. The SEMs reported for dance were 0.54 (overall), 0.56 (technical qualities), 0.67 (movement to music), 0.57 (partnering skills), and 0.54 (choreography and performance). In the present study, low SEBJ and MRDJ were found for the average of all conditions investigated, with mean values of 0.15 and 0.41, respectively. Considering the greater complexity in surfing judgment, due to the large number of factors to be observed in the athlete's performance, the lowest results of the present study were considered unusual, deserving future investigations to identify which procedural elements are practiced by the WSL that contributes to the cohesiveness of judges among themselves.

The interpretation of absolute reliability produces a measure of error between the scores of judges when compared with each other (e.g., judge 1 vs. judge 2, judge 2 vs. judge 3, etc.), serving to dimension their qualifications. For the total number of waves (n = 4095) surfed in 2021 using the SEBJ (0.14) as a criterion for the difference between the judges, when expressing the mean error, a frequency of 65% effectively different scores was observed. A frequency of 36% was observer when the MRDJ (0.40), which expresses a 95% certainty of different scores, was used. Despite the very high relative reliability scores, the analysis of the results using absolute reliability broadens the understanding of differences between the scores given by the judges.

Since reliability analysis does not describe the error of a judge's measurement in relation to CS, its practical application becomes limited for evaluating the competitive dynamics of modality using only the values of SEBJ and MRDJ. In addition, the interpretation of the SEBJ should be performed with caution in the event of heteroscedasticity in the scores (Atkinson & Nevill, 1998). Therefore, the present study produced the results of TEJCS_1W and MRDCS_1W.

Validity

To the best of our knowledge, two studies have investigated the validity of judgment by comparing the results of judges with those of a CS, and both involved artistic gymnastics (Leskošek et al., 2010; Pajek et al., 2013). In general, unclear methodological details were observed for the adequate understanding of performed analyses or aims intended for specific purposes (i.e. nationality bias, sequential bias and comparisons between equipment in gymnastics). Leskošek et al. (2010) produced validity indices considered satisfactory by the authors, whereas Pajek et al. (2013) interpreted the results as unsatisfactory. For this purpose, the authors used ANOVA and Kendall W analysis, which makes comparisons with the findings produced in the present study impossible because those techniques do not measure the degree of error.

Based on the results produced in the present study, a concern spot was identified. To define the winner of a heat, the WSL currently uses a difference of 0.01 between the sum of the best two waves surfed, which is below what is necessary to consider the natural error of its judgment process. The TEJCS and MRDCS results provide an overview of the magnitude of difference required between two surfers for a winner to be defined with a low probability (< 5%) of a random and possibly wrong and unfair result. With respect to potential practical applications of the results of this study, we present an innovative and scientifically more efficient scenario to be used in new competitions with the aim of more fairly determining the winners of heats, rounds, events and championships. An applied description of this concept utilization is available in Figure 3 as supplementary material (https://osf.io/yk2vt/) exploring the final of the event MEO Pro (Peniche, PT), which was held in March 2022. The difference between the winner (Griffin Colapinto, USA) and the second-place finisher (Filipe Toledo, BRA) was 0.14, which was lower than that of TEJCS_2W.

The results produced in the present study can be extended to all judging sport modalities, with error magnitudes that need to be established. The use of judgment error would improve the judgment process, establishing certainty (95%) in the definition of winning and losing athletes, reducing judgment bias and improving the sense of justice, resulting in an important advancement by the WSL to improve its evaluation routines. Previously, modalities such as gymnastics (Heiniger & Mercier, 2019) made changes in judgment to make the process more objective and potentially justifiable. By utilizing an innovative statistical strategy, the present study dimensioned the error of judgment in the WSL for the 2021 season.

The results produced in the present study, analyzed considering their limitations, can offer new perspectives to surf. The main limitation of our study was that the analysis used only data from the competition held in 2021, restricting a possible evaluation of the performance of judges during different years, thus enabling the production of a historical series. In addition, other population groups (i.e., women, surfers in the access and youth categories) and surfing modalities (i.e., long board, big wave, etc.) need to be investigated. Furthermore, possible reasons for the error were not investigated, which occurred previously in other sports, such as gymnastics (Heiniger & Mercier, 2021) and ski jumping (Lyngstad et al., 2020).

Conclusions

Despite the near-perfect, yet unlikely, reliability between judges, it is concluded that there is an inevitable judgment error in the surfing scores in WSL the 2021 season, expressed in validity analyses of 0.22 and 0.25 for TEJCS_1W and MRDCS_1W, respectively, this implies the need to consider the error inherent in this type of evaluation when comparing the performance of competitive surfers to reduce uncertainty and increase the fairness of these comparisons, which can define the athlete's destiny in the competition. These results suggest the need to modify the competitive dynamics for surf and, possibly, for other judging modalities, changing the way judges assign scores.

Acknowledgements

The authors of this study thank Vinicius de Oliveira Damasceno for his collaboration in data extraction. We are also grateful for the considerations and reflections provided by Maicon Rodrigues Albuquerque regarding strategy and the features of Rstudio in data analysis and Wagner Prado for overall comments.

Financing

L.E.R.S., L.C.P., and T.C.B.M. are grateful for scholarship by CAPES Brazil. I.V. and C.B. are grateful for scholarship by FAPESP Brazil (#2020/04827-0). This research did not receive any type of funding.

References

Atkinson, G., & Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine*, *26*(4), 217-238. https://doi.org/10.2165/00007256-199826040-00002

Copay, A. G., Subach, B. R., Glassman, S. D., Polly, D. W., Jr., & Schuler, T. C. (2007). Understanding the minimum clinically important difference: a review of concepts and methods. *Spine Journal*, *7*(5), 541-546. https://doi.org/10.1016/j.spinee.2007.01.008

Emerson, J. W., Seltzer, M., & Lin, D. (2009). Assessing Judging Bias: An Example From the 2000 Olympic Games. *The American Statistician*, *63*(2), 124-131. https://doi.org/10.1198/tast.2009.0026

Flessas, K., Mylonas, D., Panagiotaropoulou, G., Tsopani, D., Korda, A., Siettos, C., . . . Smyrnis, N. (2015). Judging the judges' performance in rhythmic gymnastics. *Medicine and Science in Sports and Exercise*, *47*(3), 640-648. https://doi.org/10.1249/mss.0000000000000425

Heiniger, S., & Mercier, H. (2019). Judging the judges: A general framework for evaluating the performance of international sports judges. *arXiv*, *Pre Print*(10055), 1-9. https://arxiv.org/abs/1807.10055.

Heiniger, S., & Mercier, H. (2021). Judging the judges: evaluating the accuracy and national bias of international gymnastics judges. *Journal of Quantitative Analysis in Sports*, *17*(4), 289-305. https://doi.org/https://doi.org/10.1515/jqas-2019-0113

Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine*, *30*(1), 1-15. https://doi.org/10.2165/00007256-200030010-00001

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Leandro, C., Avila-Carvalho, L., Sierra-Palmeiro, E., & Bobo-Arce, M. (2017). Judging in Rhythmic Gymnastics at Different Levels of Performance. *Journal of Human Kinetics*, *60*, 159-165. https://doi.org/10.1515/hukin-2017-0099

Leskošek, B., Čuk, I., Karácsony, I., Pajek, J., & Bučar, M. (2010). Reliability and validity of judging in men's artistic gymnastics at the 2009 university games. *Science of Gymnastics Journal*, *2*(1), 25-34.

Lyngstad, T. H., Härkönen, J., & Rønneberg, L. T. S. (2020). Nationalistic bias in sport performance evaluations: An example from the ski jumping world cup. *European Journal for Sport and Society*, *17*(3), 250-264. https://doi.org/10.1080/16138171.2020.1792628

McGregor, T., & Wills, S. (2016). *Natural Assets: Surfing a Wave of Economic Growth*. https://EconPapers.repec.org/RePEc:syd:wpaper:2016-06

Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: does self-control resemble a muscle? *Psychol Bull*, *126*(2), 247-259. https://doi.org/10.1037/0033-2909.126.2.247

Pajek, M. B., Cuk, I., Pajek, J., Kovač, M., & Leskošek, B. (2013). Is the quality of judging in women artistic gymnastics equivalent at major competitions of different levels? *Journal of Human Kinetics*, *37*, 173-181. https://doi.org/10.2478/hukin-2013-0038

Premelč, J., Vučković, G., James, N., & Leskošek, B. (2019). Reliability of Judging in DanceSport. *Front Psychol*, *10*, 1001. https://doi.org/10.3389/fpsyg.2019.01001

R Core Team. (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* https://www.R-project.org/

Revelle, W. (2021). *Procedures for personality and psychological research* https://CRAN.R-project.org/package=psych

Rice, E. L. (2021). Surfing. *J Sports Med Phys Fitness*, *61*(8), 1098-1103. https://doi.org/10.23736/s0022-4707.21.12847-6

Román, C., Borja, A., Uyarra, M. C., & Pouso, S. (2022). Surfing the waves: Environmental and socio-economic aspects of surf tourism and recreation. *Sci Total Environ*, *826*, 154122. https://doi.org/10.1016/j.scitotenv.2022.154122

Santos González, D. (2024). No vas a dejar el surf. Reflexiones sobre las motivaciones y condicionantes sociales que conducen a su práctica en entornos naturales y artificiales (You won't quit surfing. Reflections on the motivations and social conditions leading to its practice in natural and artificial environments). *Retos*, *59*, 903-911. https://doi.org/10.47197/retos.v59.108656

Ste-Marie, D. M. (1996). International Bias in Gymnastic Judging: Conscious or Unconscious Influences? *Perceptual and Motor Skills*, *83*(3), 963-975. https://doi.org/10.2466/pms.1996.83.3.963

World Surf League. (2022). *Rules and regulations*. https://www.worldsurfleague.com/pages/rules-and-regulations

Authors' and translators' details:

Tony Meireles Santos	tony.meireles@ufpe.br	Author
Lucas Eduardo Rodrigues Santos	lucas.rodriguessantos@ufpe.br	Author
Ítalo Vinícius	italovinicius@usp.br	Author
Cayque Brietzke	cayquebbarreto@alumni.usp.br	Author
Lucas Camilo Pereira	lucascamilo.edf@gmail.com	Author
Paulo Henrique Melo	paulo.hmelo2@ufpe.br	Author
Thaiene Camila Beltrão Moura	thaiene.moura@ufpe.br	Author
Taddeo De Negri	taddeodenegri@gmail.com	Author
Hassan Mohamed Elsangedy	hassan.elsangedy@gmail.com	Author
Flávio Oliveira Pires	piresfo@usp.br	Author
American Journal Experts	support@aje.com	Translator