Estimating the Replicability of Sports and Exercise Science Research

MURPHY, Jennifer, KIRK, Christopher, EDWARDS, Tobias and WARNE, Joe

Sports Medicine (2025) 55:2659–2679
https://doi.org/10.1007/s40279-025-02201-w

ORIGINAL RESEARCH ARTICLE

Estimating the Replicability of Sports and Exercise Science Research

Jennifer Murphy[1] · Aaron R. Caldwell[2] · Cristian Mesquida[1,3] · Aera J. M. Ladell[4] · Alberto Encarnación-Martínez[5] · Alexandre Tual[6] · Andrew Denys[7] · Bailey Cameron[8] · Bas Van Hooren[9] · Ben Parr[10] · Bianca DeLucia[11] · Billy R. J. Mason[12] · Brad Clark[12] · Brendan Egan[13] · Calum Brown[14] · Carl Ade[15] · Chiarella Sforza[16] · Christopher B. Taber[17] · Christopher Kirk[14] · Christopher McCrum[9] · Cian OKeeffe Tighe[1] · Ciara Byrne[1] · Claudia Brunetti[16] · Cyril Forestier[6] · Dan Martin[18] · Danny Taylor[18] · David Diggin[19] · Dearbhla Gallagher[20] · Deborah L. King[19] · Elizabeth Rogers[15] · Eric C. Bennett[4] · Eric T. Lopatofsky[17] · Gemma Dunn[20] · Gérome C. Gauchard[21] · Guillaume Mornieux[21] · Ignacio Catalá-Vilaplana[22] · Ines Caetan[18] · Inmaculada Aparicio-Aparicio[5] · Jack Barnes[14] · Jake Blaisdell[19] · James Steele[8] · Jared R. Fletcher[4] · Jasmin Hutchinson[11] · Jason Au[23] · Jason P. Oliemans[4] · Javad Bakhshinejad[11] · Joaquin Barrios[24] · Jose Ignacio Priego Quesada[5] · Joseph Rager[25] · Julia B. Capone[4] · Julie S. J. Walton[26] · Kailey Stevens[23] · Katie Heinrich[15] · Kelly Wu[27] · Kenneth Meijer[9] · Laura Richards[20] · Lauren Jutlah[23] · Le Tong[7] · Lee Bridgeman[8] · Leo Banet[21] · Leonard Mbiyu[25] · Lucy Sefton[25] · Margaux de Chanaleilles[6] · Maria Charisi[27] · Matthew Beerse[24] · Matthew J. Major[27] · Maya Caon[28] · Mel Bargh[18] · Michael Rowley[7] · Miguel Vaca Moran[27] · Nicholas Croker[11] · Nicolas C. Hanen[21] · Nicole Montague[1] · Noel E. Brick[10] · Oliver R. Runswick[25] · Paul Willems[9] · Pedro Pérez-Soriano[5] · Rebecca Blake[29] · Rebecca Jones[18] · Rebecca Louise Quinn[10] · Roberto Sanchis-Sanchis[5] · Rodrigo Rabello[16] · Roisin Bolger[10] · Roy Shohat[28] · Sadie Cotton[19] · Samantha Chua[8] · Samuel Norwood[28] · Samuel Vimeau[28] · Sandro Dias[18] · Sissel Pedersen[9] · Spencer S. Skaper[4] · Taylor Coyle[1] · Terun Desai[30] · Thomas I. Gee[18] · Tobias Edwards[14] · Torsten Pohl[31] · Vanessa Yingling[7] · Vinicius Ribeiro[28] · Youri Duchene[21] · Zacharias Papadakis[28] · Joe P. Warne[1]

Accepted: 27 February 2025 / Published online: 16 June 2025

Abstract

Background The replicability of sports and exercise research has not been assessed previously despite concerns about scientific practices within the field.

Aim This study aims to provide an initial estimate of the replicability of applied sports and exercise science research published in quartile 1 journals (SCImago journal ranking for 2019 in the Sports Science subject category; www.scimagojr.com) between 2016 and 2021.

Methods A formalised selection protocol for this replication project was previously published. Voluntary collaborators were recruited, and studies were allocated in a stratified and randomised manner on the basis of equipment and expertise. Original authors were contacted to provide deidentified raw data, to review preregistrations and to provide methodological clarifications. A multiple inferential strategy was employed to analyse the replication data. The same analysis (i.e. *F* test or *t* test) was used to determine whether the replication effect size was statistically significant and in the same direction as the original effect size. *Z*-tests were used to determine whether the original and replication effect size estimates were compatible or significantly different in magnitude.

Results In total, 25 replication studies were included for analysis. Of the 25, 10 replications used paired *t* tests, 1 used an independent *t* test and 14 used an analysis of variance (ANOVA) for the statistical analyses. In all, 7 (28%) studies demonstrated robust replicability, meeting all three validation criteria: achieving statistical significance ($p < 0.05$) in the same direction as the original study and showing compatible effect size magnitudes as per the *Z* test ($p > 0.05$).

Conclusion There was a substantial decrease in the published effect size estimate magnitudes when replicated; therefore, sports and exercise science researchers should consider effect size uncertainty when conducting subsequent power analyses. Additionally, there were many barriers to conducting the replication studies, e.g., original author communication and poor data and reporting transparency.

Key Points

This is the first large replication project in sports and exercise science due to concerns about replication in the field.

Findings showed that 28% of studies were replicated successfully.

The results raise concerns about research findings in this discipline that should be explored further.

Researchers should consider the accumulation of evidence rather than relying on standalone findings in sports and exercise science.

1 Introduction

Science is a dynamic process of discovering new effects and testing theories and their application. Yet, an "effect" that has been found once but cannot be replicated arguably does not qualify as a scientific discovery [1]. Replication studies are at the heart of the scientific process, as they advance knowledge by confirming or refuting previous findings or explore boundary conditions and the underlying variation in the true effect [2]. Although it is generally accepted that replications play an important role in science, the inability to reproduce research findings is still a long-standing problem [3]. The current research culture is largely driven by career incentives and novel research, which has given rise to poor scientific behaviours that prioritise individual researcher success (e.g. career advancement) rather than what is beneficial for science as a whole [4–6]. This motivates researchers to place emphasis on novel or "flashy" findings and hunt for statistical significance within their datasets to garner a "high impact" publication. Consequently, replication has come to the forefront of discussions, particularly in psychology, due to the observed failures to replicate well-known psychological results [7]. This exacerbated the "crisis of confidence" in scientific findings [8, 9], and there is also evidence of scientific misconduct, including fraud and questionable research practices e.g., *p*-hacking, multiple analyses and selective reporting of a desirable result [10–12]. The proclamation that most research findings are false has fuelled scepticism within the scientific community, prompting the adoption of open science practices to enhance transparency and rigour [13].

Sports and exercise science, similar to other fields, has grappled with criticisms surrounding questionable research practices, including overly optimistic statistical conclusions and a scarcity of replication studies [14]. Overall, there are concerns about study design, and statistical and reporting practices within the field, resulting in calls for more replication studies [15, 16]. A history of low-powered studies has contributed to these concerns [17], as well as the rampant misuse of null hypothesis significance testing (NHST) [18–20]. In particular, stating the null and alternative hypotheses and setting both the alpha and beta levels are prerequisites for NHST. Still, investigations show less than a quarter of studies report an a priori power analysis and 82% of studies that do not state a hypothesis use NHST anyway [20–22]. Reporting transparency is also an obvious problem in sports and exercise science, with a data sharing rate of less than 5% and almost zero studies making computer code available [23]. Furthermore, implausibly high positive result rates of 81% across sports and exercise science journals [20], and 82% across sports medicine and physiotherapy journals [24], indicate the presence of publication bias given the average observed power of studies [22, 25]. Sports and exercise scientists also rarely collaborate with statisticians [26], despite the regularity of statistical errors [27] and our awareness of our overall sub-standard statistical competency [28]. These errors are then compounded by further errors in meta-analyses [29]. Finally, the reporting of basic statistical information such as test statistics, degrees of freedom and confidence intervals is sporadic at best [22]. All of these issues raise concerns about replicability within our field.

The aforementioned concerns, coupled with limited reporting of null or trivial results and a lack of transparency, underscore the need for a reformation of research practices within sports and exercise science [15]. A subjective survey of over 500 sports and exercise scientists revealed a widespread belief in a reproducibility (using the same data to obtain the same results) and replicability (using new data to obtain a similar result) crisis within the field, and despite ongoing discussions about these issues, substantial barriers to both reproducibility and replicability persist [28]. While isolated replication studies exist [30–33], a comprehensive quantitative assessment of replicability in our field remains elusive. Most discussions on the topic rely on indirect inferences or anecdotes rather than empirical data. Thus, this study aims to provide an initial estimate of the replicability of sports and exercise science research published in quartile 1 journals between 2016 and 2021. Previous research reports that the expected replication rate was 0.61, suggesting that, if we replicate significant findings (with the same statistical power and sample size), 61% would be expected to yield another significant effect [22]. This was a pilot study of 89 studies published in the *Journal of Sport Science* and requires further research; however, in the absence of any other information, we expect a similar replication rate here.

2 Methods

2.1 Replication Study Selection

This project began in September 2020 with the completion of data collection in 2024. A formalised selection protocol for this replication project was created to minimise bias where possible and published for full transparency [34]. The key aspects of the selection protocol focused on the year of publication and citation rankings, research discipline, study type, the research question and key dependent variable, study methods and feasibility. In summary, studies were selected if they had a statistically significant main effect published between 2016 and 2021 in quartile 1 applied sports and exercise science journals (SCImago journal rank for 2019 in the Sports Science subject category; www.scimagojr.com) and they were experimental or quasi-experimental quantitative studies, whereby an independent variable was manipulated to determine the effect on the dependent variable, in pairwise, independent or crossover study designs and across two or more groups. The project leaders selected the key applied dependent variable that was first stated in the abstract, first or primary hypothesis or aim. If this dependent variable was not an applied sports and exercise science variable, or if it was not analysed using a *t*-test or analysis of variance (ANOVA) as per the selection criteria [34], this dependent variable was disregarded, and the next stated applied dependent variable was selected from the abstract, first or primary hypothesis or aim. Where this was unclear, the dependent variable was randomly selected when it met all of the other inclusion criteria. Studies to be replicated were screened by J.M., J.W. and C.M. using an online survey which was created to screen potential replication studies (https://doi.org/10.17605/OSF.IO/SFBVA). When the screening was completed, the studies were divided into research sub-disciplines (applied sports and exercise biomechanics, psychology, physiology, nutrition and injury prevention) and numbered (https://doi.org/10.17605/OSF.IO/SFBVA).

2.2 Collaborator Recruitment and Allocation

Voluntary collaborators were recruited via social media, where researchers filled out an expression of interest form, including details of expertise, equipment, software and their versions in their laboratories. Collaborators were then matched to a replication study when the equipment was readily available (e.g., Optojump™) and the study topic was in the researcher's area of expertise. If more than one study was matched to a collaborator, one study was allocated using a random number generator. Allocation was completed by J.M. only. Collaborators then reviewed their allocated study for feasibility and accepted or rejected the study. Feasibility checks included reviewing the type and number of participants to be recruited (i.e., could they access the population?), access to equipment and availability of laboratory hours at their university, etc. We emphasised that rejection was only possible based on feasibility and not because of personal preference for any study type or topic, but this required faith and trust in our collaborators to adhere to this. The process of study allocation was repeated if a different study was needed. Only a subset of studies from the larger study pool could be replicated in line with the number of collaborators who volunteered for this project. Ethical approval was obtained at each local university.

2.3 Replication Study Preparation

2.3.1 Contacting of Original Author

As we aimed to conduct close replications, methods for the replication study were based closely on the original study, with any differences being those that were unavoidable e.g., a new sample or different equipment [35]. When a study was a potential match, the corresponding original author was contacted to inform them of a potential replication attempt and to maximise replication quality. At this stage, we requested the deidentified raw data for the specific dependent variable of interest and any other materials deemed important to the replication e.g., statistical code. Another author was contacted (the last author) where possible if the corresponding author did not respond. Later, when the study was accepted for replication by the collaborator, we contacted the original authors again if raw data had not been provided to ask for full details of test statistics where they were not reported (e.g., *t*-values or *F*-values, degrees of freedom and exact *p*-values) and further methodological details where necessary. Each replication study was also individually preregistered on the *Open Science Framework* (overall project page: https://doi.org/10.17605/OSF.IO/3VUFG). We contacted the original authors again to provide them with an opportunity to review or express concerns about the preregistration. Any concerns from the original authors were discussed amongst the replication team and applied to the replication study where possible while trying to maximise replication quality and minimise deviation from the original published protocols. Our collaborating researchers were entitled to publish the individual replication study as an independent study and were required to contact the original authors to provide them with an opportunity to review the manuscript before submission.

2.3.2 Statistical Power Calculations

Multiple methods for sample size calculation were used and are fully detailed in the formalised selection protocol [34]. Briefly, we aimed to adjust for uncertainty around the original effect size point estimate and potential publication bias by using the *BUCSS R* package [36, 37]. However, this method was not always possible, as it does not give an output (infinite) if publication bias is deemed too high. As a result, we also calculated the replication sample size using the observed effect size from the original study or the lower limit of the observed effect size confidence interval at power ≥ 95%. Lastly, if the other methods could not be used, or the replication sample size was calculated as smaller than the original sample size, the original sample size was simply doubled. *R* files for all power analyses and sample size calculations (including justification of the method chosen) are available on the *Open Science Framework* (overall project page: https://doi.org/10.17605/OSF.IO/3VUFG). In addition, preregistrations for each replication study, along with supplementary materials, the study screening survey and a list of selected studies, are also available (https://doi.org/10.17605/OSF.IO/SFBVA).

2.4 Data Management and Analysis

A multiple inferential strategy was employed to analyse the replication data. Despite the ongoing debates about the strengths and weaknesses of NHST as a statistical inference procedure [38–41], it remains the most commonly used procedure in sports and exercise science. Vote counting is one of the main methods to assess replication outcomes using NHST [42]; therefore, this method was used to determine whether the replication effect size was statistically significant, and in the same direction as the original effect size. In other words, the same statistical analysis was applied in the replication study as in the original study, when statistical assumptions were met (e.g., normality), at the original study's alpha level. In cases where replication data were not normal, we visually inspected boxplots and computed the interquartile range. Any extreme outliers were removed from the dataset, and the final replication sample size (Tables 1, 2) reflects the sample size after the removal of outliers. The vote counting method can result in an exaggeration of replication failures when solely used to assess replication outcomes [42]; therefore, we also compared effect size estimates to assess the potential inflation of those estimates as a result of small sample size and bias [21, 43].

Assessing the direction of both the original and replication effect sizes is a simple technique to implement (i.e. to

Table 1 Original and replication study descriptives for the *t*-test studies

Original Study Title	Statistical Test	Original Data Availability	NHST			Statistical Power	Total Sample Size	
			Original *p*-value	Replication *p*-value	Comparison		Original	Replication
t-test								
Age-related differences in maximal and rapid torque characteristics of the hip extensors and dynamic postural balance in healthy, young and old females [45]	Independent t-test	Unavailable	0.011	0.546	Different	Completed short of target sample size (N = 26)	22	24
A single dose of oral ATP supplementation improves performance and physiological response during lower body resistance exercise in recreational resistance-trained males [46]	Paired t-test	Unavailable	0.005	0.515	Different	Completed short of target sample size (N = 22)	11	15
Effect of a compressive garment on kinematics of jump-landing tasks [47]	Paired t-test	Available	0.002	0.002	Similar	Completed with required sample size	27	54
Caffeine increases strength and power performance in resistance-trained females during early follicular phase [48]	Paired t-test	Available	< 0.001	0.028	Similar	Completed short of target sample size (N = 29)	15	21
The effects of a heel wedge on hip, pelvis and trunk biomechanics during squatting in resistance trained individuals [49]	Paired t-test	Unavailable	< 0.001	< 0.001	Similar	Completed with required sample size	14	29
Effects of preferred vs. nonpreferred music on resistance exercise performance [50]	Paired t-test	Unavailable	0.005	0.545	Different	Completed with required sample size	12	28
Acute enhancement of jump performance, muscle strength, and power in resistance-trained men after consumption of caffeinated chewing gum [51]	Paired t-test	Available	< 0.001	0.064	Different	Completed with required sample size	19	38
Passive recovery promotes superior performance and reduced physiological stress across different phases of short-distance repeated sprints [52]	Paired t-test	Unavailable	0.002	< 0.001	Similar	Completed with required sample size	9	16
Associated ACL risk factors differences during an unanticipated volleyball blocking movement [53]	Paired t-test	Unavailable	0.046	0.593	Different	Completed short of target sample size (N = 24)	12	12
Effect of consecutive jumping trials on metatarsophalangeal, ankle, and knee biomechanics during take-off and landing [54]	Paired t-test	Unavailable	< 0.001	< 0.001	Similar	Completed with required sample size	18	39
The effects of caffeine on vertical jump height and execution in collegiate athletes [55]	Paired t-test	Unavailable	0.001	0.58	Different	Completed with required sample size	25	39

ACL anterior cruciate ligament, *ATP* adenosine triphosphate, *NHST* null hypothesis significant testing

Table 2 Original and replication study descriptives for the *F*-test studies

Original Study Title	Statistical Test	Original Data Availability	NHST			Statistical Power	Total Sample Size	
			Original *p*-value	Replication *p*-value	Comparison		Original	Replication
F-test								
A real-time feedback method to reduce loading rate during running: effect of combining direct and indirect feedback [56]	1-way repeated measures ANOVA	Unavailable	< 0.01	< 0.001	Similar	Completed short of target sample size (N = 42)	20	22
Changes in bench press velocity and power after 8 weeks of high-load cluster- or traditional-set structures [57]	2 x 2 mixed ANOVA	Available	0.041	0.078	Different	Completed with required sample size	14	25
Does mental fatigue negatively affect outcomes of functional performance tests? [58]	2 x 2 repeated measures ANOVA	Unavailable	0.028	0.294	Different	Completed with required sample size	14	30
Beep test performance is influenced by 30 minutes of cognitive work [59]	1-way repeated measures ANOVA	Unavailable	< 0.01	0.021	Similar	Completed with required sample size	13	27
Voluntary muscle relaxation can mitigate fatigue and improve countermovement jump performance [60]	2 x 10 x 2 mixed ANOVA	Unavailable	0.016	0.049	Similar	Completed with required sample size	11	56
Maximizing performance: augmented feedback, focus of attention, and/or reward? [61]	1-way repeated measures ANOVA	Available	< 0.001	0.612	Different	Completed with required sample size	18	40
Altering work to rest ratios differentially influences fatigue indices during repeated sprint ability testing [62]	1-way repeated measures ANOVA	Unavailable	< 0.001	< 0.001	Similar	Completed with required sample size	8	21
Investigating the effects of maximal anaerobic fatigue on dynamic postural control using the Y-balance test [63]	1-way repeated measures ANOVA	Unavailable	0.025	0.002	Similar	Completed with required sample size	20	39
Effect of different loading conditions on running mechanics at different velocities [64]	3 x 3 repeated measures ANOVA	Unavailable	< 0.001	0.008	Similar	Completed with required sample size	13	24
Effect of attentional focus strategies on the biomechanical performance of the drop jump [65]	3 x 2 repeated measures ANOVA	Unavailable	0.018	< 0.001	Similar	Completed with required sample size	17	55
Further evidence for an external focus of attention in running: looking at specific focus instructions and individual differences [66]	1-way repeated measures ANOVA	Available	< 0.001	0.193	Different	Completed with required sample size	30	45
Enhanced expectancies facilitate golf putting [67]	2 x 5 mixed ANCOVA	Unavailable	0.032	0.614	Different	Completed with required sample size	34	68
Carbohydrate mouth rinse improves morning high-intensity exercise performance [68]	1-way repeated measures ANOVA	Available	0.002	0.007	Similar	Completed short of target sample size (N = 31)	12	18
Side-to-side differences in lower extremity biomechanics during multidirectional jump landing in volleyball athletes [69]	2 x 4 repeated measures ANOVA	Unavailable	< 0.001	< 0.001	Similar	Completed with required sample size	21	42

NHST null hypothesis significant testing

answer the question: are both effect size estimates in the same direction?), but quite limited for evaluating replicability. Consequently, we quantitatively compared the effect size estimates using *Z*-tests in the *TOSTER R* package (version 0.8.3) to measure compatibility between the original and replication study when the same underlying effect was being measured [44]. For this, the original and replication effect sizes were converted into a *z*-score, and a one-tailed *p*-value was computed to determine whether the original effect size was significantly larger than the replication (alpha = 0.05). In cases where the reported effect size was not appropriate for the study design (e.g., a Cohen's d_{av} was reported instead of Cohen's d_z for a paired design), or where we could not reproduce the original effect size, we calculated the effect size for the original study using the reported information (means, test statistics, sample size and degrees of freedom). We then computed the *Z*-test to compare the appropriate effect sizes (Eq. 1) e.g., the calculated original d_z versus the replication d_z but also compared the reported effect size to the appropriate replication effect size (see supplementary materials on https://doi.org/10.17605/OSF.IO/SFBVA).

$$z = \frac{d_{\text{original}} - d_{\text{replication}}}{\sqrt{\text{SE}^2_{d_{\text{original}}} + \text{SE}^2_{d_{\text{replication}}}}}. \quad (1)$$

In accordance with the guidelines of Brandt et al. [35], we considered a study successfully replicated when it was significantly different from the null (i.e., $p < 0.05$ for the *t*-test or *F*-test) and the effect sizes were not significantly different and were compatible ($p > 0.05$ for the *Z*-test). Otherwise, studies were classified as an informative failure to replicate (either not different from null or in the opposite direction from the original, and a replication effect size that was significantly different from the original effect size), a practical failure to replicate (both significantly different from the null and a replication effect size that was significantly different from the original effect size) or inconclusive (neither significantly different from null, and replication and original effect sizes that were not significantly different). All data, code and analyses are available online (https://doi.org/10.17605/OSF.IO/SFBVA).

Finally, we completed a post-replication recipe to report differences between the original and replication studies [35], and created tables of methodological differences between original and replication studies, all of which are available on the individual replication project pages (accessed through

the overall project here: https://doi.org/10.17605/OSF.IO/3VUFG). On the basis of these, we present short notes of interest for each replication study which inform our subjective assessment of replication quality, rated as "poor", "moderate" and "good".

3 Results

3.1 Screening

We screened 9385 studies, and 638 abstracts met our selection criteria [34]. On further review of the full text, 587 articles were included, and 51 were excluded, as they did not meet the criteria of the selection protocol (e.g. a linear mixed model was conducted, and that was not immediately obvious from the review of the abstract). The final pool of studies was divided into sub-disciplines (applied sports and exercise biomechanics, psychology, physiology, nutrition and injury prevention), and from this stratified pool of studies, a study was randomly selected for allocation to the collaborator for replication when the equipment was readily available to them, and the discipline was matched to their expertise. Our specific methods for this selection protocol have been described previously [34].

3.2 Completed Replications

Of 189 collaborators who expressed interest, 33 collaborators began the process of conducting replication studies. One dropped out due to equipment malfunction, one could not make the data collection deadline and two did not follow the original protocols exactly. Therefore, 29 finished data collection. Of these completed replications, 19 reached the requested sample size, and 6 were short of this sample size but greater than the original sample size. However, 4 replication sample sizes were smaller than the original sample size and were removed from the analysis. In total, 25 replications were analysed. The mean replication sample size was $n = 33$, while the mean original sample size was $n = 17$.

3.3 Original Author Contact, Data Sharing and Reporting

We contacted a total of 156 original authors throughout the selection process for deidentified raw data when they were a potential match to collaborators. Of those, 14% ($n = 21$) shared data. Of the 29 completed replication studies, 24% ($n = 7$) of the original study authors shared

Table 3 Effect size estimates (Cohen's *d*)

Original Study Title	Reported Effect Size		Effect Size Magnitudes			Effect Size 95% Confidence Intervals	
	Original	Type	Original	Replication	Replication vs. Original	Original	Replication
Cohen's ds							
Age-related differences in maximal and rapid torque characteristics of the hip extensors and dynamic postural balance in healthy, young and old females [45]	1.05	Cannot reproduce	1.20	0.25	smaller	0.27, 2.10	-0.56, 1.05
Cohen's dz							
A single dose of oral ATP supplementation improves performance and physiological response during lower body resistance exercise in recreational resistance-trained males [46]	0.73	dav	1.08	0.18	smaller	0.31, 1.82	-0.35, 0.70
Effect of a compressive garment on kinematics of jump-landing tasks [47]	0.58	dav	0.65	-0.44	smaller	0.23, 1.06	-0.72, -0.16
Caffeine increases strength and power performance in resistance-trained females during early follicular phase [48]	0.27	Hedges g	1.83	0.52	smaller	0.98, 2.66	0.05, 0.97
The effects of a heel wedge on hip, pelvis and trunk biomechanics during squatting in resistance trained individuals [49]	None	None	1.13	1.82	larger	0.44, 1.79	1.21, 2.41
Effects of preferred vs. nonpreferred music on resistance exercise performance [50]	0.84	dav	1.01	-0.12	smaller	0.29, 1.70	-0.49, 0.26
Acute enhancement of jump performance, muscle strength, and power in resistance-trained men after consumption of caffeinated chewing gum [51]	0.27	ds	1.51	0.31	smaller	0.83, 2.16	-0.02, 0.63
Passive recovery promotes superior performance and reduced physiological stress across different phases of short-distance repeated sprints [52]	1.17	dav	1.50	1.48	smaller	0.51, 2.45	0.75, 2.19
Associated ACL risk factors differences during an unanticipated volleyball blocking movement [53]	0.30	dav	0.65	0.16	smaller	0.01, 1.26	-0.41, 0.72
Effect of consecutive jumping trials on metatarsophalangeal, ankle, and knee biomechanics during take-off and landing [54]	1.35	dav	1.21	0.78	smaller	0.59, 1.82	0.41, 1.13
The effects of caffeine on vertical jump height and execution in collegiate athletes [55]	None	None	0.75	0.09	smaller	0.30, 1.19	-0.23, 0.40

Effect size magnitudes refer to the effect size estimates recalculated by the replication team. Cohen's d_s, d_z and d_{av} differ on the basis of the calculation of the denominator. The denominator for Cohen's d_s is the pooled standard deviation, for Cohen's d_z is the standard deviation of the difference scores and for Cohen's d_{av} is the average standard deviation [70]

ACL anterior cruciate ligament, *ATP* adenosine triphosphate

Table 4 Effect size estimates (partial eta squared)

Original Study Title	Reported Effect Size		Effect Size Magnitudes			Effect Size 95% Confidence Intervals	
	Original	Type	Original	Replication	Replication vs. Original	Original	Replication
Partial eta squared							
A real-time feedback method to reduce loading rate during running: effect of combining direct and indirect feedback [56]	0.56	Partial eta squared*	0.834	0.473	smaller	0.546, 0.913	0.241, 0.637
Changes in bench press velocity and power after 8 weeks of high-load cluster- or traditional-set structures [57]	None	Partial eta squared	0.304	0.129	smaller	0.000, 0.675	0.000, 0.423
Does mental fatigue negatively affect outcomes of functional performance tests? [58]	0.344	Partial eta squared	0.344	0.038	smaller	0.000, 0.702	0.000, 0.256
Beep test performance is influenced by 30 minutes of cognitive work [59]	0.67	Partial eta squared	0.670	0.188	smaller	0.214, 0.875	0.002, 0.471
Voluntary muscle relaxation can mitigate fatigue and improve countermovement jump performance [60]	0.496	Partial eta squared*	0.456	0.070	smaller	0.019, 0.786	0.000, 0.239
Maximizing performance: augmented feedback, focus of attention, and/or reward? [61]	0.348	Partial eta squared	0.348	0.015	smaller	0.155, 0.477	0.000, 0.075
Altering work to rest ratios differentially influences fatigue indices during repeated sprint ability testing [62]	0.751	Partial eta squared	0.751	0.409	smaller	0.374, 0.890	0.128, 0.632
Investigating the effects of maximal anaerobic fatigue on dynamic postural control using the Y-balance test [63]	None	Partial eta squared	0.376	0.159	smaller	0.000, 0.619	0.025, 0.320
Effect of different loading conditions on running mechanics at different velocities [64]	0.901	Partial eta squared*	0.814	0.205	smaller	0.605, 0.904	0.017, 0.427
Effect of attentional focus strategies on the biomechanical performance of the drop jump [65]	0.223	Partial eta squared	0.223	0.287	larger	0.007, 0.458	0.143, 0.424
Further evidence for an external focus of attention in running: looking at specific focus instructions and individual differences [66]	0.280	Partial eta squared	0.271	0.035	smaller	0.113, 0.423	0.000, 0.100
Enhanced expectancies facilitate golf putting [67]	0.140	Partial eta squared	0.140	0.010	smaller	0.000, 0.396	0.000, 0.110
Carbohydrate mouth rinse improves morning high-intensity exercise performance [68]	0.468	Partial eta squared	0.468	0.280	smaller	0.091, 0.723	0.028, 0.535
Side-to-side differences in lower extremity biomechanics during multidirectional jump landing in volleyball athletes [69]	None	Partial eta squared	0.258	0.855	larger	0.054, 0.432	0.782, 0.900

[a]Stated as partial eta squared but cannot reproduce

data. If the original authors did not provide the raw data, we asked for test statistics; however, no response was received. We additionally contacted the original authors to provide them with an opportunity to review or express concerns about the preregistration; 31% ($n=9$) reviewed the preregistrations, and of these, 56% ($n=5$) approved the preregistration, and 44% ($n=4$) expressed concerns about the replication study.

Of the 25 analysed replications, 10 used paired *t*-tests, 1 used an independent *t*-test and 14 used an ANOVA for the statistical analyses (we focused on the main effects of the factorial designs). In the original studies ($N=25$), 48% reported the test statistic; otherwise, 16% were calculated using the data provided by the original authors, and 36% were estimated on the basis of reporting information in the original study. For the degrees of freedom, 36% were reported; otherwise, 16% were calculated using the provided data, and 48% were estimated. Lastly, 68% reported effect size point estimates, 12% were calculated using original data, and 20% were estimated.

3.4 Replication Outcomes

As stated in our selection protocol [34], we selected original studies with statistically significant findings. For the replication NHST outcomes, 56% ($n=14$) were significant, similar to the original studies, and 44% ($n=11$) were not significant (Tables 1, 2).

In the 10 original studies that conducted paired *t*-tests, we calculated all of the Cohen's d_z, as 5 original studies reported a Cohen's d_{av}, 2 studies did not report an effect size estimate, 1 study reported a Cohen's d_s and 1 could not be reproduced. One study reported a Hedges *g*, a bias-corrected effect size, but we calculated Cohen's d_z for the *Z*-test comparison. We estimated the Cohen's d_s for the independent *t*-test, as we could not reproduce the original reported effect size. For the ANOVAs, we calculated 6 partial eta squared, as we could not reproduce 3 that were published, and 3 did not report any partial eta squared at all (Tables 3, 4).

For the 25 replication studies, 88% ($n=22$) of the original effect sizes regressed to smaller values when replicated. For these replication studies ($n=25$), the median percentage decrease in magnitude from the original to the replication effect size estimate was 75%. Our *Z*-test results showed that 64% ($n=16$) of the replication effect size estimates were not statistically compatible with the original, and 36% ($n=9$) were compatible (Figs. 1, 2).

Therefore, as per the classifications in Brandt et al. [35], 28% ($n=7$) of the replications were successful, 36% ($n=9$) were informative failures to replicate, 28% ($n=7$) were practical failures to replicate and 8% ($n=2$) were inconclusive (Tables 5, 6).

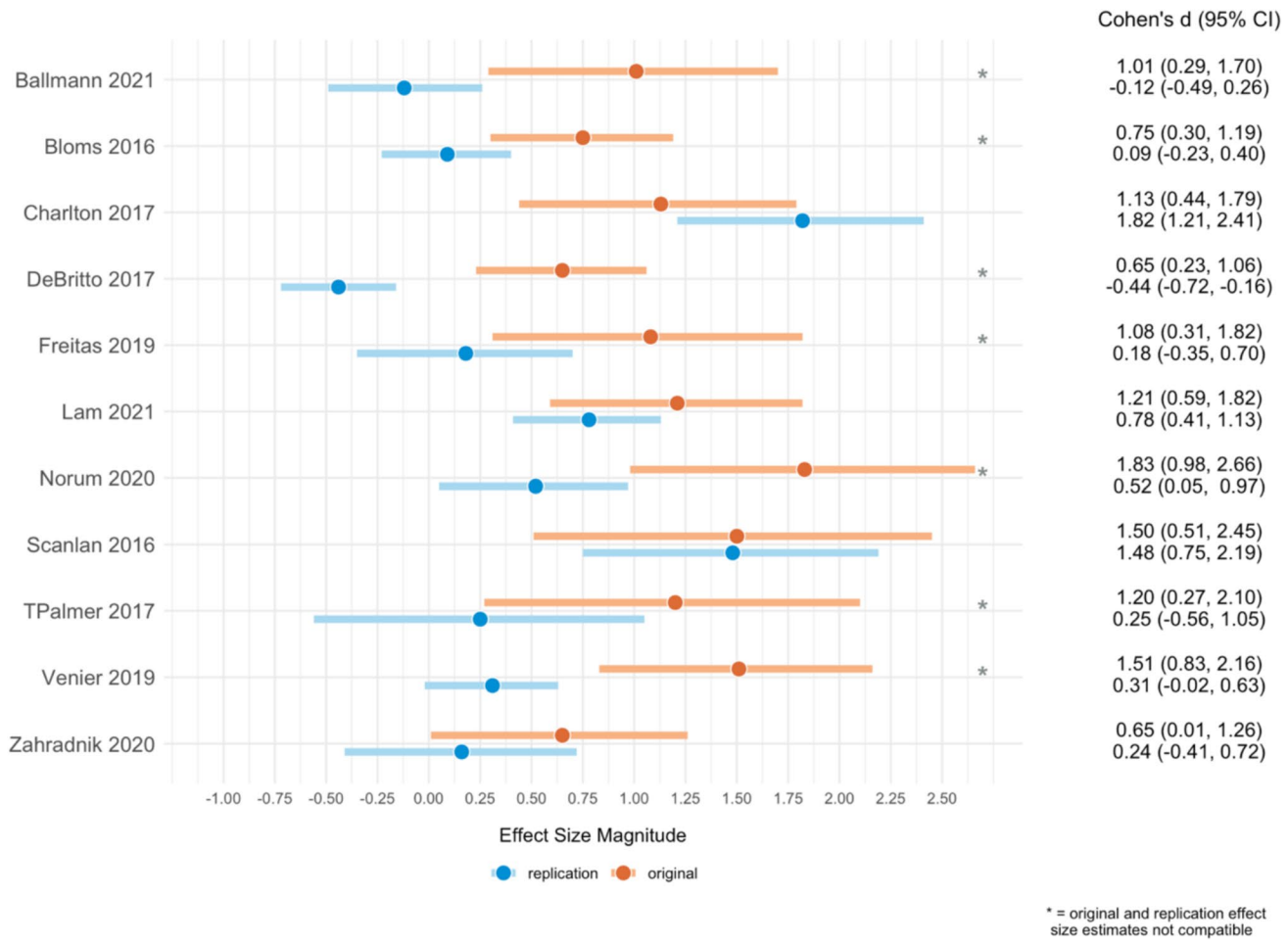

Fig. 1 Plot showing original and replication Cohen's *d* magnitude and confidence intervals

4 Discussion

This study aimed to provide an initial estimate of the replicability of sports and exercise science studies published between 2016 and 2021 in quartile 1 journals. We hypothesised that 61% of our studies could be expected to yield another significant effect given the previously published expected replication rate [22]. Considering our NHST-only results, 56% of the 25 analysed replication studies had a statistically significant *p*-value similar to the original studies. Yet, the effect size comparison via the *Z*-tests provides more context; only 36% of replication and original effect size estimates were compatible. When considering overall outcomes where they had both a significant NHST outcome and compatible effect sizes [35], 28% of replications were successful.

This pioneering project in sports and exercise science was the first large replication project in the field and the first to publish a transparent, randomised protocol for minimising bias in the selection of studies to replicate. The empirical data support subjective concerns about replication, reproducibility and transparency in sports and exercise science, as 78% of surveyed researchers believe there is a replication and reproducibility crisis in our field [28]. It also supports previous research implying we should have a healthy scepticism of our published literature, as it potentially includes a substantial number of false-positive findings [20]. Therefore, we echo calls for an immediate need to increase research transparency in our field [23, 71]. Many of the current research practices in sports and exercise science potentially contributed to the low replication rate e.g., publishing an excess of significant findings [20, 22, 25], using small sample sizes [21, 43, 72], poor reporting practices [22] and stating vague hypotheses [20], which will be discussed. Although issues with the current research practices have been identified already, replication has historically been undervalued in our field, and replication attempts are rare [20]. Perhaps the empirical data and low replication rate here will finally be the catalyst for change that is sorely needed in sports and exercise science.

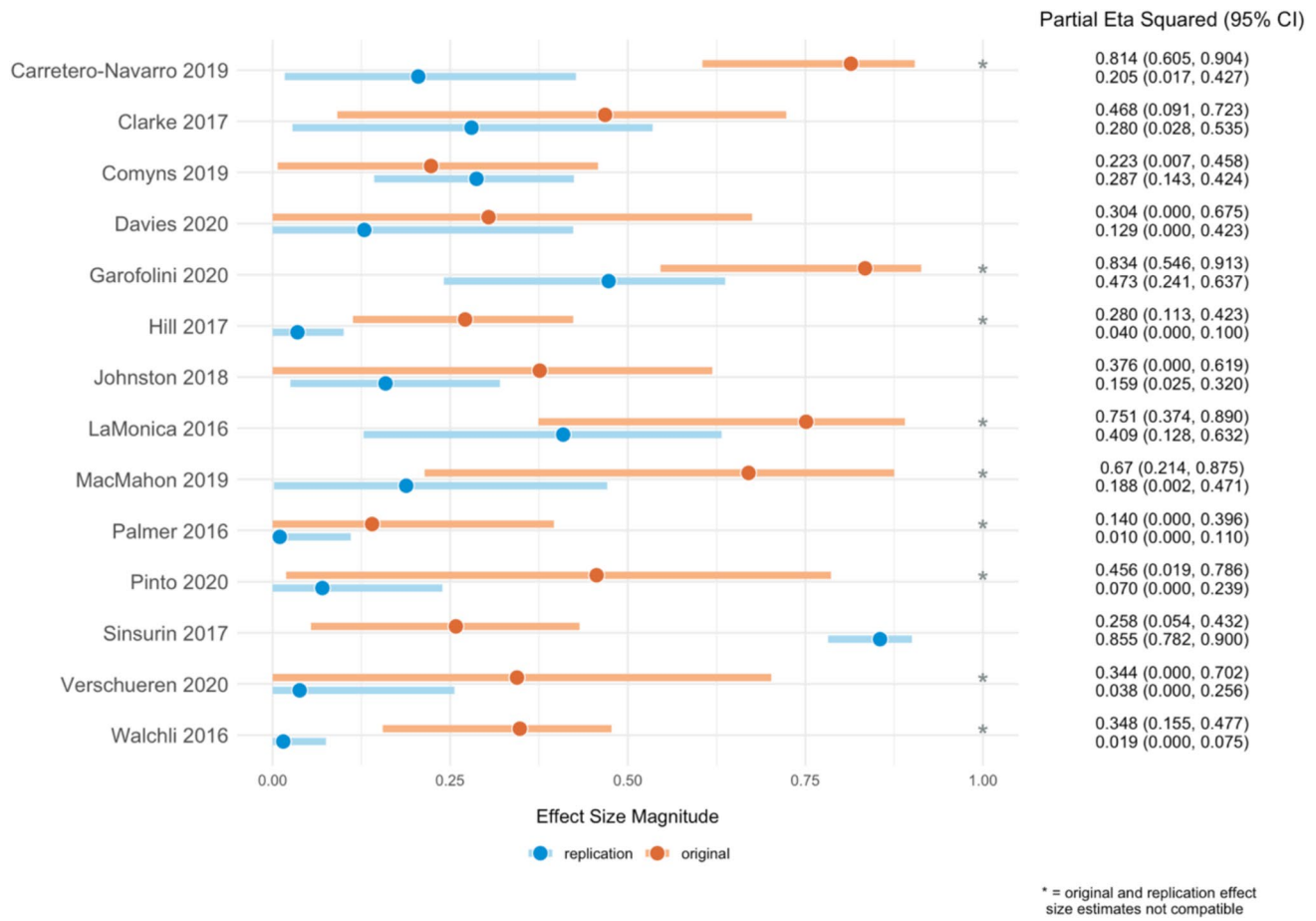

Fig. 2 Plot showing original and replication partial eta squared magnitude and confidence intervals

The decrease in the magnitude of our replicated effect sizes from the originally reported effect sizes is a major concern. Replication effect sizes are expected to regress to their true values, as many original effect sizes tend to be inflated due to small sample sizes and publication bias [43]; these sample sizes are more likely to be affected by variation in the sampled data and other moderators (noise) even when designing replication studies close to the original [73]. We found that 88% of our replication effect sizes decreased in magnitude compared with the original effect sizes, with the median percentage decrease in magnitude from the original to the replication effect size estimate equal to 75% for all of the replication effect sizes (median of 77% for Cohen's d_z, 79% for Cohen's d_s and 65% for partial eta squared). We expected some effect size regression, as the Reproducibility Project: Psychology reported the magnitude of the replication effect size estimates was approximately half that of the original effect size estimates [7]. However, the reduction in the magnitude of effect sizes was much larger in our study. The reported use of small sample sizes and the estimated low statistical power in sports and exercise science indicated that many effects might be much smaller in magnitude than originally published [21, 74]. The mean sample size was 17 participants across the original studies selected for replication, and most of the original effect sizes were considered large as per typical effect size threshold guidelines (all of the partial eta squared values were ≥ 0.14, and 73% of Cohen's d values were > 0.8). Therefore, given the magnitude of the original effect size estimates and the original sample sizes, it is unsurprising that 64% of replication effect sizes were smaller than the original and were statistically incompatible. However, the magnitude of the difference in the effect size estimates was substantial, and this should be considered when using a published effect size in a power analysis for a subsequent study. Otherwise, statistical power based on this point estimate will be much lower than intended [36]. This regression towards smaller values also affects meta-analyses, as inflated effect sizes will impact the data presented. Sports and exercise scientists should therefore assume a large degree of uncertainty in published effect size estimates rather than assuming they are fixed or certain since most replication studies indicated substantial reductions in the effect size compared with the original.

Table 5 Overall replication outcomes and interpretations (Cohen's *d*)

Original Study Title	Replication NHST Outcome	Effect Size		Z-test			Overall Outcome[l]
		Original	Replication	*z*-score	*p*-value	Outcome	
Cohen's ds							
Age-related differences in maximal and rapid torque characteristics of the hip extensors and dynamic postural balance in healthy, young and old females [45]	Non-significant	1.20	0.25	2.16	0.02	Incompatible	Informative failure to replicate (not different from the null, or in the opposite direction and significantly different from the original)
Cohen's dz							
A single dose of oral ATP supplementation improves performance and physiological response during lower body resistance exercise in recreational resistance-trained males [46]	Non-significant	1.08	0.18	1.94	0.026	Incompatible	Informative failure to replicate (not different from the null, or in the opposite direction and significantly different from the original)
Effect of a compressive garment on kinematics of jump-landing tasks [47]	Significant	0.65	-0.44	4.27	< 0.001	Incompatible	Practical failure to replicate (both significantly different from the null and from the original),
Caffeine increases strength and power performance in resistance-trained females during early follicular phase [48]	Significant	1.83	0.52	2.73	0.012	Incompatible	Practical failure to replicate (both significantly different from the null and from the original),
The effects of a heel wedge on hip, pelvis and trunk biomechanics during squatting in resistance trained individuals [49]	Significant	1.13	1.82	-1.51	0.934	Compatible	Success (different from the null, and similar to or larger than the original and in the same direction)
Effects of preferred vs. nonpreferred music on resistance exercise performance [50]	Non-significant	1.01	-0.12	2.80	0.003	Incompatible	Informative failure to replicate (not different from the null, or in the opposite direction and significantly different from the original)
Acute enhancement of jump performance, muscle strength, and power in resistance-trained men after consumption of caffeinated chewing gum [51]	Non-significant	1.51	0.31	3.20	0.001	Incompatible	Informative failure to replicate (not different from the null, or in the opposite direction and significantly different from the original)
Passive recovery promotes superior performance and reduced physiological stress across different phases of short-distance repeated sprints [52]	Significant	1.50	1.48	0.03	0.489	Compatible	Success (different from the null, and similar to or larger than the original and in the same direction)
Associated ACL risk factors differences during an unanticipated volleyball blocking movement [53]	Non-significant	0.65	0.16	1.14	0.127	Compatible	Inconclusive (neither significantly different from the null nor the original)
Effect of consecutive jumping trials on metatarsophalangeal, ankle, and knee biomechanics during take-off and landing [54]	Significant	1.21	0.78	1.21	0.114	Compatible	Success (different from the null, and similar to or larger than the original and in the same direction)
The effects of caffeine on vertical jump height and execution in collegiate athletes [55]	Non-significant	0.75	0.09	2.39	0.008	Incompatible	Informative failure to replicate (not different from the null, or in the opposite direction and significantly different from the original)

ACL anterior cruciate ligament, *ATP* adenosine triphosphate

[a]Brandt et al. [35]

Table 6 Overall replication outcomes and interpretations (partial eta squared)

Original Study Title	Replication NHST Outcome	Effect Size		Z-test			Overall Outcome[l]
		Original	Replication	*z*-score	*p*-value	Outcome	
Partial eta squared							
A real-time feedback method to reduce loading rate during running: effect of combining direct and indirect feedback [56]	Significant	0.834	0.473	2.71	0.003	Incompatible	Practical failure to replicate (both significantly different from the null and from the original),
Changes in bench press velocity and power after 8 weeks of high-load cluster- or traditional-set structures [57]	Non-significant	0.304	0.129	1.07	0.14	Compatible	Inconclusive (neither significantly different from the null nor the original)
Does mental fatigue negatively affect outcomes of functional performance tests? [58]	Non-significant	0.344	0.038	2.49	0.006	Incompatible	Informative failure to replicate (not different from the null, or in the opposite direction and significantly different from the original)
Beep test performance is influenced by 30 minutes of cognitive work [59]	Significant	0.670	0.188	2.17	0.015	Incompatible	Practical failure to replicate (both significantly different from the null and from the original),
Voluntary muscle relaxation can mitigate fatigue and improve countermovement jump performance [60]	Significant	0.456	0.070	2.43	0.007	Incompatible	Practical failure to replicate (both significantly different from the null and from the original),
Maximizing performance: augmented feedback, focus of attention, and/or reward? [61]	Non-significant	0.348	0.015	7.71	< 0.001	Incompatible	Informative failure to replicate (not different from the null, or in the opposite direction and significantly different from the original)
Altering work to rest ratios differentially influences fatigue indices during repeated sprint ability testing [62]	Significant	0.751	0.409	1.96	0.025	Incompatible	Practical failure to replicate (both significantly different from the null and from the original),
Investigating the effects of maximal anaerobic fatigue on dynamic postural control using the Y-balance test [63]	Significant	0.376	0.159	1.65	0.05	Compatible	Success (different from the null, and similar to or larger than the original and in the same direction)
Effect of different loading conditions on running mechanics at different velocities [64]	Significant	0.814	0.205	4.56	< 0.001	Incompatible	Practical failure to replicate (both significantly different from the null and from the original),
Effect of attentional focus strategies on the biomechanical performance of the drop jump [65]	Significant	0.223	0.287	-0.63	0.735	Compatible	Success (different from the null, and similar to or larger than the original and in the same direction)
Further evidence for an external focus of attention in running: looking at specific focus instructions and individual differences [66]	Non-significant	0.271	0.035	5.70	< 0.001	Incompatible	Informative failure to replicate (not different from the null, or in the opposite direction and significantly different from the original)
Enhanced expectancies facilitate golf putting [67]	Non-significant	0.140	0.010	3.73	< 0.001	Incompatible	Informative failure to replicate (not different from the null, or in the opposite direction and significantly different from the original)
Carbohydrate mouth rinse improves morning high-intensity exercise performance [68]	Significant	0.468	0.280	1.13	0.13	Compatible	Success (different from the null, and similar to or larger than the original and in the same direction)
Side-to-side differences in lower extremity biomechanics during multidirectional jump landing in volleyball athletes [69]	Significant	0.258	0.855	-7.07	1	Compatible	Success (different from the null, and similar to or larger than the original and in the same direction)

[a]Brandt et al. [35]

Assessing the variability and uncertainty of published effect size estimates would be easier when the effect size and their confidence intervals are fully reported. Unfortunately, we had to compute the appropriate standardised effect size for the study design for 16 original studies, and 5 of these were conservatively estimated using the reported statistical information. Overall, 68% of the original studies reported some type of effect size, and this is a slightly lower reporting rate than the 79% reported by Twomey et al. [20], although they had a much larger sample size of 300 articles. Additionally, only 16% of the original studies here reported confidence intervals for either the standardised or unstandardised effect. All sports and exercise science researchers should fully report effect sizes and their confidence intervals (or, at minimum, the standard errors of the stated effect size) [75, 76]. This provides crucial information about the magnitude and uncertainty of observed effects, enabling readers to fully evaluate the data [17].

Issues with the reporting of effect size estimates also extended to reporting issues with the NHST framework itself. *F*-tests and *t*-tests are common statistical tests in sports and exercise science, but these tests are littered with errors in the literature [27]. For example, in one of the original studies herein, the analysis was reported as a two-way repeated measures ANOVA, but it was clear from the experimental design that the authors had conducted a mixed ANOVA. In terms of reporting quality, we found it to be poor, with only 48% reporting the test statistic and 36% reporting associated degrees of freedom. The lack of reporting of this information directly impacted the ability to evaluate methodological quality effectively [16], and this is a wider issue in sports and exercise science [77]. The lack of reporting of this critical information is worrisome considering the frequent use of frequentist statistical tests and the importance of such statistical tests to the study's interpretation. Human error is also a factor, as many researchers, including those in the sports and exercise science field, undertake data analysis themselves without consulting other authors or statisticians [26]. The issues related to replication could be mitigated if researchers made their data, code and materials publicly available for other researchers to evaluate, which would assist with error correction in the long run. If data were made openly available, authors would be free to reduce the amount of cumbersome detail reported in the manuscript in favour of providing greater detail in online repositories of their data and analyses. Overall, the prevalence of errors and omitted information had a substantial impact on the ability to replicate studies within this project, and we encourage far better sharing of data and analyses in the future.

Our focus on replicating statistically significant findings for this project [34] was partly due to the high publication rate of significant findings in our field; Büttner et al. [24] reported an 82% positive result rate for sports medicine and physiotherapy, and Twomey et al. [20] reported an 81% for sports and exercise science. These rates could hypothetically be true if studies were appropriately planned to thoroughly inform hypotheses, the proportion of true hypotheses amongst all tested hypotheses was high, and the statistical power for each study was high. Yet, the studies would need more than 90% statistical power if all hypotheses tested were true to make this rate plausible [78]. Although many other questionable research practices and statistical errors can inflate the percentage of statistically significant effects, the high positive results rate is mostly likely facilitated by selectively reporting or "cherry-picking" desirable results, typically those results where $p < 0.05$, for maximum impact and publication potential i.e., publication bias. Publication bias is an observed phenomenon in sports and exercise science [22, 25] and is perpetuated by both researchers and journal publishers [28]. Consequently, publication bias could partially explain the low replication rate in this project. The replication of a study is difficult in cases of selective reporting or publication bias, as the original *p*-value might have passed the significance threshold at the upper tail of the *p*-value distribution [79]. Thus, a replication study with higher statistical power results in a non-significant *p*-value, thereby, non-replication. Publication bias in our field has possibly resulted in a published literature body of overinflated effect size estimates which regress by a median of 75% when replicated. However, replication is not normally attempted in sports and exercise science, and published claims are then "canonized" [80], leading to an accumulation of false claims in the literature that are considered irrefutable facts. When we combine this phenomenon with the inability to self-correct the literature because of poor data sharing, we likely create a knowledge base that fails to progress meaningfully.

This replication project sought to assess the validity of published findings in sports and exercise science, recognising that replication can either corroborate robust results or highlight those in need of re-evaluation. However, a study's validity is often jeopardised before statistical analysis even begins due to flawed study design and the absence of a well-formulated hypothesis [27]. A key principle of NHST is the statements of the null hypothesis and alternative hypothesis, where the alternative hypothesis is sufficiently detailed to make it statistically falsifiable rather than just a complement of the null hypothesis [81]. Yet, falsifiability is rarely considered in hypothesis testing within sports and exercise science, as a substantial proportion of studies fail to formulate testable hypotheses and proceed to use NHST anyway [20]. In fact, hypotheses are often so vaguely stated in our field that any result could be spun to support the hypothesis due to researcher flexibility [20, 24, 82]. While any dataset can yield a significant finding, it may be a false positive, difficult to replicate and not representative of a true phenomenon [83]. Unfalsifiable original hypotheses may contribute to the

low replication rate observed in this project. Additionally, such hypotheses make non-replications difficult to interpret, hindering knowledge advancement and theory development. The misinterpretation of the logic of hypothesis tests, their assumptions and the explicit control of type 1 and type 2 errors highlights a lack of understanding of experimental design and statistical practices amongst sports and exercise science researchers. Despite the field-wide acknowledgement of this [28], we rarely seek the help of statisticians [17] and continue to undertake flawed inferences, resulting in subsequently flawed published claims [84].

The universal norms of science include openness and rigour; therefore, sharing analytical materials (i.e., open data and code) is an essential step of an open and transparent scientific process [85], which we have already mentioned. Again, we want to emphasise that this practice allows researchers or reviewers to "trust but verify" [86], identify and correct errors, and facilitate secondary data analysis to update statistical models or extend knowledge of the topic under investigation [87]. This also helps facilitate meta-analysis and other evidence-synthesis endeavours [88]. However, most sports and exercise science manuscripts do not include open data and code. Borg et al. [23] observed that only 4% of 299 studies shared data, and no study shared any code or syntax related to the statistical analysis. Another team reported less than 1% of studies shared data [20]. The data sharing rate for this study was slightly higher albeit with a much smaller sample; we contacted 156 original authors for raw data and 14% shared original data, while 24% of the 29 completed replications shared data. When the original authors did not respond to requests for deidentified raw data, we asked them for complete statistical test details i.e., *t*-values or *F*-values, degrees of freedom, mean differences, etc., so that we could more accurately estimate effect sizes. None of the original authors provided this information. The level of data sharing in sports and exercise science appears to be considerably lower than in other fields [89, 90]. Data concealment can facilitate poor statistical analyses, manipulation, or selective reporting of results and a higher rate of statistical errors [89]. Although there can be concerns about data sharing, there are solutions to this [23]. There is also a need for a culture change to normalise that errors are part of the scientific process and should be communicated to original authors respectfully [91], but this can be delicate.

Lastly, the field of sport and exercise science faces unique challenges in advancing scientific knowledge due to its comparatively limited funding relative to other biomedical disciplines. This resource constraint magnifies the importance of research quality and methodological rigour. Enhancing transparency in research practices and eliminating questionable methodologies are not merely beneficial but essential for maximising the impact of available funding. By implementing more robust research practices, the field can ensure that limited resources are optimally utilised to generate reliable, replicable findings that meaningfully advance our understanding of human performance.

5 Limitations

Our methodology stated that we aimed for 95% power, but it is far from realistic to assume we achieved this. In many of the replication studies we doubled the original sample size, and this undoubtedly led to underpowered replication studies; e.g., if an original study had a *p*-value of 0.03, the replication power is estimated to be only 50% when using this method [92]. In addition to the statistical power being less than intended in the replication studies, there are other limitations in this project. More than half of the studies in this pool did not replicate, and although we aimed to provide an initial estimate of the replicability of sports and exercise science research, this is certainly not a representative estimate. Firstly, 25 studies from the thousands of published sports and exercise science research could never possibly result in accurate estimates of replicability for the field. Secondly, much information had to be conservatively estimated e.g., original effect size estimates or test statistics (*t*-values or *F*-values) for analyses. This is not necessarily a limitation of the replication project itself—if we only selected original studies with all available information, this might have biased us towards higher-quality studies with better reporting—yet, it does affect the overall outcomes. When we had to conservatively estimate effect sizes where they were not reported in the original study, we calculated the smallest plausible effect size on the basis of the reported relative *p*-value, meaning that the original effect may have been larger than we estimated. Additionally, the comparison between ANOVA effect sizes i.e. partial eta-squared, is only an approximation based on Fisher's *z* transformation, which simulations indicated preserved type 1 error well. Therefore, the outcome of the *Z*-tests here may be more conservative than the true original values.

In our selection protocol, we stated that we would not try to improve any methods so that we could attempt to replicate the same theoretical dimensions of the original study. This potentially means we could have taken poorly designed original studies and created poorly designed replication studies (Table 7). In some cases, a lack of methodological details in the original studies also made some aspects of the replication questionable. Critics will view this as a waste of resources, but the intention of our project was not to update knowledge on particular theories but to "simply" replicate what was published, in a way that minimised bias and maximised representativeness of applied sports and exercise science research.

Table 7 Subjective assessment of replication quality

Original Study Title	Replication Quality	Notes
Cohen's ds		
Age-related differences in maximal and rapid torque characteristics of the hip extensors and dynamic postural balance in healthy, young and old females [45]	Good	Original authors indicated that two balance tests were performed, each including three assessments, but did not indicate which test was used for analysis in the methods i.e., whether it was the mean of the tests or the best score. Mean values were used for the replication study analysis for overall stability index instead of the best value. However, the use of mean values did not change the overall replication outcome compared to the best values.
Cohen's dz		
A single dose of oral ATP supplementation improves performance and physiological response during lower body resistance exercise in recreational resistance-trained males [46]	Moderate	Very experienced lifters with excellent 1RMs in the replication study (replication half-squat 1RM = 168.2 ± 39.5kg vs. original = 127.9 ± 19.7kg). An alternative ATP supplement was used in the replication study but consisted of the same active ingredient (Peak ATP, TSI Group Co., Missoula, MT, USA) in the same dosage (400 mg denosine-5'-triphosphate disodium).
Effect of a compressive garment on kinematics of jump-landing tasks [47]	Moderate	Despite a similar sample across both studies, the SD of the difference in the original dataset for valgus angle was surprisingly small (SD = 0.09) compared to the replication (SD = 1.73). Secondly, the original study used a 15-camera motion analysis system and 100Hz force plate versus a 9 camera 60Hz system and Chronojump in the replication study. The force platform has a higher acquisition rate than a mocap system, so the original study's initial contact identification should be more accurate, but both methods of measurement are valid.
Caffeine increases strength and power performance in resistance-trained females during early follicular phase [48]	Moderate	The original study excluded all participants that were currently taking any ergogenic aid. The replication team did not think this was feasible given the difficulty of recruiting this population so instead assessed the use of ergogenic aids and only included participants if they abstained from use of ergogenic aids and ensured a suitable washout period before taking part in any of the main conditions. The replication study only measured 1RM back squat, therefore, participants may have been in a lesser state of fatigue at the point of 1RM back squat measurement compared to the original study.
The effects of a heel wedge on hip, pelvis and trunk biomechanics during squatting in resistance trained individuals [49]	Good	Participants self-reported 1RM was estimated to be 162.0 ± 31.2% of body mass in the replication study and 132.1 ± 33.4% of body mass in the original study.
Effects of preferred vs. nonpreferred music on resistance exercise performance [50]	Good	The original study mean bench press 1RM = 127.9kg while the replication mean bench press 1RM = 107.3kg.
Cohen's dz		
Acute enhancement of jump performance, muscle strength, and power in resistance-trained men after consumption of caffeinated chewing gum [51]	Poor	The replication was a single-blind study design in contrast to the original double-blind study due to the clear differences in appearance between the caffeinated and placebo chewing gum so that replicators were unable to blind the primary researcher. Thus, there was ineffective blinding in the replication study, however, this was similar to the original study's blinding procedures as they were also not fully effective. Therefore, it was unlikely that the ineffectiveness to blind replication participants affected the observed vertical jump findings.
Passive recovery promotes superior performance and reduced physiological stress across different phases of short-distance repeated sprints [52]	Moderate	The original study separated sprinting sessions by 24 hours on 2 consecutive days, but replication testing sessions were separated by 1 week to decrease the risk of injury from the maximal sprinting bouts. Dual beam timing lights (cooper fusion) were used in the original study to assess sprint times, but Witty timing gates were used in the replication study which are single beam. The original study stated that all participants must run at 95% or faster than their reference sprint. Despite participants maximally sprinting in all protocols in the replication study, seven did not meet this criterion for both of their sessions but the data was used for analyses anyway.
Associated ACL risk factors differences during an unanticipated volleyball blocking movement [53]	Moderate	The video was the same height across both studies but 1.5m away in the replication instead of 0.33m. In the original study, it was unclear whether a single force plate or two force plates was used to determine the > 20 N threshold for the first occurrence of the vertical ground reaction force. The replication analysis identified first contact as the moment when the force plate under the right foot (the limb presented in the article) was > 20 N. If the threshold included both plates being > 20 N, this would lead to potentially larger forces on the right leg if the right leg makes contact first, particularly during the tilt block condition.
Effect of consecutive jumping trials on metatarsophalangeal, ankle, and knee biomechanics during take-off and landing [54]	Moderate	For the original study, participants wore standard test shoes (Li Ning Wade Cloud, Beijing, China) which are Basketball specific with a low collar for ankle protection. The replication study used a model that was not Basketball specific but used for indoor sports like Volleyball and Handball without a collar for the ankle. The replication study included Basketball players from different schools or universities resulting in a potential difference in the playing level of participants compared to the original sample.
The effects of caffeine on vertical jump height and execution in collegiate athletes [55]	Moderate	There were differences in sporting background in the replication sample (10 Basketball, 2 Cross Country, 6 Volleyball, 14 Soccer, 5 Softball, 3 Tennis and 5 Rowing) compared to the original sample (6 Track Throwers, 4 Jumpers, 6 Sprinters, 5 Baseball, 4 Football).

6 Future Recommendations

In summary of the current state of published sports and exercise science research, we know that statistical and methodological errors and lack of reporting transparency are prevalent in sports and exercise science [27, 74, 93]. Over three-quarters (78%) of surveyed sports and exercise science researchers already believe there is a replication and reproducibility crisis in sports and exercise science and that poor experimental design, insufficient mentoring, publishing pressure and selective reporting are the factors contributing most towards failed replications [28]. In addition, we know that we overestimate our level of statistical expertise and that a better understanding of study design, and the use of different statistical techniques to analyse data, would improve reproducibility and replicability [16, 28]. Therefore, there are many avenues that we can improve in the future.

Table 7 (continued)

Original Study Title	Replication Quality	Notes
Partial eta squared		
A real-time feedback method to reduce loading rate during running: effect of combining direct and indirect feedback [56]	Good	The sample and procedures were similar across the studies. Data was calculated using the same formulas and with different but comparable hardware and software. All calculations were clarified with the original author.
Changes in bench press velocity and power after 8 weeks of high-load cluster- or traditional-set structures [57]	Good	In the original study, the traditional group had a 1RM of 64kg and the cluster group had a 1RM of 66.59kg. The replication study had higher 1RMs; 91.73kg in the traditional group and 75.50kg in the cluster group.
Does mental fatigue negatively affect outcomes of functional performance tests? [58]	Moderate	In the replication study, individuals participated in the study between 8am and 8pm, contrary to the original study where individuals only participated between 8am and 10:30am. However, participants performed their two experimental sessions at the same time each week, and the session were spaced with the same delay (i.e., one week) as the original study.
Beep test performance is influenced by 30 minutes of cognitive work [59]	Good	Only two groups in the experimental design but the original study ran a repeated measures ANOVA for the analysis. Replicators did not have original data to run a paired *t*-test and compare the effect sizes, therefore, an ANOVA was used. Replicators also assumed the original author converted the dependent variable beep time of minutes and seconds into seconds for the analysis.
Voluntary muscle relaxation can mitigate fatigue and improve countermovement jump performance [60]	Good	There were some missing values in the replication dataset over the 10 trials for some participants that were replaced using a mean substitution.
Maximizing performance: augmented feedback, focus of attention, and/or reward? [61]	Good	Jump height was assessed using a 3D motion capture system in the original study to determine the maximal displacement of the hip during the jump compared to normal stance height. In the replication study, jump height was assessed using an AMTI force plate as the predicted jump height from take-off velocity. There were fewer trials in each condition in the replication study compared to the original (2 vs 6).
Altering work to rest ratios differentially influences fatigue indices during repeated sprint ability testing [62]	Good	Original study participants had greater body mass with larger absolute VO_2 outcomes throughout all protocols. There is a possible difference in first sprint protocol between the original and replication studies. There were limited details in the original methods, the replication study started the first sprint from a stationary flywheel; overcoming this flywheel inertia resulted in lower power and oxygen consumption during the first sprint.
Investigating the effects of maximal anaerobic fatigue on dynamic postural control using the Y-balance test [63]	Moderate	Replicators were not experienced or trained at Y-balance tests leaving scope for unintentional error. There was a higher mean mass in the replication study, which could be related to increased muscle mass in this cohort, that may have influenced the level of fatigue.
Partial eta squared		
Effect of different loading conditions on running mechanics at different velocities [64]	Poor	The replication (T170 treadmill, Cosmed) and original (Excite® Run MD, Technogym SpA) treadmills potentially had a different belt stiffness affecting leg stiffness. The Optogait sensor placement was not reported in the original study for spatiotemporal variable measurement. In the replication study, the sensor placement was on slightly raised foot pads on either side of the treadmill belt. Due to this position, the device did not always record the participants' movements at the lower speeds so participants were encouraged to "lift their feet" as they ran during the 60% MAS condition. This automatically altered the participants' running mechanics, affecting spatiotemporal variables, and therefore, had a significant effect on the leg stiffness calculations for the lower speeds.
Effect of attentional focus strategies on the biomechanical performance of the drop jump [65]	Good	Participants were included from field-based sports in the replication study but not specifically Rugby and none from Gaelic football like the original study.
Further evidence for an external focus of attention in running: looking at specific focus instructions and individual differences [66]	Good	Session 1 was completed on an indoor running track in the original study and an outdoor track in the replication study. Runners in the replication study were older, more experienced, and completed a greater volume of running training during a typical week than runners in the original study.
Enhanced expectancies facilitate golf putting [67]	Good	The replication study may be limited by the level of challenge presented by the putting surface and putter itself. While replicators have been able to replicate the length of the putt and target areas faithfully, the actual green was smaller and the speed of the green, which has significant bearing on the difficulty, was not reported in the original study. The inherent limitation of the study design itself; expectation was not measured directly and instead inferred by the size of the target.
Carbohydrate mouth rinse improves morning high-intensity exercise performance [68]	Moderate	The 10m sprint occurred on a different surface in the replication study (concrete flooring) compared to the original study (athletics track). Testing occurred at 7:30am after an eleven hour overnight fast in the original study, however, six trials in the replication study were performed at 12pm following a fifteen hour overnight fast.
Side-to-side differences in lower extremity biomechanics during multidirectional jump landing in volleyball athletes [69]	Poor	The original study sample contained collegiate Volleyball players while the replication study recruited active college-aged females to reach the sample size in the replication study. Therefore, the replication sample was quite heterogenous which could account for the very large replication partial eta squared.

3D three-dimensional, *MAS* maximal aerobic speed, *RM* repetition maximums, *SD* standard deviation, VO_2 oxygen consumption

As highlighted previously, discerning the "truth" between an original and replication study solely on the basis of their results is inherently fraught. This limitation is not specific to our project but rather a fundamental reality of research. It remains exceedingly difficult to definitively disprove a theory or claim solely on the basis of a single, divergent replication [94]. Therefore, we must evaluate the quality and rigour of our published claims on the basis of the study characteristics rather than placing irrational privilege in the chronological order of studies [95]. Original studies should not be prioritised over replication studies, and a replication study cannot overturn the original study results; rather, the focus should be on the accumulation of evidence rather than each study in a standalone manner. If a replication and original study differ, an auxiliary hypothesis can be formulated to expand

Table 8 A list of recommendations and resources for improving research practices in sports and exercise science

Action	Why?	How?
Justify the sample size	Power, precision and sample size estimation in sport and exercise science research [21]: https://doi.org/10.1080/02640414.2020.1776002	Sample size justification [99]: https://doi.org/10.1525/collabra.33267
State a clear hypothesis	Degrees of freedom in planning, running, analysing and reporting psychological studies: a checklist to avoid *p*-hacking [82]: https://doi.org/10.3389/fpsyg.2016.01832	Why hypothesis testers should spend less time testing hypotheses [81]: https://doi.org/10.1177/1745691620966795
Pre-register the hypothesis and analyses	Preregistration is hard and worthwhile [100]: https://doi.org/10.1016/j.tics.2019.07.009	https://osf.io/registries https://sportrxiv.org/index.php/server
Consider a registered report	Moving sport and exercise science forward: a call for the adoption of more transparent research practices [71]: https://doi.org/10.1007/s40279-019-01227-1	https://www.cos.io/initiatives/registered-reports Sports and exercise science example: https://doi.org/10.51224/cik.v1i3.43
Conduct appropriate statistical analysis	Improving your statistical inferences [101]: https://lakens.github.io/statistical_inferences/	PsyteachR open source textbooks: https://psyteachr.github.io
Fully report all statistical results	Replication concerns in sports and exercise science: a narrative review of selected methodological issues in the field [16]: https://doi.org/10.1098/rsos.220946	Publication bias, statistical power and reporting practices in the *Journal of Sports Sciences*: potential barriers to replicability [22]: https://doi.org/10.1080/02640414.2023.2269357
Report effect sizes and their magnitudes	Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for *t* tests and ANOVAs [70]: https://doi.org/10.3389/fpsyg.2013.00863	Guide to effect sizes and confidence intervals [102]: https://doi.org/10.17605/OSF.IO/D8C4G TOSTER package: https://aaroncaldwell.us/TOSTERpkg/articles/SMD_calcs.html
Minimise statistical errors	Wish list for improving the quality of statistics in sports science [17]: https://doi.org/10.1123/ijspp.2022-0023 Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy [26]: https://doi.org/10.1136/bjsports-2020-102607	Ten common statistical errors from all phases of research, and their fixes [27]: https://doi.org/10.1002/pmrj.12395 *p*-checker: https://shinyapps.org/apps/p-checker/
Make your code and data open	Comment on: 'Moving sport and exercise science forward: A call for the adoption of more transparent research practices' [23]: https://doi.org/10.1007/s40279-020-01298-5	OSF: https://osf.io GitHub: https://github.com

OSF Open Science Framework

a theory: i.e., "sophisticated falsification" [96]. If multiple replications are undertaken and more "falsification" arises (i.e., diagnostic evidence that does not support the claim), a "strategic retreat" of the original claim is warranted [94]. Therefore, non-replications can be informative by identifying boundary conditions for a claim and leading to the generation of new or reformulated hypotheses [97]. Furthermore, sports and exercise science researchers must embrace intellectual humility and acknowledge the inherent uncertainty in their work [98]. Transparency about flaws and limitations fosters humility, while concealing them breeds overconfidence, arrogance and intellectual fragility [98]. Achieving the ideals of quality and rigour over mere outcomes requires time and patience, and while crucial, replicability is but one facet of high-quality science. To ensure the long-term health of our field, we can take immediate action. Table 8 provides a roadmap, offering recommendations and resources to guide the implementation of these practices into our research endeavours.

7 Conclusion

In the first collaborative sports and exercise science replication project, only 28% of studies were successfully replicated, and there was a substantial regression of the reported effect size estimates. The low replication rate is potentially caused by poor research practices in our field e.g., publication bias towards significant findings [20, 22, 25], the use of small sample sizes [21, 43], poor reporting practices [22] and unfalsifiable, vague hypotheses [20]. Consequently, our current practices made it challenging to conduct this large replication project in sports and exercise science. The results of this project, in combination with previous research identifying issues in our field [16, 20, 22, 28], do not alleviate any concerns about the internal validity of sports and exercise science research and suggest a need to improve our research practices moving forward. To improve sports and exercise science research, we must make changes to our scientific process and culture, and we have recommended changes and provided a list of resources to assist with this. We hope that this preliminary outcome will excite a renewed vigour into conversations around research culture and current practices.

Acknowledgements The authors would like to thank all of the participants who contributed to the replication studies in this project.

Declarations

Ethics approval Ethical approval was obtained at each local university, and all studies were conducted in accordance with the Declaration of Helsinki.

Consent to participate All participants provided consent to participate.

Funding OpenAccess funding provided by the IReL Consortium. This work was supported by the Irish Research Council's Government of Ireland Postgraduate Scholarship Programme [GOIPG/2020/1155].

Conflict of interest The authors report there are no competing interests to declare.

Availability of data and material Raw data, code and other supplementary materials are available at https://doi.org/10.17605/OSF.IO/SFBVA.

Author contributions Jennifer Murphy: conceptualisation (equal), data curation (lead), formal analysis (equal), funding acquisition (lead), investigation (lead), methodology (equal) and writing – original draft preparation (lead). Aaron R. Caldwell: conceptualisation (equal), methodology (equal), supervision, formal analysis (equal) and writing – review and editing (equal). Cristian Mesquida: conceptualisation (equal). Joe P. Warne: conceptualisation (equal), methodology (equal), supervision (lead) and writing – review and editing (equal). Aera J. M. Ladell, Alberto Encarnación-Martínez, Alexandre Tual, Andrew Denys, Bailey Cameron, Bas Van Hooren, Ben Parr, Bianca DeLucia, Billy R. J. Mason, Brad Clark, Brendan Egan, Calum Brown, Carl Ade, Chiarella Sforza, Christopher B. Taber, Christopher Kirk, Christopher McCrum, Cian O. Keeffe Tighe, Ciara Byrne, Claudia Brunetti, Cyril Forestier, Dan Martin, Danny Taylor, David Diggin, Dearbhla Gallagher, Deborah L. King, Elizabeth Rogers, Eric C. Bennett, Eric T. Lopatofsky, Gemma Dunn, Gérome C. Gauchard, Guillaume Mornieux, Ignacio Catalá-Vilaplana, Ines Caetan, Inmaculada Aparicio-Aparicio, Jack Barnes, Jake Blaisdell, James Steele, Jared R. Fletcher, Jasmin Hutchinson, Jason Au, Jason P. Oliemans, Javad Bakhshinejad, Joaquin Barrios, Jose Ignacio Priego Quesada, Joseph Rager, Julia B. Capone, Julie S. J. Walton, Kailey Stevens, Katie Heinrich, Kelly Wu, Kenneth Meijer, Laura Richards, Lauren Jutlah, Le Tong, Lee Bridgeman, Leo Banet, Leonard Mbiyu, Lucy Sefton, Margaux de Chanaleilles, Maria Charisi, Matthew Beerse, Matthew J. Major, Maya Caon, Mel Bargh, Michael Rowley, Miguel Vaca Moran, Nicholas Croker, Nicolas C. Hanen, Nicole Montague, Noel E. Brick, Oliver R. Runswick, Paul Willems, Pedro Pérez-Soriano, Rebecca Blake, Rebecca Jones, Rebecca Louise Quinn, Roberto Sanchis-Sanchis, Rodrigo Rabello, Roisin Bolger, Roy Shohat, Sadie Cotton, Samantha Chua, Samuel Norwood, Samuel Vimeau, Sandro Dias, Sissel Pedersen, Spencer S. Skaper, Taylor Coyle, Terun Desai, Thomas I. Gee, Tobias Edwards, Torsten Pohl, Vanessa Yingling, Vinicius Ribeiro, Youri Duchene, Zacharias Papadakis: data collection, drafting of preregistration, replication reports and review of final manuscript.

References

1. Popper K. The logic of scientific discovery. Classics S. London: Routledge; 1959.

2. Shrout PE, Rodgers JL. Psychology, science, and knowledge construction: broadening perspectives from the replication crisis. Annu Rev Psychol. 2018;69:487–510. https://doi.org/10.1146/annurev-psych-122216-011845.
3. Begley CG, Ioannidis JPA. Reproducibility in science: improving the standard for basic and preclinical research. Circ Res. 2015;116:116–26. https://doi.org/10.1161/CIRCRESAHA.114.303819.
4. Nosek BA, Spies JR, Motyl M. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. Perspect Psychol Sci. 2012;7:615–31. https://doi.org/10.1177/2F1745691612459058.
5. Chambers CD, Feredoes E, Muthukumaraswamy SD, Etchells PJ. Instead of 'playing the game' it is time to change the rules: registered reports at AIMS Neuroscience and beyond. AIMS Neurosci. 2014;1:4–17. https://doi.org/10.3934/Neuroscience.2014.1.4.
6. Smaldino PE, McElreath R. The natural selection of bad science. R Soc Open Sci. 2016. https://doi.org/10.1098/rsos.160384.
7. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015. https://doi.org/10.1126/science.aac4716.
8. Pashler H, Wagenmakers EJ. Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? Perspect Psychol Sci. 2012;7:528–30. https://doi.org/10.1177/1745691612465253.
9. Schmidt FL, Oh I-S. The crisis of confidence in research findings in psychology: is lack of replication the real problem? Or is it something else? Arch Sci Psychol. 2016;4:32–7. https://doi.org/10.1037/arc0000029.
10. Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS One. 2009;4:5738. https://doi.org/10.1371/journal.pone.0005738.
11. Agnoli F, Wicherts JM, Veldkamp CLS, Albiero P, Cubelli R. Questionable research practices among Italian research psychologists. PLoS One. 2017. https://doi.org/10.1371/journal.pone.0172792.
12. de Vrieze J. Large survey finds questionable research practices are common Dutch study finds 8% of scientists have committed fraud. Science. 2021;373:265. https://doi.org/10.1126/science.373.6552.265.
13. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2:2–8. https://doi.org/10.1371/journal.pmed.0020124.
14. Sainani KL, Lohse KR, Jones PR, Vickers A. Magnitude-based inference is not Bayesian and is not a valid method of inference. Scand J Med Sci Sports. 2019;29:1428–36. https://doi.org/10.1111/sms.13491.
15. Halperin I, Vigotsky AD, Foster C, Pyne DB. Strengthening the practice of exercise and sport-science research. Int J Sports Physiol Perform. 2018;13:127–34. https://doi.org/10.1123/ijspp.2017-0322.
16. Mesquida C, Murphy J, Lakens D, Warne J. Replication concerns in sports and exercise science: a narrative review of selected methodological issues in the field. R Soc Open Sci. 2022. https://doi.org/10.1098/rsos.220946.
17. Sainani K, Chamari K. Wish list for improving the quality of statistics in sport science. Int J Sports Physiol Perform. 2022;17:673–4. https://doi.org/10.1123/ijspp.2020-0934.
18. Vagenas G, Palaiothodorou D, Knudson D. Thirty-year trends of study design and statistics in applied sports and exercise biomechanics research. Int J Exerc Sci. 2018;11:239–59.
19. Kim Y, Lee JL. Common mistakes in statistical and methodological practices of sport management research. Meas Phys Educ Exerc Sci. 2019;23:314–24. https://doi.org/10.1080/1091367X.2018.1537278.
20. Twomey R, Yingling VR, Warne JP, Scheider C, McCrum C, Atkins WC, et al. The nature of our literature: a registered report on the positive result rate and reporting practices in kinesiology. Commun Kinesiol. 2021;1:1–14. https://doi.org/10.51224/cik.v1i3.43.
21. Abt G, Boreham C, Davison G, Jackson R, Nevill A, Wallace E, et al. Power, precision, and sample size estimation in sport and exercise science research. J Sports Sci. 2020;38:1933–5. https://doi.org/10.1080/02640414.2020.1776002.
22. Mesquida C, Murphy J, Lakens D, Warne J. Publication bias, statistical power and reporting practices in the *Journal of Sports Sciences*: potential barriers to replicability. J Sports Sci. 2023;41:1507–17. https://doi.org/10.1080/02640414.2023.2269357.
23. Borg DN, Bon JJ, Sainani KL, Baguley BJ, Tierney NJ, Drovandi C. Comment on: 'Moving sport and exercise science forward: a call for the adoption of more transparent research practices.' Sports Med. 2020;50:1551–3. https://doi.org/10.1007/s40279-020-01298-5.
24. Büttner F, Toomey E, McClean S, Roe M, Delahunt E. Are questionable research practices facilitating new discoveries in sport and exercise medicine? The proportion of supported hypotheses is implausibly high. Br J Sports Med. 2020;54:1–7. https://doi.org/10.1136/bjsports-2019-101863.
25. Borg DN, Barnett AG, Caldwell AR, White NM, Stewart IB. The bias for statistical significance in sport and exercise medicine. J Sci Med Sport. 2023;26:164–8. https://doi.org/10.1016/j.jsams.2023.03.0029.
26. Sainani KL, Borg DN, Caldwell AR, Butson ML, Tenan MS, Vickers AJ, et al. Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. Br J Sports Med. 2021;55:118–22. https://doi.org/10.1136/bjsports-2020-102607.
27. Borg DN, Lohse KR, Sainani KL. Ten common statistical errors from all phases of research, and their fixes. PM&R. 2020;12:610–4. https://doi.org/10.1002/pmrj.12395.
28. Murphy J, Mesquida C, Warne JP. A survey on the attitudes towards and perception of reproducibility and replicability in sports and exercise science. Commun Kinesiol. 2023. https://doi.org/10.5122/cik.2023.53.
29. Kadlec D, Sainani KL, Nimphius S. With great power comes great responsibility: common errors in meta-analyses and meta-regressions in strength and conditioning research. Sports Med. 2023. https://doi.org/10.1007/s40279-022-01766-0.
30. Pitsch W, Emrich E. The frequency of doping in elite sport: results of a replication study. Int Rev Sociol Sport. 2012;47:559–80. https://doi.org/10.1177/1012690211413969.
31. Chalmers S, Debenedictis TA, Zacharia A, Townsley S, Gleeson C, Lynagh M, et al. Asymmetry during functional movement screening and injury risk in junior football players: a replication study. Scand J Med Sci Sports. 2018;28:1281–7.
32. Morin JB, Samozino P, Murata M, Cross MR, Nagahara R. A simple method for computing sprint acceleration kinetics from running velocity data: replication study with improved design. J Biomech. 2019;94:82–7. https://doi.org/10.1016/j.jbiomech.2019.07.020.
33. Holgado D, Troya E, Perales JC, Vadillo MA, Sanabria D. Does mental fatigue impair physical performance? A replication study. Eur J Sport Sci. 2021;21:762–70. https://doi.org/10.1080/17461391.2020.1781265.
34. Murphy J, Mesquida C, Caldwell AR, Earp BD, Warne JP. Proposal of a selection protocol for replication of studies in sports and exercise science. Sports Med. 2023;53:281–91. https://doi.org/10.1007/s40279-022-01749-1.
35. Brandt MJ, Ijzerman H, Dijksterhuis A, Farach FJ, Geller J, Giner-Sorolla R, et al. The replication recipe: what makes for

a convincing replication? J Exp Soc Psychol. 2014;50:217–24. https://doi.org/10.1016/j.jesp.2013.10.005.
36. Anderson SF, Kelley K, Maxwell SE. Sample-size planning for more accurate statistical power: a method adjusting sample effect sizes for publication bias and uncertainty. Psychol Sci. 2017;28:1547–62. https://doi.org/10.1177/0956797617723724.
37. Anderson SF, Kelley K. BUCSS: bias and uncertainty corrected sample size. Version 1.2.1. 2020. Available at:https://cran.r-project.org/package=BUCSS. Accessed 20 Oct 2021
38. Trafimow D, Earp BD. Null hypothesis significance testing and type I error: the domain problem. New Ideas Psychol. 2017;45:19–27. https://doi.org/10.1016/j.newideapsych.2017.01.002.
39. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2018;2:6–10. https://doi.org/10.1038/s41562-017-0189-z.
40. Lakens D, Adolfi FG, Albers CJ, Anvari F, Apps MAJ, Argamon SE, et al. Justify your alpha. Nat Hum Behav. 2018;2:168–71. https://doi.org/10.1038/s41562-018-0311-x.
41. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. Am Stat. 2019;73:235–45. https://doi.org/10.1080/00031305.2018.1527253.
42. Valentine JC, Biglan A, Boruch RF, Castro FG, Collins LM, Flay BR, et al. Replication in prevention science. Prev Sci. 2011;12:103–17. https://doi.org/10.1007/s11121-011-0217-6.
43. Schweizer G, Furley P. Reproducible research in sport and exercise psychology: the role of sample sizes. Psychol Sport Exerc. 2016;23:114–22. https://doi.org/10.1016/j.psychsport.2015.11.005.
44. Caldwell AR. Exploring equivalence testing with the updated TOSTER R package. 2022. https://orcid.org/0000-0002-4541-6283
45. Palmer TB, Thiele RM, Thompson BJ. Age-related differences in maximal and rapid torque characteristics of the hip extensors and dynamic postural balance in healthy, young and old females. J Strength Cond Res. 2017;31:480–8. https://doi.org/10.1519/JSC.0000000000001503.
46. Freitas MC, Cholewa JM, Gerosa-Neto J, Gonçalves DC, Caperuto EC, Lira FS, et al. A single dose of oral ATP supplementation improves performance and physiological response during lower body resistance exercise in recreational resistance-trained males. J Strength Cond Res. 2019;33:3345–52. https://doi.org/10.1519/JSC.0000000000002198.
47. DeBritto MA, Lemos AL, Dos Santos CS, Strefanyshyn DJ, Carpes FP. Effect of a compressive garment on kinematics of jump-landing tasks. J Strength Cond Res. 2017;31:2480–8. https://doi.org/10.1519/JSC.0000000000001620.
48. Norum M, Risvang LC, Bjørnsen T, Dimitriou L, Rønning PO, Bjørgen M, et al. Caffeine increases strength and power performance in resistance-trained females during early follicular phase. Scand J Med Sci Sports. 2020;30:2116–29. https://doi.org/10.1111/sms.13776.
49. Charlton JM, Hammond CA, Cochrane CK, Hatfield GL, Hunt MA. The effects of a heel wedge on hip, pelvis and trunk biomechanics during squatting in resistance trained individuals. J Strength Cond Res. 2017;31:1678–87. https://doi.org/10.1519/JSC.0000000000001655.
50. Ballmann CG, McCullum MJ, Rogers RR, Marshall MR, Williams TD. Effects of preferred vs nonpreferred music on resistance exercise performance. J Strength Cond Res. 2021;35:1650–5. https://doi.org/10.1519/JSC.0000000000002981.
51. Venier S, Grgic J, Mikulic P. Acute enhancement of jump performance, muscle strength, and power in resistance-trained men after consumption of caffeinated chewing gum. Int J Sports Physiol Perform. 2019;14:1415–21. https://doi.org/10.1123/ijspp.2019-0098.
52. Scanlan AT, Madueno MC. Passive recovery promotes superior performance and reduced physiological stress across different phases of short-distance repeated sprints. J Strength Cond Res. 2016;30:2540–9. https://doi.org/10.1519/JSC.0000000000001339.
53. Zahradnik D, Jandacka D, Beinhauerova G, Hamill J. Associated ACL risk factors differences during an unanticipated volleyball blocking movement. J Sports Sci. 2020. https://doi.org/10.1080/02640414.2020.1785727.
54. Lam WK, Jia SW, Baker JS, Ugbolue UC, Gu Y, Sun W. Effect of consecutive jumping trials on metatarsophalangeal, ankle, and knee biomechanics during take-off and landing. Eur J Sport Sci. 2021;21:53–60. https://doi.org/10.1080/17461391.2020.1733671.
55. Bloms LP, Fitzgerald JS, Short MW, Whitehead JR. The effects of caffeine on vertical jump height and execution in collegiate athletes. J Strength Cond Res. 2016;30:1855–61. https://doi.org/10.1519/JSC.0000000000001280.
56. Garofolini A, Oppici L, Taylor S. A real-time feedback method to reduce loading rate during running: effect of combining direct and indirect feedback. J Sports Sci. 2020;38:2446–53. https://doi.org/10.1080/02640414.2020.1788288.
57. Davies TB, Halaki M, Orr R, Helms ER, Hackett DA. Changes in bench press velocity and power after 8 weeks of high-load cluster- or traditional-set structures. J Strength Cond Res. 2020;34:2734–42. https://doi.org/10.1519/JSC.0000000000003166.
58. Verschueren JO, Tassignon B, Proost M, Teugels A, Van Cutsem J, Roelands B, et al. Does mental fatigue negatively affect outcomes of functional performance tests? Med Sci Sports Exerc. 2020;52:2002–10. https://doi.org/10.1249/MSS.0000000000002323.
59. MacMahon C, Hawkins Z, Schücker L. Beep test performance is influenced by 30 minutes of cognitive work. Med Sci Sports Exerc. 2019;51:1928–34. https://doi.org/10.1249/MSS.0000000000001982.
60. Pinto BL, McGill SM. Voluntary muscle relaxation can mitigate fatigue and improve countermovement jump performance. J Strength Cond Res. 2020;34:1525–9. https://doi.org/10.1519/jsc.0000000000003326.
61. Wälchli M, Ruffieux J, Bourquin Y, Keller M, Taube W. Maximizing performance: augmented feedback, focus of attention, and/or reward? Med Sci Sports Exerc. 2016;48:714–9. https://doi.org/10.1249/MSS.0000000000000818.
62. La Monica MB, Fukuda DH, Beyer KS, Hoffman MW, Miramonti AA, Riffe JJ, et al. Altering work to rest ratios differentially influences fatigue indices during repeated sprint ability testing. J Strength Cond Res. 2016;30:400–6. https://doi.org/10.1519/JSC.0000000000001122.
63. Johnston W, Dolan K, Reid N, Coughlan GF, Caulfield B. Investigating the effects of maximal anaerobic fatigue on dynamic postural control using the Y-balance test. J Sci Med Sport. 2018;21:103–8. https://doi.org/10.1016/j.jsams.2017.06.007.
64. Carretero-Navarro G, Márquez G, Cherubini D, Taube W. Effect of different loading conditions on running mechanics at different velocities. Eur J Sport Sci. 2019;19:595–602. https://doi.org/10.1080/17461391.2018.1537378.
65. Comyns TM, Brady CJ, Molloy J. Effect of attentional focus strategies on the biomechanical performance of the drop jump. J Strength Cond Res. 2019;33:626–32. https://doi.org/10.1519/jsc.0000000000003009.
66. Hill A, Schücker L, Hagemann N, Strauß B. Further evidence for an external focus of attention in running: looking at specific focus

instructions and individual differences. J Sport Exerc Psychol. 2017;39:352–65. https://doi.org/10.1123/jsep.2016-0272.
67. Palmer K, Chiviacowsky S, Wulf G. Enhanced expectancies facilitate golf putting. Psychol Sport Exerc. 2016;22:229–32. https://doi.org/10.1016/j.psychsport.2015.08.009.
68. Clarke ND, Hammond S, Kornilios E, Mundy PD. Carbohydrate mouth rinse improves morning high-intensity exercise performance. Eur J Sport Sci. 2017;17:955–63. https://doi.org/10.1080/17461391.2017.1333159.
69. Sinsurin K, Srisangboriboon S, Vachalathiti R. Side-to-side differences in lower extremity biomechanics during multidirectional jump landing in volleyball athletes. Eur J Sport Sci. 2017;17:699–709. https://doi.org/10.1080/17461391.2017.1308560.
70. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for *t*-tests and ANOVAs. Front Psychol. 2013;4:1–12. https://doi.org/10.3389/fpsyg.2013.00863/abstract.
71. Caldwell AR, Vigotsky AD, Tenan MS, Radel R, Mellor DT, Kreutzer A, et al. Moving sport and exercise science forward: a call for the adoption of more transparent research practices. Sports Med. 2020;50:449–59. https://doi.org/10.1007/s40279-019-01227-1.
72. McCrum C, van Beek J, Schumacher C, Janssen S, Van Hooren B. Sample size justifications. Gait Posture. 2022;92:333–7. https://doi.org/10.1016/j.gaitpost.2021.12.0106.
73. Kenny DA, Judd CM. The unappreciated heterogeneity of effect sizes: implications for power, precision, planning of research, and replication. Psychol Methods. 2019;24:578–89. https://doi.org/10.1037/met0000209.
74. Knudson D. Confidence crisis of results in biomechanics research. Sports Biomech. 2017;16:425–33. https://doi.org/10.1080/14763141.2016.1246603.
75. Rhea MR. Determining the magnitude of treatment effects in strength training research through the use of the effect size. J Strength Cond Res. 2004;18:918–20. https://doi.org/10.1519/14403.1.
76. Drinkwater E. Applications of confidence limits and effect sizes in sport research. Open Sports Sci J. 2008;1:3–4. https://doi.org/10.2174/1875399x00801010003.
77. Weissgerber TL, Garcia-Valencia O, Garovic VD, Milic NM, Winham SJ. Why we need to report more than 'data were analyzed by *t*-tests or ANOVA'. Elife. 2018;7:1–16. https://doi.org/10.7554/eLife.36163.
78. Scheel A, Schijen M, Lakens D. An excess of positive results: comparing the standard psychology literature with registered reports. Adv Methods Pract Psychol Sci. 2021;4:1–14. https://doi.org/10.1177/25152459211007467.
79. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle *p* value generates irreproducible results. Nat Methods. 2015. https://doi.org/10.1038/nmeth.3288.
80. Nissen SB, Magidson T, Gross K, Bergstrom CT. Publication bias and the canonization of false facts. Elife. 2016. https://doi.org/10.7554/eLife.21451.
81. Scheel AM, Tiokhin L, Isager PM, Lakens D. Why hypothesis testers should spend less time testing hypotheses. Perspect Psychol Sci. 2020;16:744–55. https://doi.org/10.1177/1745691620966795.
82. Wicherts JM, Veldkamp CLSS, Augusteijn HEMM, Bakker M, van Aert RCMM, van Assen MALMLM. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid *p*-hacking. Front Psychol. 2016;7:1–12. https://doi.org/10.3389/fpsyg.2016.01832
83. Rubin M. When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. Rev Gen Psychol. 2017;21:308–20. https://doi.org/10.1037/gpr0000128.
84. Speed HD, Andersen MB. What exercise and sport scientists don't understand. J Sci Med Sport. 2000;3:84–92. https://doi.org/10.1016/S1440-2440(00)80051-1.
85. Crüwell S, Van Doorn J, Etz A, Makel MC, Moshontz H, Niebaum JC, et al. seven easy steps to open science: an annotated reading list. J Psychol. 2019;227:237–48. https://doi.org/10.1027/2151-2604/a000387.
86. Simons DJ. The value of direct replication. Perspect Psychol Sci. 2014;9:76–80. https://doi.org/10.1177/2F1745691613514755.
87. Allen C, Mehler DMA. Open science challenges, benefits and tips in early career and beyond. PLoS Biol. 2019;17:e3000246. https://doi.org/10.1371/journal.pbio.3000246.
88. Nuijten MB, Hartgerink CHJ, van Assen MALM, Epskamp S, Wicherts JM. The prevalence of statistical reporting errors in psychology (1985–2013). Behav Res Methods. 2016;48:1205–26. https://doi.org/10.3758/s13428-015-0664-2.
89. Wicherts JM, Bakker M, Molenaar D. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. PLoS One. 2011;6:1–7. https://doi.org/10.1371/journal.pone.0026828.
90. Wallach JD, Boyack KW, Ioannidis JPA. Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017. PLoS Biol. 2018;16:2015–7. https://doi.org/10.1371/journal.pbio.2006930.
91. Kahneman D. A new etiquette for replication. Soc Psychol. 2014;45:310–1. https://doi.org/10.1027/1864-9335/a000202.
92. van Zwet EW, Goodman SN. How large should the next study be? Predictive power and sample size requirements for replication studies. Stat Med. 2022;41:3090–101. https://doi.org/10.1002/sim.9406.
93. Nielsen RO, Chapman CM, Louis WR, Stovitz SD, Mansournia MA, Windt J, et al. Seven sins when interpreting statistics in sports injury science. Br J Sports Med. 2018;52:1410–2. https://doi.org/10.1136/bjsports-2017-098524.
94. Meehl PE. Appraising and amending theories: the strategy of Lakatosian defense and two principles that warrant it. Psychol Inq. 1990;1:108–41.
95. Zwaan RA, Etz A, Lucas RE, Donnellan MB, Donnellan B. Making replication mainstream. Behav Brain Sci. 2018;41:4–5. https://doi.org/10.1017/S0140525X18000638.
96. Lakatos I. Falsification and the methodology of scientific research programmes. Can theories be refuted. Springer; 1976. p. 205–59. https://doi.org/10.1007/978-94-010-1863-0_14.
97. Nosek BA, Errington TM. What is replication? PLoS Biol. 2020;18:1–8. https://doi.org/10.1371/journal.pbio.3000691.
98. Hoekstra R, Vazire S. Aspiring to greater intellectual humility in science. Nat Hum Behav. 2021;5:1602–7. https://doi.org/10.1038/s41562-021-01203-8.
99. Lakens D. Sample size justification. Collabra Psychol. 2022;8:1–31. https://doi.org/10.1525/collabra.33267.
100. Nosek BA, Beck ED, Campbell L, Flake JK, Hardwicke TE, Mellor DT, et al. Preregistration is hard, and worthwhile. Trends Cogn Sci. 2019;23:815–8. https://doi.org/10.1016/j.tics.2019.07.009.
101. Lakens D. Improving your statistical inferences. 2022. https://lakens.github.io/statistical_inferences/
102. Jané MB, Xiao Q, Yeung SK, Ben-Shachar MS, Caldwell AR, Cousineau D, et al. Guide to effect sizes and confidence intervals. 2024. https://doi.org/10.17605/OSF.IO/D8C4G

Authors and Affiliations

Jennifer Murphy[1] · Aaron R. Caldwell[2] · Cristian Mesquida[1,3] · Aera J. M. Ladell[4] · Alberto Encarnación-Martínez[5] · Alexandre Tual[6] · Andrew Denys[7] · Bailey Cameron[8] · Bas Van Hooren[9] · Ben Parr[10] · Bianca DeLucia[11] · Billy R. J. Mason[12] · Brad Clark[12] · Brendan Egan[13] · Calum Brown[14] · Carl Ade[15] · Chiarella Sforza[16] · Christopher B. Taber[17] · Christopher Kirk[14] · Christopher McCrum[9] · Cian OKeeffe Tighe[1] · Ciara Byrne[1] · Claudia Brunetti[16] · Cyril Forestier[6] · Dan Martin[18] · Danny Taylor[18] · David Diggin[19] · Dearbhla Gallagher[20] · Deborah L. King[19] · Elizabeth Rogers[15] · Eric C. Bennett[4] · Eric T. Lopatofsky[17] · Gemma Dunn[20] · Gérome C. Gauchard[21] · Guillaume Mornieux[21] · Ignacio Catalá-Vilaplana[22] · Ines Caetan[18] · Inmaculada Aparicio-Aparicio[5] · Jack Barnes[14] · Jake Blaisdell[19] · James Steele[8] · Jared R. Fletcher[4] · Jasmin Hutchinson[11] · Jason Au[23] · Jason P. Oliemans[4] · Javad Bakhshinejad[11] · Joaquin Barrios[24] · Jose Ignacio Priego Quesada[5] · Joseph Rager[25] · Julia B. Capone[4] · Julie S. J. Walton[26] · Kailey Stevens[23] · Katie Heinrich[15] · Kelly Wu[27] · Kenneth Meijer[9] · Laura Richards[20] · Lauren Jutlah[23] · Le Tong[7] · Lee Bridgeman[8] · Leo Banet[21] · Leonard Mbiyu[25] · Lucy Sefton[25] · Margaux de Chanaleilles[6] · Maria Charisi[27] · Matthew Beerse[24] · Matthew J. Major[27] · Maya Caon[28] · Mel Bargh[18] · Michael Rowley[7] · Miguel Vaca Moran[27] · Nicholas Croker[11] · Nicolas C. Hanen[21] · Nicole Montague[1] · Noel E. Brick[10] · Oliver R. Runswick[25] · Paul Willems[9] · Pedro Pérez-Soriano[5] · Rebecca Blake[29] · Rebecca Jones[18] · Rebecca Louise Quinn[10] · Roberto Sanchis-Sanchis[5] · Rodrigo Rabello[16] · Roisin Bolger[10] · Roy Shohat[28] · Sadie Cotton[19] · Samantha Chua[8] · Samuel Norwood[28] · Samuel Vimeau[28] · Sandro Dias[18] · Sissel Pedersen[9] · Spencer S. Skaper[4] · Taylor Coyle[1] · Terun Desai[30] · Thomas I. Gee[18] · Tobias Edwards[14] · Torsten Pohl[31] · Vanessa Yingling[7] · Vinicius Ribeiro[28] · Youri Duchene[21] · Zacharias Papadakis[28] · Joe P. Warne[1]

✉ Jennifer Murphy
jennymurphy1@outlook.ie

[1] School of Biological, Health and Sport Sciences, Technological University Dublin, Tallaght Campus, Blessington Rd, Dublin 9, Ireland

[2] Institute for Community Health Innovation, University of Arkansas for Medical Sciences Northwest, Springdale, AR 72762, USA

[3] Human-Technology Interaction Group, Eindhoven University of Technology, Eindhoven, The Netherlands

[4] Department of Health and Physical Education, Mount Royal University, Calgary, AB T3E 6K6, Canada

[5] Department of Physical Education and Sports, Universitat de València, València, Spain

[6] Laboratoire Motricité, Interactions, Performance, Le Mans Université, Le Mans, France

[7] Department of Kinesiology, California State University East Bay, Hayward, CA 94542, USA

[8] Department of Sport and Health, Solent University, Southampton SO14 0YN, UK

[9] Department of Nutrition and Movement Sciences, NUTRIM School for Nutrition and Translational Research in Metabolism, Maastricht University, 6211 LK Maastricht, The Netherlands

[10] School of Psychology, Ulster University, Coleraine, Co., Londonderry BT52 1SA, UK

[11] Department of Exercise Science and Athletic Training, Springfield College, Springfield, MA, USA

[12] University of Canberra Research Institute for Sport and Exercise, University of Canberra, Bruce, ACT 2617, Australia

[13] School of Health and Human Performance, Dublin City University, Glasnevin, Dublin 9, Ireland

[14] School of Sport and Physical Activity, Sheffield Hallam University, Sheffield, UK

[15] Department of Kinesiology, Kansas State University, Manhattan, KS 66506, USA

[16] Department of Biomedical Sciences for Health, Università Degli Studi Di Milano, Via Mangiagalli, 31, 20133 Milan, Italy

[17] Department of Physical Therapy and Human Movement Science, Sacred Heart University, Fairfield, CT, USA

[18] School of Sport and Exercise Science, University of Lincoln, Lincoln LN6 7TS, UK

[19] School of Health Sciences and Human Performance, Ithaca College, Ithaca, NY 14850, USA

[20] School of Human and Social Sciences, Buckinghamshire New University, High Wycombe, UK

[21] Development, Adaptation, Handicap-CARE Grand Est, Université de Lorraine, Nancy, France

[22] Department of Physical Therapy, The University of British Columbia, Vancouver, Canada

[23] Department of Kinesiology and Health Sciences, University of Waterloo, Waterloo, ON, Canada

[24] Department of Physical Therapy, University of Dayton, Dayton, OH, USA

[25] Institute of Psychiatry, Psychology and Neuroscience, Kings College London, London, UK

[26] Faculty of Kinesiology, University of Calgary, Calgary, AB T2N 1N4, Canada

[27] Department of Physical Medicine & Rehabilitation, Northwestern University, Chicago, IL, USA

[28] Human Performance Laboratory and Motion Analysis Center, Department of Health Promotion and Clinical Practice, Barry University, Miami Shores, FL 33161, USA

[29] Specialist Facilities, Solent University, Southampton SO14 0YN, UK

[30] Institute of Sport, Exercise and Health, Division of Surgery and Interventional Science, University College London, London, UK

[31] Conservative and Rehabilitative Orthopaedics, Technical University of Munich, G80992 Munich, Germany