Communications in Kinesiology

Registered Report

Please cite as:

Name et al. (2021). The Nature of Our Literature: An Observational Study of the Positive Result Rate and Reporting Practices in Kinesiology. *Communications in Kinesiology*. doi: 10.51224/XXXXXXXXX

Subject Areas:

metascience

Keywords:

key1, key2, key3

Author for correspondence:

S. Name

e-mail:

s.name@fake.edu

The Nature of Our Literature: An Observational Study of the Positive Result Rate and Reporting Practices in Kinesiology

Some Name[1], Same Name[1], Another Nombre[2]

[1]Fake University

[2]Fake Place

Scientists rely upon an accurate scientific literature in order to build and test new theories about the natural world. In the past decade, observational studies of the scientific literature have indicated that numerous questionable research practices and poor reporting practices may be hindering scientific progress. In particular, 3 recent studies have indicated an implausibly high rate of studies with positive (i.e., hypothesis confirming) results. In Sports Medicine, a field closely related to Kinesiology, studies that tested a hypothesis indicated support for their primary hypothesis ~70% of the time. However, a study of journals that cover the entire field of Kinesiology has yet to be completed, and the quality of other reporting practices, such as clinical trial registration, has not been evaluated. Therefore, in this study we retrospectively evaluated 300 original research articles from the flagship journals of America (Medicine and Science in Sport and Exercise), Europe (European Journal of Sport Science), and Australia (Journal of Science and Medicine in Sport). The hypothesis testing rate (~64%) and positive result rate (~81%) were much lower than what has been reported in other fields (e.g., psychology). However, most studies did not report trial registration, and rarely included accessible data indicating rather poor reporting practices. We also observed that 92% of studies without a hypothesis, and 96% with a hypothesis relied upon significance testing. Overall, we conclude that while the positive result rate was lower than expected, other reporting practives must improve within the Kinesiology literature.

1. Introduction

Scientists and knowledge-users who make decisions based on scientific evidence rely on the published literature to be reported transparently and to be an accurate representation of the research that scientists conduct. The ability to replicate scientific findings is vital to establish the credibility of scientific claims and to allow research to progress (Brian A. Nosek & Errington, 2019). A large-scale collaborative effort estimated the replicability of findings in psychological science and found that most replication effects are smaller than originally reported Collaboration (2015). Whilst this is a complex issue, questionable research practices (QRPs) and publication bias explain at least some of the differences between the original and replication effect sizes (Head et al., 2015; John et al., 2012; Simmons et al., 2011). Alongside psychology Collaboration (2015), other fields have struggled to replicate or reproduce findings, including neuroscience (Boekel et al., 2015; Kharabian Masouleh et al., 2019; Turner et al., 2018), cancer biology (Brian A. Nosek & Errington, 2017), human genetics ("Replicating Genotype–Phenotype Associations," 2007) and pharmacology (Prinz et al., 2011). This type of systematic replication and evaluation of previously published results has not yet been attempted in kinesiology (alternatively known as sport and exercise science). However, considering the similarities (e.g,. the study of human behaviour) and overlap (e.g. sport and exercise psychology) between psychology and kinesiology, we have reason to believe it may suffer from the same QRPs. Replication appears to be rare in kinesiology, which is perhaps surprising considering that kinesiology has been the focus of significant critique due to overly optimistic inferences (Kristin L. Sainani et al., 2019) and a history of underpowered studies (Abt et al., 2020). Furthermore, a lack of sample size estimation (Abt et al., 2020), misuse of p-values and statistical significance testing, limited collaboration with statisticians (Kristin L. Sainani et al., 2020), minimal or arbitrary use of effect sizes (Aaron R. Caldwell & Vigotsky, 2020), and other reporting issues (Borg, Lohse, et al., 2020) appear to be commonplace.

In the past few years, a community of researchers in kinesiology have been advocating for and adopting open and replicable research practices (Borg, Bon, et al., 2020; Borg, Lohse, et al., 2020; Aaron R. Caldwell et al., 2020; Aaron R. Caldwell & Vigotsky, 2020; Kristin L. Sainani et al., 2020; Vigotsky et al., 2020). Some journals in the field have started to adopt the Registered Report format for manuscripts which is commendable (see \url{www.cos.io/rr for a list of participating journals). However, such practices include openly sharing data and code, pre-registration, and using the registered reports format (for a primer, see Aaron R. Caldwell et al. (2020) for details). Some of the issues that motivated the open science movement in psychology and other fields Munafò et al. (2017) are comparatively unexplored in kinesiology, and in addition currently, the number of kinesiology researchers adopting open research practices is largely unknown. There is some indication that both pre-registration and sharing of data is uncommon (Borg, Lohse, et al., 2020; Tamminen & Poucher, 2018) and flagship journals of our field (e.g., Medicine & Science in Sport & Exercise, European Journal of Sport Science) do not include a statement encouraging data sharing in the author guidelines (Oct 2020). Evaluating a recent sample of the kinesiology literature for such practices may help draw attention to these potential issues.

Another issue that warrants consideration is the positive result rate (the rate at which a published study finds support for its hypothesis) of published kinesiology studies. Recently, Büttner et al. (2020) estimated the positive result rate in three high ranking sports journals and one high ranking sports physiotherapy journal. In line with previous research in other scientific disciplines (Fanelli, 2010; Scheel, Schijen, et al., 2020), the positive result rate exceeded 80%. What are the mechanisms for the suspiciously high positive result rates in the scientific literature? Given the assumption of a completely unbiased literature, such a high positive result rate could only occur if both statistical power and the proportion of true hypotheses that researchers have chosen to test is consistently high Scheel, Schijen, et al. (2020). The more plausible explanation perhaps, corroborated in previous work (John et al., 2012; Simmons et al., 2011), is that the literature is distorted by undisclosed flexibility in analysis and other QRPs, and the incentive to publish positive results. Registered reports are specifically designed to help mitigate these issues Chambers et al. (2015). Therefore, Scheel, Schijen, et al. (2020) assessed the positive result rate in research articles published using the traditional format in comparison to registered reports in a sample of the psychology literature. The positive result rate was an implausibly high 96% for traditional articles and a significantly lower 46% for registered reports. The increased methodological rigour inherent to the registered report format has clearly led to an increase in the reporting of null findings in the psychological literature.

The equivalent findings regarding standard and registered reports have not been reported for kinesiology, and simply would not be possible given the current literature; unlike psychological science Scheel, Schijen, et al. (2020), and to our knowledge, kinesiology has not accumulated more than 70 RRs to evaluate against traditional publication formats. Nevertheless, the adoption of registered reports in kinesiology is progressing slowly. One reason for this may be a lack of awareness regarding the replication crisis and movement towards more rigorous and transparent research practices. However, the slow adoption of registered reports could also be due to a lack of concern about the kinesiology literature given the limited evidence that a problem exists. Therefore, the primary aim of this study was to assess the positive result rate of reported hypotheses in the recent kinesiology literature, using society-affiliated flagship journals from the field. Considering the majority of scientific disciplines documented by Fanelli (2009) had a positive rate of 80%, we hypothesized that the $>$80% of the published studies in kinesiology would report positive results (i.e, support for the hypothesis) for their first stated hypothesis. Our secondary aims were to assess a number of related research practices, including whether the kinesiology literature includes replications of previous effects, the detail of statistical reporting and adoption of other transparent reporting practices.

2. Methods

(a) Sample

Research articles were sampled from three flagship kinesiology journals: Medicine and Science in Sport and Exercise (MSSE), the European Journal of Sport Science (EJSS) and the Journal of Science and Medicine in Sport (JSAMS), which represent three major kinesiology societies of North America (American College of Sports Medicine), Europe (European College of Sport Science) and Australia (Sports Medicine Australia), respectively. We selected three major societies and their official flagship journals because we believed they represent a diverse selection of research in kinesiology and provide insights into the practices of the field as a whole. In addition, we chose to focus on these three journals rather than a random sample of the entire literature because these journals should represent the best research in the field (compared to any published article which could be sampled from a possible predatory publisher). We selected 100 original research articles per journal, 300 in total, excluding study protocols, methodological tutorials/reports, opinions, commentaries, perspectives, conference proceedings, narrative reviews, systematic reviews and meta-analyses. We also excluded research articles if they have been retracted or contain insufficient information to reach coding decisions. To sample a recent selection of the literature, research articles were sampled consecutively backwards from December 31, 2019, by the data analyst (ARC).

(b) Data Extraction

We dentified nine coders who were responsible for data extraction. Coders underwent standardized training that has been designed based on the queries raised and clarification required during pilot testing (see later section). These nine coders will form three teams of three, and each team will be randomly allocated the research articles from one journal (MSSE, EJSS, or JSAMS). All coders extracted data independently and entered this directly into a Qualtrics survey. The Qualtrics survey was refined after pilot testing and a copy is available at https://osf.io/nwcx6/?view_only=a41116388e9244b7821bfb9fe5496bd2 . Each team was coordinated by a team leader trained at a doctoral level in a kinesiology discipline (RT, VY and JW). Once independent coding was complete, interrater reliability was assessed using Fleiss's Kappa. Team leaders were responsible for resolving all conflicts (any instance where there was not agreement between all group members) within their team through group review of the item and group discussion. Where conflicts could not resolved (and revised if necessary) using this process, the team leader consulted the other two team leaders. All data (original coder responses and summary decisions) is available on study's Open Science Framework repository.

(c) Measures and Coding Procedure

All articles were categorized as basic physiology (animal and cell physiology), applied exercise physiology (human), environmental physiology (heat, cold, and altitude), clinical research, biomechanics, motor learning/control/behaviour, epidemiology, sport/exercise psychology, sport performance, or other (the category that best describes the article). Research articles thtat did include explicit statements that a hypothesis was tested were not included in the analysis of the positive result rate. However, all articles (i.e., 300) were included in analysis related to replication status, statistical reporting and other reporting practices, as described in the following sections.

(d) Support for a Hypothesis in the Kinesiology Literature

From the 300 articles, we expected that approximately 60% would include explicit statements that a hypothesis was tested as part of the study (e.g., "We hypothesized that. . . ") (Büttner et al., 2020). Therefore, we expected to extract data on the positive results rate from approximately 180 research articles. The main dependent variable was whether the *first* stated hypothesis was supported or not, as reported by the authors. We planned to closely follow the coding procedure used by Fanelli (2010) and Scheel, Schijen, et al. (2020), which is described as follows: By examining the abstract and/or full text, it will be determined whether the authors of each paper had concluded to have found a positive (full or partial) or negative (null or negative) support. If more than one hypothesis was being tested, only the first one to appear in the text was considered. The coding of support for the hypothesis was based on the author's description of their results. In line with previous work (Büttner et al., 2020; Scheel, Schijen, et al., 2020), we coded a hypothesis as having received "support," "partial support," "no support" or "unclear or not stated." We added this fourth option after pilot indicated that some authors failed to state whether or not the study's hypotheses were, or were not, supported in the discussion section of the manuscript. This was re-coded into a binary "support" (full or partial) vs. "no support" variable, with "unclear or not stated" removed, for the main analysis. The language used to state a hypothesis and support for the first tested hypothesis were included in the data extraction and are included in the final dataset.

(e) Replication Status

Coders assessed whether a study is a replication of a previously published one, as reported by the authors. Coders searched the full texts of all papers for the string 'replic*' and, for papers that contained it, indicated whether the coded hypothesis was a close replication with the goal to verify a previously published result (Scheel, Schijen, et al., 2020).

(f) Statistical Reporting

Coders assessed whether authors included language related to statistical significance and if p-values were included in the results (relating to all analyses and not only the first hypothesis). If yes, coders assessed if the p-value was interpreted as significant and if the exact or relative p-value was reported (i.e., $p = 0.049$ vs. $p < 0.05$). If a relative p-value was reported, the level of the reported p-value (e.g., $p < 0.05$, $p < 0.01$) were coded. Coders also extracted whether an effect size was reported, including, but not limited to: Cohen's d, correlation coefficients, mean differences, and measures of model fit (e.g., coefficient of determination: R^2). Finally, coders assessed whether the information on sample size was provided, and if provided, the total sample size (the number of participants included in the analyses, rather than the planned sample size) will be extracted. Finally, coders assessed whether any sample size justification (e.g. power analysis) were included in the manuscript.

(g) Other Reporting Practices

Coders assessed whether the study was a clinical trial, according to the ICJME definition (https://hub.ucsf.edu/clinicaltrialsgov-definition-clinical-trial). If yes, coders assessed if a clinical trial registration was reported in the manuscript. For all other types of studies, coders assessed whether studies were pre-registered (as reported within the manuscript). Coders assessed if a manuscript provides a statement on data availability, and if yes, whether there was open access to the original data and/or code via a link or supplementary file.

(h) Pilot Testing

To ensure that our questionnaire for our raters accurately and consistently reflects the above-detailed information from relevant articles, we conducted pilot testing before submission of the Stage 1 manuscript. Fifteen original research articles published in 2018, five from each of our three chosen journals, were selected to be used for pilot testing. One team of naive coders (i.e., were not trained prior to coding) extracted all data from these articles and entered this into Qualtrics. Independent coding was checked for disagreements, and this was used to inform training procedures. Pilot aggregated data were generated, and further adjustments were made to refine the planned extraction and analysis process. A summary report of the pilot work can be found at https://osf.io/nwcx6/?view_only=a41116388e9244b7821bfb9fe5496bd2. Overall, our pilot work indicated minimally acceptable agreement among the raters on the questions essential to our study such as whether a hypothesis was tested (κ= 0.903; complete agreement = 14/15) and if the authors found support for this hypothesis (κ= 0.586; complete agreement = 6/9). For all items with rater disagreement, at least two coders were in agreement on the rating. After the conclusion of pilot testing, a forum among the team was completed in order to appropriately adjust the questionnaire and refine future instructions/training for the coding teams in the full study. Prior to coding, all coding team members underwent formal training and were presented with example articles (not from the study sample) in order to improve consistency in the coding process.

(i) Statistical Analysis

First, we estimated the rate at which kinesiology research finds support for the first tested hypothesis (as reported by the authors). Further, we planned to compare this to the majority of disciplines surveyed in Fanelli (2010) which reported a positive result rate in excess of 80% (16 of 20 disciplines). We believed it unlikely that kinesiology would have a positive result rate lower than 80%, and believe that the actual rate is closer to the social sciences at approximately 85% (Fanelli, 2010). Considering we had a good prior information, and a belief we wanted to test, we opted to use a Bayesian analysis to test our hypothesis. Therefore, we planned to test our hypothesis that the positive result rate is greater than 80% using a generalized Bayesian regression model (Bürkner, 2017). We assumed a prior of $\beta(17, 3)$ on the intercept of the model (i.e., the rate of positive results). Evidence for our hypothesis is reported as the posterior probability, $pr(Intercept > .8|data)$, of our hypothesis and the Bayes Factor (BF), the ratio of evidence for our hypothesis versus the null. We performed a Monte Carlo simulation assuming we obtained 150 studies which contained hypotheses from a population where 85% will contain a positive result for the first stated hypothesis. This simulation indicated that our model would have an 87% chance of being able to obtain some evidence (BF in favor of our hypothesis > 3) for our hypothesis. All other data is summarized descriptively and as frequencies and proportions with chi-squared and binomial proportions tests where appropriate. A detailed summary of the planned hypothesis test and "power" analysis can be found at our OSF repository. Brackets indicate a 95% compatibility interval (confidence or posterior for frequentist and Bayesian approaches respectively).

3. Results

(a) Confirmatory Results

There was only weak support for our hypothesis that manuscripts would find some support for their hypothesis 80% of the time. There was only a 70.82% posterior probability of our hypothesis with it being 2.43 times more likely than the null hypothesis. However, the data did favor our secondary hypothesis that at least 60% of manuscripts perform hypothesis testing with it being 9.72 times more likely than the null (Posterior Probability: 90.67%). Overall, we estimate that the positive result is 81.43% [75.78, 86.3], and there is a 63.58% [58.12, 68.97] rate of hypotheses being tested in manuscripts (Figure 1A). Interestingly, we did find a substantial rate (6.8%) of manuscripts not reporting whether or not a hypothesis was supported (Figure 1B).

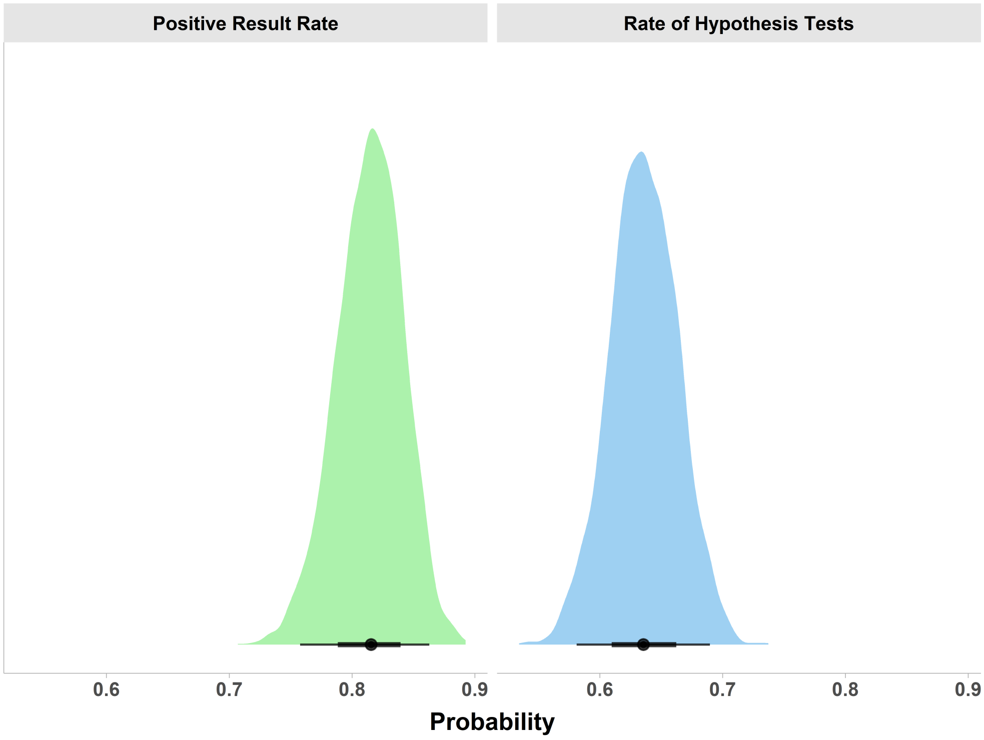

Figure 1. A) Posterior distributions from Bayesian model with the 50% and 95% percent compatibility intervals represented by the error bars at the bottom and B) Relative frequencies of the level of support reported for hypotheses.

(b) Exploratory Results

(i) Statistics Reporting

Nearly all manuscripts, 90% [86.03, 93.15], reported some form of significance testing. Even when a hypothesis was not stated or tested, significance testing was utilized in 81.65% [73.09, 88.42] of manuscripts (N = 89). Most manuscripts, 79.33% [74.3, 83.77], also reported some form of effect size to accompany the results. In addition, 33.7% [28.09, 39.68] of manuscripts reported exact p-values (p = .045) versus relative p-values ($p < .05$). Though, 89.63% [85.36, 93] of manuscripts reported at least *some* exact p-values (e.g., p = .045) versus relative p-values (e.g., $p < .05$).

(ii) Other Important Reporting Practices

The rate of study registration/preregistration was poor with 9% [6.01, 12.82] of manuscripts reporting preregistration or clinical trial registration information. Sample size information was often well reported and 97.67% [95.25, 99.06] of manuscripts reported all the required sample size information. However, sample size justification information (e.g., power analysis) only appeared in 22.67% [18.05, 27.83] of manuscripts. None of the manuscripts analyzed for this study were considered a replication attempt by the original study authors. Only 2.33% [0.94, 4.75] of manuscripts had a data accessibility statement. Further, a meager 0.67% [0.08, 2.39] of manuscripts reported some form of data sharing or open data.

(iii) Analysis by Journal

The degree of support for the main hypothesis, $\chi^2(6) = 2.4$; p=0.879, was consistent between journals (Figure 2A) with "Full support" occurring in all manuscripts >45% for all journals. However, there were significant differences, $\chi^2(2) = 20.43$; p<0.001, in the rate of hypotheses being tested (Figure 2B). The majority of MSSE and EJSS had hypothesis tests (74% and 71% respectively), but JSAMS had a much lower rate of hypothesis tests (46%). Effect sizes were often reported in manuscripts, but EJSS (90%) had a much better reporting rate, $\chi^2(2) = 10.9$; p=0.004, than JSAMS (72%) or MSSE (76%; Figure 2C). While sample size justifications were rare (Figure 2D), MSSE (35%) had a higher rate of reporting sample size justification, $\chi^2(2) = 13.73$; p=0.001, compared to EJSS (19%) or JSAMS (14%). The rate of reporting significance tests in all journals was high, but slightly lower, $\chi^2(2) = 6.22$; p=0.045, in JSAMS (84%) than EJSS (92%) or MSSE (94%).

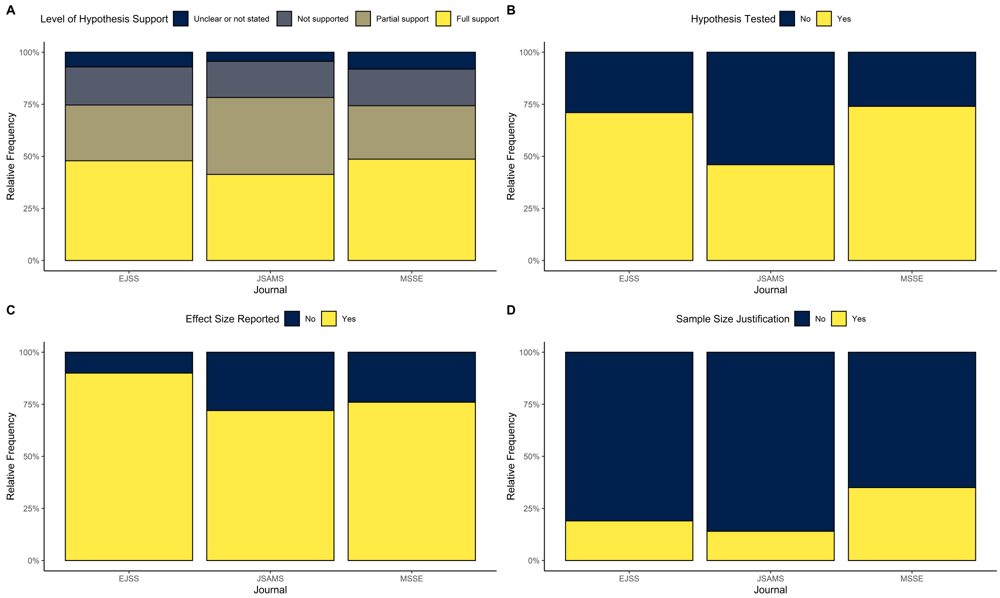

Figure 2. Relative frequencies, by journal, for A) level of reported support for hypotheses, B) indication of whether a hypothesis was tested, C) indication of whether an effect size was reported, or D) indication of if sample size was justified by the authors. Journals included the Journal of Science and Medicine in Sport (JSAMS), Medicine and Science in Sport and Exercise (MSSE), and the European Journal of Sport Science (EJSS).

(iv) Analysis by Discipline

Disciplines varied greatly in the degree of support found for the proposed hypothesis, $\chi^2(27) = 40.02$; p=0.051. In fact, motor behavior and environmental physiology studies all found full or partial support within the sample of manuscripts (Figure 3A). Basic physiology was the worst at not reporting whether or not a hypothesis was supported with 37.5% of the studies never making a clear statement of support (Figure 3A). The rate of hypothesis testing differed greatly between disciplines, $\chi^2(9) = 28.44$; p<0.001 (Figure 3B). The extremes of the spectrum ranged from epidemiology (25.9%) to basic physiology (88.9%). Sample size, evaluated using a linear model with a natural log transformation of the total sample size, differed between disciplines, F(9, 285) = 21.81, p = $2.2 \cdot 10^{-16}$, η_g^2 = 0.408. The estimated average sample size per discipline ranged from the lowest in environmental physiology, N = 16 [7, 37], to the highest in epidemiology, N = 1162 [691, 1952] (Figure 2C).

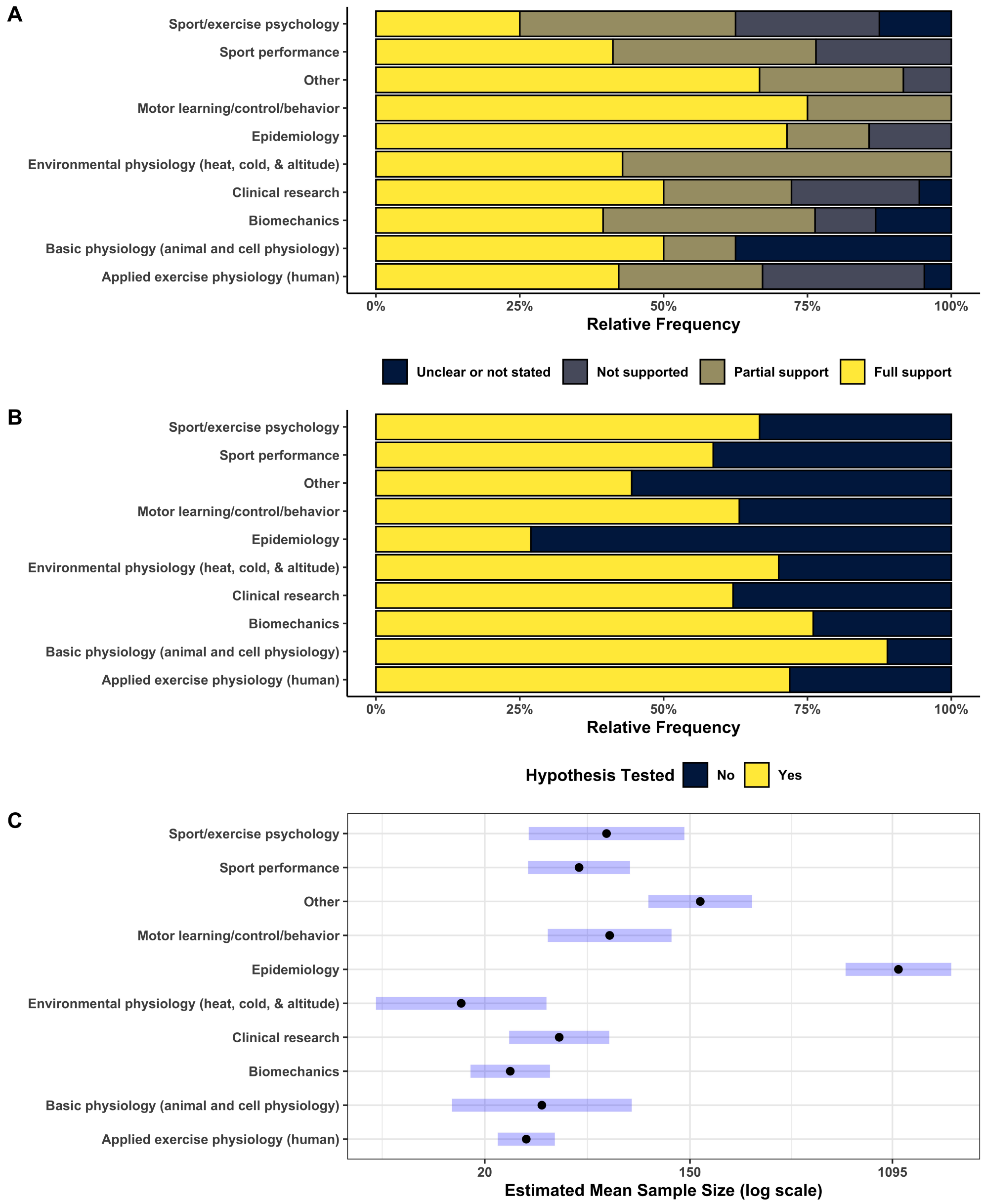

Figure 3. The breakdown, by dispcipline, for A) level of reported support for hypotheses, B) indication of whether a hypothesis was tested and C) the estimated total sample size (grey bands indicate 95% confidence intervals).

(v) Analysis of RCT and Clinical Trials

Clinical trials (N = 40) had lower rates of reported support for the hypothesis, 64% [42.5, 82], but similar hypothesis testing rates, 67.5% [50.8, 81.4], compared to the rest of the analyzed manuscripts. Despite guidelines requiring sample size justifications, only 62.5% [45.8, 77.3] reported a sample size justification. In addition, despite regulations that require clinical trial registration, only 57.5% [40.9, 72.9] reported clinical trial registration or preregistration documentation.

Another category of studies that requires particular reporting are RCTs (N = 64). Overall, the manuscripts including RCTs had similar rates of supporting the hypothesis, 75% [59.7, 86.8] and a high rate, 73.4% [60.9, 83.7], of testing hypotheses. Like clinical trials, RCTs often lacked sample size justifications, 50% [37.2, 62.8], and lacked pre-registrations, 28.1% [17.6, 40.8].

4. Discussion

We performed a systematic evaluation of the 300 journal articles published in the flagship journals of three major sport and exercise science societies. Our primary hypothesis that the proportion of studies finding support for their first hypothesis would be more than 80% was only weakly corroborated. However, we estimate that the positive result in the kinesiology literature is approximately 81%, and while much lower than most other disciplines Fanelli (2010) including psychology (96%; Scheel, Schijen, et al. (2020)), it is still excessive. Our secondary hypothesis that more than 60% of articles would explicitly report a hypothesis was corroborated, though our estimate of approximately 64% is relatively low when considering that ~90% of articles used null hypothesis significance testing. The low proportion of null results, lack of sample size justifications (based on a calculation or resources), low numbers of pre-registrations (even in the case of clinical trials), the near absence of open data, and the complete absence of replication studies cast doubt on the credibility of the scientific record of sport and exercise science.

Assuming no bias in the scientific record, the positive result rate of a sample of articles would depend on the statistical power and proportion of true hypotheses tested in the included studies (Ioannidis, 2005; Scheel, Schijen, et al., 2020). The proportion of true hypotheses being tested may be higher in sport and exercise science because studies can be resource-intensive due to the use of specialist equipment and techniques or the time and personnel required for specific study designs (for example, training studies with multiple laboratory visits). Studies can also be demanding or invasive for participants. Therefore, to ensure a favourable result given the resource investment, kinesiology researchers may design studies and test trivial hypotheses where a positive result is largely foreseeable (and potentially unimportant) ~80% of the time. Arguably, the most resource-intensive discipline is environmental physiology, and, in our sample, 100% of studies found some support for their hypothesis. However, we find it unlikely that such a high rate of true hypotheses in literature explains the high positive result rate because this also depends on the vast majority of studies having a high statistical power (~80%). Considering that sample size justifications of any type (not only a sample size calculation based on the predicted effect size, but also on resource limitations) were included in less than 25% of articles (and for example, in athletes, researchers are often interested in small effects), it is unlikely that statistical power was sufficiently high to explain the positive result rate, and is especially implausible given the 100% positive result rate in environmental physiology articles where the average sample size was lowest (N = 16). Therefore, rather than a consistently high proportion of true hypotheses being tested and consistently high statistical power, it is more reasonable to suggest that a combination of factors including bias and QRPs explain the excessive positive result rate in the kinesiology literature.

QRPs can be intentional or unintentional; some researchers may simply lack awareness, and consider QRPs to be a normal part of the research process rather than detrimental practices that inflate the Type 1 error rate and lead to a biased literature. Unconscious biases may cause a tendency for researchers to confirm tested hypotheses (confirmation bias) and can influence participants to meet researcher expectations. Similarly, researchers are not immune to publication bias and may be influenced by the perception or reality that a compelling "story" will be more publishable. Despite worldwide initiatives (Cagan, 2013), there are also clear academic incentives for arriving at positive results because publication quantity and journal-based metrics can be rated above societal impact in funding, appointment, and promotion decisions, and therefore impact career advancement. RRs offer one solution because articles are peer-reviewed before data collection, so poorly designed research does not progress to an in-principal acceptance, and the format is designed to prevent several QRPs and a bias (whether from the researchers, reviewer, or editor) towards findings that support the hypothesis. RRs also prevent the findings from being suppressed by peer reviewers (e.g., in the case that the findings refute previous work) an in-principal acceptance is based on the rationale and methods alone. The effect of RRs is clear in psychology, where the format moves the positive result rate closer to 50% and introduces adequately powered studies with null results into the scientific record (Allen & Mehler, 2019). This data is then available to other researchers rather than in the "file drawer" (an analogy for a researcher's negative results that were either not submitted or not accepted for publication), who may have otherwise wasted valuable resources towards testing a hypothesis that may be false.

Because only 9% of the studies were pre-registered and none of our selected journals offer the RR format, it is not possible to know if hypotheses presented as a priori were generated a priori or resulted from undisclosed post hoc hypothesizing (or HARKing; hypothesizing after the results are known). Similarly, it is not possible to know if undisclosed analytic flexibility, and selective outcome reporting, was used to obtain the most favorable results (for example, $p<0.05$ in the direction of the hypothesis). In other words, the high positive result rate may be due to nonconfirmatory research (exploratory or hypothesis-generating research that investigates problems that are not clearly defined) being presented as confirmatory (hypothesis-testing) research and a lack of awareness of the distinction between the two. This is unfortunate because nonconfirmatory research is no less essential and lays the necessary groundwork that leads to informative confirmatory tests (Scheel, Tiokhin, et al., 2020). Our data indicate that JSAMS may be more accepting of articles that do not explicitly test a hypothesis. However, the more stringent word limit at JSAMS (maximum of 3500 words for original research) may also explain the lower proportion of hypothesis-testing articles (46%) simply due to authors removing the langauge regarding hypothesis tests. In contrast, MSSE states that it does not publish preliminary research, demonstrating a clear preference for confirmatory tests.

It is particularly disconcerting that less than two-thirds of clinical trials were not pre-registered, considering that since 2008, the Declaration of Helsinki has stated that every clinical trial must be registered in a publicly accessible database *before* recruitment of the first participant (Krleža-Jerić & Lemmens, 2009). It is possible that clinical trials involving exercise that comply with international standards are accepted to more rigorous or disease-specific journals. However, recent findings suggest that a lack of pre-registration (and selective outcome reporting) may be an issue with clinical exercise science more broadly (Singh et al., 2021). Although not extracted, coders also noted that very few (if any) supplementary files included contained completed checklists for the relevant EQUATOR reporting guidelines, and very few (if any) statements were included about the use of reporting guidelines in the articles. No RCTs reporting using CONSORT guidelines, despite JSAMS explicitly including this in author instructions. JSAMS also included X unregistered clinical trials despite explicitly including this in author instructions, and MSSE included X unregistered clinical trials despite purporting to adhere to the Declaration of Helsinki. None of the nine animal studies reported using the ARRIVE guidelines, despite MSSE explicitly including this in author instructions. In summary, reporting of kinesiology research in our society journals does not meet international standards for the reporting of health or animal research.

Also disappointing was the lack of dating sharing, with only two articles (<1%) including a link to the data that support study findings (Dalecki et al., 2019; Harris et al., 2018). This was driven by the two sets of authors because our selected journals do not require authors to provide a data availability statement (though EJSS and JSAMS advise that datasets can be uploaded as a supplement and linked to the article). A data availability statement asks authors to report where data supporting the results reported is available, links to the publicly archived dataset, or conditions under which data can be accessed (e.g., for sensitive clinical data). Open data is part of a broad global open science movement that is advancing science and scientific communication (Huston et al., 2019), and the current literature shows that kinesiology is behind in creating a culture that embraces open research practices. The more encouraging findings are that the majority of studies included an effect size measure, though we used a broad definition of effect sizes, and reporting was not always considered best practice by coders (e.g., only reporting percent changes). Still, ~20% of studies did not provide any indication of the magnitude of the effect and relied only on p-values, without consideration of the practical or clinical significance of an intervention or experimental manipulation.

Statistical inference in almost all papers relied upon "significance" testing and all papers reported p-values. Even papers that did not include hypothesis tests almost always reported "significant" p-values despite significance testing being a hypothesis testing procedure. The practice of significance testing has been widely criticized by the statistical community (Wasserstein & Lazar, 2016). While the authors do not have a problem with using p-values or significance testing per se, it is troubling that these have become a *sine qua non* of publishing in the peer reviewed literature for sport and exercise science. As Gigerenzer (2018) eloquently pointed out, when these practices become ingrained, to the point of becoming a requirement for publication, statistical thinking is discarded in favor of statistical rituals. This does necessarily mean the often maligned p-value is to blame, as McShane et al. (2019) noted other statistical hypothesis tests can be misused. Instead, many manuscripts, especially those without hypothesis tests, can adopt a continuous and unconditional interpretation of statistics (Rafi & Greenland, 2020). Studies that are exploratory, or at least are not focused on hypothesis tests, should spend more time describing the statistical results within the manuscript and avoid placing emphasis on statistical significance. Generally, we would recommend that sport and exercise scientists adopt a more diverse set of statistical tools and for journals to encourage manuscripts submissions that do not rely upon significance testing.

(a) Limitations

We chose to use the flagship scholarly journals run by scientific societies that have the largest memberships worldwide and represent large continental regions (North America, Europe, and Australia). Journal subscription is included with membership with the society, and the official journal of the society is often considered a leading multidisciplinary journal within the field by society members. Our decision was also based on the high proportion of original investigations published in MSSE, EJSS, and JSAMS. MSSE states that "seeks to publish only the very highest quality science." Nevertheless, these journals may not provide a representative sample of the highest quality research in our field and may not have the most progressive editorial policies and reporting standards. Many articles that fall under the broad umbrella of kinesiology are submitted to sub-discipline specific journals (e.g., for sport and exercise physiology or psychology). Assessing the highest-ranked journals may be of interest in future work, though we note that citation data and journal prestige are not necessarily a surrogate of research quality or methodological rigour. Furthermore, our findings are similar to those of Büttner et al. (2020), who found a similar positive result rate of 82.2% in sports medicine/physical therapy journals, so we doubt that a different selection of journals would alter our conclusions.

A possible limitation is that support for the hypothesis was based on the author's language rather than inspection of the data and statistical analysis by our coders. This was necessary because the latter was not feasible; equivocal hypotheses and limited reporting were common, and different analytic choices influence results (Silberzahn et al., 2018). Although our interest was in the author's interpretation of the data as a reflection of how often authors claim support for the hypotheses in the peer-reviewed literature, the extent to which support for the hypothesis was warranted based on the data and statistical analysis is unknown. Another possible limitation in the hypothesis coding is that the first stated hypothesis may not have always been the primary hypothesis. Finally, there were other considerations to our coding procedures that we list here for transparency: although coders reached agreement on the single category that best described an article, many categorizations required discussion, and often two were suitable; many articles did not include explicit statements of support/no support for the hypothesis, but all coders reached consensus following review and discussion; we coded the number of participants (human or animal), and not the number of observations; although we found no articles that were described as replication studies by the authors, it's possible that some did involve a replication attempt, but were not labelled as such due to the perception or reality that a lack of novelty would preclude publication.

(b) Conclusion

A moderate proportion (~64%) of scientific articles published by society-led kinesiology journals are confirmatory (hypothesis testing), and the vast majority of these (~81%) report partial or full support for their first hypothesis. Although clearly lower than reported for disciplines with human behavioral experiments (such as psychology), the positive result rate in kinesiology is still questionably high. This cannot convincingly be explained by a consistently high statistical power coupled with an oddly high number of true hypotheses being tested. Instead, the high positive result rate is more likely a reflection of a scientific record that includes many false research findings. Indeed, we found a general lack of transparency, replication, adherence to established reporting standards, and an overreliance on statistical significance testing (even in articles with no stated hypothesis). Therefore, it is more plausible that the high positive result rate is due to a combination of questionable research practices, driven by publication bias and traditional academic incentives. Overall, we conclude that while the positive result rate was lower than expected, other reporting standards must improve within the kinesiology literature. Adoption of improved reporting practices should help increase the credibility of the sport and exercise literature.

5. Additional Information

(a) Data Accessibility

The authors agree to share the raw data, digital study materials and analysis code. All study materials can be found on our OSF repository: https://osf.io/nwcx6/?view_only=a41116388e9244b7821bfb9fe5496bd2

(b) Author Contributions

- Contributed to conception and design: TBA
- Contributed to acquisition of data: TBA
- Contributed to analysis and interpretation of data: TBA
- Drafted and/or revised the article: TBA
- Approved the submitted version for publication: TBA

(c) Funding

This research was not funded activity. All necessary support was provided by the authors institutions. This study is an analysis of published research and does not require ethical approval.

(d) Acknowledgments

We would like to thank John Mills for his assistance in setting up our Qualtrics survey for coding process. We would also like to thank Anne Scheel and Fionn Buttner for their early feedback on this project's design.

(e) Preregistration

Following Stage 1 in-principle acceptance, the authors agreed to pre-registration of the approved protocol on the Open Science Framework. The IPA registration can be found here: https://osf.io/3pqr7.

(f) Conflicts of Interest

ARC, RT, and VRY currently serve as executive committee members for the Society of Transparency, Openness, and Replication in Kinesiology (STORK). VRY is a section editor and ARC is on the Steering Board for Registered Reports in Kinesiology. Neither will be involved in any aspect of handling this manuscript except as authors. The opinions or assertions contained herein are the private views of the author(s) and are not to be construed as official or reflecting the views of the Army or the Department of Defense. Any citations of commercial organizations and trade names in this report do not constitute an official Department of the Army endorsement of approval of the products or services of these organizations. No authors have any conflicts of interest to disclose. Approved for public release; distribution is unlimited.

6. References

Abt, G., Boreham, C., Davison, G., Jackson, R., Nevill, A., Wallace, E., & Williams, M. (2020). Power, precision, and sample size estimation in sport and exercise science research. *Journal of Sports Sciences*, *38*(17), 1933–1935. https://doi.org/10.1080/02640414.2020.1776002

Allen, C., & Mehler, D. M. A. (2019). Open science challenges, benefits and tips in early career and beyond. *PLOS Biology*, *17*(5), e3000246. https://doi.org/10.1371/journal.pbio.3000246

Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, J., Brown, S., & Forstmann, B. U. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*, *66*, 115–133. https://doi.org/10.1016/j.cortex.2014.11.019

Borg, D. N., Bon, J. J., Sainani, K. L., Baguley, B. J., Tierney, N. J., & Drovandi, C. (2020). Comment on: 'Moving sport and exercise science forward: A call for the adoption of more transparent research practices.' *Sports Medicine*, *50*(8), 1551–1553. https://doi.org/10.1007/s40279-020-01298-5

Borg, D. N., Lohse, K. R., & Sainani, K. L. (2020). Ten common statistical errors from all phases of research, and their fixes. *PM&R*, *12*(6), 610–614. https://doi.org/10.1002/pmrj.12395

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Büttner, F., Toomey, E., McClean, S., Roe, M., & Delahunt, E. (2020). Are questionable research practices facilitating new discoveries in sport and exercise medicine? The proportion of supported hypotheses is implausibly high. *British Journal of Sports Medicine*. https://doi.org/10.1136/bjsports-2019-101863

Cagan, R. (2013). *San francisco declaration on research assessment*. The Company of Biologists Ltd.

Caldwell, Aaron R., & Vigotsky, A. D. (2020). *Does one effect size fit all? The case against default effect sizes for sport and exercise science*. SportRxiv. https://doi.org/10.31236/osf.io/tfx95

Caldwell, Aaron R., Vigotsky, A. D., Tenan, M. S., Radel, R., Mellor, D. T., Kreutzer, A., Lahart, I. M., Mills, J. P., Boisgontier, M. P., Boardley, I., Bouza, B., Cheval, B., Chow, Z. R., Contreras, B., Dieter, B., Halperin, I., Haun, C., Knudson, D., Lahti, J., … Consortium for Transparency in Exercise Science (COTES) Collaborators. (2020). Moving sport and exercise science forward: A call for the adoption of more transparent research practices. *Sports Medicine*, *50*(3), 449–459. https://doi.org/10.1007/s40279-019-01227-1

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: Realigning incentives in scientific publishing. *Cortex*, *66*, A1–2. https://doi.org/10.1016/j.cortex.2015.03.022

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Dalecki, M., Gorbet, D. J., Macpherson, A., & Sergio, L. E. (2019). Sport experience is correlated with complex motor skill recovery in youth following concussion. *European Journal of Sport Science*, *19*(9), 1257–1266. https://doi.org/10.1080/17461391.2019.1584249

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS One*, *4*(5), e5738. https://doi.org/10.1371/journal.pone.0005738

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLOS ONE*, *5*(4), e10068. https://doi.org/10.1371/journal.pone.0010068

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, *1*(2), 198–218.

Harris, D. J., Vine, S. J., & Wilson, M. R. (2018). An external focus of attention promotes flow experience during simulated driving. *European Journal of Sport Science*, *19*(6), 824–833. https://doi.org/10.1080/17461391.2018.1560508

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLOS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Huston, P., Edge, V., & Bernier, E. (2019). Reaping the benefits of open data in public health. *Canada Communicable Disease Report*, *45*(10), 252–256. https://doi.org/10.14745/ccdr.v45i10a01

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kharabian Masouleh, S., Eickhoff, S. B., Hoffstaedter, F., & Genon, S. (2019). Empirical examination of the replicability of associations between brain structure and psychological variables. *eLife*, *8*. https://doi.org/10.7554/elife.43464

Krleža-Jerić, K., & Lemmens, T. (2009). 7th revision of the declaration of helsinki: Good news for the transparency of clinical trials. *Croatian Medical Journal*, *50*(2), 105–110. https://doi.org/10.3325/cmj.2009.50.105

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*(sup1), 235–245.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021

Nosek, Brian A., & Errington, T. M. (2017). Making sense of replications. *eLife*, *6*. https://doi.org/10.7554/elife.23383

Nosek, Brian A., & Errington, T. M. (2019). *What is replication?* Center for Open Science. https://doi.org/10.31222/osf.io/u4g6t

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*(9), 712–712. https://doi.org/10.1038/nrd3439-c1

Rafi, Z., & Greenland, S. (2020). Semantic and cognitive tools to aid statistical science: Replace confidence and significance by compatibility and surprise. *BMC Medical Research Methodology*, *20*(1), 1–13.

Replicating genotype–phenotype associations. (2007). *Nature*, *447*(7145), 655–660. https://doi.org/10.1038/447655a

Sainani, Kristin L., Borg, D. N., Caldwell, A. R., Butson, M. L., Tenan, M. S., Vickers, A. J., Vigotsky, A. D., Warmenhoven, J., Nguyen, R., Lohse, K. R., & al., et. (2020). Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. *British Journal of Sports Medicine*, bjsports-2020-102607. https://doi.org/10.1136/bjsports-2020-102607

Sainani, Kristin L., Lohse, K. R., Jones, P. R., & Vickers, A. (2019). Magnitude-based inference is not bayesian and is not a valid method of inference. *Scandinavian Journal of Medicine & Science in Sports*, *29*(9), 1428–1436. https://doi.org/10.1111/sms.13491

Scheel, A. M., Schijen, M., & Lakens, D. (2020). An excess of positive results: Comparing the standard psychology literature with registered reports. *PsyArXiv*. https://doi.org/10.31234/osf.io/p6e9c

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2020). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahnik, S., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Rosa, A. D., Dam, L., Evans, M. H., Cervantes, I. F., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Singh, B., Fairman, C. M., Christensen, J. F., Bolam, K. A., Twomey, R., Nunan, D., & Lahart, I. M. (2021). *Outcome reporting bias in exercise oncology trials (OREO): A cross-sectional study*. https://doi.org/10.1101/2021.03.12.21253378

Tamminen, K. A., & Poucher, Z. A. (2018). Open science in sport and exercise psychology: Review of current approaches and considerations for qualitative inquiry. *Psychology of Sport and Exercise*, *36*, 17–28. https://doi.org/10.1016/j.psychsport.2017.12.010

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, *1*(1). https://doi.org/10.1038/s42003-018-0073-z

Vigotsky, A. D., Nuckols, G. L., Heathers, J., Krieger, J., Schoenfeld, B. J., & Steele, J. (2020). *Improbable data patterns in the work of barbalho et al*. SportRxiv. https://doi.org/10.31236/osf.io/sg3wm

Wasserstein, R. L., & Lazar, N. A. (2016). *The ASA statement on p-values: Context, process, and purpose*. Taylor & Francis.