***Kinesiology Review*, (Ahead of Print)**
https://doi.org/10.1123/kr.2022-0003

First Published Online: Aug. 5, 2022

Human Kinetics
WORKSHOP SESSION

Extraordinary Claims in the Literature on High-Intensity Interval Training: II. Are the Extraordinary Claims Supported by Extraordinary Evidence?

Panteleimon Ekkekakis[1] and Nicholas B. Tiller[2]
[1]Department of Kinesiology, Michigan State University, East Lansing, MI, USA; [2]Institute of Respiratory Medicine and Exercise Physiology, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA

Dishman challenged kinesiologists to seek a compromise between “the ideal physiological prescription and a manageable behavioral prescription.” High-intensity interval training (HIIT) is the first exercise modality that has been claimed to meet this challenge, combining substantial benefits for fitness and health with pleasure and enjoyment. If true, these claims may revolutionize the science and practice of exercise. In this paper, four claims are critically appraised: (a) HIIT lowers the risk of mortality more than moderate-intensity continuous exercise, (b) HIIT doubles endurance performance after only 15 min of training over 2 weeks, (c) 1 min of HIIT is equivalent to 45 min of moderate-intensity continuous exercise, and (d) HIIT is more pleasant and enjoyable than moderate-intensity continuous exercise. The evidence for these claims appears questionable. Kinesiology should heed the principle endorsed by Hume, Laplace, and Sagan, namely that extraordinary claims should be supported by commensurate evidence.

***Keywords**:* mortality, health-related quality of life, endurance performance, Type I error, Type II error, statistical power

Increasing the rates of physical activity in the population has proven one of the most challenging and persistent problems in the chronicles of public health (Barreto, 2013; Holtermann et al., 2021; Pratt et al., 2020; Rütten et al., 2013). For the past half-century, researchers have been striving to understand the motivational processes underlying the behaviors of engaging in physical activity, adhering to it, and dropping out to return to a state of hypoactivity. This extensive search has yielded few reliable answers beyond the consensus that the problem is multifaceted and complex. Accordingly, efforts to encourage a change from hypoactivity to activity (targeting individuals, schools, workplaces, communities, etc.) have had limited success (e.g., Guthold et al., 2018; Hoffman et al., 2017; Love et al., 2019). In the latest installment of *The Lancet* series on physical activity and health, the editors wrote: “Since 2001, there has been no improvement in global levels of physical activity” (The Lancet, 2021, p. 365). Researchers are conceding that, despite “exploding recognition and research” regarding the health benefits of physical activity, “population impact has been elusive” (Pratt et al., 2020, p. 760) and progress in increasing the rates of public participation “remains illusory” (p. 761).

Against the backdrop of researchers grappling with this highly complex problem, high-intensity interval training (HIIT), namely the modality of exercise in which several short periods of high-intensity exercise are interspersed with periods of low-intensity active recovery or passive rest, quickly transcended the boundaries of basic research and gained exceptional popularity in the mass media and the fitness industry (Thompson, 2013). At least in part, this popularity may be attributable to the combination of (a) a central message characterized by simplicity, definitiveness, and optimism appearing against the backdrop of messages from science characterized by complexity, uncertainty, and a disheartening lack of breakthrough insights; and (b) a series of extraordinarily bold and intriguing claims. For example, while the problem of physical inactivity is described elsewhere in the exercise-science literature as the result of “interconnected complex, multifaceted cognitive, neurological and physiological processes” (Rhodes et al., 2019, p. 104), its portrayal in the HIIT literature is reassuringly simple, albeit inaccurately so. The low level of physical activity participation is described as being mainly due to “lack of time,” which is described unambiguously as “the primary reason for [the] failure to exercise on a regular basis” (Gibala & McGee, 2008, p. 61). Accordingly, researchers have suggested that the problem of physical inactivity can be addressed, to a large extent, by offering “time-efficient” forms of exercise, namely HIIT or its variants (Coyle, 2005; Gibala, 2007; Gillen & Gibala, 2018). As an editorialist wrote in the prestigious *Journal of Physiology*, “for those looking to do the least possible work to be fit,” research on HIIT “presents hope” (Baar, 2006, p. 690).

A claim or promise can be deemed “extraordinary” if it falls outside the boundaries of current understanding or scientific consensus (Deming, 2016). For example, a well-publicized claim in the HIIT literature is that a total of only 15 min of high-intensity interval exercise performed in six sessions over a period of 2 weeks “doubled endurance time to fatigue during cycling at 80% $\dot{V}O_{2peak}$ in recreationally active subjects” without changing maximal aerobic capacity (Burgomaster et al., 2005, p. 1988). This qualifies as an “extraordinary” claim for at least three reasons. First, there is a striking lack of proportionality between the unusually large training effect ($d = 1.13$) compared with the minimalist intervention. Second, there is no readily apparent mechanistic explanation that can convincingly account for the full magnitude of this effect. Third, in the past nearly two decades, there have been no direct replications. Yet, this “extraordinary” result has been cited hundreds of times, suggesting that within kinesiology, as in other

Ekkekakis (ekkekaki@msu.edu) is corresponding author, https://orcid.org/0000-0003-4260-4702

scientific fields (Serra-Garcia & Gneezy, 2021), "extraordinary" results may have broader appeal or hold greater intrigue than "ordinary" results.

Four decades ago, Dishman (1982) issued a grand challenge to kinesiology. He wrote that, while much is being discovered about the "doses" of exercise that lead to favorable adaptations in multiple bodily systems, not nearly as much is being learned about the "dose" that encourages exercise adherence, even though one is not meaningful without the other. So, his challenge was to seek an elusive "compromise between the ideal physiological prescription and a manageable behavioral prescription" (p. 248). Dishman argued that finding this compromise "may be necessary to allow adherence to be sufficient for desired biological changes to occur" (p. 248). Specifically regarding intensity, he warned that, although higher intensity may prompt faster or larger adaptations in multiple bodily systems, "it is clear ... that an exercise intensity that exceeds tolerance for physical stress will force an individual to stop exercising" (p. 248).

Arguably, HIIT may represent the first instance in which researchers have claimed to have devised an exercise training method that approximates an "ideal physiological prescription" and, at the same time, "a manageable behavioral prescription," thus approaching the elusive "compromise" that Dishman (1982) had envisioned. The literature contains numerous studies claiming to have demonstrated that HIIT is a powerful stimulus for a wide range of physiological adaptations and health benefits. For example, a review on the benefits of HIIT concluded that "higher exercise intensities may be superior to moderate intensity for maximizing health outcomes" (Karlsen et al., 2017, p. 67). In addition, despite early warnings that "given the extreme nature of the exercise, it is doubtful that the general population could safely or practically adopt [HIIT]" (Gibala, 2007, p. 212; also see Gibala & McGee, 2008, p. 62), more recently, reviewers have asserted that HIIT is accompanied by a higher level of enjoyment than moderate-intensity continuous exercise: "participants report highest enjoyment for HIIT conditions" (Kilpatrick et al., 2014, p. 14).

The following sections examine four of the most noteworthy claims that have appeared in the HIIT literature. Specifically, the following dubious claims are subjected to critical appraisal: (a) HIIT lowers the risk of mortality more than moderate-intensity continuous exercise, (b) HIIT doubles endurance performance after a total of only 15 min of training over 2 weeks, (c) 1 min of HIIT is equivalent to 45 min of moderate-intensity continuous exercise, and (d) HIIT is more pleasant and enjoyable than moderate-intensity continuous exercise. By adopting a critical perspective and invoking basic, uncontroversial, and well-known principles of statistics and research methods, findings that might have seemed "extraordinary" upon initial examination appear doubtful upon closer inspection.

The working hypothesis adopted here is that the main driver of extraordinary claims about HIIT is "spin" (Boutron et al., 2010; Boutron & Ravaud, 2018). Spin often stems from "a burning desire to demonstrate that treatments researchers believe in are effective" (Harvey, 2015, p. 417). Because spin usually does not amount to outright fabrication or falsification, it is commonly practiced and, importantly, widely tolerated, "even, sometimes, presented as the way to get ahead in science" (Bailar, 2006, p. 217). In the case of HIIT, it could be argued that the intense worldwide fascination with the subject has led to spin becoming more frequent, more brazen, and more widely tolerated. The common denominator of "spun" articles, editorials, and news items appears to be a shared desire among authors, reviewers, editors, journalists, and social-media personalities to amplify the virtues of HIIT, even in the absence of strong supporting empirical evidence. In addition, "spin" may be fueled by the fact that the popularity of HIIT within the fitness industry has added a financial dimension, resulting in lucrative book deals, media appearances, contracts with speaker bureaus, and increased opportunities for grant funding for HIIT researchers. The narrative underscores the importance of critical appraisal in an informational landscape that contains claims ranging from those whose scientific basis is obscure (e.g., 10,000 steps a day; Tudor-Locke et al., 2011) to those that are blatantly pseudoscientific (Tiller et al., 2022).

"It Protects Against Premature Death"

On December 23, 2020, an article in *The New York Times*, one of the highest-circulation newspapers in the United States, was entitled: "The secret to longevity? 4-minute bursts of intense exercise may help"[1] and subtitled "Including high-intensity training in your workouts provided better protection against premature death than moderate workouts alone." Several similar statements can also be found in the scientific literature: (a) "one large-scale randomized controlled trial ($n = 1{,}567$) investigated the influence of different aerobic exercise modalities on overall mortality in healthy older adults [and] provides prospective evidence that HIIT is superior to continuous training with moderate intensity and standard training recommendations" (Joisten et al., 2022, p. 85); (b) "high-intensity exercise has been shown to lower the risk of mortality to a greater extent than moderate exercise through greater improvements in aerobic capacity" (Herbert, 2022, p. 48); and (c) "recent randomized trial data suggest even lower all-cause mortality in older patients performing HIIT" (Wernhart et al., 2021, p. 2). These articles referred to a publication in the *British Medical Journal*, reporting results from the "Generation 100" study, one of the largest randomized controlled trials ever conducted investigating the effects of exercise (Stensvold et al., 2020).

The trial, conducted in Norway, was designed with the extremely ambitious goal of attempting "to determine the effects of regular exercise training over a 5-year period on overall mortality in elderly people (70–76 years of age)" (Stensvold et al., 2015, p. 2). Even in a sample of septuagenarians, investigating mortality as the primary outcome in an experimental study is "extremely difficult" (Bahls et al., 2022, p. e266) because (a) the sample size must be sufficiently large and (b) the follow-up period must be sufficiently long to result in a high enough number of deaths that would give the trial a meaningful level of statistical power (Kujala, 2018). This challenge was exacerbated in the case of the "Generation 100" trial by the fact that there was no physically inactive control group; instead, 78% of the participant-volunteers allocated to the control group met the national physical activity guideline at baseline. The study was powered ($1 - \beta = 90\%$) to detect a 50% reduction in mortality, from the 10% suggested by national statistics for the age group to 5%. However, these power calculations referred to a comparison between the control group (mostly active but without supervision, $N = 780$) and a combined group of individuals participating in partially supervised exercise, either in the form of HIIT or in the form of moderate-intensity continuous exercise ($N = 400 + 387 = 787$). In other words, the trial was neither originally designed nor adequately powered to compare HIIT to moderate-intensity continuous exercise in terms of mortality (the power achieved for a 5% difference in mortality with sample sizes of 400 and 387 was 75%, below the 90% target set in power calculations).

Comparisons between HIIT and moderate-intensity continuous exercise were made considerably more complicated (or less meaningful) by several factors. First, all-cause mortality as an outcome is too unspecific, containing variance that may be unrelated to the experimental treatments (e.g., deaths due to accidents). As a result, when the total number of deaths recorded during the follow-up period is low, even a small number of deaths from causes unrelated to health, if distributed unevenly between the groups (which is possible given their small number), may weaken the resultant effect size.

Second, in the "Generation 100" trial, all-cause mortality after the 5-year follow-up period was much lower than the 10% that was initially assumed (4.6% overall, 4.7% in the control group), possibly because the sample consisted of volunteers who were already mostly physically active. Whatever the case, the inaccurate initial assumptions increased the sample size required to maintain the desired level of statistical power (assuming a 50% reduction in mortality from 4.7% would have required $N = 1{,}346 + 1{,}346 = 2{,}692$ to maintain 90% power).

Third, because the trial was not originally designed as a comparison between HIIT and moderate-intensity continuous exercise, no safeguards were put in place to minimize treatment crossover (or "cross-contamination"). Specifically, the supervised sessions occurred only once every 6 weeks, at which time the participants wore heart rate monitors to ensure compliance with the instructions given to their respective groups (i.e., either four 4-min intervals at 90% of peak heart rate or 16 on the 6–20 rating scale of perceived exertion versus 50 min of continuous exercise at 70% of peak heart rate or 13 on the 6–20 rating scale of perceived exertion). Importantly, however, when the participants were unsupervised, the instructions they were given specified that they were expected to comply with their prescribed type of exercise for only two out of the recommended five exercise sessions per week. As a result, only 47%–50% of the participants allocated to the group labeled as "HIIT" chose to do HIIT (27%–28% chose to do moderate-intensity continuous exercise) and 33% had dropped out by the fifth year. In contrast, 51%–63% of the participants assigned to the moderate-intensity continuous exercise group adhered to their given prescription (only 11%–14% chose to do HIIT) and only 26% had dropped out by the fifth year. This extensive crossover between treatment groups further decreased the likelihood of detecting any differential effects on mortality between the two exercise treatments. Moreover, this situation created considerable interpretational problems for intention-to-treat analyses, because, when the investigators were referring to "the HIIT group," they were referencing a group of participants, most of whom either decided to abandon HIIT and switch to moderate-intensity continuous exercise (27%–28%) or dropped out of the study altogether (33%). The investigators had announced in the trial protocol that, in addition to the standard intention-to-treat analysis, they would also "perform a per protocol analysis, based on adherence to the intervention" (Stensvold et al., 2015, p. 6). However, a per protocol analysis was not included in the final report (Stensvold et al., 2020).

Given these challenges, it is unsurprising that the "Generation 100" trial found no difference in mortality between the control group (37 deaths) and the combined (HIIT + moderate) "supervised exercise" group (35 deaths). The investigators then also performed comparisons between the HIIT subgroup (12 deaths) and the moderate-intensity continuous exercise subgroup (23 deaths). The hazard ratio adjusted for sex, cohabitation status, and age at baseline was 0.51, but the confidence interval (CI) was wide and included 1.0 (95% CI [0.25, 1.02]), indicating that it was not statistically significant. Moreover, an examination of the causes of death showed no consistent pattern. In particular, the deaths from cardiovascular causes were too few to be meaningful (two deaths in HIIT and four deaths in the moderate-intensity continuous exercise group). The largest imbalance between HIIT (one death) and the moderate-intensity continuous exercise group (four deaths) was in the "other" category (i.e., noncardiovascular, noncancer deaths). Large imbalances in nonfatal clinical events were in favor of the moderate-intensity continuous exercise group, since HIIT had, for example, 3.2 times as many acute myocardial infarctions (16 vs. 5) and 3.3 times as many diagnoses of prostate cancer (13 vs. 4). The two groups exhibited nearly identical numbers of all cardiovascular events combined and all cancers combined (moderate-intensity continuous training: 58 and 43; HIIT: 61 and 48, respectively), with adjusted hazard ratios of 0.98 (95% CI [0.69, 1.41]) and 1.05 (95% CI [0.70, 1.58]), respectively.

Besides the primary outcome of all-cause mortality, in the article published in the *British Medical Journal*, the "Generation 100" investigators chose to report results on three secondary outcomes (out of a much larger set), namely (a) $\dot{V}O_{2peak}$, (b) physical component of health-related quality of life, and (c) mental component of health-related quality of life. For all three of these variables, the investigators reported statistically significant differences between the HIIT group and the MICT group at the 5-year follow-up, all in favor of HIIT (for each comparison, $p = .04$). However, all three of these differences should be viewed cautiously for the following reasons.

First, there is no apparent explanation for choosing to report this particular subset of secondary outcomes. For example, since these outcomes clearly do not constitute complete reporting (the list of secondary outcomes, according to the trial protocol, was much longer; see Stensvold et al., 2015), it is possible that these variables were selected because they happened to show differences in favor of HIIT.

Second, neither the statistical test (i.e., linear mixed models) nor the mix of covariates (i.e., sex, cohabitation status, age at baseline) were specified in the trial protocol. In their instructions for avoiding false-positive results, Simmons et al. (2011) wrote that not having prespecified the set of covariates or experimenting with various combinations of covariates represents an example of what they called "researcher degrees of freedom," namely questionable statistical practices that researchers can use to produce the desired result: "If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate. Reporting covariate-free results makes transparent the extent to which a finding is reliant on the presence of a covariate" (p. 1363). While the results for all-cause mortality included both unadjusted and adjusted statistics, this was not the case for the secondary outcomes of $\dot{V}O_{2peak}$ and quality of life.

Third, no adjustments were made to prevent the inflation of the Type I error rate due to multiple comparisons. If such adjustments were made, none of the comparisons would have remained statistically significant given the reported p values of .04. In general, unless the analyses of secondary outcomes are adjusted for multiplicity, any secondary outcomes should be discussed as exploratory and, as such, subject to further confirmatory trials (Li et al., 2017).

Fourth, the reported difference in $\dot{V}O_{2peak}$ of 0.7 $ml \cdot kg^{-1} \cdot min^{-1}$ at the fifth year of follow-up cannot be regarded as statistically reliable or clinically meaningful. It should be noted that (a) approximately half of the difference in favor of HIIT over the moderate-intensity continuous exercise group already existed at baseline (0.3 $ml \cdot kg^{-1} \cdot min^{-1}$: 28.9 ± 6.4 vs. 28.6 ± 6.6) and (b) 41% of participants did not meet the criteria for maximal oxygen uptake, so most tests were terminated when participants decided to stop (reached "exhaustion"; see Stensvold et al., 2017). A point of "exhaustion" that could not be confirmed by conventional physiological criteria (e.g., a respiratory exchange ratio of 1.05, as specified in the trial's protocol) should be considered susceptible to social environmental influences such as encouragement (e.g., Andreacci et al., 2002), especially since there is no indication that the outcome assessors were blinded to group allocation.

Fifth, differences in the physical and mental components of health-related quality of life were reported as statistically significant ($p = .04$), but the differences between the HIIT group and the moderate-intensity continuous exercise group were consistently lower than 2 units on a 100-point scale (see Figure 1). Although the information on the "minimal clinically important difference" of the shortened scale that was used (Medical Outcomes Study Short-Form [8-item] Health Survey [SF-8]) is sparse, data obtained with the unabridged original version (Medical Outcomes Study [36-item] Health Survey [SF-36]), which is also scored on a *T*-score scale with a mean of 50 and a *SD* of 10, suggest that the minimal clinically important difference is 3–5 units (Samsa et al., 1999). It should also be noted that approximately half of the 1.7-unit difference for the physical component ($d = 0.22$) and the 1.2-unit difference for the mental component ($d = 0.20$) of health-related quality of life found at the fifth year of follow-up was already present at baseline (delta = 0.7, $d = 0.11$ and delta = 0.6, $d = 0.10$, respectively). These baseline differences apparently occurred as a result of imperfect randomization and cannot, therefore, be causally attributed to the interventions.

The report on the "Generation 100" trial was published in the *British Medical Journal* on October 7, 2020. The following day, a press release was issued by the Norwegian University of Science and Technology, which appeared under the title "High intensity training best for older people."[2] The subtitle read: "Five years of high-intensity interval training increased quality of life, improved fitness and might very well have extended the lives of participants in the Generation 100 study."

This description of the results of the trial, however, was misleading. An examination of the pattern of changes in $\dot{V}O_{2peak}$ ("fitness") in a figure within the Supplementary Material (available online) does not show an "improvement" per se but rather an initial increase at the first year of follow-up, followed by a steady decline at the third and fifth years (although it is important to note that the absence of a decline from baseline by the fifth year of follow-up in a sample of septuagenarians should be interpreted as a positive outcome). Likewise, Table S6 in the supplementary material published with the study shows small declines in the physical and mental components of health-related quality of life over time. Moreover, as noted earlier, the phrase "five years of high-intensity interval training" is a misrepresentation, insofar as most participants initially allocated to "the HIIT group" either switched to moderate-intensity continuous exercise or dropped out of the study. Finally, the expression "might very well have extended the lives of participants" was arguably a questionable way of stating that the difference was not statistically significant. In the text of the press release, the lead author was quoted as saying the following on this issue:

> In the interval training group, 3% of the participants had died after five years. The percentage was 6% in the moderate group. The difference is not statistically significant, but the trend is so clear that we believe the results give good reason to recommend high-intensity training for the elderly.

When the results were summarized for American readers in *The New York Times*, the important caveat "not statistically

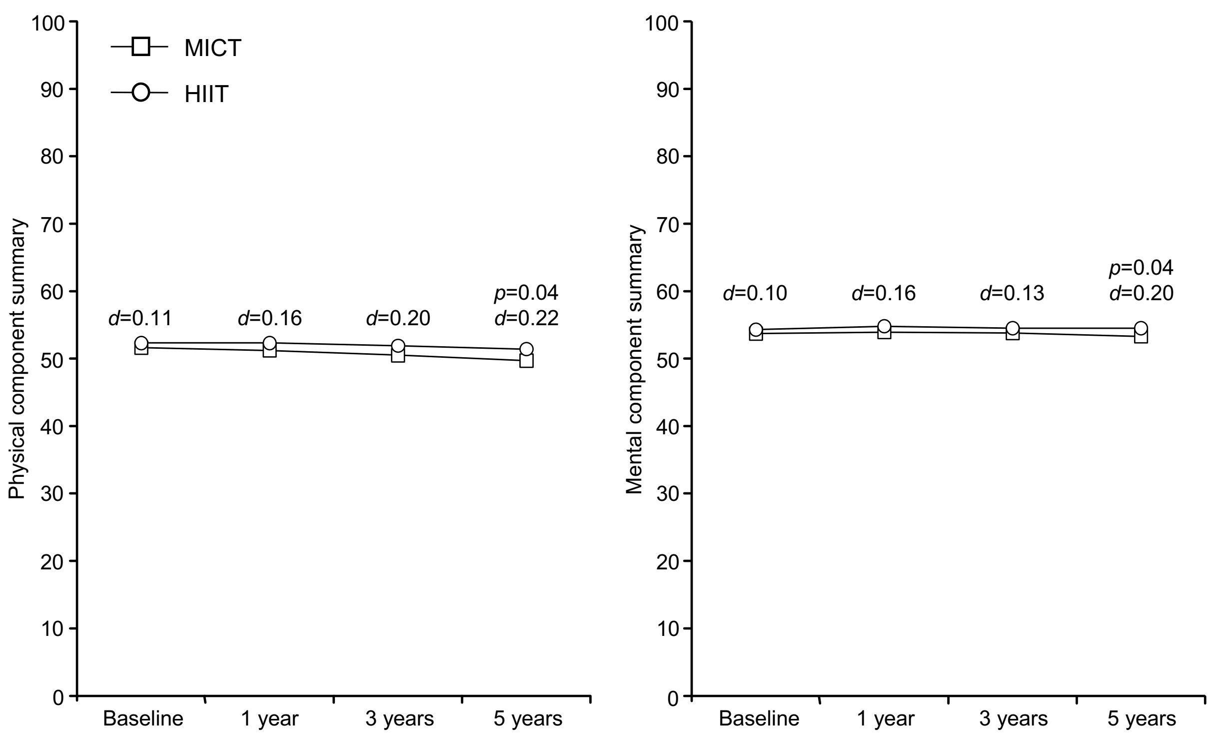

Figure 1 — Differences between the HIIT group and the MICT group in the Generation 100 trial on the physical component (left panel) and mental component (right panel) of health-related quality of life. It can be seen that half of the differences deemed statistically significant ($p = .04$) at the fifth year were already present at baseline as a result of imperfect randomization. HIIT = high-intensity interval training; MICT = moderate-intensity continuous training.

significant" and the equally important caveat "we believe" were omitted. The subtitle of the article ("Including high-intensity training in your workouts provided better protection against premature death than moderate workouts alone") insinuated that "better protection" was a definitively established result. No reference was made to the fact that the differences found were not statistically significant: "the men and women in the high-intensity-intervals group were about 2 percent less likely to have died than those in the control group, and 3 percent less likely to die than anyone in the longer, moderate-exercise group." The results on $\dot{V}O_{2peak}$ and health-related quality of life were likewise misrepresented: "The men and women in the interval group also were more fit now and reported greater gains in their quality of life than the other volunteers." As explained earlier, participants were not "more fit" than they were at baseline and did not experience any "gains in their quality of life" compared to baseline. Therefore, it could be argued that the "spin," ambiguity, and absence of crucial caveats found in the university press release influenced the subsequent newspaper coverage (Schwartz et al., 2012; Woloshin et al., 2009; Yavchitz et al., 2012).

It is noteworthy that, in this case, both the press release issued by the *British Medical Journal* on October 7, 2020 ("Exercise intensity not linked to mortality risk in older adults, finds trial"[3]) and a summary published in the December 22, 2020 issue of the *Journal of the American Medical Association* ("Exercise intensity unrelated to older adults' mortality risk"[4]; Slomski, 2020) were both accurate and unambiguous in their representations of the results.

In summary, despite assertions that the Generation 100 trial demonstrated that "HIIT is superior to continuous training with moderate intensity and standard training recommendations" with regard to reducing mortality risk (Joisten et al., 2022, p. 85), this conclusion is unsupported by a critical review of the data. What remains as an interesting, albeit unanswered (and perhaps unanswerable), question is what is causing some researchers to argue otherwise.

"It Doubles Endurance Performance in Just 15 Minutes of Training"

On June 1, 2005, the press office of McMaster University in Canada issued a press release entitled "A few 30 second sprints as beneficial as hour-long jog."[5] The study in question, which helped launch the HIIT phenomenon, was a study of only six men and two women, and it did not involve a comparison to an "hour-long jog." The participants, described as "recreationally active" university students ($\dot{V}O_{2peak}$: 44.6 ± 9 ml·kg^{-1}·min^{-1}), after taking part in only six training sessions over a period of 14 days (Monday, Wednesday, and Friday), reportedly doubled their cycling endurance performance (time to fatigue while pedaling at 80% $\dot{V}O_{2peak}$), from 26 ± 14 to 51 ± 31 min, without a change in maximal aerobic capacity (Burgomaster et al., 2005). The authors noted that the improvement would have been even larger (six of the eight participants more than doubled their performance) if it was not for a single participant who showed a slight decline. The training consisted of four to seven 30-s Wingate tests of maximal anaerobic capacity, performed back-to-back at "all out" intensity, interspersed with 4-min periods of rest or low-intensity pedaling (<50 rpm, <30 W). Thus, the total duration of high-intensity training that was needed to produce this dramatic improvement in endurance was only about 15 min.

A critical reading of the study by Burgomaster et al. (2005) reveals several potential sources of bias, including the absence of random allocation to groups, the assessment of outcomes (including endurance performance) by nonblinded assessors, uncertainty regarding the possibility of selective outcome reporting due to the lack of preregistration, the nonreporting of statistical details (e.g., t values, degrees of freedom), the lack of adjustment for the inflation of the Type I error rate due to multiple comparisons (each test was performed at $p < .05$), and the extremely low level of statistical power (6.6% to investigate a small effect, 15.4% for a medium effect, 32.0% for a large effect, assuming no measurement error and no adjustment for multiplicity).

These problems, while serious and likely consequential, even if taken collectively, may still not account for the astounding ~100% improvement in endurance performance after only 15 min of training. In an insightful critical commentary, psychophysiologist Paul Grossman (2005), characterizing the "doubling of endurance capacity" as "remarkable," identified certain "behavioral variables" that, if left uncontrolled, "could conceivably lead to false inferences" (p. 2473). For example, Grossman asked about the participants:

> Did [the participants] derive from a pool of interested exercise physiology students, did they have any knowledge of the hypotheses of the study during the training period that could have influenced their exercise performance during or between the protocol laboratory trainings, and/or were subjects randomly assigned to control and experimental groups (or possibly explicitly or implicitly selected according to criteria that could optimize positive findings, e.g., the experimental subjects may have had more prior experience with exercise studies)? (p. 2473)

Grossman (2005) explained that "awareness of the core hypotheses of the study among subjects or laboratory research staff could affect those exercise measures that are modifiable by means of variations in motivation (effort-dependent variables, e.g., cycle endurance time to fatigue)" (p. 2473). Even when the participants are not aware of the research hypotheses a priori, Grossman explained that participants, being active agents in the process, always try to infer the expectations of researchers, and often succeed, given that researchers tend to tacitly broadcast them through subtle differences in instructions, nonverbal cues, or communication styles:

> Differences in expectations can be explicitly or implicitly communicated to individual groups. When effort-dependent measures are employed (e.g., spirometric evaluations), slight differences in instruction or expectation may lead to significant effects. Subjects may be made aware of the major hypotheses of the study, and this could influence their behavior inside or outside the laboratory. Unstandardized procedures in the experimental settings may also occur, for example, haphazard and varying instruction and behavior of experimenters, flurries of extraneous experimenter activity and disturbance while carrying out a protocol, or varying numbers of experimenters and/or observers during measurements. These factors may create extra "noise," or error variance, if applied unsystematically or clearly bias experimental hypotheses if applied systematically. (p. 2474)

The researchers largely dismissed these concerns, reassuring readers that they are "keenly aware of the potential for experimenter bias and/or sloppy procedures to create error variance and/or prejudice study outcomes" and "judiciously guard against the various pitfalls in subject recruitment and other parameters that Dr. Grossman describes" (Gibala et al., 2005, p. 2475). Among the specific examples they gave for "judiciously guarding against the various pitfalls," the researchers noted that they employed a "control group for the performance test that was drawn from the same subject population" and applied "appropriate and rigorous statistical analyses" (p. 2475). However, as noted earlier, the control group was not formed via random allocation and the analyses greatly raised the risk of both Type I (i.e., false positives, or rejecting the null hypothesis when it is true) and Type II errors of statistical inference (i.e., false negatives, or failing to reject the null hypothesis when it is false).

Of greater concern was a new description of the experimental procedures that was published several years later in a book written for the general public (Gibala, 2017). This new description revealed that the sources of bias Grossman had intuitively suspected were, in fact, present. The principal investigator explained that the individuals who conducted the intervention and outcome assessments were "talented and enthusiastic graduate students" who had participated in "brainstorming sessions" and were fully aware of the hypotheses and expectations of the study, having participated in its design. The study participants, who were generically described in the journal article as "recreationally active individuals from the . . . student population" (Burgomaster et al., 2005, p. 1986), were described in the book, more specifically, as "the sort of athletic young adults, men and women, who tend to be around [a] Kinesiology Department" (Gibala, 2017, p. 18).

Since the participants were students in the same department as the investigators, it is possible that they were made aware of the hypotheses of the study and could deduce the expectations of the investigators. There are several reasons that strengthen this possibility. First, the "study purpose was disclosed to all subjects before their participation" as required by "institutional Research Ethics Board guidelines" (Gibala et al., 2005, p. 2475). Second, "all subjects performed extensive familiarization trials before testing" (Burgomaster et al., 2005, p. 1989), thus having ample opportunity to interact with and develop interpersonal bonds with the investigators. Third, this is suggested by the description of the training sessions in the book, which differed considerably from the mundane description in the journal article ("Subjects were verbally encouraged to continue pedaling as fast as possible"; Burgomaster et al., 2005, p. 1986):

> The training sessions were pretty intense. In our write-up of the experiment, we noted that the subjects were "verbally encouraged" during their sprints. That's a staid depiction of what actually happened. The atmosphere was as loud and enthusiastic as any I've seen in a laboratory setting. Rock music blared. As each participant shifted into the sprint, a half-dozen grad students gathered around to offer encouragement: High-volume shouts. Lots of "Go! Go! GO!" and "YOU CAN DO IT!" Then, after the sprint, there were high-fives and pats on the back. (Gibala, 2017, pp. 18–19)

The day of outcome assessment was described as "the most exciting day" (Gibala, 2017, p. 19). During testing, "the lab was quiet," no "encouragement or feedback" was given, and the investigators "tried to keep [their] expressions blank, so as not to affect the subject's performance" (p. 19). However, once the testing got underway, it became "difficult to maintain that façade of impartiality" (p. 19) because "the numbers were crazy" (p. 19), "amazing" (p. 20), "incredible" (p. 20), "remarkable" (p. 20), "powerful" (p. 20), "near magic" (p. 20), "enormous" (p. 20), and "miraculous" (p. 21).

To readers with training in psychology, these descriptions are probably reminiscent of examples they encountered during courses on psychological research methods and, in particular, on the importance of controlling for threats to internal validity. A large set of highly influential articles, published mainly in the 1960s and 1970s, raised awareness among psychologists about the potential sources of bias associated with both participants and investigators (e.g., Mahoney, 1979; Orne, 1962; Rosenthal, 1963). Among the most relevant sources of bias during the testing sessions in Burgomaster et al. (2005), based on the descriptions provided above, were "outcome-orientation effects" (Rosenthal, 1964), namely (usually, unintentional) cuing or signaling that convey to participants the expectations and wishes of the experimenters, and the so-called "early data returns effect," whereby positive results arising from testing the first batch of participants cause mood improvements and subtle changes in the behavior of the investigators (Rosenthal et al., 1965). Collectively, these factors are known to raise the likelihood of results that are in line with the expectations and wishes of the investigators.

Similarly, during the training sessions, the descriptions make it clear that several elements were introduced (without disclosing them in the research article), besides the high-intensity interval exercise per se, thus inviting possible alternative explanations for the intervention effects (also see Halperin et al., 2015). Loud, high-energy music has been extensively documented to have an independent ergogenic effect, in addition to lowering perceptions of exertion and increasing ratings of pleasure during exercise (Terry et al., 2020). Moreover, following early research on the social facilitation of task performance (Zajonc, 1965), exercise researchers have published substantial evidence that such factors as praise and strong verbal encouragement can also have a powerful, independent ergogenic effect (Edwards et al., 2018; Midgley et al., 2018). Warnings about the potential biasing effects of social interactions in the exercise laboratory have been issued for decades. Peter Karpovich (1937), a student of Ivan Pavlov and one of the pioneers of exercise physiology in North America, noted that "cheering had the most marked effect" (p. 628):

> In performing these tests special care should be taken to eliminate any factor which may affect the subject, otherwise the results will not be reliable. The following experiment illustrates this. A subject was working on the arm ergograph. The writer was sitting behind a screen, hidden from the subject yet able to watch him. An assistant was pretending to do some work at a table ten feet away from the ergograph. Occasionally the assistant would rise, walk to the ergograph without paying any attention to the subject and look at the kymograph. Each time he did this the height of the contractions immediately increased. The same thing happened also when the writer rose and looked at the subject. (pp. 628–629)

Remarkably, despite their extraordinary nature, the results reported by Burgomaster et al. (2005) have not attracted the critical scrutiny of researchers in kinesiology besides the early incisive commentary by Grossman (2005). However, one study that partly replicated the procedures of Burgomaster et al. and used a sample with similar characteristics failed to reproduce the original results (Bertschinger et al., 2020).

In summary, although Burgomaster et al. (2005) expressed the belief that it is "unlikely that this finding is a spurious result"

(p. 1989), the doubling of endurance performance after only 15 min of training (in six sessions over 2 weeks) is such an extraordinary result that it should have prompted close scrutiny and replication attempts by the global exercise-science-research community. Critical appraisal of the research methods, as summarized in the previous paragraphs, would have revealed several plausible alternative explanations for this result other than the effect of training alone. The mechanisms we outlined here are uncontroversial and have long been established as representing potentially powerful threats to internal validity. Nevertheless, it is noteworthy that, while the study by Burgomaster et al. has received over one thousand citations and became the impetus for the HIIT phenomenon, the insightful, incisive, and accurate critique by Grossman (2005) has been cited only once. This represents a remarkable demonstration of the phenomenon described by Serra-Garcia and Gneezy (2021), namely that, although experts (e.g., peer reviewers, journal editors) can generally recognize methodological and statistical weaknesses, and therefore can identify studies with a low reproducibility, they "may apply lower standards" (p. 4) when considering its suitability for publication "when the paper is more interesting" (defined as the potential to attract more attention).

"One Minute Is as Effective as 45 Minutes of Moderate-Intensity Continuous Exercise"

On April 27, 2016, an article in *The New York Times* presented an astounding claim, namely that "1 minute of all-out exercise may have the benefits of 45 minutes of moderate exertion."[6] This claim was based on a study, according to which only 1 min of HIIT per session (within 10-min sessions that also included warm-up, cooldown, and recovery periods), performed three times per week, resulted in "strikingly similar 19% improvement in $\dot{V}O_{2peak}$ after 12 weeks" (Gillen et al., 2016, p. 8) as 50-min sessions (including warm-up and cooldown) of moderate-intensity continuous exercise in previously inactive men. This finding is "extraordinary" because, over the past 70 years, research has yielded no indication that the "exchange rate" between high-intensity exercise and moderate-intensity exercise could be near 1-to-45 ("3x20-second 'all-out' cycling efforts" vs. "45 minutes of continuous cycling"; see Gillen et al., 2016, p. 4) or 1-to-50 ("3 minutes of intense intermittent exercise per week" vs. "150 minutes per week of moderate-intensity continuous training"; see Gillen et al., 2016, p. 10). If true, this claim would warrant the overhaul of both the science and the practice of exercise.

The study was a comparison of three nonrandomized groups, one engaging in interval training ($n = 9$ but 8 for some analyses), the second engaging in moderate-intensity continuous exercise ($n = 10$ but 9 for some analyses), and the third being a nonexercise control ($n = 6$ but 5 for some analyses). Interval training involved three weekly sessions of 3×20-s "all-out" cycling against a resistance equal to 5% of body mass (~500 W), interspersed with 2-min active-recovery periods at 50 W. Moderate-intensity continuous exercise involved 45 min of continuous cycling at ~70% of maximal heart rate (~110 W). Both interval training and moderate-intensity continuous exercise were preceded by a 2-min warm-up and followed by a 3-min cooldown at 50 W.

The researchers examined a wide array of outcomes, including variables from a maximal exercise test, a body composition analysis, an intravenous glucose tolerance test, a resting muscle biopsy, and an arterial ultrasound imaging test (the latter data reported in a different publication). In the Gillen et al. (2016) report alone, there were 23 dependent variables (measured by nonblinded assessors), each analyzed with a criterion of $p < .05$ (i.e., unadjusted for multiplicity). Some of the analyses found that, while interval training and moderate-intensity continuous exercise differed from no-exercise control, they did not differ significantly from each other. For example, Gillen et al. reported "strikingly similar 19% improvements" in maximal oxygen uptake between the groups (although, at least at the level of group means, moderate-intensity continuous exercise resulted in 19% improvement or 0.5 L/min, from 2.7 to 3.2 L/min, whereas interval exercise resulted in 15% improvement or 0.4 L/min, from 2.6 to 3.0 L/min). On the basis of these findings, the researchers concluded that (a) "12 weeks of [sprint-interval training] in previously inactive men improved insulin sensitivity, cardiorespiratory fitness, and skeletal muscle mitochondrial content *to the same extent* as [moderate-intensity continuous exercise], despite a five-fold lower exercise volume and training time commitment" and (b) "a [sprint-interval training] protocol involving 3 minutes of intense intermittent exercise per week, within a total time commitment of 30 minutes, is *as effective as* 150 minutes per week of moderate-intensity continuous training for increasing insulin sensitivity, cardiorespiratory fitness and skeletal muscle mitochondrial content in previously inactive men" (p. 10, emphasis added).

This study exemplifies statistical problems that can raise the risk of both Type I and Type II errors of statistical inference to unacceptably high levels. Analyzing 23 dependent variables without any correction for the inflation of α can raise the risk of Type I error up to 69.3%, which entails that at least some of the analyses could reach the $p < .05$ criterion of statistical significance by chance. The likelihood of such events is enhanced by the extremely small sample sizes for the groups engaged in interval training and moderate-intensity continuous exercise (i.e., nine and 10 initially, but eight and nine for some analyses). Such small sample sizes entail extreme volatility of the sample means (e.g., due to presence of one or two unusual scores). At the same time, claims of a treatment being "as effective as" or inducing changes "to the same extent as" another treatment on the basis of the $p > .05$ criterion for a comparison between the group means is statistically inappropriate (Lakens, 2017; Parkhurst, 2001; Smith, 2020). Sample sizes of nine and 10 entail only 13.0% statistical power for intergroup comparisons for an effect on maximal oxygen uptake of $d = 0.40$ (as per the meta-analysis by Mattioni Maturana et al., 2021), raising the risk of a Type II error to 87.0%. For example, in regard to the finding highlighted by the researchers, namely that interval training and moderate-intensity continuous exercise were "as effective" in improving cardiovascular fitness or changed cardiovascular fitness "to the same extent," the v-statistic (Davis-Stober & Dana, 2014; Lakens & Evers, 2014) was zero. This indicates that a model of estimation in which both the direction and the magnitude of the effect were determined at random (and, therefore, the model was, by definition, entirely uninformative) can be expected to be consistently more accurate than a model based on the observed data (i.e., estimation based on ordinary least squares).

Besides these fundamental statistical concerns, the study by Gillen et al. (2016) underscores the need for readers to remain critical and to attempt to scrutinize and independently verify, to the extent possible, all claims made in research articles. For example, in reviewing previous work in their introduction, the researchers made a series of striking claims. They claimed that, according to one study (Jung et al., 2014), "subjects reported greater enjoyment of, and a preference to engage in [HIIT]" compared with moderate-intensity continuous exercise (Gillen et al., 2016, p. 10). However, this assertion is inconsistent with the findings in the original report. According to the data reported by Jung et al. (2014), "[HIIT] was not statistically more enjoyable than [moderate-intensity continuous exercise] at $p = .08$," there were "no significant differences in enjoyment between the [HIIT] and [moderate-intensity continuous exercise] trials ($p = .74$)," and "there was no significant difference in anticipated enjoyment between [HIIT] and [moderate-intensity continuous] exercise ($p = .60$)" (p. 12). Likewise, preference for exercise modality was assessed with two questions but neither intergroup comparison reached statistical significance ($p = .07$ and $p = .59$, respectively).

Likewise, Gillen et al. (2016) asserted that, according to another study (Jung et al., 2015), "adherence to a 4-week high-intensity interval training program, assessed by self-report in free-living conditions, was greater than for moderate-intensity continuous exercise in people with prediabetes" (p. 10). However, this statement was also inconsistent with the data in the original report. Jung et al. (2015) found that, out of 12 sessions prescribed over 4 weeks, participants in HIIT reported missing 1.32 of their 25-min-long sessions (i.e., 89 ± 11% adherence), whereas participants in moderate-intensity continuous exercise reported missing 3.48 of their 50-min-long sessions (i.e., 71 ± 31% adherence). This comparison did not reach the criterion of statistical significance ($p = .05133$), but Jung et al. (2015) characterized it as "a significant difference in adherence rates" (p. 5). More importantly, (a) the higher adherence did not translate to more physical activity, since the researchers found no difference in moderate-to-vigorous physical activity based on accelerometers (either overall or in 10-min bouts) and (b) significantly more participants dropped out of the HIIT group (five of 15 or 33%) than the moderate-intensity continuous exercise group (one of 17 or 6%), $\chi^2 = 3.94$, $p = .047$.

In summary, given the concerns outlined here, extraordinary claims that "1 minute of intense intermittent exercise" can improve indices of cardiometabolic health "*to the same extent as* … 50 minutes of continuous exercise at a moderate pace" (Gillen et al., 2016, p. 8, emphasis added) or "3 minutes of intense intermittent exercise per week" are "*as effective as* 150 minutes per week of moderate-intensity continuous training" (p. 10, emphasis added) should be viewed cautiously. The high risk of both Type I (i.e., inflation of α up to nearly 14 times the conventional rate of 5%) and Type II errors (i.e., inflation of β to more than four times the conventional rate of 20%) suggests that the results cannot be considered credible. In addition, several of the assertions about the benefits of HIIT included in the article diverge from the original evidence and are, therefore, misleading.

"HIIT Is More Pleasant and Enjoyable Than Moderate-Intensity Continuous Exercise"

The authors of a meta-analysis (Oliveira et al., 2018) comparing pleasure and enjoyment responses to HIIT versus moderate-intensity continuous exercise concluded that "most of the comparisons performed presented positive effects for HIIT" (p. 1), and therefore, "HIIT exercise may contribute to obtaining psychological responses that are equal to or more positive than [moderate-intensity continuous exercise] sessions" (p. 8). Thus, the authors proposed that "HIIT exercise may be recommended for obtaining positive psychological responses" (p. 1). These assertions, which turned decades of prior research on the negative relation between high exercise intensity and pleasure on its head (e.g., Ekkekakis et al., 2011), have had a considerable influence on the literature. For example, citing this meta-analysis as evidence, researchers have stated that "several studies have reported that HIIT usually offers more enjoyment and affective responses both during and immediately after exercise. Therefore, HIIT may be an alternative to [moderate-intensity continuous training] for inducing positive physiological adaptations and doing so in a more enjoyable way" (Serrablo-Torrejon et al., 2020, p. 2).

A closer inspection of the details of the meta-analysis conducted by Oliveira et al. (2018), and some of the articles included therein, reveals several irregularities. First, three studies contributed multiple effect sizes, with the same samples used repeatedly. This approach violates a fundamental assumption underpinning the random-effects model of meta-analysis, namely the assumption of the independence of effect sizes (Cheung, 2019; Gucciardi et al., 2022; Senn, 2009). Violating the assumption of independence falsely attenuates estimates of standard error, thereby resulting in deceptively narrow confidence intervals and thus an inflation of the Type I error rate. Second, claims in favor of HIIT should have been tempered by the fact that the effect sizes were strongly heterogeneous ($I^2 = 79\%$ for pleasure, $I^2 = 69\%$ for enjoyment) and, importantly, the pooled effect size for pleasure was not significantly different from zero.

As consequential as these issues are, there were even more serious interpretational problems. Although the stated purpose of the meta-analysis was to compare "the acute effects of HIIT and [moderate-intensity continuous training] on affective and enjoyment responses" (p. 2), some of the effect sizes that were pooled referred, instead, to comparisons between HIIT and continuous exercise sessions performed at intensities that were as high as or *higher* than those used in HIIT but were misleadingly characterized as "moderate-intensity." Specifically, Oliveira et al. (2018) noted that "the studies of [Jung et al., 2014], [Kilpatrick et al., 2015] and [Martinez et al., 2015] showed in general beneficial effects of HIIT on affective responses compared to [moderate-intensity continuous training]" (p. 5). These were the same three studies that contributed multiple effect sizes and were, therefore, collectively assigned more weight in the meta-analysis. We next review the methodologies of these three studies in more detail, to explain why several of the effect sizes derived from them did not represent comparisons between HIIT and moderate-intensity continuous exercise, contrary to the claim by Oliveira et al.

Jung et al.'s Study

This study included, in addition to a HIIT condition, two continuous-intensity conditions, one of which was labeled "continuous vigorous intensity" ("CVI"). In this condition, the participants had to maintain 80% of peak power for 20 min, a very demanding task. Although Jung et al. (2014) did not disclose the full details of the heart rate responses during their experimental conditions, they did report that, at minute 18:30 of the 20-min "CVI" condition, participants reached 169.40 ± 14.24 beats/min. This intensity approximately corresponded to 90% of maximal heart rate, which is near the top of the "vigorous" range (i.e., 77%–95% of maximal heart rate), and just short of the range considered "near maximal" (American College of Sports Medicine, 2022). The heart rate reached at the same time during the HIIT condition was slightly *lower* (167.93 ± 14.62 beats/min). It should be emphasized, however, that, during the HIIT session, participants were exposed to this level of intensity intermittently and only for a total of 10 min (during the ten 1-min intervals), whereas participants in the "CVI" condition exercised continuously for 20 min. Nevertheless, Oliveira et al. (2018) analyzed the comparison between HIIT and

"CVI" as a comparison between HIIT and "moderate-intensity continuous exercise."

Kilpatrick et al.'s Study

This study included four experimental conditions: (a) "moderate continuous," (b) "heavy continuous," (c) "heavy interval," and (d) "severe interval." The domains of "moderate," "heavy," and "severe" intensity have profound implications for metabolic strain. In theory, intensities labeled "heavy" are supposed to be higher than those labeled "moderate," and intensities labeled "severe" are supposed to be higher than those labeled "heavy." Kilpatrick et al. (2015) set "moderate" intensity as 20% below the ventilatory threshold, "heavy" intensity at the ventilatory threshold, and "severe" intensity at 20% above the ventilatory threshold. However, they did not take into account that physiological indices of metabolic strain (a) do not rise instantly but rather over a period of minutes (with the rate depending on the workload) and (b) tend to exhibit an upward drift over time (i.e., do not remain steady) during continuous exercise, especially at high workloads. The combination of these two factors crucially altered the meaning of the labels given to the intensity conditions. The so-called "moderate continuous" condition culminated in 161 ± 18 beats/min, whereas the highest heart rate recorded during the so-called "heavy interval" condition was only 154 ± 16 beats/min. Likewise, the so-called "heavy continuous" condition culminated in 174 ± 17 beats/min, whereas the highest heart rate recorded during the so-called "severe interval" condition was only 165 ± 17 beats/min. Moreover, while the participants were exposed to supposedly "heavy" and "severe" intensities for only a total of 10 out of 20 min during the interval conditions, they were exposed to the "moderate" and "heavy" intensities for 20 out of 20 min during the continuous conditions (i.e., resulting in a combination of higher physiological intensities and double duration of exposure). Despite the fact that the lead researcher of the Kilpatrick et al. (2015) study was also a coauthor of the meta-analysis by Oliveira et al. (2018), the meta-analysis included one comparison between the (higher intensity) "moderate continuous" and the (lower intensity) "heavy interval" conditions, and another comparison between the (higher intensity) "heavy continuous" and the (lower intensity) "severe interval" conditions, both of which were analyzed, inappropriately, as representing comparisons between "moderate-intensity continuous exercise" and "HIIT."

Martinez et al.'s Study

This study did not include a moderate-intensity continuous exercise condition at all. It involved a comparison between a 20-min continuous exercise condition performed at "heavy" intensity (defined as 10% of the difference between the ventilatory threshold and maximal capacity) and three 24-min HIIT conditions performed at "severe" intensity (defined as 60% of the difference between the ventilatory threshold and maximal capacity), the difference between them being the duration of the high-intensity intervals and rest periods (30, 60, and 120 s). Although Martinez et al. (2015) did not disclose the full details of the heart rate responses during their experimental conditions, they did report that the overall average during the "heavy-intensity continuous" condition was the highest of all experimental conditions (158 ± 14 beats/min, approximately corresponding to 84 ± 7% of peak heart rate). This intensity is within what is considered the range of "vigorous" intensity (i.e., 77%–95% of maximal heart rate; American College of Sports Medicine, 2022) rather than "moderate" (i.e., 64%–76% of maximal heart rate). Data reported in the original Master's thesis by Martinez, based on 14 of the eventual 20 participants, show that heart rate in the "continuous heavy" condition progressively drifted upward, from 141 ± 14 to 167 ± 17 beats/min (i.e., approximately equivalent to a drift from 75% to 88% of peak heart rate), and was within the "vigorous" range (rather than "moderate") for almost the entire 20-min session. In contrast, only one of the three HIIT conditions (i.e., involving six 120-s intervals) exhibited heart rate peaks that were higher (from 151 ± 21 to 180 ± 10 beats/min). During the other two HIIT conditions, heart rate peaked at levels lower than (30-s intervals: from 137 ± 20 to 149 ± 18 beats/min) or similar to (60-s intervals: from 140 ± 22 to 168 ± 14 beats/min) the levels recorded during the "continuous heavy" condition. Furthermore, while participants engaged in exercise continuously for 20 min during the "continuous heavy" condition, they were only exposed to high intensities for a total of 12 out of the 24 min in the three interval conditions (recovery periods were at only 5 ± 4% of peak workload or 10%–20% of peak capacity). In other words, the condition labeled "continuous heavy" involved intensity that was higher than or similar to two of the three interval conditions and exposure to high exercise intensity for a total duration that was 40% longer than all three interval conditions. Despite the fact that the senior investigator of Martinez et al.'s study was also a coauthor of the meta-analysis by Oliveira et al. (2018), the "continuous heavy" condition was analyzed as being representative of "moderate-intensity continuous exercise" and the two interval conditions that involved lower intensity than the "continuous heavy" condition (i.e., 30- and 60-s intervals) were considered as "HIIT."

In summary, one of the two effect sizes associated with Jung et al. (2014), two of the four effect sizes associated with Kilpatrick et al. (2015), and two of the three effect sizes associated with Martinez et al. (2015) represented comparisons in which the conditions that were mislabeled "moderate-intensity continuous exercise" involved intensities that were similar to or higher than those in the "HIIT" conditions in terms of heart-rate responses (in addition to involving, in all cases, longer-duration exposures). It should be apparent that, if participants engage in exercise that is (a) intermittent, (b) performed at lower intensity, and (c) for a shorter duration, then they will report more pleasure and enjoyment compared with exercise that is (a) continuous, (b) performed at higher intensity, and (c) for a longer duration. Remarkably, the mislabeling of the experimental conditions described in the preceding paragraphs has been overlooked not only by Oliveira et al. (2018) but also in other reviews (e.g., Niven et al., 2021; Stork et al., 2017), which have continued to present the studies by Jung et al. (2014), Kilpatrick et al. (2015), and Martinez et al. as containing evidence supporting the affective and enjoyment benefits of HIIT.

Taken together, the problems identified thus far suffice to invalidate the conclusions of the meta-analysis by Oliveira et al. (2018). However, the validity of the meta-analysis is further weakened by additional serious concerns. Specifically, it can be easily demonstrated that some of the reported effect sizes could not have been derived from the original studies. The authors reported that "for the studies with several measurements of … affective responses (pre, during and post exercise), [they] calculated mean and standard deviation values reducing the data to only one value in each exercise condition" (p. 4). Despite experimenting with numerous possibilities, however, it has not been possible to verify the numbers that were reported. For example, participants in Jung et al. (2014) reported consistently lower levels of pleasure throughout the HIIT condition compared with the moderate-intensity continuous exercise condition (labeled "CMI"). Accordingly, Jung et al. (2014) themselves unambiguously concluded that "affect was significantly less positive in the [HIIT] . . . trial than in the [continuous moderate-intensity] trial" (p. 10). Nevertheless, the effect size associated with the comparison between the HIIT and CMI conditions in the meta-analysis by Oliveira et al. was statistically significant in favor of HIIT (0.62, 95% CI [0.19, 1.04]). Similarly, participants reported only slightly less pleasure during the vigorous-intensity continuous exercise condition (labeled "CVI") than during HIIT. However, Oliveira et al. reported a large and significant effect size of 1.38 in favor of HIIT (95% CI [0.91, 1.85]).

The meta-analysis by Oliveira et al. (2018) continues to be cited, even by some of its coauthors, as providing evidence that HIIT is pleasant and enjoyable (e.g., Fleming et al., 2020). Likewise, the authors of the Jung et al. (2014) study, who could presumably detect that their study was portrayed as having found the opposite of what they themselves had reported, have also been echoing the conclusions of the meta-analysis without drawing attention to the aforementioned irregularities. For example, citing Oliveira et al.'s meta-analysis as supporting evidence, Eather et al. (2020) stated: "it has been suggested that HIIT is unlikely to have utility as a public health intervention as programs will not be adopted or maintained by the general public; however, new findings suggest that HIIT sessions . . . can positively influence affective valence" (p. 115). Similarly, citing Oliveira et al., Sabag et al. (2022) stated: "there is a relative lack of data to confirm that HIIT leads to poor affective responses or that these supposed responses impede exercise adoption and/or adherence; in fact, a recent meta-analysis showed that HIIT was superior to [moderate-intensity continuous training] for improving affective and enjoyment responses to exercise" (p. 10). Likewise, citing Oliveira et al., Mastrofini et al. (2022) wrote: "Concerns regarding the tolerability of HIIT have been raised, but research makes clear that a wide variety of HIIT protocols produce psychological responses that are perhaps equal to and in some cases more positive than the continuous moderate-intensity exercise that is most often prescribed for health and wellness" (pp. 1–2).

How Should Scientists Treat Extraordinary Claims? HIIT in Epistemological Context

In 2011, psychologist Daryl Bem, an experienced researcher affiliated with one of the most prestigious universities in the United States (Cornell University), published a paper in the flagship journal of the American Psychological Association, the *Journal of Personality and Social Psychology*. In it, Bem reported the results of nine experiments, comprising more than 1,000 participants, which provided evidence for "psi effects," such as precognition and preemotion (i.e., having better-than-chance conscious awareness of and congruent emotional reactions to events that have not yet occurred and cannot be anticipated). For example, Bem (2011) found that undergraduate students who were told that they were participating in an experiment investigating extrasensory perception predicted the future position (left or right) of an erotic image on the computer screen at a rate that was significantly higher than the 50% that could be expected by chance (i.e., 53.1%, $p = .01$, $d = 0.25$). Bem also found evidence of retroactive priming, with students rating positive images as pleasant and negative images as unpleasant, 15.0 ms faster when the images were followed, not preceded, by subliminally presented primes (i.e., the words "beautiful" or "ugly," respectively), $p = .006$, $d = 0.25$. Bem insisted that he "[does] not advocate believing impossible things" (p. 423) but suggested that, based on the findings he reported, individuals may "use psi information implicitly and nonconsciously to enhance their performance in a wide variety of everyday tasks" (p. 422), which is an ability that "might have been acquired through evolution by conferring survival and reproductive advantage on the species" (p. 422).

Commenting on their decision to publish the extremely controversial article by Bem, editors Judd and Gawronski (2011) wrote that they did so after the manuscript successfully passed through "a rigorous review process, involving a large set of extremely thorough reviews by distinguished experts," despite the fact that "the reported findings conflict with [their] own beliefs about causality and that [they] find them extremely puzzling" (p. 406). Echoing a timeless scientific principle famously endorsed by Scottish philosopher David Hume (1748/1921), French polymath Pierre-Simon de Laplace (1814/1951), and American astrophysicist Carl Sagan (1979), Judd and Gawronski argued that "strong claims, running strongly counter to expectations and existing understandings, require exceptionally compelling evidence" (p. 406). So, in their effort to offer readers a more balanced perspective on the article by Bem, they also decided to publish a critique by Wagenmakers et al. (2011).

Wagenmakers et al. (2011) expressed the view that, rather than proving the existence of psi, the article by Bem proved that "the field of psychology currently uses methodological and statistical strategies that are too weak, too malleable, and offer far too many opportunities for researchers to befuddle themselves and their peers" (p. 426). Also echoing Hume, Laplace, and Sagan, Wagenmakers et al. expressed their agreement with the principle that "extraordinary claims require extraordinary evidence" (p. 428). When trying to decide whether an "anomalous phenomenon" is true or not, one must take into account whether the claimed phenomenon "conflicts with what we know to be true about the world" (p. 426). In Bayesian terms, they argued that "the prior probability attached to a given hypothesis affects the strength of evidence required to make a rational agent change his or her mind" (p. 428). In the case of a phenomenon like "psi," the prior probability should be considered very low because (a) "there is no mechanistic theory of precognition" and (b) "there is no real-life evidence that people can feel the future," since casinos still make profits (p. 428).

It should be noted that some philosophers of science have objected to the idea of two types of evidence, namely ordinary and extraordinary, sounding the alarm that what is often hiding under these terms is the resistance of the entrenched paradigm against innovation. Deming (2016) wrote that "extraordinary claims require extraordinary evidence" is an argument that can be misused "to suppress innovation and maintain orthodoxy" (p. 1329). He emphasized that "ideas, theories, or observations that are merely novel are not 'extraordinary,' nor do they require an 'extraordinary' amount of evidence for corroboration" (p. 1329). While this warning can be said to apply to numerous examples in the history of science, the argument can also be reversed: purported novelty can also be used as a defense, to justify the use of questionable research practices on the basis of the claim that they are widespread or commonly tolerated or overlooked.

In the case of Bem (2011), the critique by Wagenmakers et al. (2011) resembled Occam's razor: before one accepts "psi" and starts grappling with its dizzying implications for our current understanding of the laws of the universe, one must first ensure that the research methods complied with certain fundamental principles. Taking this approach, Wagenmakers et al. offered a series of elementary but compelling explanations that are taught as part of every introductory course in statistics or research methodology. For example, Bem noted that "all significance levels reported in this article are based on one-tailed tests" (p. 409); likely conducted multiple tests of probability without adjusting for the inflation of the Type I error rate (e.g., the experiment showing prediction of the location of an erotic image also included sets of neutral images, negative images, positive images, and images that were romantic but nonerotic); performed analyses on both transformed and untransformed data when the need for the transformations was not always established; and performed analyses of moderators (e.g., by sex) when it was not always clear whether these analyses were hypothesis-driven. Wagenmakers et al. concluded their critique by stating: "It is easy to blame Bem . . . . However, Bem played by the implicit rules that guide academic publishing . . . . Instead, our assessment suggests that something is deeply wrong with the way experimental psychologists design their studies and report their statistical results" (p. 431). Along similar lines, other authors devised the term "researcher degrees of freedom" to describe the "common (and accepted) practice for researchers to explore various analytic alternatives, to search for a combination that yields 'statistical significance,' and to then report only what 'worked'" (Simmons et al., 2011, p. 1359).

Since its publication, the article by Bem (2011) is widely regarded as the singular event that initiated a phase of self-reflection in psychology, as the discipline was forced to confront the deeply disconcerting fact that it was facing a replication crisis (Colling & Szucs, 2021). Others have argued that the article by Bem should not be seen as an indication that "science is broken," insofar as the problems were detected and widely exposed in the literature:

> This is precisely how science should work (i.e., science is supposed to correct its inevitable mistakes in due course). Instead, the problem is that, unlike the provisional findings reported by Bem, other high-profile findings sometimes appeared to pass the test of having been independently replicated, yet they ultimately turned out to have effect sizes very close to zero. Thus, something went wrong with the self-correction mechanism. (Wilson et al., 2020, p. 5561)

The history of science dealing with extraordinary claims contains some potentially useful exemplars and guidelines for how a self-correcting kinesiology should respond to the extraordinary claims in the HIIT literature: The level of evidence provided should be commensurate with the extraordinary nature of the claims. As Deming (2016) warned, it is possible that the "visceral" skepticism triggered by these extraordinary claims represents nothing more than the resistance of orthodoxy against novel ideas that threaten the paradigmatic status quo. However, reasonable people can probably agree that, before the textbooks are shredded and rewritten anew, it is essential to first ensure that the studies producing these fascinating results have adhered to fundamental methodological and statistical principles.

Kinesiology should let history be its guide. It is not the first scientific field that was infatuated by extraordinary claims, and it will not be the last. Extraordinary claims, "spin," and beguiling narratives are part of scientific discourse, just as they are part of marketing and the news industry. However, society has higher expectations of science, and science itself has traditionally aspired to be "self-correcting" (Alberts et al., 2015; Cofnas, 2016; Ioannidis, 2012; Jamieson, 2018). Failure to adhere to these ideals can undermine public trust in science and raise the risk of exacerbating anti-science movements, as seen around the world in recent years (Hotez, 2020). Confronted with extraordinary claims, kinesiology should heed the Hume–Laplace–Sagan principle, just as scientists have done throughout the history of science.

Notes

1. https://www.nytimes.com/2020/12/23/well/move/high-intensity-exercise-workouts.html
2. https://norwegianscitechnews.com/2020/10/high-intensity-training-best-for-older-people
3. https://www.bmj.com/company/newsroom/exercise-intensity-not-linked-to-mortality-risk-in-older-adults-finds-trial/
4. https://jamanetwork.com/journals/jama/article-abstract/2774401
5. https://dailynews.mcmaster.ca/articles/a-few-30-second-sprints-as-beneficial-as-hour-long-jog/
6. https://well.blogs.nytimes.com/2016/04/27/1-minute-of-all-out-exercise-may-equal-45-minutes-of-moderate-exertion/

References

Alberts, B., Cicerone, R.J., Fienberg, S.E., Kamb, A., McNutt, M., Nerem, R.M., Schekman, R., Shiffrin, R., Stodden, V., Suresh, S., Zuber, M.T., Pope, B.K., & Jamieson, K.H. (2015). Self-correction in science at work. *Science, 348*(6242), 1420–1422. https://doi.org/10.1126/science.aab3847

American College of Sports Medicine. (2022). *ACSM's guidelines for exercise testing and prescription* (11th ed.). Wolters Kluwer.

Andreacci, J.L., LeMura, L.M., Cohen, S.L., Urbansky, E.A., Chelland, S.A., & Von Duvillard, S.P. (2002). The effects of frequency of encouragement on performance during maximal exercise testing. *Journal of Sports Sciences, 20*(4), 345–352. https://doi.org/10.1080/026404102753576125

Baar, K. (2006). To perform your best: Work hard not long. *Journal of Physiology, 575*(Pt. 3), 690. https://doi.org/10.1113/jphysiol.2006.117317

Bahls, M., Baurecht, H., Hanssen, H., & Van Craenenbroeck, E.M. (2022). How to establish causality between physical inactivity and mortality? *European Journal of Preventive Cardiology, 29*(8), e266–e267. https://doi.org/10.1093/eurjpc/zwac021

Bailar, J.C. (2006). How to distort the scientific record without actually lying: Truth, and the arts of science. *European Journal of Oncology, 11*(4), 217–224.

Barreto, P. (2013). Why are we failing to promote physical activity globally? *Bulletin of the World Health Organization, 91*(6), 390–390A. https://doi.org/10.2471/BLT.13.120790

Bem, D.J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*(3), 407–425. https://doi.org/10.1037/a0021524

Bertschinger, R., Giboin, L.S., & Gruber, M. (2020). Six sessions of sprint-interval training did not improve endurance and neuromuscular performance in untrained men. *Frontiers in Physiology, 10,* 1578. https://doi.org/10.3389/fphys.2019.01578

Boutron, I., Dutton, S., Ravaud, P., & Altman, D.G. (2010). Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *Journal of the American Medical Association, 303*(20), 2058–2064. https://doi.org/10.1001/jama.2010.651

Boutron, I., & Ravaud, P. (2018). Misrepresentation and distortion of research in biomedical literature. *Proceedings of the National Academy of Sciences of the United States of America, 115*(11), 2613–2619. https://doi.org/10.1073/pnas.1710755115

Burgomaster, K.A., Hughes, S.C., Heigenhauser, G.J., Bradwell, S.N., & Gibala, M.J. (2005). Six sessions of sprint interval training increases muscle oxidative potential and cycle endurance capacity in humans. *Journal of Applied Physiology, 98*(6), 1985–1990. https://doi.org/10.1152/japplphysiol.01095.2004

Cheung, M.W. (2019). A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review, 29*(4), 387–396. https://doi.org/10.1007/s11065-019-09415-6

Cofnas, N. (2016). Science is not always "self-correcting": Fact–value conflation and the study of intelligence. *Foundations of Science, 21,* 477–492. https://doi.org/10.1007/s10699-015-9421-3

Colling, L.J., & Szucs, D. (2021). Statistical inference and the replication crisis. *Review of Philosophy and Psychology, 12*(1), 121–147. https://doi.org/10.1007/s13164-018-0421-4

Coyle, E.F. (2005). Very intense exercise training is extremely potent and time efficient: A reminder. *Journal of Applied Physiology, 98*(6), 1983–1984. https://doi.org/10.1152/japplphysiol.00215.2005

Davis-Stober, C.P., & Dana, J. (2014). Comparing the accuracy of experimental estimates to guessing: A new perspective on replication and the "crisis of confidence" in psychology. *Behavior Research Methods, 46*(1), 1–14. https://doi.org/10.3758/s13428-013-0342-1

Deming, D. (2016). Do extraordinary claims require extraordinary evidence? *Philosophia, 44*(4), 1319–1331. https://doi.org/10.1007/s11406-016-9779-7

Dishman, R.K. (1982). Compliance/adherence in health-related exercise. *Health Psychology, 1*(3), 237–267. https://doi.org/10.1037/0278-6133.1.3.237

Eather, N., Beauchamp, M.R., Rhodes, R.E., Diallo, T., Smith, J.J., Jung, M.E., Plotnikoff, R.C., Noetel, M., Harris, N., Graham, E., & Lubans, D.R. (2020). Development and evaluation of the High-Intensity Interval Training Self-Efficacy Questionnaire. *Journal of Sport and Exercise Psychology, 42*(2), 114–122. https://doi.org/10.1123/jsep.2019-0166

Edwards, A.M., Dutton-Challis, L., Cottrell, D., Guy, J.H., & Hettinga, F.J. (2018). Impact of active and passive social facilitation on self-paced endurance and sprint exercise: Encouragement augments performance and motivation to exercise. *BMJ Open Sport and Exercise Medicine, 4*(1), Article e000368. https://doi.org/10.1136/bmjsem-2018-000368

Ekkekakis, P., Parfitt, G., & Petruzzello, S.J. (2011). The pleasure and displeasure people feel when they exercise at different intensities: Decennial update and progress towards a tripartite rationale for exercise intensity prescription. *Sports Medicine, 41*(8), 641–671. https://doi.org/10.2165/11590680-000000000-00000

Fleming, A.R., Martinez, N., Collins, L.H., Ashley, C.D., Chiodini, M., Waddell, B.J., & Kilpatrick, M.W. (2020). Psychological responses to high-intensity interval training: A comparison of graded walking and ungraded running at equivalent metabolic loads. *Journal of Sport & Exercise Psychology, 42*(1), 70–81. https://doi.org/10.1123/jsep.2019-0200

Gibala, M. (2017). *The one-minute workout: Science shows a way to get fit that's smarter, faster, shorter*. Avery.

Gibala M.J. (2007). High-intensity interval training: A time-efficient strategy for health promotion? *Current Sports Medicine Reports, 6*(4), 211–213.

Gibala, M.J., Burgomaster, K.A., & Heigenhauser, G.J.F. (2005). Reply. *Journal of Applied Physiology, 99*(6), 2474–2475. https://doi.org/10.1152/japplphysiol.00702.2005

Gibala, M.J., & McGee, S.L. (2008). Metabolic adaptations to short-term high-intensity interval training: A little pain for a lot of gain? *Exercise and Sport Sciences Reviews, 36*(2), 58–63. https://doi.org/10.1097/JES.0b013e318168ec1f

Gillen, J.B., & Gibala, M.J. (2018). Interval training: A time-efficient exercise strategy to improve cardiometabolic health. *Applied Physiology, Nutrition, and Metabolism, 43*(10), iii–iv. https://doi.org/10.1139/apnm-2018-0453

Gillen, J.B., Martin, B.J., MacInnis, M.J., Skelly, L.E., Tarnopolsky, M.A., & Gibala, M.J. (2016). Twelve weeks of sprint interval training improves indices of cardiometabolic health similar to traditional endurance training despite a five-fold lower exercise volume and time commitment. *PLoS One, 11*(4), Article e0154075. https://doi.org/10.1371/journal.pone.0154075

Grossman, P. (2005). A comment on Burgomaster et al. and a general plea to consider behavioral influences in human physiology studies. *Journal of Applied Physiology, 99*(6), 2473–2475. https://doi.org/10.1152/japplphysiol.00702.2005

Gucciardi, D.F., Lines, R.L.J., & Ntoumanis, N. (2022). Handling effect size dependency in meta-analysis. *International Review of Sport and Exercise Psychology, 15*(1), 152–178. https://doi.org/10.1080/1750984X.2021.1946835

Guthold, R., Stevens, G.A., Riley, L.M., & Bull, F.C. (2018). Worldwide trends in insufficient physical activity from 2001 to 2016: A pooled analysis of 358 population-based surveys with 1.9 million participants. *Lancet Global Health, 6*(10), e1077–e1086. https://doi.org/10.1016/S2214-109X(18)30357-7

Halperin, I., Pyne, D.B., & Martin, D.T. (2015). Threats to internal validity in exercise science: A review of overlooked confounding variables. *International Journal of Sports Physiology and Performance, 10*(7), 823–829. https://doi.org/10.1123/ijspp.2014-0566

Harvey, L.A. (2015). Spin kills science. *Spinal Cord, 53*(6), 417. https://doi.org/10.1038/sc.2015.89

Herbert, P. (2022). Promoting exercise in older people to support healthy ageing. *Nursing Standard, 37*(4), 46–50. https://doi.org/10.7748/ns.2022.e11787

Hoffman, S.A., Warnick, J.L., Garza, E., & Spring, B. (2017). Physical activity: A synopsis and comment on "community-wide interventions for increasing physical activity." *Translational Behavioral Medicine, 7*(1), 39–42. https://doi.org/10.1007/s13142-016-0419-3

Holtermann, A., Straker, L., Lee, I.M., Stamatakis, E., & van der Beek, A.J. (2021). Workplace physical activity promotion: Why so many failures and few successes? The need for new thinking. *British Journal of Sports Medicine, 55*(12), 650–651. https://doi.org/10.1136/bjsports-2020-103067

Hotez, P.J. (2020). Combating antiscience: Are we preparing for the 2020s?. *PLoS Biology, 18*(3), Article e3000683. https://doi.org/10.1371/journal.pbio.3000683

Hume, D. (1748/1921). *An enquiry concerning human understanding and selections from a treatise of human nature*. Open Court. (Original work published 1748)

Ioannidis, J.P.A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7*(6), 645–654. https://doi.org/10.1177/1745691612464056

Jamieson K.H. (2018). Crisis or self-correction: Rethinking media narratives about the well-being of science. *Proceedings of the National Academy of Sciences of the United States of America, 115*(11), 2620–2627. https://doi.org/10.1073/pnas.1708276114

Joisten, N., Gehlert, S., & Zimmer, P. (2022). Is high-intensity interval training harmful to health? *Trends in Endocrinology and Metabolism, 33*(2), 85–86. https://doi.org/10.1016/j.tem.2021.07.003

Judd, C.M., & Gawronski, B. (2011). Editorial comment. *Journal of Personality and Social Psychology, 100*(3), 406. https://doi.org/10.1037/0022789

Jung, M.E., Bourne, J.E., Beauchamp, M.R., Robinson, E., & Little, J.P. (2015). High-intensity interval training as an efficacious alternative to moderate-intensity continuous training for adults with prediabetes. *Journal of Diabetes Research, 2015,* Article 191595. https://doi.org/10.1155/2015/191595

Jung, M.E., Bourne, J.E., & Little, J.P. (2014). Where does HIT fit? An examination of the affective response to high-intensity intervals in comparison to continuous moderate- and continuous vigorous-intensity exercise in the exercise intensity-affect continuum. *PLoS One, 9*(12), Article e114541. https://doi.org/10.1371/journal.pone.0114541

Karlsen, T., Aamot, I.L., Haykowsky, M., & Rognmo, Ø. (2017). High intensity interval training for maximizing health outcomes. *Progress in Cardiovascular Diseases, 60*(1), 67–77. https://doi.org/10.1016/j.pcad.2017.03.006

Karpovich, P. (1937). Physiological and psychological dynamogenic factors in exercise. *Arbeitsphysiologie, 9,* 620–629.

Kilpatrick, M.W., Greeley, S.J., & Collins, L.H. (2015). The impact of continuous and interval cycle exercise on affect and enjoyment. *Research Quarterly for Exercise and Sport, 86*(3), 244–251. https://doi.org/10.1080/02701367.2015.1015673

Kilpatrick, M.W., Jung, M.E., & Little, J.P. (2014). High-intensity interval training: A review of physiological and psychological responses. *ACSM's Health & Fitness Journal, 18*(5), 11–16. https://doi.org/10.1249/FIT.0000000000000067

Kujala, U.M. (2018). Is physical activity a cause of longevity? It is not as straightforward as some would believe: A critical analysis. *British Journal of Sports Medicine, 52*(14), 914–918. https://doi.org/10.1136/bjsports-2017-098639

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D., & Evers, E.R. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science, 9*(3), 278–292. https://doi.org/10.1177/1745691614528520

Laplace, P.S.M. (1814/1951). *A philosophical essay on probabilities* (F.W. Truscott & F.L. Emory, Trans.). Dover. (Original work published 1814)

Li, G., Taljaard, M., Van den Heuvel, E.R., Levine, M.A., Cook, D.J., Wells, G.A., Devereaux, P.J., & Thabane, L. (2017). An introduction to multiplicity issues in clinical trials: The what, why, when and how. *International Journal of Epidemiology, 46*(2), 746–755. https://doi.org/10.1093/ije/dyw320

Love, R., Adams, J., & van Sluijs, E. (2019). Are school-based physical activity interventions effective and equitable? A meta-analysis of cluster randomized controlled trials with accelerometer-assessed activity. *Obesity Reviews, 20*(6), 859–870. https://doi.org/10.1111/obr.12823

Mahoney, M.J. (1979). Psychology of the scientist: An evaluative review. *Social Studies of Science, 9*(3), 349–375. https://doi.org/10.1177/030631277900900304

Martinez, N., Kilpatrick, M.W., Salomon, K., Jung, M.E., & Little, J.P. (2015). Affective and enjoyment responses to high-intensity interval training in overweight-to-obese and insufficiently active adults. *Journal of Sport & Exercise Psychology, 37*(2), 138–149. https://doi.org/10.1123/jsep.2014-0212

Mastrofini, G.F., Collins, R.P., Rosado, A.P., Tauran, R.C., Fleming, A.R., & Kilpatrick, M.W. (2022). The impact of variation and autonomy on psychological responses to high intensity interval training exercise. *Psychology of Sport and Exercise, 60,* Article 102142. https://doi.org/10.1016/j.psychsport.2022.102142

Mattioni Maturana, F., Martus, P., Zipfel, S., & Nieß, A.M. (2021). Effectiveness of HIIE versus MICT in improving cardiometabolic risk factors in health and disease: A meta-analysis. *Medicine & Science in Sports & Exercise, 53*(3), 559–573. https://doi.org/10.1249/MSS.0000000000002506

Midgley, A.W., Marchant, D.C., & Levy, A.R. (2018). A call to action towards an evidence-based approach to using verbal encouragement during maximal exercise testing. *Clinical Physiology and Functional Imaging, 38*(4), 547–553. https://doi.org/10.1111/cpf.12454

Niven, A., Laird, Y., Saunders, D.H., & Phillips, S.M. (2021). A systematic review and meta-analysis of affective responses to acute high intensity interval exercise compared with continuous moderate- and high-intensity exercise. *Health Psychology Review, 15*(4), 540–573. https://doi.org/10.1080/17437199.2020.1728564

Oliveira, B., Santos, T.M., Kilpatrick, M., Pires, F.O., & Deslandes, A.C. (2018). Affective and enjoyment responses in high intensity interval training and continuous training: A systematic review and meta-analysis. *PLoS One, 13*(6), Article e0197124. https://doi.org/10.1371/journal.pone.0197124

Orne, M.T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*(11), 776–783. https://doi.org/10.1037/h0043424

Parkhurst, D.F. (2001). Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation. *BioScience, 51*(12), 1051–1057. https://doi.org/10.1641/0006-3568(2001)051[1051:SSTEAR]2.0.CO;2

Pratt, M., Ramirez Varela, A., Salvo, D., Kohl, H.W., III, & Ding, D. (2020). Attacking the pandemic of physical inactivity: What is holding us back? *British Journal of Sports Medicine, 54*(13), 760–762. https://doi.org/10.1136/bjsports-2019-101392

Rhodes, R.E., McEwan, D., & Rebar, A.L. (2019). Theories of physical activity behaviour change: A history and synthesis of approaches. *Psychology of Sport and Exercise, 42,* 100–109. https://doi.org/10.1016/j.psychsport.2018.11.010

Rosenthal, R. (1963). On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. *American Scientist, 51*(2), 268–283.

Rosenthal, R. (1964). Experimenter outcome-orientation and the results of the psychological experiment. *Psychological Bulletin, 61*(6), 405–412. https://doi.org/10.1037/h0045850

Rosenthal, R., Kohn, P., Greenfield, P.M., & Carota, N. (1965). Psychology of the scientist: XIV. Experimenters' hypothesis-confirmation and mood as determinants of experimental results. *Perceptual and Motor Skills, 20*(Suppl. 3), 1237–1252. https://doi.org/10.2466/pms.1965.20.3c.1237

Rütten, A., Abu-Omar, K., Gelius, P., & Schow, D. (2013). Physical inactivity as a policy problem: Applying a concept from policy analysis to a public health issue. *Health Research Policy and Systems, 11*(1), Article 9. https://doi.org/10.1186/1478-4505-11-9

Sabag, A., Little, J.P., & Johnson, N.A. (2022). Low-volume high-intensity interval training for cardiometabolic health. *Journal of Physiology, 600*(5), 1013–1026. https://doi.org/10.1113/JP281210

Sagan, C. (1979). *Broca's brain: Reflections on the romance of science*. Ballantine Books.

Samsa, G., Edelman, D., Rothman, M.L., Williams, G.R., Lipscomb, J., & Matchar, D. (1999). Determining clinically important differences in health status measures: A general approach with illustration to the Health Utilities Index Mark II. *PharmacoEconomics, 15*(2), 141–155. https://doi.org/10.2165/00019053-199915020-00003

Schwartz, L.M., Woloshin, S., Andrews, A., & Stukel, T.A. (2012). Influence of medical journal press releases on the quality of associated newspaper coverage: Retrospective cohort study. *British Medical Journal, 344,* Article d8164. https://doi.org/10.1136/bmj.d8164

Senn, S.J. (2009). Overstating the evidence: Double counting in meta-analysis and related problems. *BMC Medical Research Methodology, 9,* Article 10. https://doi.org/10.1186/1471-2288-9-10

Serrablo-Torrejon, I., Lopez-Valenciano, A., Ayuso, M., Horton, E., Mayo, X., Medina-Gomez, G., Liguori, G., & Jimenez, A. (2020). High intensity interval training exercise-induced physiological changes and their potential influence on metabolic syndrome clinical biomarkers: A meta-analysis. *BMC Endocrine Disorders, 20*(1), Article 167. https://doi.org/10.1186/s12902-020-00640-2

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances, 7*(21), Article eabd1705. https://doi.org/10.1126/sciadv.abd1705

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

Slomski A. (2020). Exercise intensity unrelated to older adults' mortality risk. *Journal of the American Medical Association, 324*(24), 2476. https://doi.org/10.1001/jama.2020.23342

Smith, R.J. (2020). *P* > .05: The incorrect interpretation of "not significant" results is a significant problem. *Physical Anthropology, 172*(4), 521–527. https://doi.org/10.1002/ajpa.24092

Stensvold, D., Bucher Sandbakk, S., Viken, H., Zisko, N., Reitlo, L.S., Nauman, J., Gaustad, S.E., Hassel, E., Moufack, M., Brønstad, E., Aspvik, N.P., Malmo, V., Steinshamn, S.L., Støylen, A., Anderssen, S.A., Helbostad, J.L., Rognmo, Ø., & Wisløff, U. (2017). Cardiorespiratory reference data in older adults: The Generation 100 study. *Medicine & Science in Sports & Exercise, 49*(11), 2206–2215. https://doi.org/10.1249/MSS.0000000000001343

Stensvold, D., Viken, H., Rognmo, Ø., Skogvoll, E., Steinshamn, S., Vatten, L.J., Coombes, J.S., Anderssen, S.A., Magnussen, J., Ingebrigtsen, J.E., Fiatarone Singh, M.A., Langhammer, A., Støylen, A., Helbostad, J.L., & Wisløff, U. (2015). A randomised controlled study of the long-term effects of exercise training on mortality in elderly people: Study protocol for the Generation 100 study. *BMJ Open, 5*(2), Article e007519. https://doi.org/10.1136/bmjopen-2014-007519

Stensvold, D., Viken, H., Steinshamn, S.L., Dalen, H., Støylen, A., Loennechen, J.P., Reitlo, L.S., Zisko, N., Bækkerud, F.H., Tari, A.R., Sandbakk, S.B., Carlsen, T., Ingebrigtsen, J.E., Lydersen, S., Mattsson, E., Anderssen, S.A., Fiatarone Singh, M.A., Coombes, J.S., Skogvoll, E., Vatten, L.J., . . . Wisløff, U. (2020). Effect of exercise training for five years on all cause mortality in older adults—The Generation 100 study: Randomised controlled trial. *British Medical Journal, 371,* Article m3485. https://doi.org/10.1136/bmj.m3485

Stork, M.J., Banfield, L.E., Gibala, M.J., & Martin Ginis, K.A. (2017). A scoping review of the psychological responses to interval exercise: Is interval exercise a viable alternative to traditional exercise? *Health Psychology Review, 11*(4), 324–344. https://doi.org/10.1080/17437199.2017.1326011

Terry, P.C., Karageorghis, C.I., Curran, M.L., Martin, O.V., & Parsons-Smith, R.L. (2020). Effects of music in exercise and sport: A meta-analytic review. *Psychological Bulletin, 146*(2), 91–117. http://dx.doi.org/10.1037/bul0000216

The Lancet. (2021). A sporting chance: Physical activity as part of everyday life. *The Lancet, 398*(10298), 365. https://doi.org/10.1016/S0140-6736(21)01652-4

Thompson, W.R. (2013). Worldwide survey of fitness trends for 2014. *ACSM's Health & Fitness Journal, 17*(6), 10–20.

Tiller, N.B., Sullivan, J.P., & Ekkekakis, P. (2022). Baseless claims and pseudoscience in health and wellness: A call to action for the sports, exercise, and nutrition-science community. *Sports Medicine*. Advance online publication. https://doi.org/10.1007/s40279-022-01702-2

Tudor-Locke, C., Craig, C.L., Brown, W.J., Clemes, S.A., De Cocker, K., Giles-Corti, B., Hatano, Y., Inoue, S., Matsudo, S.M., Mutrie, N., Oppert, J.M., Rowe, D.A., Schmidt, M.D., Schofield, G.M., Spence, J.C., Teixeira, P.J., Tully, M.A., & Blair, S.N. (2011). How many steps/day are enough? For adults. *International Journal of Behavioral Nutrition and Physical Activity, 8,* Article 79. https://doi.org/10.1186/1479-5868-8-79

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H.L.J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*(3), 426–432. https://doi.org/10.1037/a0022790

Wernhart, S., Hedderich, J., Wunderlich, S., Schauerte, K., Weihe, E., Dellweg, D., & Siemon, K. (2021). The feasibility of high-intensity interval training in patients with intensive care unit-acquired weakness syndrome following long-term invasive ventilation. *Sports Medicine Open, 7*(1), Article 11. https://doi.org/10.1186/s40798-021-00299-6

Wilson, B.M., Harris, C.R., & Wixted, J.T. (2020). Science is not a signal detection problem. *Proceedings of the National Academy of Sciences of the United States of America, 117*(11), 5559–5567. https://doi.org/10.1073/pnas.1914237117

Woloshin, S., Schwartz, L.M., Casella, S.L., Kennedy, A.T., & Larson, R.J. (2009). Press releases by academic medical centers: Not so academic? *Annals of Internal Medicine, 150*(9), 613–618. https://doi.org/10.7326/0003-4819-150-9-200905050-00007

Yavchitz, A., Boutron, I., Bafeta, A., Marroun, I., Charles, P., Mantz, J., & Ravaud, P. (2012). Misrepresentation of randomized controlled trials in press releases and news coverage: A cohort study. *PLoS Medicine, 9*(9), Article e1001308. https://doi.org/10.1371/journal.pmed.1001308

Zajonc, R.B. (1965). Social facilitation. *Science, 149*(3681), 269–274. https://doi.org/10.1126/science.149.3681.269