INTERNATIONAL JOURNAL OF -EXERCISE SCIENCE- Inclusive

Editorial

Editorial: A Step-by-step Statistical Decision Framework for a Gender-inclusive Approach in Sport and Exercise Science Research

James W. Navalta‡[1], Dustin W. Davis†[1], Jafra D. Thomas‡[2], Whitley J. Stone‡[3]

[1]Department of Kinesiology and Nutrition Sciences, University of Nevada, Las Vegas, Las Vegas, NV, USA; [2]Department of Kinesiology and Public Health, California Polytechnic State University, San Luis Obispo, CA, USA; [3]School of Kinesiology, Recreation, and Sport, Western Kentucky University, Bowling Green, KY, USA

†Denotes early-career investigator, ‡Denotes established investigator

Abstract

International Journal of Exercise Science 18(1): 1010-1029, 2025. https://doi.org/10.70252/ITKQ9186 The conflation of sex and gender in sport and exercise science research has led to gaps in representation and understanding of gender-inclusive outcomes. This invited editorial presents a framework to guide researchers in conducting statistical analyses that account for multiple gender identities beyond the traditional binary classification. The invited editorial guides deliberation on whether to test for sex or gender differences in sport or exercise science research. It prompts investigators to answer the question, "Is there a valid, literature-supported rationale for investigating sex or gender differences?" If "yes", we propose approaches that may help investigators plan a study for two sex or gender groups, or in situations with three or more sex or gender groups. The editorial provides a valid, step-by-step statistical decision framework to ensure a robust, and ethical, research design while addressing the limitations of current sex- and gender-based classifications in sport and exercise science. By adopting gender-inclusive research practices, the field can better support equitable exercise prescriptions, rehabilitation strategies, and training periodization for diverse populations.

Keywords: Kinesiology, research design and statistical approach, diversity and inclusion, representation and underrepresentation, publication bias

Introduction

Historically, the concepts of sex and gender have been conflated and used interchangeably.[1] Sex refers to sets of biological attributes in humans and animals associated with physical and physiological features including chromosomes, gene expression, hormone function and reproductive or sexual anatomy.[2] Gender refers to the socially constructed (or expected) roles, behaviors and identities that are most often represented by feminine, masculine and gender-diverse people.[2,3] Within the gender-diverse identity, individuals can further identify as transgender or non-binary. The term transgender is "an umbrella term used to describe the full

range of people whose gender identity and/or gender role do not conform to what is typically associated with their sex assigned at birth."[4] For example, a transgender person might identify and live in ways culturally associated with femininity while having been assigned male at birth. More specifically, transgender is the expression of a gender other than the one traditionally associated with a particular sex assigned at birth (e.g., a transgender person might have primarily feminine roles, behaviors, or identities while having been assigned the male sex at birth).[5] A non-binary gender identity falls outside the traditional binary categories of female and male. The terms non-binary and genderqueer refer to people who "have a gender which is neither male nor female and may identify as both male and female at one time, as different genders at different times, as no gender at all, or dispute the very idea of only two genders."[3] A non-binary gender (sometimes understood as androgynous) means that a person's roles, behaviors, or identities do not exclusively fall within traditional female or male categories, and they may intentionally minimize any particular gender label.[6]

The current estimation is that 1.6% of the total U.S. population identifies as transgender or non-binary.[6] This reported value is likely an undercount, given the stigma that sex and gender minorities have historically faced.[7] Reported percentages in transgender and non-binary populations could increase in future surveys as more inclusive methods of gathering and disseminating sex and gender data become accepted and utilized. To highlight this development, it has been reported that 5% of people under 30 years of age identify as a gender not traditionally associated with their sex assigned at birth.[6] There is a movement for biomedical research to become personalized (that is, individualized healthcare),[8] which has long been suggested for sport and exercise science.[9] To facilitate continued personalization of sport and exercise prescriptions in lifestyle medicine, rehabilitation, and training periodization, there is a need to reexamine and update the way sport and exercise scientists approach inclusivity when collecting sex and gender data.[10-12] Sex and gender underrepresentation is indicative of barriers previously identified in kinesiology and allied health education programs, where institutional norms and curricular design have historically marginalized gender-diverse populations.[13]

Few sport and exercise science investigations are gender-inclusive. A self-study by the *International Journal of Exercise Science* found that, of 151,043 participants evaluated across 851 published original research articles, only one participant identified as transgender, three identified as other, and one declined to identify their gender.[14] Because 1.6% to 5% of the U.S. population identifies as transgender or non-binary,[6] between 2,417 and 7,552 participants could have been classified as the incorrect gender (that is, misgendered). The theoretical consequences of such misgendering have been recently reported.[11] Using NHANES data, it was found that statistical and effect size results for anthropometric measurements differ when individuals are theoretically misgendered and are compared to a non-inclusive data set.[11] This highlights the importance of using inclusive methods for obtaining sex and gender data in research studies. These data are crucial for drawing valid conclusions from statistical comparisons of sex and gender groups, which can inform personalized exercise prescriptions.

The *All of Us* Research Program and data set has provided such an example of being inclusive.[15] The program recruits people across demographic categories including those who are

underrepresented in biomedical research, accounting for race, ethnic group, age, sex, gender, sexual orientation, disability status, access to care, income, educational attainment, and geographic location. In a similar vein, a leading organization in sport and exercise science research, the American College of Sports Medicine (ACSM), has noted a need to be more inclusive in sport and exercise science.[16,17] The most recent edition of the ACSM's *Guidelines for Exercise Testing and Prescription*, however, contains one paragraph of guidance to researchers and practitioners on testing individuals who identify as a gender other than the one traditionally associated with their sex assigned at birth.[18] It was concluded that, due to a lack of evidence, the ACSM could not provide recommendations for gender-inclusive exercise testing and prescriptions, and that future research is required to establish gender-inclusive normative data.[18]

It seems apparent that a barrier prevents the generation of normative data required to overcome the currently acknowledged lack of evidence by the ACSM and other scientific organizations.[14,18] Thus, our working hypothesis is that many sport and exercise scientists and practitioners may not have the knowledge or confidence to design gender inclusive investigations.[10] Such investigations are necessary to collect sex and gender data to support gender-inclusive recommendations for individualized exercise testing, prescription and other applications from sport and exercise science subdisciplines. The purpose of this editorial was to demonstrate a practical framework for conducting gender-inclusive research. By describing how to analyze physical activity metrics across diverse gender identities, we provide a replicable approach to improve inclusivity in sport and exercise science research.

Detailed Framework for an Approach to Conducting Statistical Testing for Sex or Gender Differences

Presented in Figure 1 is the full framework for addressing considerations and checkpoints when conducting statistical testing for differences among people of different sexes or genders. The flow of this framework is presented as an example of an approach researchers may take, with appropriate methodological considerations described. For enhanced visibility, Figure 1 was divided into parts shown in Figures 1a, 1b and 1c. A streamlined version is presented in Figure 2.

- **Figure 1a** guides deliberation on whether to test for sex or gender differences in a specific sport or exercise science research study. It prompts investigators to answer the question, "Is there a valid, literature-supported rationale for investigating sex or gender differences?"

- **Figure 1b** presumes the investigators evidenced a yes-response to the question in Figure 1a. It helps the investigators plan a study for two sex or gender groups.

- **Figure 1c**, like Figure 1b, presumes investigators evidenced a yes-response to the question in Figure 1a. Figure 1c helps the investigators plan a study for at least three sex or gender groups.

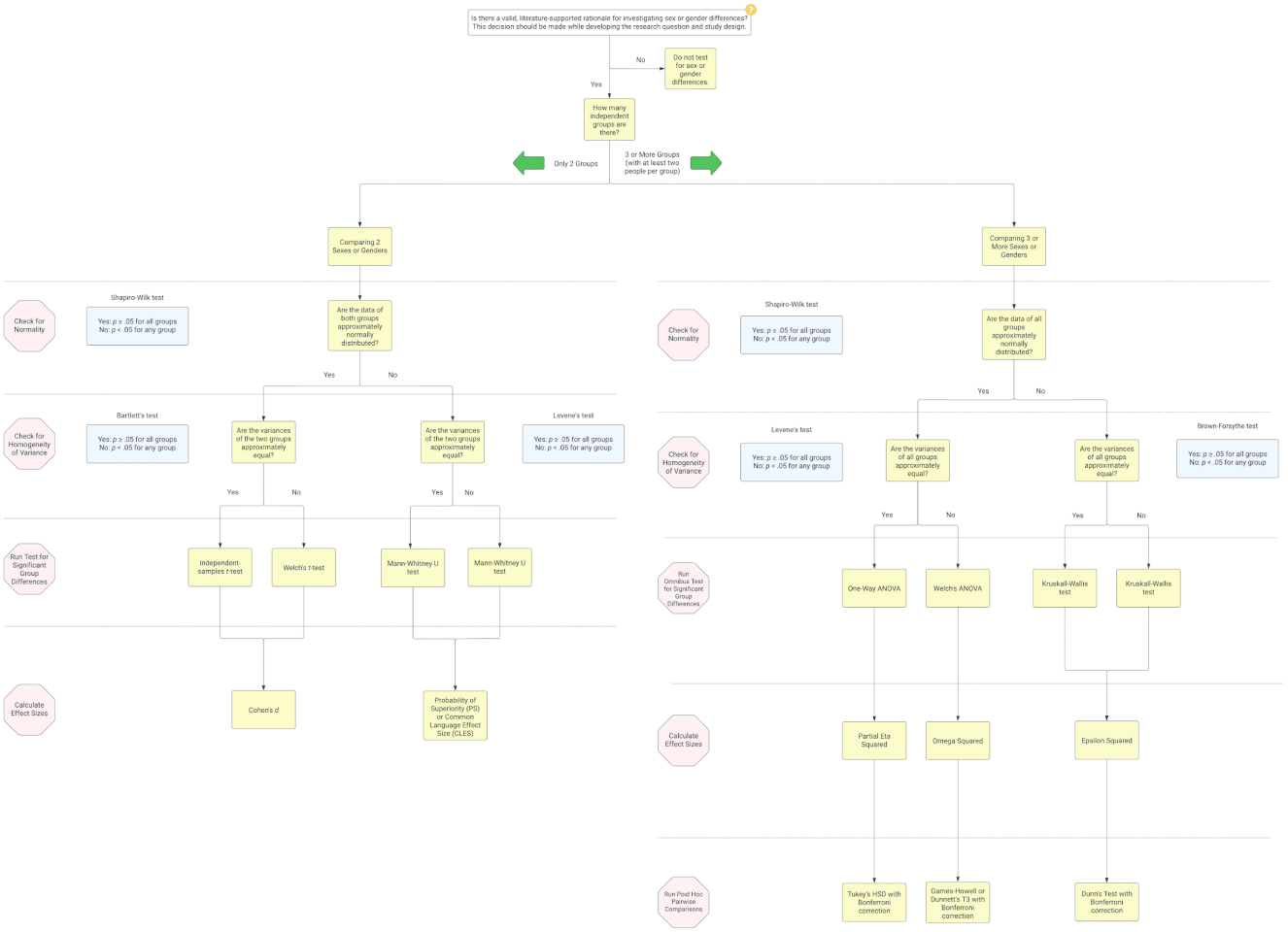

Figure 1. Framework for an approach to conducting statistical testing for sex or gender differences in sport and exercise science research studies.

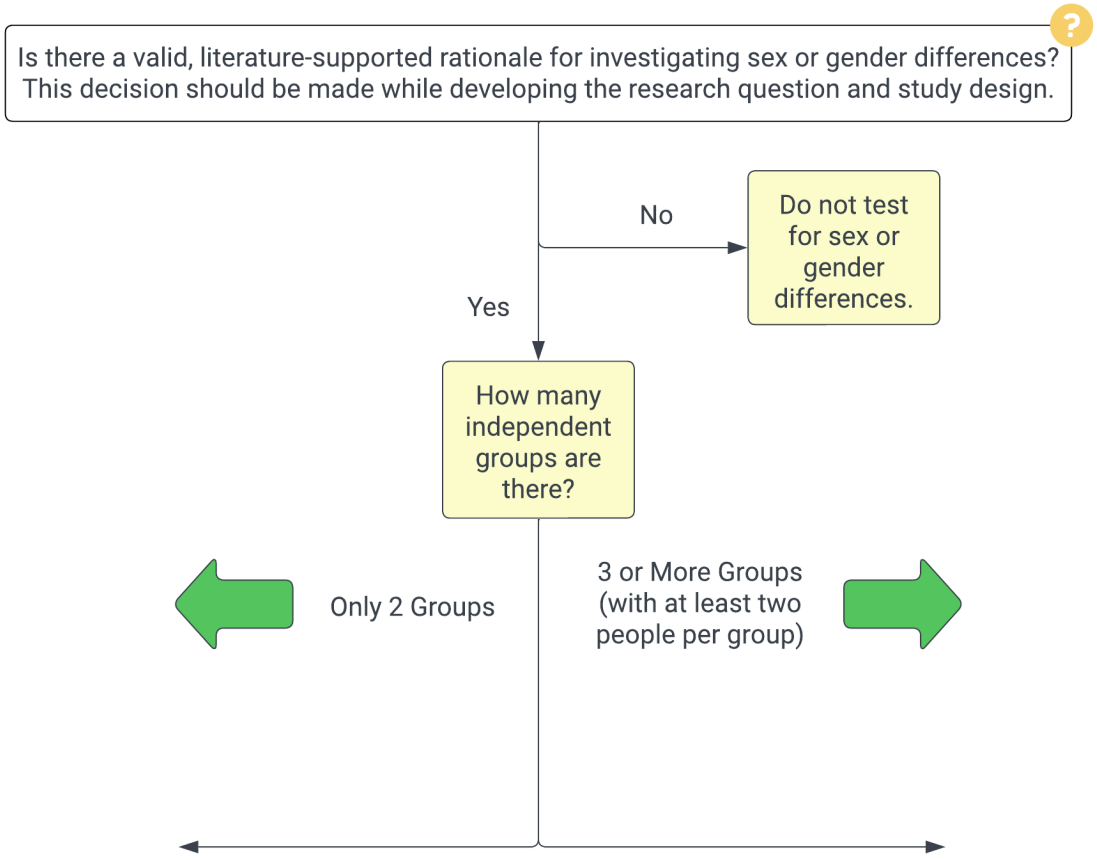

Figure 1a. Illustrates using the framework to decide whether to test for sex or gender differences in sport and exercise science research studies.

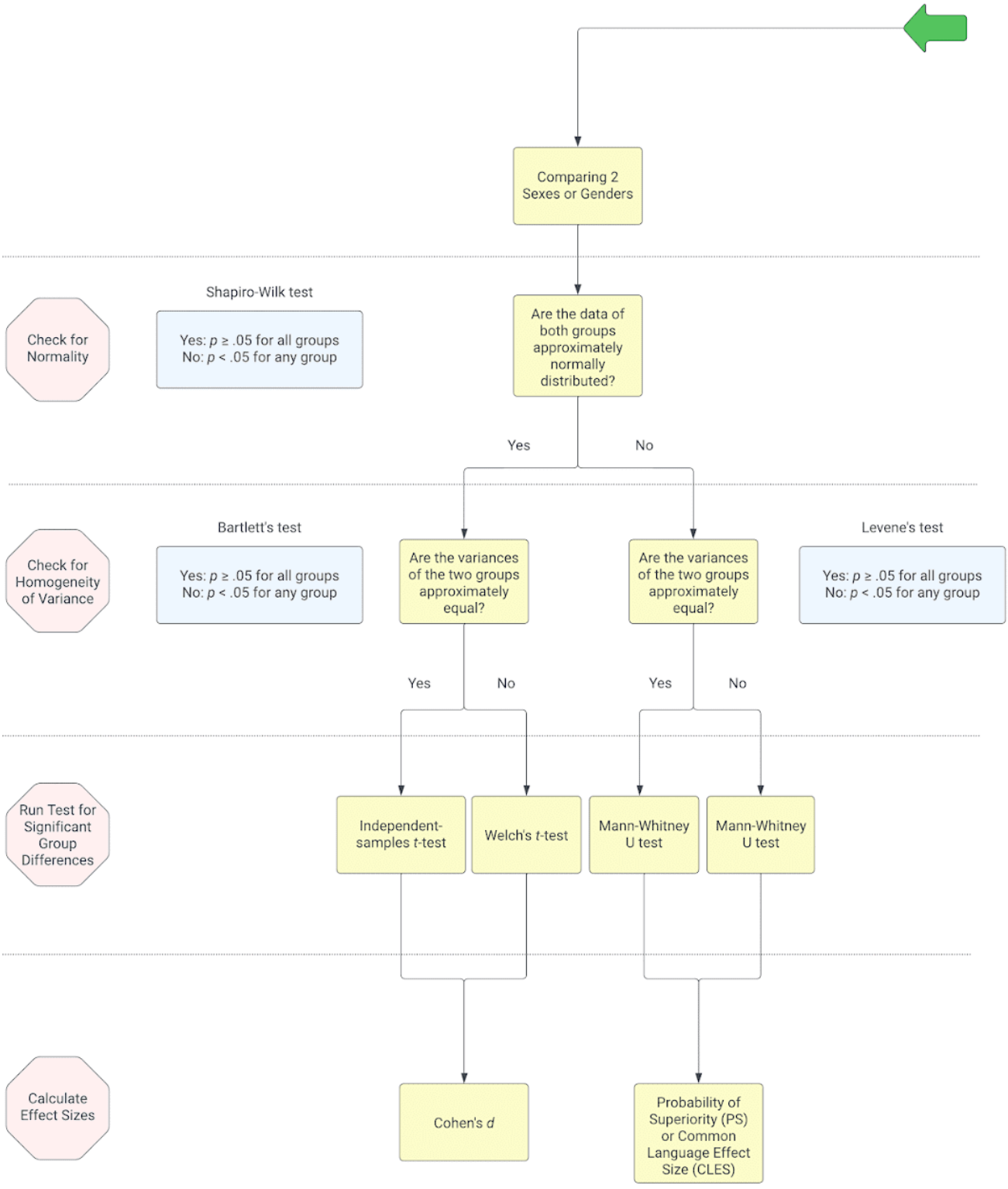

Figure 1b. Illustrates using the framework to test for sex or gender differences in sport and exercise science research studies with two groups.

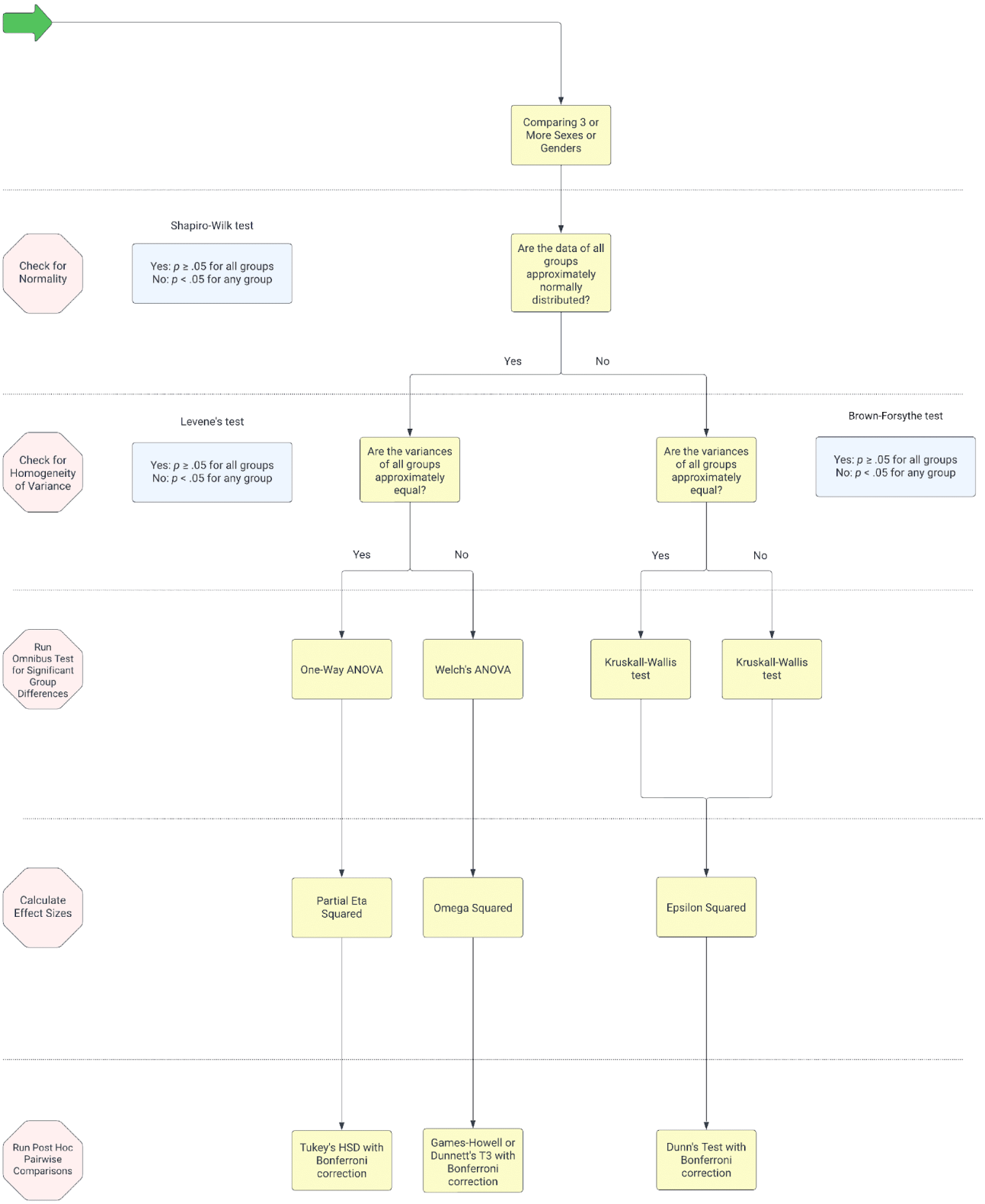

Figure 1c. Illustrates using the framework to test for sex or gender differences in sport and exercise science research studies with at least three groups.

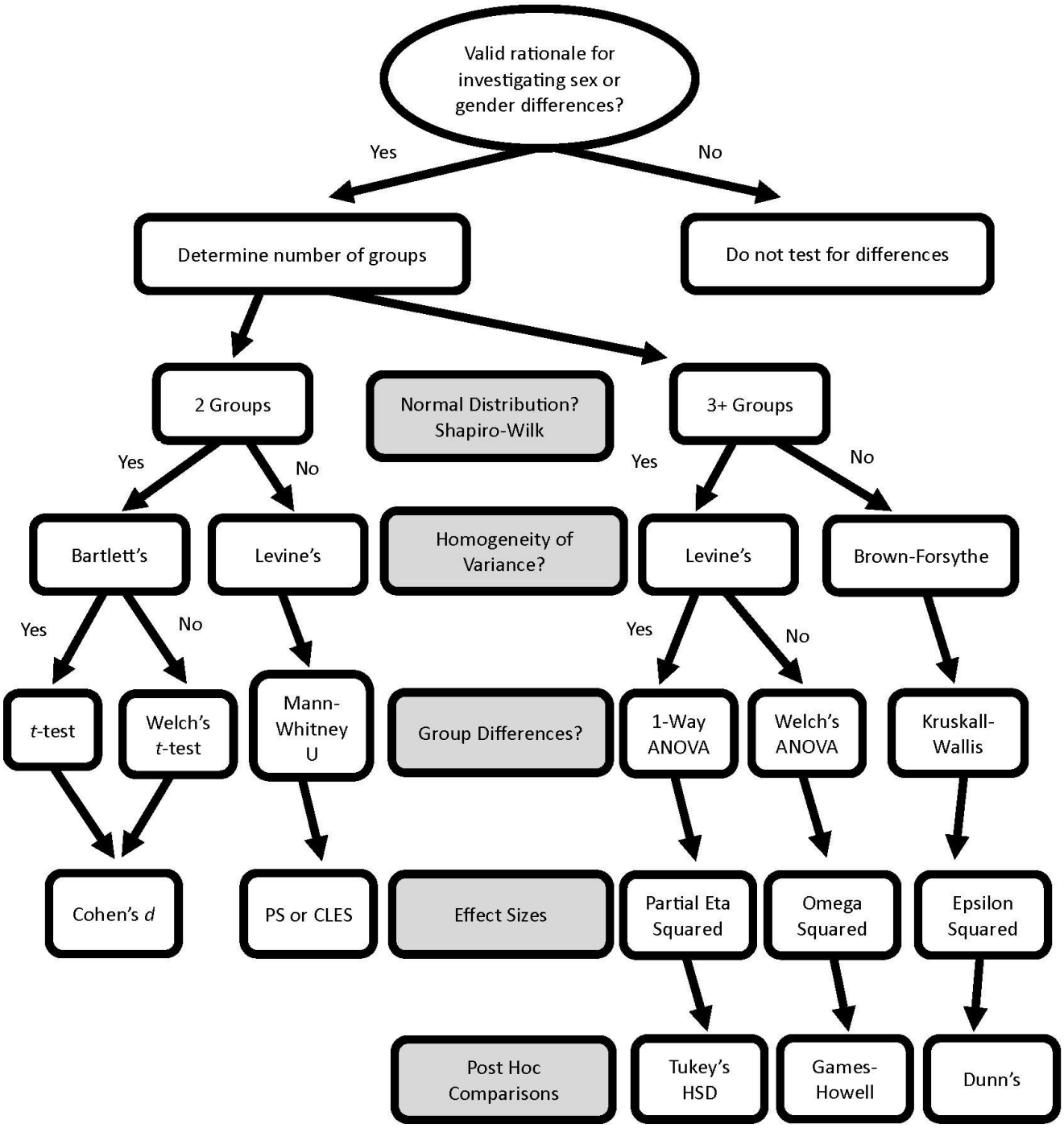

Figure 2. Condensed framework for an approach to conducting statistical testing for sex or gender differences in sport and exercise science research studies.

Below, we will discuss the need for 1) a valid, literature-supported rationale for testing for sex or gender differences, 2) identifying the number of independent groups evaluated, and 3) conducting an appropriate statistical analysis.

A Priori Rationale to Conduct Statistical Testing for Sex or Gender Differences

An important part of any study design is to have an established and justified purpose for conducting tests for sex or gender differences (that is, an a priori rationale). It is well established that many sport and exercise science investigations are underpowered.[19,20] Conducting inference testing for sex or gender differences when the study is not powered for such an analysis exacerbates the likelihood of Type II errors, leading to false negatives or underestimating true differences, which can result in misleading conclusions and inappropriate generalizations.[21,22]

It is important to note that a literature-supported rationale should also be presented for excluding certain populations from the investigation. An important part of the study design is to have an established purpose for conducting tests for sex or gender differences. In the mid-2010s, the National Institutes of Health, recognizing that sex and gender affected health and disease processes differently across individuals, began to require that researchers account for sex as a biological variable when developing research questions and study designs.[23] Assumptions about sex and gender may influence hypotheses, how data is collected, and how findings are interpreted.[24] Flawed assumptions can have a trickle-down effect, limiting the scope of research and overlooking potentially important findings.[24] Acknowledging a bias toward testing cisgender male individuals in research, the Sex and Gender Equity in Research (SAGER) guidelines were developed, detailing a comprehensive approach for the reporting of sex and gender information in study design, data analyses, results and interpretation.[25] While our purpose is not to summarize these guidelines, we reiterate that scientists should appropriately account for sex or gender differences in their study design from the outset of the investigation. A bias towards testing cisgender male individuals has been repeatedly reported in sport and exercise science research.[14,26,27] Even if researchers do not have cause to test for differences among sex or gender groups, we propose adhering to the SAGER guideline of reporting disaggregated data by sex or gender to align with best practices.[25]

Reviews demonstrate a common trend in the exercise and sport sciences literature, where tests for statistical significance are conducted at times without an a priori hypothesis.[28] The concern with testing if outcomes differ significantly without a priori justification is this method of investigation is a form of hypothesizing after the results are known, called "*p*-HARking."[29] The practice of *p*-HARking refers to strategies investigators employ to generate statistically significant results,[29] such as not disclosing when significant results are from secondary study aims rather than the original primary aims, conducting statistical tests on numerous outcomes without a clear rationale or theoretical argument, or testing for significant associations or differences not within their original plan.[29] Any *p*-HARKing is problematic because it raises the risk of publishing false-positive results or results with a lower likelihood of being replicated than presumed by reported *p*-values.[30]

To mitigate implicit bias or insufficiently powered studies, authors should provide an a priori rationale, supported by validated research literature before they conduct any test for sex or gender differences.[12,25]

This need for an a priori rationale supported by credible research literature also applies to any subsequent analysis of the same data that is exploratory in nature, including those based on unexpected results or observations.[12] Several journals in the kinesiology discipline allow for authors to submit brief reports presenting findings derived from exploratory study designs or novel lines of research, including the present journal (the *International Journal of Exercise Science*)[31] and the *International Journal of Kinesiology in Higher Education*.[32]

Identifying the Number of Independent Groups to be Evaluated

The second consideration is appropriate study design, specifically that researchers should identify the number of independent groups to be evaluated prior to beginning the investigation. This allows for a sufficient target number of participants to be recruited to test for differences between or among groups, and to find a difference if one is present (avoiding a Type II error).[33] There is a strong case for performing an a priori power analysis before conducting a research study, and many leading journals require it.[34,35] In short, a power analysis is an effort to determine how many participants a study needs in order to detect a true difference or association of a certain size. Effect size is the size of the difference or association (e.g., small, moderate, large). However, Albers and Lakens have detailed how using pilot data to estimate the effect size future studies may observe (and therefore the necessary sample size future studies need to detect such effects) may often lead to inaccurate and underpowered main studies.[36] Even so, exercise and sport science researchers may be reliant upon pilot or small sample studies due to the cost of one-time use supplies, the amount of resources needed to test a large number of participants, time constraints, or the invasiveness of some study procedures. While pilot studies have their advantages, the draw backs of conducting large scale studies could motivate exercise and sport science researchers to rely upon pilot tests or small sample studies to investigate their research questions or hypotheses, rather than implement a full-scale follow-up study.[37] This further constrains the ability to investigate sex or gender differences in exercise and sport science research. Researchers should be aware of these added challenges to investigating group differences, and they should proactively seek to mitigate them using principles for sound, incremental research.[38,39]

We also acknowledge that power analyses may be "gamed" after the fact (post hoc) to align with the number participants that were actually recruited and tested. Althouse writes, "The important thing to understand is that using the observed effect size to compute observed power means that every nonsignificant result will appear to have low observed power [almost always because of insufficient sample size]" (p. A4); but inadequate sample size is not the only explanation for null results.[40] As mentioned previously, effect size metrics are used to determine practical significance, by gauging the magnitude (i.e., size) of a difference or association (e.g., small, moderate, or larges); effect size cut-points are used to make practical decisions using research results.[41] However, it should be noted cut-points for interpreting effect size are not absolute and may vary by discipline; investigators should use effect size interpretations that are discipline-specific and in conjunction with theoretical and applied significance.[42]

Conducting an Appropriate Statistical Analysis

The third consideration is choosing and applying the appropriate statistical analysis based on the research question, study design, and data type. We aim to facilitate this process via our proposed framework, detailed in Figures 1, 1a, 1b, and 1c. We suggest that researchers use the framework as a guide, whether they are evaluating sex or gender differences or not.

Step 1: Determining Normality

Regardless of the number of groups, the first step is to evaluate data for normality. Readers can dive deeper into normality testing in movement sciences with Yagin et al's publication.[43] Many parametric statistics (e.g., *t*-test, ANOVA) assume that sample data come from a normal distribution in order to make inferences about population parameters based on sample statistics. When normality is not attained from a sampled population, researchers may consider non-parametric tests as they do not rely on the normality assumption, depending on sample size and outcomes. Running parametric tests on data that are not normally distributed may increase the risks of a Type I error (i.e., incorrectly rejecting the null hypothesis—false positive) and of a Type II error (i.e., incorrectly failing to reject the null hypothesis—false negative). In both cases (Figure 1b and 1c), testing for normality is commonly accomplished through the Shapiro-Wilk test, where data are considered normally distributed if the *p*-value is ≥ 0.05 and not normal if the *p*-value is < 0.05. It should be noted that the appropriateness of the Shapiro-Wilk test and the validity of its result depend on sample size. For samples with fewer than 50 observations, the Shapiro-Wilk test can accurately detect deviations from normality. For larger sample sizes, the Shapiro-Wilk test can become overly sensitive and detect minor statistical deviations from normality that may not be practically significant.[44] In these cases, even if the Shapiro-Wilk test indicates a departure from normality, it may not be of practical concern. It might be more appropriate to rely on visual inspection of histograms or normal probability plots and to consider the robustness of the subsequent statistical test used.

Step 2: Determining Homogeneity of Variance

The second step is to evaluate homogeneity of variance. This assumption for parametric testing refers to the concept that the variability (or variance) between or among groups or samples is similar. Referring to the framework, the specific test depends on the decision made for normality of data. If the analysis is conducted on two sexes or genders, and if the data are considered normal, Bartlett's test can be used to check for homogeneity of variance because it gives a more reliable assessment of the data (Figure 1b).[45] If the analysis is conducted on two sexes or genders and the data are not normally distributed, Levene's test can be used because it is more robust (Figure 1b).[46] Because of its robustness, we suggest that Levene's test also be used when evaluating three or more sexes or genders and the data are considered normally distributed (Figure 1c). Finally, when evaluating three or more sexes or genders and the data are not normally distributed, the Brown-Forsythe test can be used because it is less affected by violations of the assumptions of normality and equal group sizes (Figure 1c).[47]

Step 3a: Inferential Statistics with Two Sexes or Genders

After determining normality and homogeneity of variances, the ensuing step is to run the appropriate inferential statistical test for significant group differences. If the analysis is conducted on two sexes or genders (Figure 1b), there are four possibilities:

1. Normality and homogeneity confirmed: Evaluate group differences using an independent sample *t*-test.
2. Normality confirmed but not homogeneity: Welch's *t*-test is suggested. There is greater confidence in the validity of Welch's *t*-test than in the independent sample *t*-test.[55] The increased confidence comes from a lower risk of a Type I or a Type II error.
3. Homogeneity confirmed but not normality: The Mann-Whitney U is suggested regardless of whether the variance between the groups on the dependent variable is considered homogeneous.
4. Neither normality nor homogeneity confirmed: Mann-Whitney U.

After testing for group differences via an omnibus test, best practices suggest that researchers determine effect sizes associated with each test described above.[35] For ease of use, we will list the test and then the suggested effect size calculation.

Independent t-test, Welch's t-test.

An effect size for two independent groups is Cohen's *d*. Cohen's *d* shows the difference between the means of the two groups relative to the standard deviation. Cohen's *d* is usually interpreted as: negligible effect ≤ 0.2, small effect = 0.2 (meaning that the means of the two groups are separated by 0.2 standard deviations), medium effect = 0.5, large effect = 0.8.

Mann-Whitney U.

The Mann-Whitney U test is a non-parametric test that compares the centers (i.e., the midpoint) of two independent groups. The test does not require the data to be normally distributed or for the variances to be equal, but it does assume the majority of data points will cluster around a group's midpoint.[48] It ranks all the data points from both groups together, then calculates a U-statistic based on the ranks. Because the Mann-Whitney U test is based on ranks rather than raw data values, it is less affected by outliers and non-normality in the data.[48]

The probability of difference (PD) represents the likelihood that a randomly chosen person from one group will have a higher score or outcome than a randomly chosen person from the other group (also known as the probability of superiority,[49] however we present alternative phrasing similar to how Vaske advocates for a more context-aware approach to interpreting effect sizes).[41] Interpreting the PD involves understanding the direction and magnitude of the effect between the two groups. A probability of difference close to 0.5 suggests that there is little difference between the groups, while a value close to 1 indicates a strong likelihood that one group is different than the other.

The Common Language Effect Size (CLES) is another measure used to quantify the practical significance of the difference between two groups in a study. Like the PD, it represents the probability that a randomly selected person from one group will have a higher score than a randomly selected person from the other group. The CLES is a transformation of PD and is expressed as a percentage, ranging from 0% to 100%. A CLES of 50% indicates that there is little difference between the groups, while a CLES of 100% indicates a strong likelihood that one group is different than the other.

Step 3b: Inferential Statistics with Three or More Sex or Gender Groups

Switching over to when three or more sex or gender groups are evaluated (Figure 1c), the same four possibilities exist when testing for significant group differences:

1. Normality and homogeneity confirmed: Evaluate group differences using a one-way analysis of variance (ANOVA). This test is generally considered to be robust to violations of the normality assumption, especially when group sizes are equal or approximately equal. However, when group sizes are very unequal, or when the data are heavily skewed or have extreme outliers, the robustness of the ANOVA test to the normality assumption may be compromised. This will likely often be the case with gender-inclusive studies, such as is provided in the accompanying Brief Report.[50]
2. Normality confirmed but not homogeneity: Welch's ANOVA is suggested. While Welch's ANOVA requires the assumption of normality, the advantage of the test is robustness to violations of homogeneity of variance. Therefore, the test can be used when there are differences in variance among groups or if group size is unequal.
3. Homogeneity confirmed but not normality: The Kruskal-Wallis test is suggested regardless of whether the variance among groups is homogeneous (that is, when data is not normally distributed but has homogeneity) or not homogeneous (when data is not normally distributed and lacks homogeneity). The Kruskal-Wallis test compares the medians or centers of three or more independent groups.
4. Neither normality nor homogeneity confirmed: Kruskal-Wallis test.

After testing for group differences via an omnibus test and calculating effect sizes, the last step is to conduct post hoc pairwise comparisons and calculate their associated measures of effect size. While it is beyond the scope of this paper to review all of the possible options, we will suggest commonly used tests below, similar to how effect sizes were presented. Whichever tests are utilized, authors should have an appropriate justification for why they are employed.

One-Way ANOVA.

A common measure of effect size for a one-way ANOVA is partial eta squared (η_p^2). Partial eta squared is a proportion from 0 to 1, showing the variance attributable to groups. Zero indicates that the independent variable explains none of the variance in the dependent variable, and 1

indicates that the independent variable explains all of the variance. Usually, η_p^2 is interpreted as: small = 0.01, medium = 0.06, large = 0.14.

Tukey's Honestly Significant Difference (HSD) test is a post-hoc test commonly used after conducting a one-way ANOVA to determine which specific groups differ from each other. It compares all pairs of group means and calculates a critical value based on the overall significance level and the number of groups. If the difference between the means of two groups is greater than this critical value, the difference is considered statistically significant. Tukey's HSD controls the family-wise error rate across all comparisons, controlling the overall Type I error rate when making multiple comparisons. An alternative approach to controlling this error rate is the Bonferroni correction. This correction is used to adjust the significance level (alpha-level) by dividing the desired alpha-level (e.g., 0.05) by the number of pairwise comparisons. This adjusted alpha-level is then used as the more stringent threshold to determine statistical significance for each individual comparison.

Welch's ANOVA.

Omega squared (ω^2) is a measure of effect size used in the context of ANOVA, interpreted similarly to η_p^2. Omega squared gives a less-biased estimate of the population effect size than η_p^2, which tends to overestimate the population effect size when there are many groups or group sizes and variances are unequal. This is why some researchers prefer ω^2 after Welch's ANOVA, which is used when the assumption of equal variances is violated. Usually, ω^2 is interpreted as: small = 0.01, medium = 0.06, large = 0.14.

The Games-Howell and Dunnett's T3 tests are both post hoc tests used in ANOVA to compare multiple groups when the assumption of homogeneity of variance is violated and/or group sizes are unequal. Games-Howell is used when the variances are unequal and the group sizes are different. It is considered more conservative than other post hoc tests like Tukey's HSD, making it suitable for situations with unequal variances and group sizes. Dunnett's T3 is similar to Games-Howell but is used specifically when comparing each treatment group to a control group. It is also suitable for unequal variances and group sizes, providing a more conservative approach to controlling the family-wise error rate compared to other post hoc tests.

Kruskal-Wallis Test.

Epsilon squared (ε^2) is a measure of effect size used in non-parametric tests. Epsilon squared estimates the proportion of variance in the dependent variable that is explained by the independent variable in the population. It is calculated as the sum of ranks variance explained by the independent variable divided by the total sum of ranks variance. The interpretation of ε^2 is similar to η_p^2 in ANOVA, with values close to 0 indicating a small effect and values close to 1 indicating a large effect. There are no universally agreed-upon thresholds for interpreting ε^2 in the context of the Kruskal-Wallis test. However, in general, the following guidelines can be used: small = 0.01, medium = 0.06, large = 0.14.

Dunn's test is a non-parametric post hoc test used after obtaining a significant result from the Kruskal-Wallis test. Dunn's test identifies which groups differ from each other according to their medians, utilizing the same pooled rankings as the Kruskal-Wallis test. Because the test does not rely on an underlying assumption of data normality, Dunn's test is suitable for analyzing data with non-normal distributions or when parametric test assumptions are violated. Researchers should be aware that a limitation of Dunn's test is that it can be less powerful when there are many tied ranks in the data.

Closing Thoughts

We feel that one item needs to be acknowledged before the conclusion. It is apparent that research needs to be more inclusive of individuals who identify as sex and gender minorities. The lack of guidance by sport and exercise science flagship organizations[18] may stem from a dearth of research in the area. We hope the information presented here will help sport and exercise science researchers with practical considerations around study design and implementation, data analysis, and interpretation of the results for sex-and gender-inclusive research.

While more research is needed, there is an equivalent need to protect the confidentiality of individuals who identify as sex and gender minorities, and the data that could potentially identify them. It is likely that at least one participant recruited in a typical exercise or sport science investigation identifies as a sex or gender minority (assuming an $n \geq 20$).[11,14] In this case, we urge researchers to err on the side of caution by not singling out the individual in a way that could allow others to identify them (i.e., demographic information used alongside the institutional affiliation). We make this statement acknowledging the rise in legislation targeting Lesbian, Gay, Bisexual, Transgender, Queer or Questioning, Intersex, Asexual, and Two-Spirit individuals (LGBTQIA2S+) in many U.S. states, and the concurrent increase in attacks on sex and gender minorities[51] which mirrors what has been observed in other countries.[52] The *All of Us* research program only allows reporting of data if 20 or more participants are grouped together.[15] This sample size minimum may not be feasible for most sport and exercise science investigations, so we encourage researchers to use their best judgement in presenting much-needed data that has previously been overlooked or gone underreported. For imbalanced data sets, one strategy would be to treat the sex and gender categories as populations, generate a random sample from those subgroups,[53] and then systematically reduce the numbers to approach a more counter-balanced group.[54] An example of this approach using communication medium can be found in the investigation by Thomas and Cardinal.[55] Whatever the approach, data should be reported in line with the participant protections outlined in the *Declaration of Helsinki*, which emphasizes respect for privacy and confidentiality, honest reporting, and avoiding harm to participants.[56] Additionally, researchers must comply with any relevant requirements set by institutional review boards and funding agencies.

In conclusion, we have presented an approach that researchers may use to be more inclusive when designing investigations across several sexes or genders. A recent Position Stand in the *International Journal of Exercise Science* encouraged the following: involving team members whose identities align with marginalized groups when conducting research, using correct pronouns to

foster a respectful environment, being accommodating to LGBTQIA2S+ participants by offering flexible timelines and considering options for participant safety and comfort, not assuming heterosexuality, and explaining to participants why the data are being collected, how it will be used, and how it will benefit the community.[12] We extend these guidelines by stating that there should be a valid, literature-supported rationale for testing for sex or gender differences. Once the number of independent groups to be evaluated has been identified, following the statistical analysis framework presented here will assist researchers in analyzing and interpreting their data appropriately.

Acknowledgements

The authors wish to acknowledge and thank Dr. Trevor Dean Ruiz (Assistant Professor, Department of Statistics, California Polytechnic State University, San Luis Obispo) for providing a review, particularly with regards to statistical approach and interpretation.

University of Nevada, Las Vegas (UNLV) is situated on the traditional homelands of Indigenous groups, including the Nuwu or Nuwuvi, Southern Paiute People, descendants of the Tudinu, or Desert People.

Western Kentucky University (WKU) honors and acknowledges the Indigenous peoples' land on which this University was built. All land in the state of Kentucky was once Indigenous territory, which is why it is our duty to acknowledge that WKU exists on Native land. The particular region of Kentucky wherein WKU sits was home to both the Shawnee (Shawandasse Tula) and Cherokee East (ᏣᎳᎩᏗᏣᏓ Tsalaguwetiyi) tribes.

The California Polytechnic State University in San Luis Obispo, California (Cal Poly) sits on the traditional lands of the Yak titʸu titʸu yak tiłhini Northern Chumash Tribe of San Luis Obispo County and Region. The Yak titʸu titʸu yak tiłhini have a documented presence for over 10,000 years. The Tiłhini Peoples have stewarded their ancestral and unceded homelands which include all of the cities, communities, federal and state open spaces within the San Luis Obispo County region. These homelands extend East into the Carrizo Plains toward Kern County, South to the Santa Maria River, North to Ragged Point, and West beyond the ocean's shoreline in an unbroken chain of lineage, kinship, and culture.

References

1. Webster L. "Ties that bind" The continued conflation of sex, sexuality and gender. *J Lang Sex*. 2021;10(1):63–70. https://doi.org/10.1075/jls.00015.web

2. Coen S, Banister E. *What a difference sex and gender make: a gender, sex and health research casebook*. CIHR Institute of Gender and Health; 2012.

3. Richards C, Bouman PW, Seal L, Barker JM, Nieder OT, T'Sjoen G. Non-binary or genderqueer genders. *Int Rev Psychiatry*. 2016;28(1):95–102. https://doi.org/10.3109/09540261.2015.1106446

4. American Psychological Association. Transgender. In: APA Dictionary of Psychology. May 22, 2025. https://dictionary.apa.org/transgender

5. American Psychological Association. Defining transgender terms. *Monitor on psychology*. 2018;49(8):32.

6. Brown A. About 5% of young adults in the US say their gender is different from their sex assigned at birth. *Pew Research Center*. 2022;7

7. Meerwijk EL, Sevelius JM. Transgender population size in the United States: a meta-regression of population-based probability samples. *Am J Public health*. 2017;107(2):e1–e8. https://doi.org/10.2105/AJPH.2016.303578

8. Hamburg MA, Collins FS. The path to personalized medicine. *N Engl J Med*. 2010;363(4):301–304. http://doi.org/10.1056/NEJMp1006304

9. Tolson H, Chevrette JM. Changes in attitudes toward physical activity as a result of individualized exercise prescription. *J Psychology*. 1974;87(2):203–207. https://doi.org/10.1080/00223980.1974.9915691

10. Davis DW, Garver MJ, Thomas JD, et al. How an IJES Working Group grappled with the complexities of three letters-DEI-with the goal to broaden inclusion and representation in exercise science research. *Int J Exerc Sci*. 2024;17(8):852–860. https://doi.org/10.70252/VYXH2713

11. Navalta JW, Davis DW, Stone WJ. Implications for cisgender female underrepresentation, small sample sizes, and misgendering in sport and exercise science research. *PLoS One*. 2023;18(11):e0291526. https://doi.org/10.1371/journal.pone.0291526

12. Navalta JW, Davis DW, Thomas JD, et al. The 2024 International Journal of Exercise Science Position Stand on Inclusion. *Int J Exerc Sci*. 2024;17(8):730–749. https://doi.org/10.70252/TVIW9464

13. Judge L, Livergood K, Smith A, Razon S. Advancing diversity, equity, and inclusion in kinesiology departments and allied health professions: Barriers, facilitators, and effective strategies. *J Equity Soc Justice Educ n*. 2024;3:1–18. https://doi.org/10.62889/2024/jkas1127

14. Garver MJ, Navalta JW, Heijnen MJH, et al. IJES self-study on participants' sex in exercise science: Sex-data gap and corresponding author survey. *Int J Exerc Sci*. 2023;16(6):364–376. https://doi.org/10.70252/DZZC8088

15. Investigators All of Us Research Program. The "All of Us" research program. *N Engl J Med*. 2019;381(7):668–676. https://doi.org/10.1056/NEJMsr1809937

16. Soto Y, Muñoz MA. Advancing BIPOC Diversity, Equity & Inclusion within Kinesiology. American College of Sports Medicine. Sept. 21, 2022, 2022. Accessed November 20, 2024, 2024. https://www.acsm.org/docs/default-source/member-hub-documents/dei-newsletters/advancing-bipoc-dei-within-kinesiology.pdf?sfvrsn=df5e33ce_2

17. American College of Sports Medicine. The American College of Sports Medicine Releases Official Statement on Diversity, Equity and Inclusion. American College of Sports Medicine. June 9, 2020, Accessed November 20, 2024, 2024. https://www.acsm.org/news-detail/2020/06/09/the-american-college-of-sports-medicine-releases-official-statement-on-diversity-equity-and-inclusion

18. American College of Sports Medicine, Riebe D, Ehrman JK, Liguori G, Magal M. *ACSM's Guidelines for Exercise Testing and Prescription*. Tenth edition. ed. Wolters Kluwer Health; 2018:, 472 pages.

19. Christensen JE, Christensen CE. Statistical power analysis of health, physical education, and recreation research. *Res Q Am Alliance Health Phys Educ Recreat.* 1977;48(1):204–208. https://doi.org/10.1080/10671315.1977.10762173

20. Speed HD, Andersen MB. What exercise and sport scientists don't understand. *J Sci Med Sport*. 2000;3(1):84–92. https://doi.org/10.1016/S1440-2440(00)80051-1

21. Knudson DV, Lindsey C. Type I and Type II errors in correlations of various sample sizes. *Compr Psychol*. 2014;3. https://doi.org/10.2466/03.CP.3.1

22. Mesquida C, Murphy J, Lakens D, Warne J. Publication bias, statistical power and reporting practices in the Journal of Sports Sciences: potential barriers to replicability. *J Sports Sci*. 2023;41(16):1507–1517. https://doi.org/10.1080/02640414.2023.2269357

23. National Institutes of Health. Consideration of sex as a biological variable in NIH-funded Research. Accessed October 1, 2024, https://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-102.html

24. Miller LR, Marks C, Becker JB, et al. Considering sex as a biological variable in preclinical research. *FASEB J*. 2017;31(1):29–34. https://doi.org/10.1096/fj.201600781R

25. Heidari S, Babor FT, Castro DP, Tort S, Curno M. Sex and Gender Equity in Research: rationale for the SAGER guidelines and recommended use. *Res Integr Peer Rev*. 2016;1(1). https://doi.org/10.1186/s41073-016-0007-6

26. Costello JT, Bieuzen F, Bleakley CM. Where are all the female participants in Sports and Exercise Medicine research? *Eur J Sport Sci*. 2014;14(8):847–851. https://doi.org/10.1080/17461391.2014.911354

27. Cowley ES, Olenick AA, McNulty KL, Ross EZ. "Invisible Sportswomen": The sex data gap in sport and exercise science research. *Women Sport Phys Act J*. 2021;29:146–151. https://doi.org/10.1123/wspaj.2021-0028

28. Twomey R, Yingling V, Warne J, et al. The nature of our literature: A registered report on the positive result rate and reporting practices in kinesiology. *Commun Kinesiol*. 2021;1(3). https://doi.org/10.51224/cik.v1i3.43

29. Stefan AM, Schonbrodt FD. Big little lies: a compendium and simulation of p-hacking strategies. *R Soc Open Sci*. 2023;10(2):220346. https://doi.org/10.1098/rsos.220346

30. Bleakley C, Reijgers J, Smoliga J. Many high quality RCTs in sports physical therapy are making false positive claims of treatment effect: a systematic survey. *J Orthop Sports Phys Ther*. 2020;50(2):104–109. https://doi.org/10.2519/jospt.2020.9264

31. International Journal of Exercise Science. Manuscript Types. Open Access Text. Accessed February 7, 2025, 2025. https://intjexersci.com/manuscript-types/

32. International Journal of Kinesiology in Higher Education. Instructions for authors. Accessed February 7, 2025, 2025. https://www.tandfonline.com/action/authorSubmission?show=instructions&journalCode=ukhe20

33. Cohen J. *Statistical power analysis for the behavioral sciences*. Routledge; 2013. https://doi.org/10.4324/9780203771587

34. Beck TW. The importance of a priori sample size estimation in strength and conditioning research. *J Strength Cond Res*. 2013;27(8):2323–2337. https://doi.org/10.1519/JSC.0b013e318278eea0

35. Johnson SL, Stone WJ, Bunn JA, Lyons TS, Navalta JW. New author guidelines in statistical reporting: Embracing an era beyond p<. 05. *Int J Exerc Sci*. 2020;13(1):1-5. https://doi.org/10.70252/HMZN3851

36. Albers C, Lakens D. When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *J Exp Soc Psychol*. 2018;74:187–195. https://doi.org/10.1016/j.jesp.2017.09.004

37. Eisenmann J. Translational Gap between Laboratory and Playing Field: New Era to Solve Old Problems in Sports Science. *Transl J Am Coll Sports Med*. 2017;2(8):37–43. https://doi.org/10.1249/tjx.0000000000000032

38. Ditroilo M, Mesquida C, Abt G, Lakens D. Exploratory research in sport and exercise science: Perceptions, challenges, and recommendations. *J Sports Sci*. 2025;43(12):1108–1120. https://doi.org/10.1080/02640414.2025.2486871

39. Murphy J, Caldwell AR, Mesquida C, et al. Estimating the replicability of sports and exercise science research. *Sports Med*. 2025; https://doi.org/10.1007/s40279-025-02201-w

40. Althouse AD. Post hoc power: Not empowering, just misleading. *J Surg Res*. 2021;259:A3–A6. https://doi.org/10.1016/j.jss.2019.10.049

41. Vaske JJ. Communicating judgments about practical significance: Effect size, confidence intervals and odds ratios. *Hum Dimens Wildl*. 2010;7(4):287–300. https://doi.org/10.1080/10871200214752

42. Caldwell A, Vigotsky AD. A case against default effect sizes in sport and exercise science. *PeerJ*. 2020;8:e10314. https://doi.org/10.7717/peerj.10314

43. Yagin FH, Yagin B, Pinar A. Normality Distributions Commonly Used in Sport and Health Sciences. *J Exerc Sci Phys Act Rev*. 2024;2(1):124–131.

44. Razali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J Stat Model Anal*. 2011;2(1):21–33.

45. Legendre P, Borcard D. Statistical comparison of univariate tests of homogeneity of variances. *J Stat Comput Simul*. 2008;514.

46. Lim T-S, Loh W-Y. A comparison of tests of equality of variances. *Comput Stat Data Anal*. 1996;22(3):287–301. https://doi.org/10.1016/0167-9473(95)00054-2

47. Brown MB, Forsythe AB. The small sample behavior of some statistics which test the equality of several means. *Technometrics*. 1974;16(1):129–132. https://doi.org/10.1080/00401706.1974.10489158

48. Pagano M, Gauvreau K, Mattie H. *Principles of biostatistics*. Chapman and Hall/CRC; 2022. https://doi.org/10.1201/9780429340512

49. Grissom RJ. Probability of the superior outcome of one treatment over another. *J Appl Psychol*. 1994;79(2):314. https://doi.org/10.1037/0021-9010.79.2.314

50. Navalta JW, Davis DW, Thomas DJ, Stone JW. An example analysis for a gender-inclusive approach in sport and exercise science research using Fitbit outcomes from the All of Us Research Program data set. *Int J Exerc Sci*. 2025;18(1).

51. Brightman S, Lenning E, Lurie KJ, DeJong C. Anti-transgender ideology, laws, and homicide: An analysis of the trifecta of violence. *Homicide Stud*. 2024;28(3):251–269. https://doi.org/10.1177/10887679231201803

52. Katsuba S. The decade of violence: A comprehensive analysis of hate crimes against LGBTQ in Russia in the era of the "Gay Propaganda Law"(2010–2020). *Victims Offenders*. 2024;19(3):395–418. https://doi.org/10.1080/15564886.2023.2167142

53. Krejcie R, Morgan D. Determining sample size for research activities. *Educ Psychol Meas*. 1970;30:607–610. https://doi.org/10.1177/001316447003000308

54. Zhu W. Making bootstrap statistical inferences: a tutorial. *Res Q Exerc Sport*. 1997;68(1):44–55. https://doi.org/10.1080/02701367.1997.10608865

55. Thomas JD, Cardinal BJ. How credible is online physical activity advice? The accuracy of free adult educational materials. *Transl J Am Coll Sports Med*. 2020;5(9):82–91. https://doi.org/10.1249/tjx.0000000000000122

56. World Medical Association. World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human participants. *JAMA*. 2025;333(1):71–74. https://doi.org/10.1001/jama.2024.21972

Corresponding author: James Navalta; james.navalta@unlv.edu