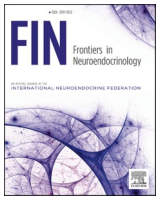
Frontiers in Neuroendocrinology 82 (2026) 101257

Contents lists available at ScienceDirect

Frontiers in Neuroendocrinology

journal homepage: www.elsevier.com/locate/yfrne



# From significant to meaningful: ATOMizing the study of sex differences and similarities

Carla Sanchis-Segura [a,*], Cristina Forn [a], Rand R. Wilcox [b]

[a] Departament de Psicologia bàsica, clinica i psicobiologia, Universitat Jaume I, Castelló, Spain
[b] Department of Psychology, University of Southern California, Los Angeles, USA

ARTICLE INFO

*Keywords:*
Sex
Gender
Means
Robust statistics
Precision medicine

ABSTRACT

The sex differences field often relies on an implicit and flawed heuristic: defining a "difference" by any statistically significant gap between group averages. This practice yields findings that are often not useful, even counterproductive, for understanding sex-related variation and for advancing personalized healthcare.

Following current statistical consensus, we argue that sex differences should be defined not by significance alone but by context-dependent criteria prioritizing informativeness. Drawing on the ATOM principles (Accept uncertainty, be Thoughtful, be Open, be Modest), we call for explicit, justified definitions and for a shift from group averages to meaningful individual variation.

To illustrate this methodological and philosophical shift, we introduce Thresholded Probability of Superiority (TPS), a method that treats sex differences as probability distributions rather than overgeneralized, fixed abstractions. Thus, TPS allows for a more nuanced, relevant, and actionable understanding of sex-related variation, with greater potential to inform precision medicine.

## 1. Introduction: Rethinking what counts as a sex difference

"A difference is a very peculiar and obscure concept. It is certainly not a thing or an event.[…] A difference, then, is an abstract matter". *Bateson, G. (1972)* (Bateson, 1999).

The study of sex[1] is distinctive among the behavioral and biomedical sciences (BBMs) in that its object is not a naturally bounded entity, but a difference—one that is conceptually defined and statistically constructed. This is why the field is ordinarily referred to as the "sex differences field," and its work, in effect, is a sustained (re)search for differences between sex categories (i.e., male and female). Yet, strikingly, very seldom (if ever) do we—researchers in this field—ask ourselves *what a sex difference is*, or more precisely, what should count as a sex difference. Instead, we rely—implicitly or explicitly—on a simple heuristic: if a difference between the group averages of males and females reaches statistical significance, we consider it a sex difference.

This convention is intrinsically problematic: it is statistically flawed, rooted in widespread misunderstanding and misinterpretation of *p*-values. It is also impractical—often counterproductive—because reliance on averages produces conclusions that are imprecise and, for a considerable though unquantified number of individuals, demonstrably inaccurate or even false. Ultimately, it is harmful: it obstructs the field's ability to achieve its core objectives (accurately describing sex-related variation and advancing precision medicine), while the way its

* Corresponding author at: Departament de Psicologia bàsica, clinica i psicobiologia, Facultat de Ciències de la Salut, Universitat Jaume I, Avda Sos Baynat, SN, 12071 Castelló, Spain.
*E-mail address:* csanchis@uji.es (C. Sanchis-Segura).

[1] We believe that the inclusion of "sex" in biomedical and behavioral sciences should be approached from a broad perspective that simultaneously considers aspects typically associated with both "sex" and "gender." Accordingly, it seems to us more accurate to refer to this as "Sex/Gender," thereby making explicit that these both sets of aspects are both important and inherently intertwined. At the same time, we think it is important to acknowledge that sex itself is not a causal agent—that is, not a factor that can be experimentally manipulated. This can be made clear by including the qualifier "related." Finally, we think that the goals of studying these sex/gender-related phenomena should be broadened beyond simple comparison (i.e., the identification of differences and similarities) to also include description, explanation, and prediction. Moreover, even when focusing on comparison, similarities should receive the same consideration as differences. Therefore, we feel it would be more appropriate to speak of "variation" rather than "differences."

Nevertheless, because this commentary aims to provide a critical perspective, we will use the commonly adopted terms "sex" and "differences" when describing current approaches, employing more nuanced terminology only occasionally and strategically.

https://doi.org/10.1016/j.yfrne.2026.101257
Received 4 May 2026; Received in revised form 19 May 2026; Accepted 20 May 2026
Available online 25 May 2026

findings are typically framed risks reinforcing stereotypes, perpetuating inequities, and contributing to discrimination.

This comment is born of the conviction that if the so-called "sex differences field" is to mature—conceptually, pragmatically, and socially—we must both reflect on what it means to call something a "sex difference" and rethink what should count as a sex difference. This piece focuses on the second of these challenges. It does not aim to be comprehensive; rather, it seeks to raise awareness of the need to shift our focus from the detection of any sex difference to the interpretation of how often and how meaningfully these differences arise. The comment also introduces the Thresholded Probability of Superiority (TPS), a novel statistical approach conceived within this alternative philosophy.

## 2. Problems with the current approach

The preceding section briefly introduced the conceptual and practical shortcomings inherent in the field's implicit definition of sex differences. Building on this, in this section we examine the specific methodological pillars that underpin this problematic framework: the overreliance on statistical significance and the uncritical use of group means. As detailed below, these deeply entrenched practices do not only lead to flawed conclusions but also actively impede a nuanced understanding of sex-related variation and obstruct the advancement of personalized healthcare (Fig. 1).

### 2.1. *The problem of "significant" sex differences*

"The textbooks are wrong. The teaching is wrong. The seminar you just attended is wrong. The most prestigious journal in your scientific field is wrong."

Ziliak and McCloskey (2008, p. 250) (Zilliak and McCloskey, 2008).

"What is wrong with NHST? Well among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that out of desperation, we nevertheless believe that it does!"

Cohen, J. (1994) (Cohen, 1994).

In the BBMs, statistical significance—and its applied framework, null hypothesis significance testing (NHST)—is often treated as a fundamental requirement for scientific validity. It forms the backbone of most researchers' statistical training (Perezgonzalez, 2015; Kline, 2013a) and is, therefore, a familiar procedure: set a threshold (usually $p < 0.05$), run a test to obtain a *p*-value, and declare as "real" (not due to chance) only those findings with *p*-values below the threshold. In the "sex differences" field, this binary pass–fail logic is even more entrenched (see Fig. 2), with statistical significance routinely used as the very definition of a sex difference (Maney, 2016). Yet this very framework is what critics now argue is fundamentally flawed, leading to a distorted scientific record (Nuzzo, 2014; Wasserstein et al., 2019; Gigerenzer et al., 2004). As Gerd Gigerenzer has argued (Gigerenzer et al., 2004), NHST has evolved into a "null ritual" that replaces genuine statistical thinking with mechanical, threshold-checking. This approach fosters a false sense of certainty and conflates what is merely "statistically significant" with what is truly significant (Zilliak and McCloskey, 2008; Cohen, 1994; Kline, 2013a; McShane et al., 2019).

NHST does not test the researchers' hypotheses, it just tests a null hypothesis—one assuming no difference. This poses a non-quantitative

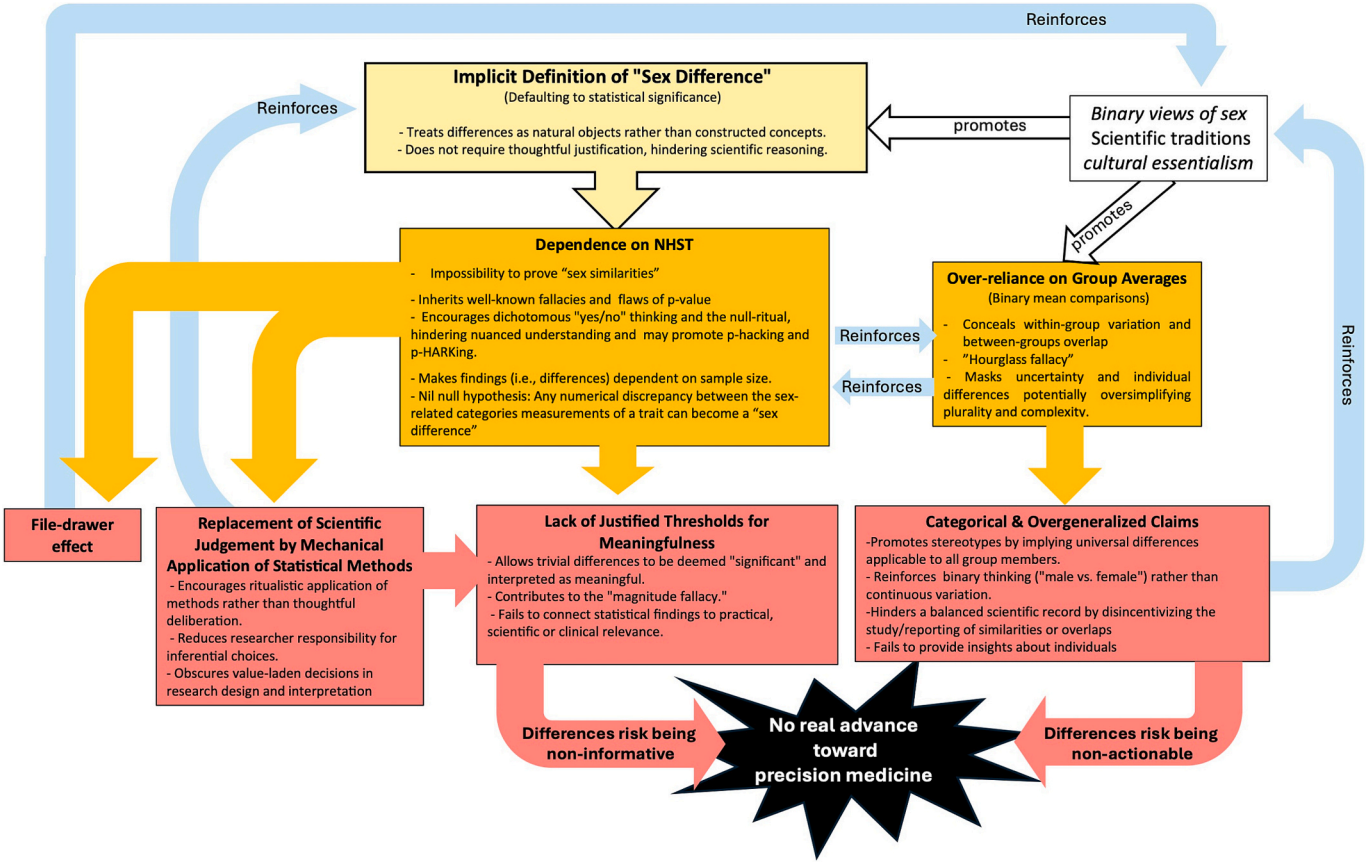



**Fig. 1.** Foibles, perils, and pitfalls of current approaches to the study of sex differences. This diagram summarizes the foibles, perills and pitfalls stemming from an implicit definition of "sex difference" and the over-reliance on Null Hypothesis Significance Testing (NHST) and binary group averages. These practices replace scientific judgment by mechanical application of statistical methods, leading to a lack of justified thresholds for meaningfulness and promoting overgeneralized, categorical claims derived from group averages. Ultimately, this framework yields findings that can be both non-informative and non-actionable, reinforcing binary thinking and hindering any real advance toward individual-level insights necessary for precision medicine.

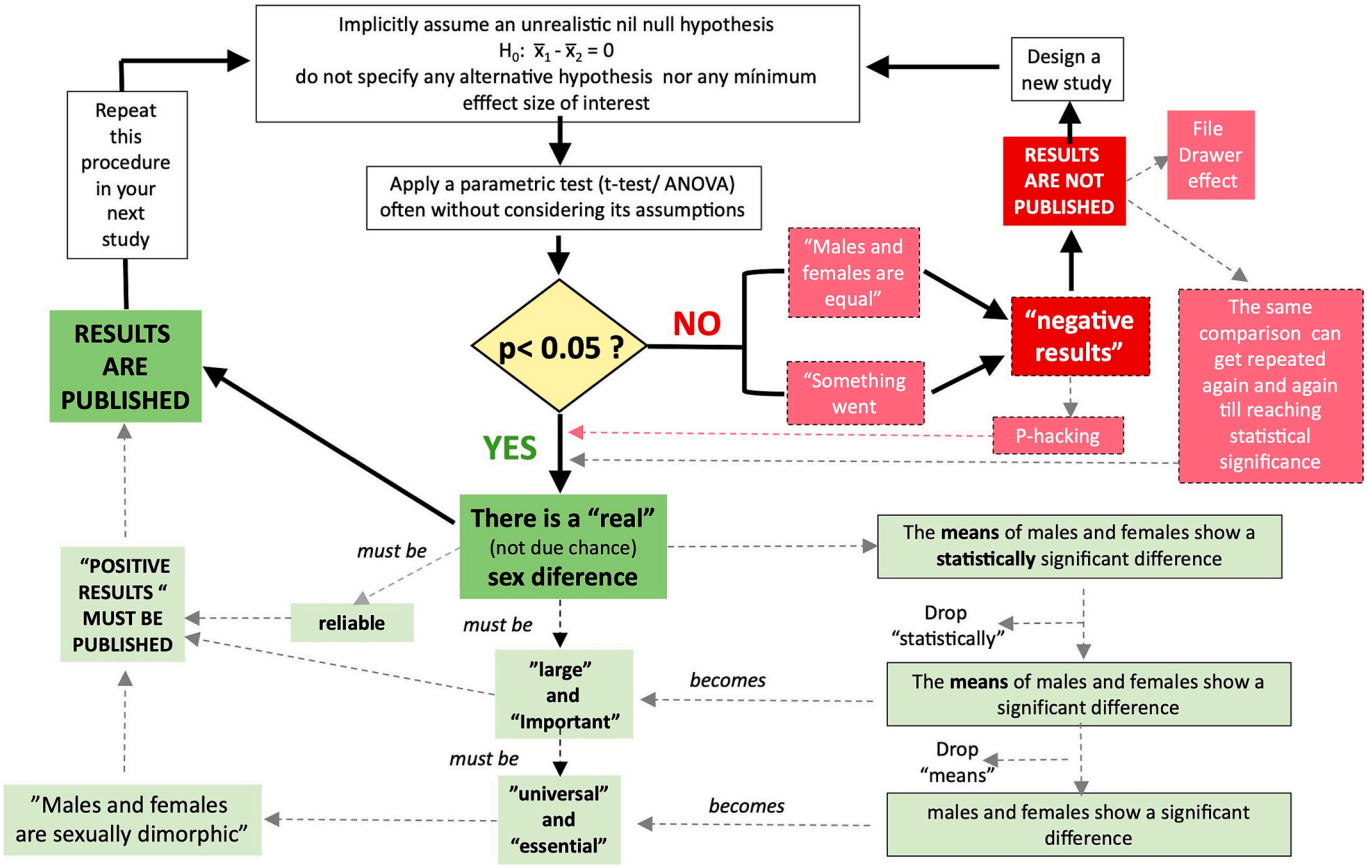



**Fig. 2.** The (re)search of sex differences as a null, methodological, and epistemic ritual. The diagram visualizes how Gigerenzer's "null ritual" (Perezgonzalez, 2015) is paradigmatically expressed in the field of sex differences research. Elements connected by continuous arrows represent the methodological circuitry that reproduces the main features of the ritual as originally described. In this field, the very object of inquiry—a difference—is defined solely by statistical significance. This definition remains implicit and is operationalized through a nil null hypothesis in classic average-comparison tests, without offering substantive alternative hypotheses or thresholds of meaningfulness. As in other areas, only findings associated with $p < 0.05$ are deemed "real," "positive," and "publishable," creating a distorted record, a persistent confusion between statistical significance and scientific importance, and further conceptual and practical distortions (illustrated by pale red boxes and their links). Unlike what happens in other domains, however, in the sex differences field this process extends beyond methodology and becomes an epistemic ritual. Through the automatic repetition of tests and the elevation of the *p*-value as a truth criterion, the practice turns into a circular gesture that legitimizes what it already presupposes: the existence of difference. This methodological automatism, sustained by institutional policies that reward "differences" through publication, transforms an analytical tool into a cultural convention. Arising from and reinforced by pre-existing essentialist notions of males and females as inherently distinct or even opposite, this process fosters a culture in which findings are uncritically accepted, leading directly to the conceptual and interpretive pitfalls (green boxes connected by discontinuous arrows) discussed in the main text.

question —*does a difference exist?* — that has been described as metaphysical rather than scientific (Zilliak and McCloskey, 2008; Lambdin, 2012). The tested hypothesis of no difference is, moreover, almost always false: Any two conditions will always differ at some decimal place, and with large enough samples, even trivial differences will become statistically significant (Tukey, 1991; Wasserstein and Lazar, 2016). Thus, what counts as a (sex) difference —and, by extension, whether (sex) differences are numerous or scarce—depends as much on sample size as on the actual effects magnitude. This renders results partly arbitrary and, with massive datasets now widely available, especially prone to the inflation of tiny variations into "significant" effects.

NHST does not only test the "wrong" hypothesis, it also tests it in the wrong way (Cohen, 1994; Kline, 2013a; Gigerenzer et al., 2004; Carver, 1993; Nickerson, 2000). The null hypothesis can be rejected but never accepted; so, the existence of a (sex) difference can be proven, but its absence cannot. This is the exact opposite of the proper epistemology of scientific hypothesis testing and falsification (Cohen, 1994; Heene and Ferguson, 2017a). Together with other biases, this inverted logic produces a distorted and incomplete scientific record favoring "significant differences" (Kline, 2013a; Amrhein et al., 2019) —many of which may be false (Ioannidis, 2005; Simmons et al., 2011; Heene and Ferguson, 2017b).

Most researchers, however, overlook these limitations and treat *p*-values below the threshold as confirmation of their hypotheses -a phenomenon known as the the *"Bayesian Id's wishful thinking"* (Cohen, *1994;* Haller et al., *2002;* Lecoutre et al., *2003;* Greenland et al., *2016;* Cumming, *2012*). This implies forgetting that NHST just test the null hypothesis, and also that *p*-values represent the probability of data under repeated sampling—not in a particular experiment, which can only be 0 or 1 (Cohen, 1994; Kline, 2013a; Wasserstein and Lazar, 2016; Greenland et al., 2016). Misinterpretation escalates when researchers, falling in the "*slippery slope of significance*" (Cumming, 2012), drop the qualifier "statistically" and report their findings simply as "significant" —a term which, in everyday language, implies large, categorical, and absolute distinctions (Kline, 2013a; Carver, 1993). This "*magnitude fallacy" (*Kline, *2013a)* make researchers misinterpret small *p*-values as indicative of a large or practically important effects (Greenland et al., 2016; Cumming, 2014), and it is particularly pronounced and more likely to emerge when essentialist beliefs about categories are already strong (as often occurs in sex-comparison studies) or when effect sizes are absent, ignored, or misunderstood.

Concerns about the widespread misuse and misinterpretation of

NHST prompted the American Statistical Association (ASA) to issue an authoritative statement clarifying the proper use and interpretation of *p*-values in 2016 (Wasserstein and Lazar, 2016). Crucially for our argumentation here, the third principle of this statement made clear that "*Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold*" and that "*The widespread use of 'statistical significance' (generally interpreted as 'p ≤ 0.05') as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process*." Specifically, "*Practices that reduce data analysis or scientific inference to mechanical 'bright-line' rules (such as 'p < 0.05') for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making. A conclusion does not immediately become 'true' on one side of the divide and 'false' on the other.*"

This explicit warning about the perils and pitfalls of using thresholds when interpreting *p*-values and dichotomous thinking has led some scholars (Wasserstein et al., 2019) to advocate for "*to stop using the term 'statistically significant' entirely. Nor should variants such as "significantly different," "p < 0.05," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way*". Their proposition –using *p*-values as continuous measures of evidence within contextual assumptions, not of importance, and abandoning any dichotomous thinking, terminology, and marking– has several precedents (for overviews, see (Zilliak and McCloskey, 2008; Kline, 2013a)), but is not shared by all scholars and ASA members (Harlow, 2014; Benjamini et al., 2021). The latter maintain that "*Thresholds are helpful when actions are required. Comparing P-values to a significance level can be useful*", but also acknowledge that "*P- values themselves provide valuable information*" and that "*If thresholds are deemed necessary as a part of decision-making, they should be explicitly defined based on study goals*" (Benjamini et al., 2021). Therefore, beneath the ongoing disputes, the core consensus established in the 2016 Statement (Wasserstein and Lazar, 2016) remains intact: Surpassing statistical significance thresholds must not be mechanically equated with scientific truth.

Here we argue that, regardless of the positioning adopted in this ongoing debate, the message of the ASA's 2016 statement strikes at the very nucleus of the "sex differences field" and we, as researchers, can no longer ignore it. In our opinion, persisting in a definition of sex differences *solely* based on automatically checking whether the observed difference surpasses a convention-based statistical significance threshold undermines the very foundations of this field. This practice inadvertently rests the discipline's central object of study on an implicit and methodologically fragile, sample-dependent criterion that is incompatible with the current statistical consensus summarized in the ASA statement (Wasserstein and Lazar, 2016) compromising both replicability and relevance. Moreover, this definition imports all NHST's flaws into a domain already prone to dichotomous thinking and, combined with NHST's reversed logic, fuels a research literature where differences are ordinarily celebrated, similarities are sidelined, and nuance is lost.

### 2.2. The problem of using group means as proxies of sex-related groups

"[…] *variation itself is nature's only irreducible essence. Variation is the hard reality, not a set of imperfect measures for a central tendency*".

Stephen J. Gould.

Concerns about defining sex differences through statistical significance—and the associated misuse and misinterpretation of NHST and *p*-values—are not the only methodological issues affecting the "sex differences field." An equally important problem is the heavy reliance on group averages. This practice is particularly problematic when the compared metrics are—as is typically the case—arithmetic means. This is because means are truly representative only when the data follow a normal distribution (a scenario rarely, if ever, encountered (Micceri, 1989)) and when the compared groups do not differ in dispersion. This latter assumption of homogeneity of variance is highly unlikely to hold and remains a matter of intense debate within the "sex differences field" (Halsey et al., 2023). These issues have been examined in detail elsewhere (Sanchis-Segura and Wilcox, 2024); here, only some central points are summarized.

Groups are not single individuals repeated *n* times. This simple fact should prompt us to question how often—and to what extent—averages, especially means, provide an adequate summary of group data

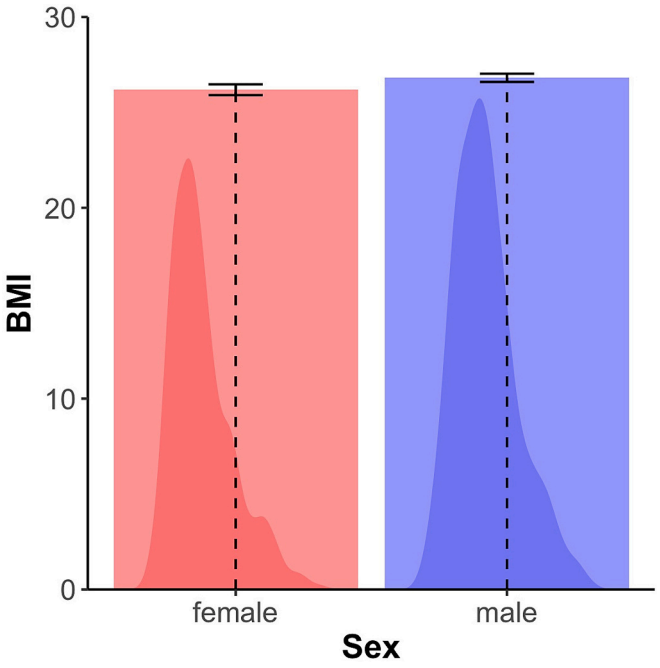



**Fig. 3.** Means hide more than they reveal. The figure depicts the means (bars), standard errors, and distribution (density plots in the background) of the females and males' BMI scores used in our worked example (see, Section 3.2). Bar plots depicting mean values ± standard errors (often misinterpreted as measures of spread) are commonly used to represent group comparisons based on means, then imposing a mirage of within-group homogeneity over truly existing individual variation information. As also illustrated here by dotted vertical lines, means may not represent what they are expected to (the most common and centered value), can provide a very misleading summary of group scores, and may hide the inadequacy of parametric comparisons.

(Speelman and McGann, 2013). Yet this question is rarely asked. By nature, means are abstractions, and they often hide more than they reveal, replacing complex and diverse realities with a mirage of homogeneity (Fig. 3). Even in very unlikely but most favorable conditions (i.e., when normality assumptions are fulfilled), compressing all observations into a single value sacrifices relevant information about variability, skewness, and distributional shape. Yet, outside the statistical fiction of the normal distribution, means may describe no one at all (Micceri, 1989).

Comparing two populations by their means not only assumes that the means are representative, but also that a single value can represent the general standing of one entire population compared to the other (Sanchis-Segura and Wilcox, 2024; Delaney and Vargha, 2002). Differences between two means may be driven by particular subsets of individuals, or even obscure conflicting trends (Speelman and McGann, 2013; Anastasi, 1981; Jacklin, 1981)—a problem captured by the

Simpson's paradox[2] (Kievit et al., 2013) and other similar phenomena (Speelman and McGann, 2013). Such distortions become more likely when groups are not formed by random assignment, as in the case of sex-related categories (Anastasi, 1981). Nevertheless, *t*-tests and other parametric mean-comparison methods remain the default in sex-differences research. This is not merely a technical habit; it reflects—and reinforces—a conceptualization of sex that privileges categorical over continuous variation, leading to what may be called the *hourglass fallacy* (Sanchis-Segura and Wilcox, 2024).

As fig. 4 illustrates, the hourglass fallacy operates through two steps. First, groups are reduced to their averages, erasing individual differences, and creating both the false impression of within-group homogeneity and the illusion of sharp categorical boundaries. Second, mean differences are overgeneralized as if they apply equally to all individuals. This fallacious generalization is made explicit —and gets reinforced— by the omission of key qualifiers (e.g., "on average," "means") when reporting results. Together with the previously outlined magnitude fallacy, it encourages generic statements (e.g., "*There is a significant difference between the heights of men and women*") and hyperbolic, essentialist expressions (e.g., "*sexual dimorphism*") that fuel stereotypes and the erroneous belief of males and females as fundamental "opposites."

Stephen Jay Gould (Gould, 2013) traced the misuse of averages to our "*Platonic heritage*" which "*seeks sharp essences and definite boundaries*" and that "*with its emphasis in clear distinctions and separated immutable entities, leads us to view statistical measures of central tendency wrongly, indeed opposite to the appropriate interpretation in our actual world of variation, shadings, and continua*". Gould also warned that such misinterpretations, can cause real harm.

In this light, overreliance on average comparisons undermines the two most compelling aims of the "sex differences field": first, to describe and explain variation in human physiology and behavior; and second, to apply this knowledge to improve and personalize healthcare (Sanchis-Segura and Wilcox, 2024; DuBois et al., 2025; Karp et al., 2025; Kmetz, 2019). However, when average comparisons replace truly more complex existing variation, "precision" medicine risks collapsing into a "two-sizes-fit-all" model—slightly less crude than "one-size-fits-all" but still poorly matched to human diversity.

### 2.3. The problem of abandoning current defaults

"There are no quick fixes. […] A critical first step forward is to begin accepting uncertainty and embracing variation in effects and recognizing that we can learn much (indeed, more) about the world by forsaking the false promise of certainty".

*BB.McShane* et al., *(2019,* 10].

Efforts to better define sex differences and improve research practices in this field involves what has been memorably described as "*moving to a world beyond* $p < 0.05$" (Wasserstein et al., 2019). As argued earlier, this requires abandoning not only bright-line significance thresholds but also escaping what might be called Plato's cave of averages and its resulting "hourglass fallacy" (Sanchis-Segura and Wilcox, 2024). This is not a single or minor step; it is a challenging journey. The "sex differences" field routinely uses *p*-values —obtained from parametric methods comparing means (i.e., *t*-tests and ANOVAs)— to sieve for "true sex differences." Therefore, removing both can feel like pulling the floor from under the discipline's feet.

For this reason, concerns about p-values and averages are often dismissed as technical nuances of limited practical relevance. Yet, many researchers reconsider their stance when confronted with the unequivocal position of the ASA statement (Wasserstein and Lazar, 2016) and when engaging more deeply with what group averages actually represent. At that point, common questions arise—sometimes framed with skepticism but also genuine curiosity: *If not statistical significance, then how should we define sex differences? And if not averages, how should we interpret the scores of dozens, hundreds, or even thousands of individuals?*

These questions highlight a discomfort that is understandable. The ASA urged researchers to abandon the use of bright-line thresholds for decision-making (Wasserstein and Lazar, 2016). Yet the ASA did not prescribe a direct substitute and critics have explicitly noted that there is no single alternative (Cohen, 1994; Wasserstein et al., 2019; McShane et al., 2019; McShane and Gal, 2017). There is no universal, plug-and-play method that can simply replace statistical significance in all contexts, and very much the same can be said for means. For a field in which NHST and mean comparisons have become almost synonymous with "analysis" itself, this absence of a ready-made replacement can feel like an unsettling void.

Yet, the absence of a single replacement for NHST is not an invitation to despair; it is a call to think differently. The ASA statement recommends principles over prescriptions (Wasserstein and Lazar, 2016). These principles and other similar recommendations (Kline, 2013a; McShane et al., 2019) have been incorporated and extended in those captured by the ATOM acronym: Accept uncertainty, be Thoughtful, be Open, be Modest (Wasserstein et al., 2019). These principles shift the focus from mechanical applications to transparent reasoning and epistemic humility. They are a compass for exploration, not a paved highway to a new promised land of definitive answers.

Here, we take the position that the ATOM principles—together with abandoning the mechanical usage of NHST to search for Platonic "sex essences" through binary mean comparisons—must guide attempts to arrive at workable definitions of sex differences. Our choice to speak of *definitions* in the plural is deliberate: there is no single, universally valid replacement, whether methodological or conceptual.

Specifically, we argue that the question "what is a sex difference?" —or more precisely, "*what should be counted as a sex difference?*" —may not have a universal answer. Instead, answers must probably be reworked and justified anew in each research context. This is a major philosophical shift —and also a demanding task— that requires:

- Accepting uncertainty;
- Thoughtfully defining meaningful differences and choosing appropriate methods;
- Openly justifying methodological choices and transparently reporting results;
- Modestly communicating balanced conclusions.

In this way, statistical thinking is restored to its proper role—a tool for judgment rather than a ritual—and sex differences can move beyond being merely "significant" and become genuinely informative.

This "ATOMized" approach must also be applied to the field's

[2] The so-called Simpson's paradox is a statistical phenomenon that occurs when a relationship observed in a population vanishes or reverses within all its subpopulations, often because a confounding variable is distributed differently across groups or not considered at all. Simpson's paradox may have significant consequences in several contexts, including efforts to incorporate sex- and gender-related variation in advancing precision medicine. For example, a higher dosage of medicine X might be associated with greater improvement rates at the population level, yet actually correlate with lower improvement rates within both females and males.

Although, as in the previous example, Simpson's paradox is often described by comparing the whole population with certain groups, the same phenomenon can occur within these supposedly homogeneous groups, which may inadvertently consist of several unidentified subgroups. This is more likely to occur when groups are large and not formed through randomized assignment, as is the case with sex-related categories. For instance, the effect of drug X observed among females as a group could be the opposite of that seen within two or more subgroups of women (e.g., pre- and post-menopausal women).

For a detailed discussion of the Simpson's paradox in behavioral and biomedical sciences see (Kievit et al., 2013), for a recent discussion of the complexity of sex/gender-related categories as source of non-fixed set of known and unknown co-variates, see (Maney et al., 2025). For a technical statistical discussion of the problems of using arithmetic means to summarize non-homogeneous (mixed normal) distributions, see (Wilcox, 1998).

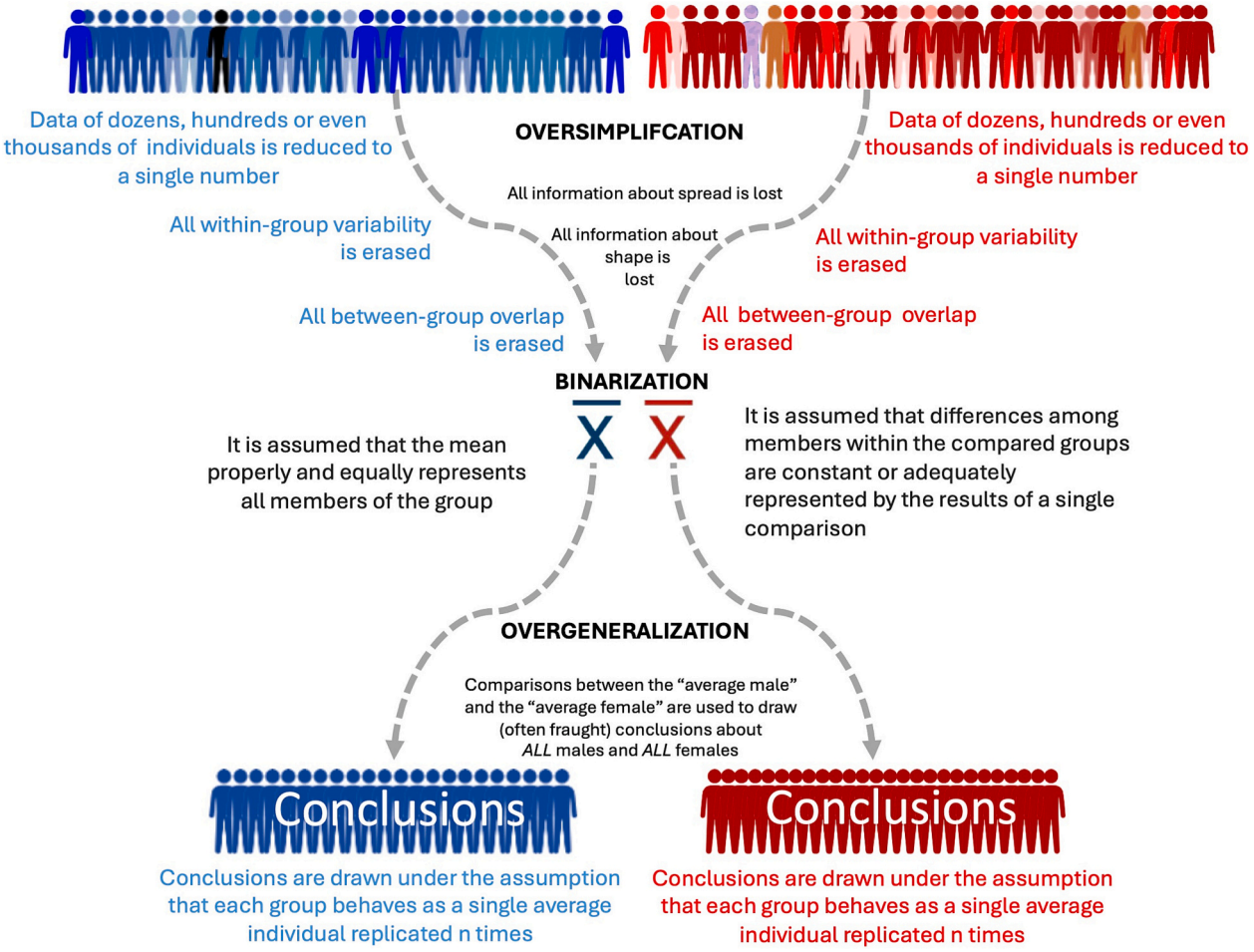



**Fig. 4.** The "hourglass fallacy". The diagram summarizes the main tenets of the so-called "hourglass fallacy". This term does not refer to a logic error but to a flawed and yet generalized analytical strategy that operates in two main steps: First, complex data from many and diverse individuals are reduced to a single value (oversimplification), usually the group mean. Through this averaging process, complex sex-related data become binary, not because of initial categories but because of the imposed mean (binarization). In the second step, the difference between these means is overgeneralized, applied not only to every individual in the sample but even to entire populations they are thought to represent. The term "hourglass" reflects the imbalance between the effort invested and the limited, often distorted, information produced. It also highlights the fragility of relying on averages as the core of research comparing broad, heterogeneous groups, particularly when such groups—like those defined by sex—are not formed by random assignment. For a detailed discussion of this concept, see (Sanchis-Segura and Wilcox, 2024).

overreliance on averages. The routine use of average comparisons often stems from convention, not thoughtful reflection. By concealing within-group variation and the true extent of between-group overlap, simple averages and their comparisons hide uncertainty and variation, violate the core principle of openness, and encourage immodest, categorical claims (Sanchis-Segura and Wilcox, 2024; Speelman and McGann, 2013). Therefore, we propose shifting from single-point averages to distributional methods. These include methods managing several location measures of each group (for a detailed discussion, see (Sanchis-Segura and Wilcox, 2024)) but also methods estimating the probability of events within group members (see next section). This move is not merely technical; it is both philosophical and pragmatic, yielding insights about individuals, not idealized abstractions, and making findings truly actionable for personalized healthcare and policy.

## 3. Toward better approaches: Many paths ahead and one step forward

"The problem of inductive inference has no single best solution—it has many good solutions. Statistical thinking involves analyzing the problem at hand and then selecting the best tool in the statistical toolbox or even constructing such a tool. No tool is best for all problems".

G. Gigerenzer et al., 2004 (Gigerenzer et al., 2004).

The previous section concluded by emphasizing the need for the "sex differences field" to embrace several philosophical shifts. The first is to move from treating methods as authoritative arbiters of truth to regarding them instead as tools that serve informed judgment. The second is to shift from seeking universal definitions to accepting that answers are contextual, plural, and revisitable. The third is to move beyond the limitations of approaches grounded in NHST and averages, developing methods better able to capture the complex patterns of sex-related variation in nature and, consequently, to support more robust knowledge and useful concepts.

This is an open landscape, and different methodological routes can be seen stretching ahead (for overviews, see (Wasserstein et al., 2019; Harlow, 2014; Kline, 2013b)). Some remain as close as possible to the familiar NHST road but recommending better practices (Hurlbert and Lombardi, 2009; Simonsohn et al., 2014) or modest revisions to its parameters. For instance, pre-registration, increasing transparency, and lowering the significance threshold from 0.05 to 0.005 have been proposed to reduce false positive findings and improve replicability (Johnson, 2013; Johnson, 2019; Benjamin et al., 2018). However, these conservative suggestions preserve the rigid binary logic of significance testing and its unrealistic nil null hypothesis (i.e., the hypothesis of zero difference or zero effect). They, therefore, would not help to provide more conceptually satisfying or practically relevant accounts of sex

differences. Moreover, making the significance criterion more stringent would very much reduce statistical power for typical sample sizes in experimental research (such as animal studies), increasing the risk of missing smaller, but potentially meaningful effects. In contrast, for large observational or survey datasets often used in human sex differences research, this change may yield little real improvement. Consequently, these proposals appear more as useful complements than genuine alternatives (McShane et al., 2019; Betensky, 2019; Gelman and Robert, 2014)).

More radically, some scholars argue that NHST cannot be fundamentally repaired, and that no amount of guidance will prevent the widespread misinterpretation of *p*-values. They propose moving much farther afield, to the Bayesian world (Clayton, 2022; Wagenmakers, 2007). This is an entirely different framework. Here, probability is not the long-run frequency of outcomes but a measure of belief that departs from specification of prior knowledge or expectancies and that gets updated as evidence accumulates (Kruschke, 2013; Wagenmakers et al., 2008). For sex differences research, this opens the door to asking how probable it is that an effect lies within a range of practical relevance, shifting focus from arbitrary significance thresholds to the plausibility of effects of a given magnitude.[3]

However, adopting Bayesian methods implies entering into an entirely new terrain that brings its own demands: redesigning the standard statistical curriculum, specifying priors grounded in substantive knowledge, retraining researchers to interpret probabilistic statements, and developing computational tools that are not yet widespread. Despite these challenges, Bayesian approaches are becoming increasingly accessible and offer a richer, more direct way to quantify and communicate uncertainty—a property particularly valuable in fields (like "sex differences research") where the magnitude and meaning of observed differences are at the core of the scientific question.

Between these extremes, there is a wide territory still to be explored. Proposals in this middle ground draw from both frequentist and Bayesian traditions, often seeking to combine the best of both worlds. These include second-generation *p*-values (Blume et al., 2019) and other tests for interval hypotheses (Wellek, 2010), confidence intervals for effect sizes coupled with "estimation/meta-analytical thinking" (Cumming, 2012), statistical decision theory (Berger, 1993), likelihoods (Royall, 2000), Bayes factor functions (Johnson et al., 2023), bootstrap-based procedures (Efron and Tibshirani, 1998; Efron, 2005), and many others.

While this intermediate territory offers many valuable tools, a full exploration is beyond the scope of this paper. Therefore, we will focus on one method that seems particularly well-suited to the specific challenges of sex-difference research: Thresholded Probability of Superiority (TPS). TPS is not a panacea, but is a powerful example of how to operationalize the ATOM principles in this context. It is especially suited for this discussion because it treats groups as distributions of individuals —then avoiding the 'hourglass fallacy'—and requires researchers to specify in advance the minimum effect size that they would deem as relevant—then compelling thoughtful reflection and openness. Most importantly, TPS reframes sex differences as distributions of potential values—rather than as fixed facts—thereby acknowledging and quantifying uncertainty, a necessary step toward greater scientific modesty. In this way, it helps shift the field from merely identifying "significant" sex differences to describing informative and actionable ones.

In the next two sections, we introduce the philosophy and operation of TPS (3.1) and illustrate its advantages over traditional average comparisons with a worked example (3.2). Some readers may find it easier to read these sections in the order presented, while others may prefer to reverse the order.

### 3.1. Introducing thresholded probability of superiority-based methods

"When asked what question a study is supposed to answer, an investigator very often answers, 'to see if people in this group (or in this condition) score higher than they do in that.' If that response is pursed by the query 'Would you like to quantify the extent to which that is true?' The answer is often an enthusiastic yes. […] It is quite difficult for mean-comparison methods to provide quantification of the sort which is needed to answer the research question in a quantified way".

*Cliff, N (1996,* 60].

"If the proponents of a psychological theory propose neither a minimum effect size consistent with their theory, nor a maximum difference from their original finding that would cause them to question it, then it is not clear what could disprove their hypothesis. This is perhaps the greatest problem that faces psychological science, and one that has received almost no attention in recent discussions about better research practices."

*Morey & Lakens (2016,* 61].

Thresholded Probability of Superiority (TPS) is one framework that can be applied to many scenarios involving two groups,[4] but that was originally conceived as one possible way to incorporate ATOM principles in the study of sex-related variation. It redirects attention from statistical significance toward informativeness, and from abstraction and overgeneralization toward individual variation and actionability. Specifically, TPS replaces the abstract question posed by classical procedures like the *t*-test—*"Is the difference between the group means exactly zero?"*—with a simpler, more informative and actionable one centered on individuals: *"If we randomly select a male and a female, how likely is it that they will differ meaningfully?"*

As this reframing illustrates, TPS is very different from classic *t*-tests

[3] Bayesian methods offer what many researchers mistakenly expect from *p*-values: an estimate of how likely their hypotheses are to be true (Cohen, 1994; Gigerenzer et al., 2004; Wagenmakers, 2007; Kruschke, 2013; Wagenmakers et al., 2008; Colquhoun, 2019). This widespread misinterpretation of the *p*-values—sometimes described as researchers' "Bayesian Id's wishful thinking" (Cohen, *1994;* Haller et al., *2002;* Lecoutre et al., *2003;* Greenland et al., *2016;* Cumming, *2012)* —reflects a reversed understanding of the relationship between data and hypotheses. A *p*-value expresses the probability of the observed data (or more extreme data) assuming the null hypothesis is true. By contrast, Bayesian methods yield posterior probabilities about competing hypotheses being true given the observed data and prior information. This makes it possible to directly express how probable specific effect sizes are. As will be discussed later, although TPS is rooted in a frequentist framework, it estimates the probability of a difference being at least as large as a user-defined threshold and thus serves as a reasonable surrogate.

[4] Here the interest is in comparing sex-related categories. The same procedure can be extended to any other two independent groups. Moreover, one of the authors (RW) has developed R functions that allow to extend comparisons to related groups and to two repeated measures of a single group of individuals. These functions are available in the *Rallfun-v45* file and can be downloaded from https://osf.io/xhe8u/. For an example of the application of these functions in clinical research conducted by CF and CSS, see (Zaragoza-Mezquita et al., 2025).

On the other hand, TPS does not allow comparisons of more than two groups and is therefore unsuitable for analyzing classical 2 × 2 designs that involve two between-group factors (e.g., sex by treatment) as a whole. However, TPS can still be applied to all dyadic comparisons arising from such designs, much like conventional post hoc tests, and the resulting effects roughly compared by eye through their confidence intervals (Cumming and Finch, 2005) without falling into the so-called "difference in sex-specific significance" error (that involves comparing *p*-values of independent comparisons; see (Gelman and Stern, 2006; Rich-Edwards and Maney, 2023)). In 2 × 2 designs with repeated measures (e.g., sex × time), the difference between the two time points can first be computed for each individual, and the resulting scores subsequently analyzed using TPS.

**Table 1**
A comparison of Student's *t*-test, Cliff's delta, and TPS.

| Feature | Student's t-test | Cliff's delta | TPS |
|---|---|---|---|
| Nature | Parametric, | Nonparametric, | Nonparametric, |
| Calculated from | Marginal distributions (the distribution of scores of each group) are used to calculate averages | Pair-wise differences are calculated by subtracting one value in a pair from the other, so for each possible pair of values in a dataset). Frequencies within this differences' distribution is used | Pair-wise differences are calculated by subtracting one value in a pair from the other, so for each possible pair of values in a dataset). Frequencies within this differences' distribution is used |
| Assumptions | Requires normality and equal variances (or adjustments needed) | None for effect size; inference methods may assume normality | None |
| What is compared | Group means | Probability of *any difference* between randomly paired individuals | Probability of meaningful differences between randomly chosen pairs of individuals |
| Thresholding differences | No; significance based on a nil null hypothesis | No; all differences equally counted regardless of size | Explicit use of one or more thresholds to exclude trivial differences |
| Null hypothesis tested | $H_o$: $\mu_1 - \mu_2 = 0$ | $H_o$: $\delta = 0$, where $\delta = P(X > Y) - P(Y > X)$ | $H_o$: $\omega_1 - \omega_2 = 0$ where $\omega_1 = P(X - Y \geq SESOI)$, $\omega_2 = P(Y - X \geq SESOI)$ |
| Effect size | Not inherently included (e.g., magnitude, mainly Cohen's *d* often added) | Probabilistic | Probabilistic and magnitude-based |

but closely related to Cliff's delta[5] (Cliff, 1996; Cliff, 1993) (see also Table 1). Thus, in contrast with t-tests, which rely on group means, TPS evaluates the full distribution of all possible pairwise comparisons between two groups. Moreover, akin to Cliff's delta (Cliff, 1996; Cliff, 1993), TPS is both an effect size and a non-parametric test for significance.

At its core, TPS is a probabilistic effect size that quantifies how often members of one group (e.g., males; M) score higher than members of another group (e.g., females; F); for a technical discussion, see method T in (Wilcox and Sanchis-Segura, 2025). Thus, TPS shares with other effect size indices, such as Cliff's delta (Cliff, 1993) or the Common Language Effect Size (CLES; (McGraw and Wong, 1992)), an individual-centered perspective that moves beyond potentially misleading aggregate measures such as means. These metrics describe group differences as the probability that a randomly selected individual from one group will have a larger outcome than a randomly selected individual from the other — represented as P(M > F) or P(F > M).

However, TPS differs from these and other probabilistic effect sizes by explicitly incorporating one or more predefined effect thresholds, known as the smallest effect sizes of interest (SESOI) (Riesthuis, 2024; Lakens et al., 2018). This feature allows TPS to evaluate whether one group tends to score higher than another only when the observed difference exceeds a meaningful, predefined magnitude. Formally, TPS estimates the probability that the difference between the scores of two randomly selected individuals from independent groups (e.g., males and females) will be equal to or greater than the SESOI; that is, P(M − F ≥ SESOI) or P(F − M ≥ SESOI), depending on the direction of comparison.

This threshold-based approach is consistent with practices in equivalence testing (Wellek, 2010; Morey and Lakens, 2016; Riesthuis, 2024; Lakens et al., 2018; Riesthuis and Cribbie, 2025) and directly aligns with several calls for "being thoughtful" in statistical practice (Wasserstein et al., 2019; Betensky, 2019; Blume et al., 2019; Wellek, 2010; Riesthuis, 2024; Lakens et al., 2018). Moreover, it ensures true informativeness—in Bateson's words, "*information is a difference that makes a difference*" (Bateson, 1999) — and, combined with TPS intrinsic probabilistic perspective, helps to avoid common interpretative biases:

-avoids abstraction by quantifying differences between individuals rather than between summary statistics.

- avoids overgeneralization by expressing the effect of interest as an estimated probability, not a universal claim.

-Ensures meaningful interpretation by using a threshold that verifies whether the observed trend toward higher scores in a group (or, more technically, its stochastic dominance) reflects a substantively important difference rather than a trivial one.

However, TPS is more than an effect size. It also functions as a test for both differences and similarities between two groups.

When testing for a difference, TPS compares the probability that the score of a randomly chosen male exceeds that of a randomly chosen female by the SESOI value or more — P (M-F ≥ SESOI) or $\omega_1$[6] — to the probability that the score of a randomly chosen female exceeds that of a randomly chosen male by the SESOI value or more — P (F-M ≥ SESOI) or $\omega_2$. In other words, it evaluates the null hypothesis of no difference between two probabilities ($H_o$: $\omega_1 - \omega_2 = 0$; see Table 1) . By explicitly calculating these two probabilities TPS highlights that, even when the average of one group exceeds that of the other by a large amount, some individuals in the lower-scoring group surpass some individuals in the higher-scoring group. Comparing these probabilities enables a rich and balanced inference about how the members of these groups differ at the individual rather than the average level.

Crucially, TPS also tests for similarity through equivalence testing procedures (Wellek, 2010; Morey and Lakens, 2016; Riesthuis, 2024; Lakens et al., 2018; Riesthuis and Cribbie, 2025) (see Fig. 5). Formally, this involves pre-defining a desired equivalence margin (Δ), and testing whether the Confidence Interval for observed difference ($\omega_1 - \omega_2$) is fully contained within the predefined equivalence bounds [−Δ, +Δ]. This dual testing framework (differences and equivalences) allows researchers to validate the absence of a meaningful difference with the same rigor applied to finding one, directly countering publication bias.

Of note, TPS was developed to accommodate several SESOIs (Wilcox and Sanchis-Segura, 2025). When doing so, TPS develops its full potential and allows researchers to estimate several effects or test several hypotheses in sequence. This flexibility responds to calls from critics of NHST (Gigerenzer et al., 2004; McShane et al., 2019; Blume et al., 2019) for greater transparency and richer inferential practice. The use of multiple thresholds results in highly informative graphical displays of

[5] Cliff's Delta is primarily a non-parametric effect size measure that quantifies the dominance or probability that a randomly chosen value from one group will be larger than a randomly chosen value from the other, minus the reverse probability (therefore, ranging between −1 and 1).Cliff's Delta can also be used as a statistical test for stochastic equivalence between two distributions. This means it tests whether one distribution tends to have larger values than the other, even if their medians or means are equal. The test aspect of Cliff's Delta is closely related to the Mann-Whitney *U* test and can be derived from its statistic. It provides a way to quantify not just whether there is a difference, but the magnitude and direction of that difference without any distributional assumption (Cliff, 1996).

[6] Here we use the same nomenclature as in our original publication describing these methods (Wilcox and Sanchis-Segura, 2025). However, as we show in the next section, it might be more practical the use labels stating "TPS", followed by the name of the dominant group, and the SESOI magnitude. Thus, for example, to refer to the probability of a randomly chosen female exceeding the scores of randomly chosen male by 3 or more points, one can use TPSf≥3.

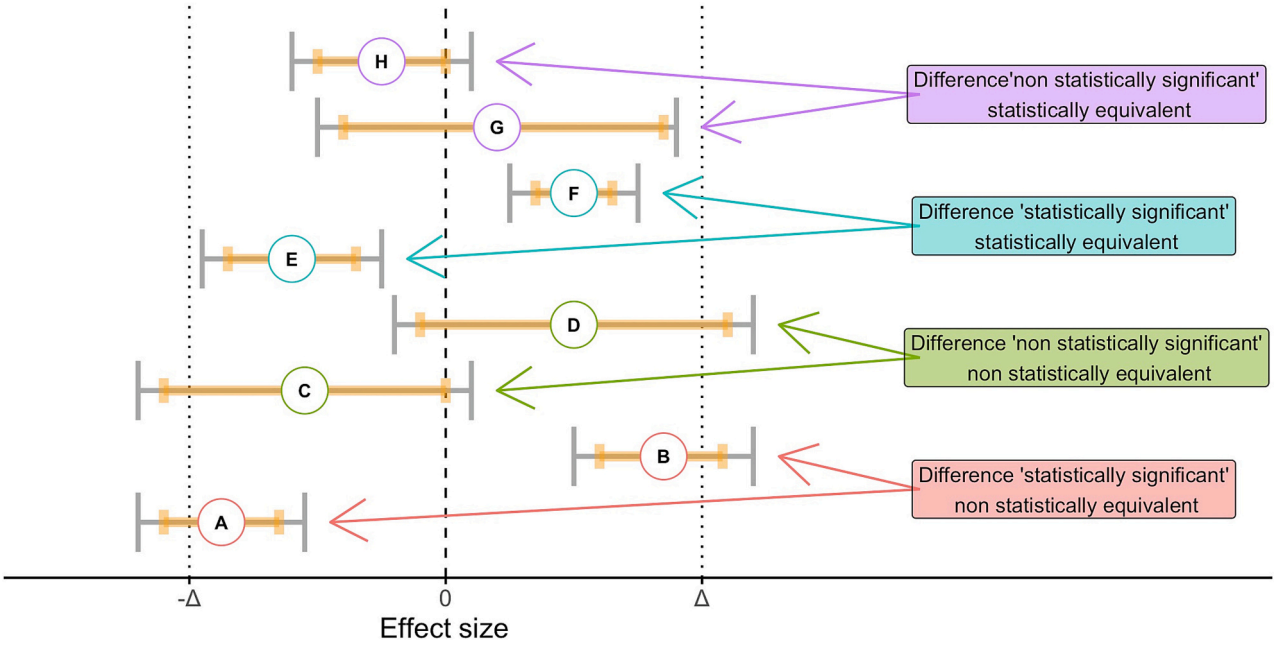



**Fig. 5.** Equivalence and difference testing. Horizontal T-shaped bars represent the 95% (gray) and 90% (orange) confidence interval (CI) for theoretical effect sizes and together they illustrate the four different scenarios that can be found when testing differences with NHST and equivalence with the Two One-Sided Tests (TOST) procedure (for a detailed tutorial, see (Lakens et al., 2018)). Traditional tests of differences (NHST) evaluate a *nil* null hypothesis stating that the effect size of the statistic of interest (e.g., difference between-means, probability of success, etc.) is zero; This null hypothesis is rejected when the 95% CI for delta excludes zero, or equivalently when the *p*-value is below 0.05 (as it occurs for the effects A, B, E, and F). For testing equivalence, a reversal of the traditional null and alternative hypotheses is required. In equivalence testing, the null hypothesis states that the effect size falls outside the predefined equivalence bounds ([−Δ, Δ], dotted lines) around zero; this null hypothesis is rejected (and equivalence, statistically confirmed) when the entire 90% CI for the observed effect size lies within these bounds [−Δ, Δ] (as it occurs for the effects E, F, G, and H), thus indicating the difference is practically negligible or "not meaningfully different from zero." The change in the level of confidence in the intervals between NHST and TOST (95% vs. 90%, respectively) is due to the fact TOST consists of two one-sided hypothesis tests, each with its own *p*-value (with the larger p-value one governing the overall conclusion of the procedure). Note that although both they are ordinarily implemented using bright-line thresholds to judge *p*-values and make dichotomous decisions, nothing impedes to interpret the same p-values as a continuous index quantifying the extent to which it is reasonable to make a decision about the direction of an effect and/or transform them in "surprise-values" (see Table 2). Note also that, because we subscribe the recommendation of avoiding entirely the distinction between "significant" and "non-significant" effects, we use these terms between quotes and just for illustrative purposes.

between-group differences. These are two strengths that TPS shares with another distributional method, the shift function (Wilcox and Rousselet, 2023; Wilcox, 2006), which we have previously advocated (Sanchis-Segura and Wilcox, 2024) . However, the shift function operates directly on marginal distributions, whereas TPS draws on the distribution of all possible pairwise differences. The two methods address distinct questions—the former concerning the comparison of individuals across groups and the second the distribution of their differences—thus providing complementary perspectives about between group variation (on these distinctions, see (Wilcox and Rousselet, 2023; Rousselet et al., 2017)).

To quantify the evidence for all these richer inferences, TPS uses robust bootstrap-based *p*-values. These p-values are not dichotomized to make a simple yes/no decision; instead, they are treated as continuous measures of evidence, where they serve to quantify the degree of incompatibility between the observed data and the null hypothesis. Thus, the closer the p-value is to zero, the stronger the empirical evidence supporting a decision about the direction of an effect. This approach is rooted in Tukey's three-decision rule (Rice and Krakauer, 2023; Jones and Tukey, 2000; Wilcox and Serang, 2017) and aligned with the ASA's official consensus against mechanical thresholds (Wasserstein and Lazar, 2016). Moreover, to enhance interpretability, these *p*-values are transformed into "surprise values" or "Shannon information values" (Greenland, 2019) (see Table 2).

Taken together, these features give TPS three conceptual advances that directly respond to long-standing problems in sex difference research.

- *First*, by requiring explicitly defined SESOIs or thresholds, TPS avoids the "magnitude fallacy," bypasses reliance on the nil null hypothesis" (i.e., the hypothesis of zero difference or zero effect), and compels transparent justification of which differences count as "sex differences" and why.
- *Second*, by focusing on the expected frequency of meaningful differences between pairs of individuals, TPS resists generalizing mean differences to all individuals, and, therefore, preserves within group variation and prevents essentialist views of sex-based categories.
- *Third*, TPS integrates description, estimation, and hypothesis testing: as an effect size, it merges magnitude-based and probabilistic perspectives; as a test, it enables explicit evaluation of both sex differences and sex equivalences (Fig. 5) without dichotomizing evidence.

In summary, TPS allows researchers to explicitly define what constitutes a sex difference—the effect warranting attention in their study—and estimates its expected frequency in randomly encountered individuals. This stands in contrast to traditional methods, which let sample size inadvertently determine what is labeled as a sex difference and yield inferences about average differences and through probabilities (*p*-values) that reference not the effect itself, but outcomes from a theoretical set of replications that were never performed. In the following section, we delve deeper into TPS and illustrate its practical and conceptual advances using an example.

### 3.2. *What TPS can provide: A worked example*

To illustrate TPS in practice, we apply it to a worked example using

**Table 2**
Surprise-values: What they are and how they improve inferences based on *p*-values.

| What are *P*-values | What are S-values |
|---|---|
| P-values are conditional probabilities. Although often forgotten, these probabilities refer to a frequentist context of repeated sampling, not to the results of a single experiment. Thus: | S-values are transformations of p-values. Specifically: $S = -\log2(p\text{-value})$ |
| A *p*-value is the probability (expected frequency) of obtaining a result (e.g., a between means difference) as large or more than the one actually observed in a particular study *if the null hypothesis were true and the experiment would be replicated many times under the same conditions.* | S-values quantify the amount of information provided by the data against the whole testing model (null hypothesis and the test assumptions) in bits. This can be interpreted as the number of fair coin tosses in a row one would need to observe for an equally surprising event to occur under the null hypothesis. For instance, a surprise value of 4 bits is as surprising as getting four heads in a row with a fair coin, while a surprise value of 20 is as surprising as getting 20 heads in a row. |
| **(some) Problems of P-values** | **S-values Solutions** |
| P-values are often misinterpreted as indicative of the probability that the observed result is due to chance alone | S-values are not probabilities and, because they reach values higher than 1, they cannot be confused with any measure of chance. |
| P-values are ordinarily dichotomized using arbitrary thresholds like 0.05, leading to misleading yes/ no answers. | S-values have an intuitive, continuous interpretation as information measured in bits— "equivalent to n heads in coin tosses." Unlike *p*-values, they do not rely on cut-offs; however, to avoid biased inferences, it may be useful to consider *in advance* how many "heads" one would require to make a decision, depending on its consequences. |
| P-values are often misinterpreted as evidence measures. However, they are probabilities, and they do not measure evidence for alternatives, just compatibility with the null hypothesis. | S-values quantify the amount of refutational evidence against the whole model (the null hypothesis and the test assumptions), allowing us to quantify how surprising the result is if the null was true and the test assumptions satisfied. |
| Because of their low values, typical thresholds (e.g., $p < 0.05$) create a false feeling of certainty when rejecting the null. | When transformed to S-value, it is clear the amount of refutational information provided by p-values close to the conventional threshold of 0.05 is actually low (~ 4.3 bits; as "surprising" as observing 4 consecutive heads when tossing a fair coin) |
| P-values cannot be linearly interpreted. | Are linearly interpretable. |
| A p-value of 0.05 and 0.01 do not represent the same incremental change of incompatibility with the null as moving from 0.10 to 0.05. | Equal intervals in S-value represent equal changes in the refutational evidence against the null hypothesis |

BMI data from 800 individuals (400 males, 400 females).[7] For this purpose, we also use the dedicated R functions developed to implement TPS methods that are explained further in the supplementary material.

If analyzed following standard practice (Weissgerber et al., 2018), these data would be subjected to a Student's *t*-test for independent samples. The group means here are 26.81 for males and 26.19 for females, and with the sample size used the small difference between them does not achieve statistical significance [$t(798) = 1.75$, $p = 0.080$]. By conventional NHST reasoning, this result would be often dismissed with a statement such as: *"Males and females do not differ in BMI"* (see (Sanchis-Segura and Wilcox, 2024; Hoekstra et al., 2006)). However, this conclusion is wrong because: 1) males and females were not compared (only their means were); 2) looking at the mean values one can see that their difference is small, but not zero; and, 3) *p*-values larger than 0.05 do not mean "no difference" but "no sufficient information to make a yes/no decision". Moreover, looking at Fig. 3, it can be readily seen that BMI sco res are not normally distributed and that groups differ in spread and skewness. Consequently, means do not properly summarize group scores and data should not be analyzed using a *t*-test.

TPS does not use summary statistics nor is it affected by any sample characteristics, so its use is correct in any distributional scenario. Moreover, it shows adequate performance even when applied to relatively small samples ($n = 20$ per group) and regardless of whether the data include tied values (see the performance of methods T and R in Section 3 of (Wilcox and Sanchis-Segura, 2025)). Yet, it requires researchers to specify in advance what constitutes a threshold of practical or theoretical relevance—and by extension, what should count as a "sex difference" and why.

Accustomed to the supposed objectivity of NHST, this requirement may feel uncomfortable: whatever threshold is chosen will inevitably reflect a subjective judgment. Yet this is precisely one of TPS's conceptual strengths. It compels researchers to abandon the illusion that statistical tools can substitute for scientific reasoning, and instead to acknowledge, justify, and communicate the rationale behind analytic choices. Far from being a drawback, this reflects —and requires!— the ATOM principles. Specifically, TPS accepts uncertainty from the outset, since no threshold is self-evident; necessitates thoughtful deliberation about what size of difference matters; promotes openness by requiring transparency about threshold choice; and cultivates modesty, since conclusions must recognize the value-laden character of these decisions. In contrast to the automatic, ritualized logic of NHST, TPS therefore encourages researchers to take responsibility for their inferential choices, yielding richer and more honest accounts of sex-related variation.

There are several strategies for defining a threshold or SESOI (for a discussion and examples, see (Lakens et al., 2018)). Researchers may draw from prior literature, adopt a proportion of the maximum possible score, or—if no field-specific guidance exists—use quantile shifts or convert conventional benchmarks ("small"/ "medium"/ "large") of standardized effects into raw-score thresholds. Some may resist this, arguing that in certain areas—especially in very new research—it is "impossible" to propose a threshold. Yet if no plausible SESOI can be specified, one must also admit that there is no clear basis for any substantive interpretation of findings, regardless of statistical outcome (Morey and Lakens, 2016). Under these conditions, the research is, at best, merely descriptive and lacks a testable hypothesis, thereby invalidating the use of hypothesis testing and any other inferential method.

However, paradoxical as it may seem, the easiest—and probably the best—way to choose a threshold is not to choose one at all. Instead, it is often more effective to define a reasonable range of potential SESOIs and calculate TPS across them. When paired with appropriate graphical displays, this approach yields a genuine map of effects probabilities that allows readers to formulate their own questions about the data.

To illustrate TPS in practice, the primary SESOI is set at 5 BMI units, and then extended to a range spanning 1–10 BMI units (four below and five above the reference). This choice is evidence-based: meta-analyses identify a 5-unit BMI difference as clinically meaningful, marking a point where BMI variation has tangible health implications (Jayedi et al., 2022; Di Angelantonio et al., 2016). Moreover, given the value of the pooled standard deviation of our sample (5.04), the lower bound of the selected range (one BMI unit), corresponds to a conventional

[7] These data were extracted form the 1200 Subject Release of the Human Connectome Project, and they are exactly the same employed in a previous publication of two of the authors (CSS and RW, see (Sanchis-Segura and Wilcox, 2024)). This allows fully illustrating how the information provided by the TPS is coherent and complementary to that of the analyses performed in this previous publication using the shift-function.

"small" statistical effect (Cohen's $d \approx 0.2$).[8] Correspondingly, the largest threshold (10 BMI units) would quantify a "very large effect".

Regarding results (Fig. 6 and Table 3), meaningful differences of five or more BMI units are expected in about 45–46% of male–female pairs. While the point estimates suggest these differences are more likely to favor males (26%) than females (19%), acknowledging uncertainty in both these estimates and their difference is crucial. The estimated gap favoring males is 7%, but the 95% confidence interval is wide, ranging from nearly zero (0.01%) to as high as 13%. This finding is key: while the data point toward a small male advantage in the frequency of clinically large BMIs, the considerable uncertainty indicates we cannot confidently rule out that the true difference is negligible. Indeed, the formal test result is below the conventional $p$-values threshold ($p = 0.02$, or $p = 0.044$ when adjusted for multiple comparisons), but corresponds to an S-value of only ~5.5 (i.e., the result is as surprising as getting 5 or 6 heads in a row with a fair coin), indicating just moderate evidence against the null hypothesis of no difference.

At the less stringent SESOI of 1 BMI unit, differences would be observed in about 88% of male–female pairs, and in roughly 58% of these cases (51% of the total), the male will have the higher BMI (TPSm$\geq$1 = 0.51 vs. TPSf$\geq$1 = 0.37, $p < 0.001$, S-value>10). Conversely, at the most stringent SESOI (10 BMI points), differences would be observed in just 16% of male-female pairs, the difference between TPSf$\geq$10 (0.080) and TPSm$\geq$10 (0.079) would be virtually zero ($p = 0.946$, S-value<1) and, indeed, between groups "equivalence" could be statistically substantiated within a 5% precision ($p = 0.002$, S-value = 9).

To better contextualize these results and extract appropriate conclusions, it is worth contrasting them with those drawn from classical (though often flawed) comparisons of means. In this regard, a $t$-test suggested that the small difference between means (0.66 BMI units) was "not significant" and thus would be often dismissed as "unimportant" or even lead to the incorrect conclusion that *"males and females do not differ in BMI"*. However, TPS reveals a more accurate perspective: *some* males and females do differ in their BMI scores. TPS also reveals that the chance of seeing such differences depends on the threshold used to define what counts as meaningful, but that they appear between *some* individuals even when using SESOIs of 1, 5, and 10 BMI points (that is, 1.5, 7.6, or 15.5 times greater than the difference between group averages, respectively). Thus, in *some* cases, sex differences in BMI favoring males are "statistically large" and may be clinically meaningful. Importantly, TPS also demonstrates that *for yet some other individuals*, equally meaningful differences may occur in the opposite direction and, by quantifying the risk associated with each scenario, enables recipients of these results to make informed judgments and interpretations.

These results illustrate what TPS uniquely provides: it allows researchers to define which differences truly matter and to quantify how often such differences occur between individuals. In practical terms, TPS enables researchers to answer the question most relevant for applied disciplines—such as psychology, education, or medicine—namely, how frequently a difference of meaningful size occurs between members of two groups chosen at random. This approach bridges the gap between abstract group statistics and the variability experienced in real populations, making TPS results directly useful for real-world decisions in ways that traditional, average-based methods cannot achieve.

The limitations of average comparisons (and the advantages of TPS) are not restricted to this example or to situations in which classical tests "fail" to identify a between group difference. To illustrate this fact, we employ the data of the heights and weights of the same individuals included in our BMI sample, which show sex-related differences favoring males that are "statistically significant" and "large" [$t(798) = 28.37$, $p < 0.001$,Cohen's d = 2.00; $t(798) = 13.24$, $p < 0.001$,Cohen's d = 0.93, respectively]. The results of these $t$-tests would often lead to the conclusion that *"males and females significantly differ in height and weight"*, and the observed between means differences would be implicitly interpreted as *if appliable to/ representative of* those existing between *all* males and *all* females (see Fig. 4). However, as the results displayed in Table 4 reveal, this conclusion is inaccurate and misrepresents the existing differences among the males and females of this sample. In fact, even when using a quite liberal criterion (± 10%) to define differences "similar" in size and direction to those existing between the group averages in height and weight (14.25 cm and 14.95 kg), such "similar-to-average differences" (ranging between 12.83 and 15.68 cm and 13.46–16.45 kg) are expected to be observed in just ~10 and 6% of the cases, respectively. Differences are expected to be smaller in ~49 and 46% of the cases and larger in ~39 and 22% of the cases, respectively. Moreover, differences with opposite direction —with females showing higher height or weight than males— are expected in the 6 and 23% of the cases, respectively.

Together, these results and observations make apparent that relying solely on average differences provides an incomplete and sometimes misleading picture. For many individuals, conclusions based on group averages are inaccurate or even incorrect. This is a particularly critical issue in applied contexts, as when studying sex differences to advance personalized medicine. After all, no healthcare provider treats averages—only individual patients walk into clinics. It is for these individuals that diagnoses and treatments must be tailored, making individual-level inference essential.

In summary, and as just a quick inspection of Figs. 3 and 6 requires, TPS provides more accurate, nuanced, and actionable information than traditional approaches based on averages. In addition, TPS encourages researchers to define expectations thoughtfully, communicate them openly, acknowledge and quantify uncertainty, and interpret results with appropriate modesty. In this way, it enables a more accurate and honest account of sex variation than that derived from current practices and can make the "sex differences field" better approach its main goals (see next section).

## 4. Conclusion

«Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise».

John Tukey. The future of data analysis (1962).

If research on sex differences is to contribute meaningfully to precision medicine while avoiding the reinforcement of stereotypes, some rethinking of current practices in this research field seems necessary. In particular, three considerations appear especially important. First, acknowledge that, as an object of study, sex differences are statistically-defined constructs, not natural objects. Second, defining them as currently done (i.e., through significance thresholds and limited to averages) is conceptually limited, potentially misleading, and, for some individuals, even counterproductive. Third, there is no single replacement for current definitions and practices, just trial-and-error inspired by the ATOM principles (Wasserstein et al., 2019) and/or other potential sources of transparency, rigor, and modesty in statistical inference (e.g., (Kline, 2013a; Gigerenzer et al., 2004; McShane et al., 2019)).

The kind of reorientation envisaged here is less about substituting one dogma for another than about adopting a pragmatic stance at both the conceptual and methodological levels. From this perspective, there may be no single, universally satisfactory definition of what qualifies as a meaningful sex difference. Instead, in line with the ATOM principles, researchers are encouraged to aprioristically and explicitly define what they consider a sex difference in the specific context of their study and why. Framed in this way, differences should be regarded as useful—that

[8] We agree with Lakens et al. (2018) when saying that relying on a benchmark is the weakest possible justification of a SESOI and should be avoided (Lakens et al., 2018). However, in this case, it is convenient as it helps to contrast its effects with the results obtained when using SESOIs that could be of clinical relevance (5 or more BMI points).

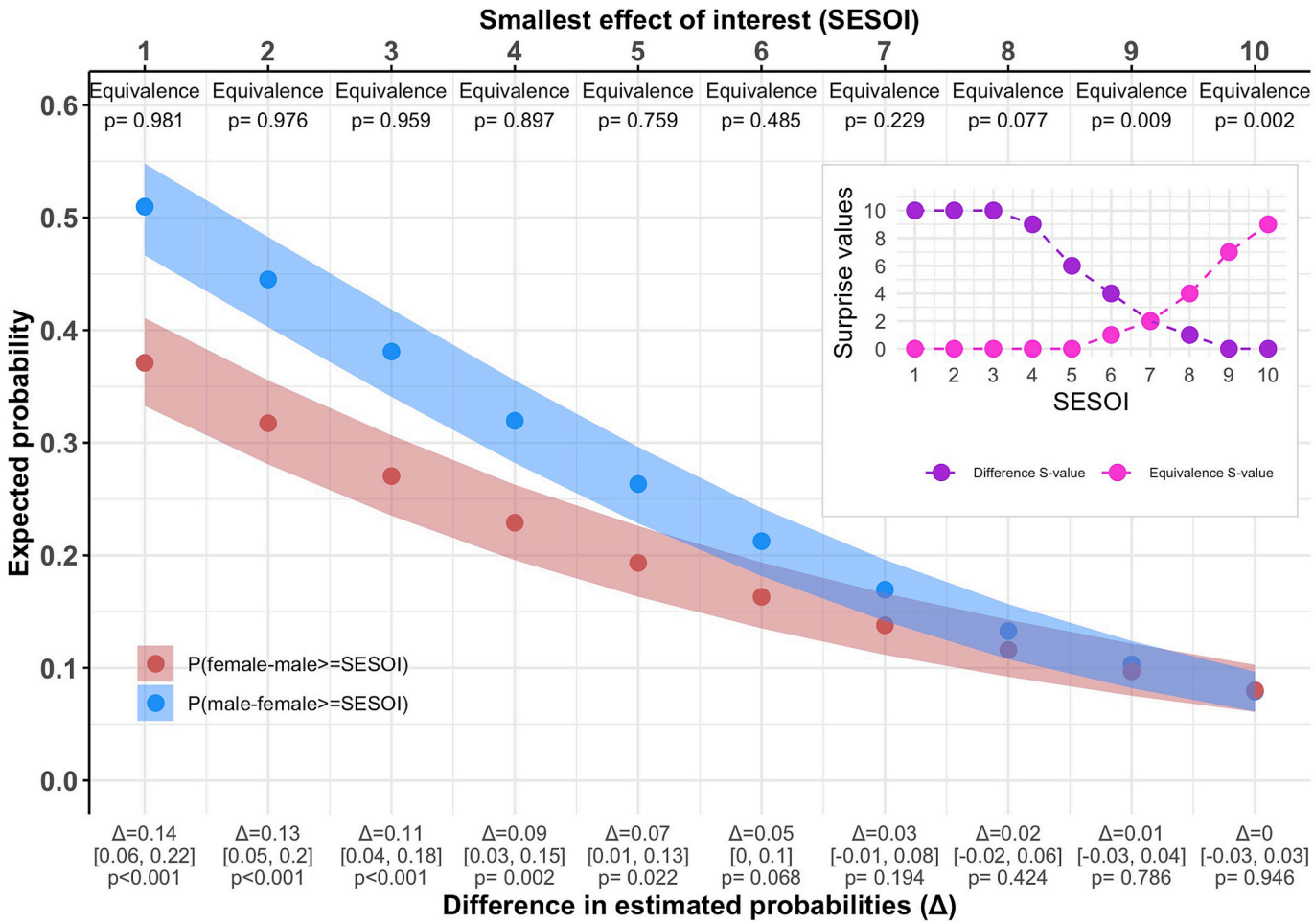



**Fig. 6.** TPS analysis of the sex differences and equivalences in BMI. The main panel depicts the estimates (points) and 95% confidence intervals (CIs; shaded bands) of the probability of finding a difference larger than several possible threshold values (Smallest Effect Size of Interest; SESOI, top X axis) favoring females (red) or males (blue). Above each SESOI value on the top X axis, the *p*-values associated with equivalence tests based on TOST (equivalence bounds [−0.05, 0.05]) are provided. Along the bottom X axis, the values of the groups' differences in the estimated probabilities, their 95% CIs, and p-values for comparisons at each threshold are shown. The inset panel depicts the S-values ("surprise values" (Greenland, 2019)) derived from the p-values obtained in the between-group comparisons ("difference S-value", purple) and in the equivalence tests ("equivalence S-value", magenta).

**Table 3**
Numerical results of the TPS analysis of the sex differences and equivalences in BMI.

| | | | | Difference | | | | | | Equivalence (equivalence bound ±0.05) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SESOI | Expected total frequency | TPSm | TPSf | Estimate | 95%CI | p-value | S-Value | Adjusted p-value | S-Value | p-value | S-value |
| 1 | 0.88 | 0.51 | 0.37 | 0.14 | [0.06, 0.22] | <0.001 | >10 | <0.001 | >10 | 0.981 | <1 |
| 2 | 0.72 | 0.46 | 0.32 | 0.13 | [0.05, 0.20] | <0.001 | >10 | <0.001 | >10 | 0.976 | <1 |
| 3 | 0.65 | 0.38 | 0.27 | 0.11 | [0.04,0.18] | <0.001 | >10 | <0.001 | >10 | 0.959 | <1 |
| 4 | 0.55 | 0.32 | 0.23 | 0.09 | [0.03, 0.15] | 0.002 | 9 | 0.005 | 8 | 0.897 | <1 |
| 5 | 0.45 | 0.26 | 0.19 | 0.07 | [0.01, 0.13] | 0.022 | 6 | 0.044 | 5 | 0.759 | <1 |
| 6 | 0.37 | 0.21 | 0.16 | 0.05 | [<−0.01, 0.1] | 0.068 | 4 | 0.113 | 3 | 0.485 | 1 |
| 7 | 0.31 | 0.17 | 0.14 | 0.03 | [−0.01, 0.07] | 0.194 | 2 | 0.277 | 2 | 0.229 | 2 |
| 8 | 0.24 | 0.13 | 0.11 | 0.02 | [−0.02, 0.06] | 0.424 | 1 | 0.530 | 1 | 0.077 | 4 |
| 9 | 0.20 | 0.10 | 0.98 | 0.02 | [−0.03, 0.04] | 0.786 | <1 | 0.873 | <1 | 0.009 | 7 |
| 10 | 0.016 | 0.08 | 0.08 | <0.01 | [−0.03, 0.03] | 0.964 | <1 | 0.946 | <1 | 0.002 | 9 |

The results displayed in the shadowed rows are discussed in the main text. (SESOI, smallest effect size of interest; expected total frequency refers to the total proportion of male-female pairs showing a difference as large or larger than the SESOI; TPSm, thresholded probability of superiority for males; TPSf, thresholded probability of superiority for females; CI, confidence interval; p-values adjusted for multiple comparisons were obtained with the Benjamini, Hochberg and Yekuteli method; precision of P-values is restricted to 0.001; S-values rounded to the nearest integer).

is, both informational, by ensuring that they reflect substantively relevant contrasts rather than trivialities, and actionable, by producing probabilistic and/or multiple claims grounded in individual variation rather than categorical averages. This orientation also underscores that statistical methods are intended to support scientific judgment rather than replace it and that, therefore, new methods need be adopted or developed.

In this methodological quest, there are many possible paths forward.

**Table 4**
Differences between group means do not necessarily reflect the differences that exist between randomly selected individuals.

| | Observed means' difference (MD) | MD boundaries criteria | MD bounds | P between MD bounds | C I 90% between MD bounds | P larger upper MD bound | CI 90% larger upper MD bound | P smaller lower MD bound | CI 90% smaller lower MD bound | P difference equals 0 | CI 90% equals 0 | P reverse direction | CI reverse direction |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BMI | 0.62 | MD 90% CI | [0.009, 1.198] | 0.07 | [0.069, 0.079] | 0.50 | [0.46, 0.54] | 0.000 | [−−] | 0.0009 | [NC, NC] | 0.43 | [0.39, 0.47] |
| | | MD ± 5% | [0.59, 0.65] | 0.004 | [0.003, 0.004] | 0.53 | [0.49, 0.57] | 0.036 | [0.033, 0.039] | 0.0009 | [NC, NC] | 0.43 | [0.39, 0.47] |
| | | MD ± 10% | [0.56, 0.68] | 0.01 | [0.007, 0.008] | 0.53 | [0.49, 0.57] | 0.034 | [0.032, 0.037] | 0.0009 | [NC, NC] | 0.43 | [0.39, 0.47] |
| | | MD ± 15% | [0.53, 0.72] | 0.01 | [0.011, 0.013] | 0.53 | [0.49, 0.57] | 0.032 | [0.03, 0.035] | 0.0009 | [NC, NC] | 0.43 | [0.39, 0.47] |
| Height* | 14.26 | MD 90% CI | [13.45, 15.05] | 0.000 | [−−] | 0.51 | [0.47, 0.55] | 0.39 | [0.36, 0.42] | 0.04 | [0.032, 0.044] | 0.06 | [0.047, 0.078] |
| | | MD ± 5% | [13.54, 14.97] | 0.000 | [−−] | 0.51 | [0.47, 0.55] | 0.39 | [0.36, 0.42] | 0.04 | [0.032, 0.044] | 0.06 | [0.047, 0.078] |
| | | MD ± 10% | [12.83, 15.68] | 0.10 | [0.09, 0.11] | 0.41 | [0.37, 0.45] | 0.39 | [0.36, 0.42] | 0.04 | [0.032, 0.044] | 0.06 | [0.047, 0.078] |
| | | MD ± 15% | [12.12, 16.39] | 0.20 | [0.19, 0.21] | 0.41 | [0.37, 0.45] | 0.29 | [0.27, 0.32] | 0.04 | [0.032, 0.044] | 0.06 | [0.047, 0.078] |
| Weight | 14.95 | MD 90% CI | [13.06, 16.78] | 0.08 | [0.072, 0.082] | 0.47 | [0.43, 0.51] | 0.22 | [0.2, 0.23] | 0.006 | [0.005, 0.007] | 0.23 | [0.2, 0.26] |
| | | MD ± 5% | [14.2, 15.7] | 0.03 | [0.024, 0.028] | 0.49 | [0.45, 0.54] | 0.24 | [0.22, 0.26] | 0.006 | [0.005, 0.007] | 0.23 | [0.2, 0.26] |
| | | MD ± 10% | [13.46, 16.45] | 0.06 | [0.058, 0.067] | 0.48 | [0.44, 0.52] | 0.22 | [0.21, 0.24] | 0.006 | [0.005, 0.007] | 0.23 | [0.2, 0.26] |
| | | MD ± 15% | [12.71, 17.19] | 0.08 | [0.075, 0.086] | 0.47 | [0.43, 0.51] | 0.22 | [0.2, 0.23] | 0.006 | [0.005, 0.007] | 0.23 | [0.2, 0.26] |

Differences between group means do not necessarily reflect the differences that exist between randomly selected individuals. The table presents point estimates (and 95% confidence intervals) of the probability that a randomly chosen male–female pair shows a difference "similar to" (shadowed column), larger, or smaller than that observed between group means in BMI, height, and weight of the males and females of our sample. Similarity was defined as a value falling within the boundaries of the 90% confidence interval of the observed mean difference (MD) or within the ranges of MD ± 5%, 10%, or 15% of the observed MD. For completeness, the table also reports the estimated probabilities of finding no difference (0) or a difference in the opposite direction (F > M) relative to the group means.

* Note that height data in the Human Connectome Project is rounded to the nearest interval, which introduces a discretization effect (i.e., the sample of 800 individuals contain only 22 unique values). This rounding may have affected estimations by producing more discrete, step-wise values with larger 'steps' rather than a smooth distribution.

In a previous piece (Sanchis-Segura and Wilcox, 2024) we highlighted one aimed to describe males and females as distributions, not as averages' incarnated Platonic ideals: The shift-function. Here we describe another, the so-called Thresholded Probability of Superiority (TPS), which more narrowly focuses on male-female differences, but also treating them as a distribution —instead of a single value.

As we see it, TPS offers three interrelated additional strengths over more commonly used methods. Specifically:

First, *it provides practical and clinical relevance*. The requirement to pre-specify a SESOI ensures that findings are substantively meaningful, compelling researchers to thoughtfully and openly justify what should count as a "sex difference" in their context. This focus on meaningfulness is complemented by the method's shift from abstract group averages to person-centered inferences. In doing so, TPS delivers more actionable, probabilistic insights that are directly relevant for fields like personalized medicine, where the individual—not the idealized group average—is the primary concern.

Second, *it provides a framework to "ATOMize" the field*. Building on the principles of being Thoughtful and Open by requiring a justified SESOI, TPS also directly operationalizes the need to Accept uncertainty. Its probabilistic nature quantifies the likelihood of meaningful differences rather than presenting a single, deterministic conclusion. This probabilistic framing inherently combats categorical thinking and, by clearly visualizing the overlap between groups, should promote Modesty in scientific claims. This modesty is further supported by the method's ability to statistically substantiate equivalence, helping to create a more complete and balanced scientific record.

Third, *it holds philosophical and heuristic value*. The method provides an intuitive, probabilistic measure of effect that, while firmly grounded in the frequentist framework, is reminiscent of Bayesian-style insights. In this way, TPS captures the very probability that researchers often seek—and mistakenly attribute to *p*-values (the so-called "Bayesian Id's wishful thinking" (Cohen, 1994; Haller et al., 2002; Lecoutre et al., 2003; Greenland et al., 2016; Cumming, 2012)). By providing a sound way to answer the question researchers often *want* to ask (Cohen, 1994), TPS is not only a useful tool in its own right but can also serve a pedagogical purpose, helping to prevent the widespread misuse of traditional p-values. In the specific case of "sex differences research" it also forces us, as researchers, to mindfully consider, define, and substantiate our very object of study.

Before finishing, it remains important to stress again that TPS is not the only possible route forward. It is not even a "route", but simply one method that illustrates how sex-related research may gain in clarity and relevance when ritualized practices and categorical formulations are abandoned. Ultimately, advancing this field will require embracing sex-related variation in its full complexity—plural, contextual, and revisitable. Studying such variation is not only about detecting or quantifying "differences," but about developing richer ways of describing, interpreting, and situating sex-related diversity. No single method can accomplish this on its own, and the required shifts are not just methodological, but conceptual and cultural. Yet, we hope that the method and the perspective outlined here can help the study of sex differences to

achieve its scientific and clinical goals.

## CRediT authorship contribution statement

**Carla Sanchis-Segura:** Writing – review & editing, Writing – original draft, Conceptualization. **Cristina Forn:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Rand R. Wilcox:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization.

## Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work one of the authors (CSS) used Gemini in order to improve the English in the initial versions of the manuscript. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.yfrne.2026.101257.

## Data availability

Data from the human connectome project are used in a example. An equivalent but self-constructed dataset is provided in the R markdown provided as supplementary material

## References


Amrhein, V., Trafimow, D., Greenland, S., 2019. Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication. Am. Stat. 73.
Anastasi, A., 1981. Sex differences: historical perspectives and methodological implications. Dev. Rev. 1 (3), 187–206.
Bateson, G., 1999. Steps to an Ecology of Mind : Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology. University of Chicago Press.
Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.J., Berk, R., et al., 2018. Redefine statistical significance. Nat. Hum. Behav. 2.
Benjamini, Y., De Veaux, R.D., Efron, B., Evans, S., Glickman, M., Graubard, B.I., et al., 2021. The ASA president's task force statement on statistical significance and replicability. Ann. Appl. Stat. 15.
Berger, J.O., 1993. Statistical Decision Theory and Bayesian Analysis. Springer-Verlag.
Betensky, R.A., 2019. The p-value requires context, not a threshold. Am. Stat. 73.
Blume, J.D., Greevy, R.A., Welty, V.F., Smith, J.R., Dupont, W.D., 2019. An introduction to second-generation p-values. Am. Stat. 73.
Carver, R.P., 1993. The case against statistical significance testing, revisited. J. Exp. Educ..61 (4),, 287–292.
Clayton, A., 2022. Bernoulli's Fallacy : Statistical Illogic and the Crisis of Modern Science. Columbia University Press.
Cliff, N., 1993. Dominance statistics: ordinal analyses to answer ordinal questions. Psychol. Bull..114 (3),, 494.
Cliff, N., 1996. Ordinal Methods for Behavioral Data Analysis. Erlbaum.
Cohen, J., 1994. The Earth Is Round ($p < .05$). Am. Psychol. 49 (12), 997–1003.
Colquhoun, D., 2019. The false positive risk: a proposal concerning what to do about p-values. Am. Stat. 73.
Cumming, G., 2012. Understanding the New Statistics: Effect Sizes, Confidence Intervals and Meta-Analysis, 1st ed. Routledge, New York.
Cumming, G., 2014. The new statistics: why and how. Psychol. Sci. 25 (1), 7–29.
Cumming, G., Finch, S., 2005. Inference by eye. confidence intervals how to read Pictures of data. Am. Psychol.60 60 (2), 170–180.
Delaney, H.D., Vargha, A., 2002. Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. Psychol. Methods 7.
Di Angelantonio, E., Bhupathiraju, S.N., Wormser, D., Gao, P., Kaptoge, S., de Gonzalez, A.B., et al., 2016. Body-mass index and all-cause mortality: individual-participant-data meta-analysis of 239 prospective studies in four continents. Lancet 388.
DuBois, L.Z., Kaiser Trujillo, A., McCarthy, M.M., 2025. Sex and Gender : Toward Transforming Scientific Practice. Springer.
Efron, B., 2005. Bayesians, frequentists, and scientists. J. Am. Stat. Assoc. 100.
Efron, B., Tibshirani, R., 1998. The problem of regions. Ann. Stat. 26.
Gelman, A., Robert, C.P., 2014. Revised evidence for statistical standards. Proc. Natl. Acad. Sci. U. S. A. 111 (19), E1933.
Gelman, A., Stern, H., 2006. The difference between "significant" and "not significant" is not itself statistically significant. Am. Stat. 60.
Gigerenzer, G., Krauss, S., Vitouch, O., 2004. The Null Ritual. In: Sage Handb Quant Methodol Soc Sci.
Gould, S.J., 2013. The median isn't the message. AMA J. Ethics 15, 77–81.
Greenland, S., 2019. Valid P-values behave exactly as they should: some misleading criticisms of P-values and their resolution with S-values. Am. Stat. 73.
Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., et al., 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur. J. 31 (4), 337–350.
Haller, H., Krauss, S., Kraus, S., 2002. Misinterpretations of significance : a problem students share with their teachers ? Methods Psychol. Res.7 (1),, 1–20.
Halsey, L.G., Esteves, G.P., Dolan, E., 2023. Variability in variability: does variation in morphological and physiological traits differ between men and women? R. Soc. Open Sci. 10.
Harlow, L.L., 2014. What if there Were no Significance Tests? What if there Were no Significance Tests?.
Heene, M., Ferguson, C.J., 2017a. Psychological science's aversion to the null, and why many of the things you think are true, Aren't. In: Lillienfield, Scott, Waldman, Irwin D. (Eds.), Psychol Sci under Scrut, 1st. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 34–52.
Heen, M. & Ferfuson, CJ (2017) Psychological science's aversion to the null and why many of the thik are true, aren't in Psychological science under scrutiny: Recent challenges and proposed solutions. Edited by Llienfeld SO & Waldman ID, pages 34–52.
Hoekstra, R., Finch, S., Johnson, A., 2006. Probability as certainty: dichotomous thinking and the misuse of p values. Psychon. Bull. 13 (6), 1033–1037.
Hurlbert, S.H., Lombardi, C.M., 2009. Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. Ann. Zool. Fenn. 46 (5), 311–349.
Ioannidis, J.P.A., 2005. Why most published research findings are false. PLoS Med 2 (8), e124.
Jacklin, C.N., 1981. Methodological issues in the study of sex-related differences. Dev Rev. 1, 266–273.
Jayedi, A., Soltani, S., Motlagh, S.Z.T., Emadi, A., Shahinfar, H., Moosavi, H., et al., 2022. Anthropometric and adiposity indicators and risk of type 2 diabetes: systematic review and dose-response meta-analysis of cohort studies. BMJ 376, 3067516.
Johnson, V.E., 2013. Revised standards for statistical evidence. Proc. Natl. Acad. Sci. U. S. A. 110.
Johnson, V.E., 2019. Evidence from marginally significant t statistics. Am. Stat. 73.
Johnson, V.E., Pramanik, S., Shudde, R., 2023. Bayes factor functions for reporting outcomes of hypothesis tests. Proc. Natl. Acad. Sci. U. S. A. 120.
Jones, L.V., Tukey, J.W., 2000. A sensible formulation of the significance test. Psychol. Methods 5.
Karp, N.A., Berdoy, M., Gray, K., Hunt, L., Jennings, M., Kerton, A., et al., 2025. The sex inclusive research framework to address sex bias in preclinical research proposals. Nat. Commun. 16, 3763.
Kievit, R.A., Frankenhuis, W.E., Waldorp, L.J., Borsboom, D., 2013. Simpson's paradox in psychological science: a practical guide. Front. Psychol. 4.
Kline, R.B., 2013a. Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, 2nd. American Psychological association, Washington DC.
Kline, R.B., 2013b. Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, 2nd ed. Beyond significance Test. Stat. reform Behav. Sci.
Kmetz, J.L., 2019. Correcting corrupt research: recommendations for the profession to stop misuse of p-values. Am. Stat. 73.
Kruschke, J.K., 2013. Bayesian estimation supersedes the t test. J. Exp. Psychol. Gen. 142.
Lakens, D., Scheel, A.M., Isager, P.M., 2018. Equivalence testing for psychological research: a tutorial. Adv. Methods Pract. Psychol. Sci. 1.
Lambdin, C., 2012. Significance tests as sorcery: science is empirical—significance tests are not. Theory. Psychol.22 (1),, 67–90.
Lecoutre, M.P., Poitevineau, J., Lecoutre, B., 2003. Even statisticians are not immune to misinterpretations of null hypothesis significance tests. Int. J. Psychol.38(1) 37–45.
Maney, D.L., 2016. Perils and pitfalls of reporting sex differences. Philos. Trans. R. Soc. B Biol. Sci.371 ((1688),), 201501119.
Maney, D.L., Duchesne, A., Grossi, G., 2025. Sex/gender entanglement: a problem of knots and buckets. Biol. Sex Differ. 16, 85.
McGraw, K.O., Wong, S.P., 1992. A common language effect size statistic. Psychol 111 (2), 361.
McShane, B.B., Gal, D., 2017. Statistical significance and the dichotomization of evidence. J. Am. Stat. Assoc.112 (519),, 885–895.
McShane, B.B., Gal, D., Gelman, A., Robert, C., Tackett, J.L., 2019. Abandon statistical significance. Am. Stat. 73.
Micceri, T., 1989. The unicorn, the Normal curve, and other improbable creatures. Psychol. Bull. 105.
Morey, R., Lakens, D., 2016. Why most of psychology is statisticallyunfalsifiable [internet]. Zenodo. https://doi.org/10.5281/zenodo.838685. Available from:
Nickerson, R.S., 2000. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol. Methods 5 (2), 241–301.

Nuzzo, R., 2014. Scientific method: statistical errors. Nature506 150–152.
Perezgonzalez, J.D., 2015. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Front. Psychol 6, 223.
Rice, K.M., Krakauer, C.A., 2023. Three-decision methods: a sensible formulation of significance tests-and much else. Annu. Rev. Stat. 10, 525–546.
Rich-Edwards, J.W., Maney, D.L., 2023. Best practices to promote rigor and reproducibility in the era of sex-inclusive research. eLife 12, e90623.
Riesthuis, P., 2024. Simulation-based power analyses for the smallest effect size of interest: a confidence-interval approach for minimum-effect and equivalence testing. Adv. Methods Pract. Psychol. Sci. 7.
Riesthuis, P., Cribbie, R.A., 2025. When are scientific findings deemed practically or theoretically relevant? A literature review on minimum-effect testing. Quant. Methods Psychol. 21, 82–94.
Rousselet, G.A., Pernet, C.R., Wilcox, R.R., 2017. Beyond differences in means: robust graphical methods to compare two groups in neuroscience. Eur. J. Neurosci.46 (2),, 1738–1748.
Royall, R.M., 2000. Statistical Evidence : A Likelihood Paradigm. Chapman & Hall/CRC.
Sanchis-Segura, C., Wilcox, R.R., 2024. From means to meaning in the study of sex/gender differences and similarities. Front. Neuroendocrinol. [Internet] 73, 101133. Available from: https://www.sciencedirect.com/science/article/pii/S0091302224000013X.
Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22.
Simonsohn, U., Nelson, L.D., Simmons, J.P., 2014. P-curve: a key to the file-drawer. J. Exp. Psychol. Gen. 143.
Speelman, C.P., McGann, M., 2013. How mean is the mean? Front. Psychol. 4.
Tukey, J.W., 1991. The philosophy of multiple comparisons. Stat Sci.(6),, 100–116.
Wagenmakers, E.J., 2007. A practical solution to the pervasive problems of p values. Psychon. Bull. 14, 779–804.
Wagenmakers, E.-J., Lee, M., Lodewyckx, T., Iverson, G.J., 2008. Bayesian Versus Frequentist Inference. In: Bayesian Eval Inf Hypotheses. Springer New York, New York, NY, pp. 181–207.
Wasserstein, R.L., Lazar, N.A., 2016. The ASA'S statement on p-values: context, process, and purpose. Am Stat 70 (2), 129–133.
Wasserstein, R.L., Schirm, A.L., Lazar, N.A., 2019. Moving to a world beyond "$p < 0.05$.". Am. Stat. 73, 1–19.
Weissgerber, T.L., Garcia-Valencia, O., Garovic, V.D., Milic, N.M., Winham, S.J., 2018. Why we need to report more than 'data were analyzed by t-tests or ANOVA'. eLife 7.
Wellek, S., 2010. Testing Statistical Hypotheses of Equivalence and Noninferiority, 2nd ed. Chapman and Hall/CRC.
Wilcox, R.R., 1998. How many discoveries have been lost by ignoring modern statistical methods? Am. Psychol. 53.
Wilcox, R.R., 2006. Graphical methods for assessing effect size: some alternatives to Cohen's d. J. Exp. Educ. 74 (4), 353–367.
Wilcox, R.R., Rousselet, G.A., 2023. An updated guide to robust statistical methods in neuroscience. Curr Protoc. 3.
Wilcox, R.R., Sanchis-Segura, C., 2025. Comparing two independent groups: inferences about the lower and upper tails of the distribution of $D = X - Y$. Int. J. Stat. Probab. 14, 24.
Wilcox, R.R., Serang, S., 2017. Hypothesis testing, p values, confidence intervals, measures of effect size, and Bayesian methods in light of modern robust techniques. Educ. Psychol. Meas. 77.
Zaragoza-Mezquita, M., Felix-Esbrí, S., Sebastián-Tirado, A., Guinot, P., Melero, C., Forn, C., et al., 2025. Exploring the potential of immersive virtual reality (VR) as a tool to enhance cognitive functions and alleviate clinical symptoms in multiple sclerosis (MS). Mult. Scler. Relat. Disord. 93, 106235.
Zilliak, Stephen T., McCloskey, Deirdre N., 2008. The Cult of Statistical Significance, 1st ed. The University of Michigan Press.