Quality & Quantity
https://doi.org/10.1007/s11135-026-02909-5

Neither the *t*-test nor Welch's test: a case for robust alternatives in two-sample comparisons with non-ideal data

Joost de Winter[1]

Received: 27 December 2025 / Accepted: 8 June 2026

Abstract
Student's *t*-test and Welch's *t*-test are the most common defaults for comparing two independent samples, but each rests on often-violated assumptions. We compared the two tests in simulations varying sample-size imbalance, variances, and skewness, plus an empirical sampling study of gender differences on two psychological scales; further simulations compared both classical tests with the Yuen-Welch test, the *t*-test on ranks, Welch's *t*-test on ranks, a permutation-based Welch's test, the Brunner-Munzel test, the Kolmogorov–Smirnov test, and the Anderson–Darling test. Under unequal sample sizes, the *t*-test failed to maintain nominal Type I error when standard deviations also differed (the classical Welch motivation), while Welch's test inflated Type I error when distributions were skewed, with the false positive rate reaching approximately 6%, 7.5%, and 9% at population skewness 1, 2, and 3 (nominal 5%). When sample sizes were equal, both classical tests held nominal Type I error; under unequal sample sizes with skewness, both lost power to robust alternatives, so the *t*-test was no remedy for Welch's inflation. Among the alternatives, the permutation-based Welch's test held the nominal Type I error across the factorial design while preserving the original measurement scale, making it a defensible default in the present simulations when the research question concerns equality of means. The Anderson–Darling test attained relatively high power when the two skewed populations differed simultaneously in mean and variability, and is a strong candidate when the research question concerns whether two distributions differ rather than whether their means differ specifically.

Keywords *t*-test · Welch's test · False positives · Skewness · Sample size · Statistical power

✉ Joost de Winter
j.c.f.dewinter@tudelft.nl

[1] Delft University of Technology, Delft, Netherlands

Published online: 04 July 2026

1 Introduction

The independent-samples *t*-test, often referred to as Student's *t*-test following the seminal work by Gosset under the pseudonym 'Student' (Student 1908), and Welch's *t*-test are key statistical methods. A core assumption differentiating these two methods concerns the equality of variances between the populations being compared. The *t*-test assumes that the variances of the two groups are equal (homoscedasticity), while Welch's test, developed by Welch (1947) as an adaptation, does not rely on this assumption, making it applicable even when variances are unequal (heteroscedasticity).

The debate has recently been invigorated by work from Delacre et al. (2017; see also Lakens 2015), who argued that Welch's test, not the *t*-test, should be the default option for psychologists. Their recommendation, supported by simulation studies which demonstrated the superior control of false positive rates of Welch's test under heteroscedasticity, has gained considerable traction, as shown by its high citation count. This perspective builds on earlier investigations into the robustness of the *t*-test under assumption violations (e.g., Boneau 1960).

Despite their strong recommendation, a body of counterevidence has emerged suggesting the choice is not so clear-cut. Concerns have been raised that under certain conditions, such as non-normal distributions combined with unequal sample sizes, Welch's test may yield inflated false positive rates (Algina et al. 1994; Curtis 2024; De Winter 2015; Sakai 2016). Based on a theoretical outline and subsampling method of empirical data, Curtis (2024) concluded: "*if the assumption of normality is violated and observations follow a Poisson distribution, then with unequal sample sizes Welch's t test has an inflated Type I error rate, is systematically biased and is prone to produce extremely low p values.*"

One way to explain the inflated Type I error rate is through the mathematical collapse of the test statistic under specific distributions. The two-sample *t*-statistic for Welch's test is defined as:

$$t_{Welch} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Suppose the second sample contains a very large number of observations. Its sample mean then approximates its population mean μ_2, and its variance contribution becomes negligible. Entering those limits into the formula reduces the *t*-statistic to a one-sample comparison:

$$t_{Welch} \rightarrow \frac{\overline{x}_1 - \mu_2}{\sqrt{s_1^2/n_1}} = \sqrt{n_1}\frac{\overline{x}_1 - \mu_2}{s_1}$$

Applying the same limit to the traditional *t*-test yields a different result. Because the pooled variance is dominated by the large second sample when $n_2 \rightarrow \infty$, the *t*-statistic converges to:

$$t \rightarrow \sqrt{n_1}\frac{\overline{x}_1 - \mu_2}{\sigma_2}$$

The key difference lies in the denominator: in Welch's test it contains s_1, a random variable computed from the small sample, whereas in the traditional *t*-test the denominator converges to σ_2, essentially a population constant.

The normal-theory exactness of Student's *t*-test, and the usual approximation behind Welch's test, depend on the sample mean and sample variance being independent, a property that holds under normality. However, in right-skewed conditions, these two quantities are positively correlated, and the consequences can be especially important for Welch's test because its denominator contains a random estimate of the standard error. When a sample from a right-skewed distribution misses the heavy tail, it can simultaneously yield a small sample mean ($\bar{x}_1$) and a small sample variance (s_1^2). This can result in an overly extreme *t*-value and a correspondingly small *p*-value. Consequently, the test may reject the null hypothesis more often than the nominal significance level would suggest. Skewness and the noisy behaviour of the estimated standard error are therefore not two separable causes of the inflation but a single mechanism: the dependence of the variance estimate on the sample mean is the channel through which skewness operates. This mechanism has an exact finite-sample expression: under a common population skewness, Algina et al. (1994) derived the sampling covariance between the numerator and the squared denominator of the Welch statistic as $\gamma_3(\sigma_1^3/n_1^2 - \sigma_2^3/n_2^2)$, where γ_3 denotes the population skewness. In the balanced equal-variance case this term cancels, but when sample sizes or variances differ, the cancellation is no longer guaranteed. Skewness can therefore induce dependence between Welch's numerator and denominator, providing a mechanism for Type I-error distortion.

While previous work has examined the robustness of Welch's test to unequal variances, the interaction with other assumption violations, such as non-normality in combination with highly unequal sample sizes, has been less examined. The influential study by Delacre et al. (2017) provided supplementary results regarding false positives and statistical power for various statistical distributions, but tested only modest sample size ratios (1:2 and 2:1); a subsequent Correction (Delacre et al. 2022) acknowledged a script error and noted that the conclusions could not be generalized to scenarios combining unequal sample sizes with highly skewed distributions. Scenarios involving highly unequal sample sizes are, however, common in many research areas. For example, in clinical trials, some treatments may be hard or expensive to administer, resulting in smaller sample sizes, while control groups can be much larger. Similarly, in observational studies, certain subpopulations might be harder to reach, which in turn can lead to a difference in sample sizes.

Different software packages expose the choice between Student's *t*-test and Welch's test in different ways: some use Welch's test as the default, some use the equal-variance *t*-test as the default, some report both, and others require the user to choose. These defaults are consequential because they encode different assumed trade-offs between statistical power and safeguards against false positives. In principle, researchers should inspect group standard deviations and the shapes of the raw-score distributions (or residuals) and treat strong non-normality as a warning sign that conventional *t*-based *p*-values may be inaccurate. Moreover, selecting the *t*-test versus Welch's test via an automatic test of equality of variances has well-known drawbacks (Delacre et al. 2017), and 'trying both' and choosing the smaller *p*-value is evidently poor practice. Hence, the debate over the *t*-test versus Welch's test warrants further investigation, especially for unbalanced designs where the procedures differ most.

Although many planned experiments aim for equal group sizes to maximize power, unequal sample sizes are common in practice due to missing data (including attrition), unequal recruitment costs, or naturally unbalanced groups. Because differences between *t*-test and Welch's test are amplified when sample sizes are unequal, the simulations below include highly unbalanced designs as a stress test; accordingly, the results primarily target such cases. More fundamentally, the choice is not limited to Student's *t*-test and Welch's test: the robust-statistics literature (Wilcox and Keselman 2003) has long advocated for alternatives such as trimmed-mean tests, rank transformations, and permutation methods that handle non-ideal data without classical-test assumptions. We accordingly extend the comparison to include the Yuen-Welch test, the *t*-test on ranks, Welch's *t*-test on ranks, a permutation-based Welch's test, the Brunner-Munzel test, and distributional equality tests (Kolmogorov–Smirnov, Anderson–Darling).

2 Methods

2.1 Simulations with equal and unequal variances

Simulations were carried out with two vectors of scores, x_1 and x_2. The sample size of x_1 was set to 50 ($n_1 = 50$), and the sample size of x_2 was varied from very low ($n_2 = 2$) to equal to n_1 ($n_2 = 50$). For each sample size, the *t*-test and Welch's test were applied across 1,000,000 replications (i.e., independently drawn pairs of samples). The *t*-test and Welch's test are fast to compute, so 1,000,000 replications were used to obtain precise estimates of false positive and true positive rates.

The simulations were conducted under eight conditions:

Condition A: *Unequal variances & equal means.* The first population's *SD* was twice as large as the second ($SD_1 = 2$, $SD_2 = 1$). The two populations were normally distributed with means of 0.

Condition B: *Unequal variances & equal means.* The second population's *SD* was twice as large as the first ($SD_1 = 1$, $SD_2 = 2$). The two populations were normally distributed with means of 0.

Condition C: *Equal variances & equal means.* The means and the standard deviations of the two populations were equal ($SD_1 = SD_2 = 1$). The two populations were normally distributed with means of 0.

Condition D: *Equal variances & equal means, skewed distribution*. This scenario involves two populations with equal variances and means, both following an identical lognormal distribution. Each population has a mean of 0.8 and a standard deviation of 1. This specific distribution was previously used by De Winter (2013). The lognormal populations exhibited a skewness of 5.70 and an excess kurtosis of 90.5. Throughout this paper, sample skewness and excess kurtosis are computed with the bias-corrected (Fisher-Pearson) estimators, obtained in Python via scipy.stats.skew and scipy.stats.kurtosis with bias = False. These high values are attributable to the presence of extreme positive values, which are rare in a normal distribution but characteristic of lognormal distributions. Note that sample size affects these measures. For example, with a small sample size of $n = 10$, the mean sample skewness is 1.45, and the mean sample excess kurtosis is 2.35, both substantially below the population values of 5.70 and 90.5.

Condition E: *Unequal variances & unequal means.* Same as Condition A, but now the population means of the two distributions are 1 and 0, respectively.

Condition F: *Unequal variances & unequal means.* Same as Condition B, but now the population means of the two distributions are 1 and 0, respectively.

Condition G: *Equal variances & unequal means.* Same as Condition C, but now the population means of the two distributions are 1 and 0, respectively.

Condition H: *Equal variances & unequal means, skewed distribution*. Same as Condition D, but with a mean shift of 1 applied to the first population (resulting in population means of 1.8 and 0.8, respectively).

For the normal-distribution conditions with a mean shift of 1, the above *SD*s yield Cohen's $d = 0.63$ for Conditions E and F (unequal *SD*s) and $d = 1.00$ for Condition G (equal *SD*s). For the lognormal Condition H, the same calculation yields $d = 1.00$, though its interpretation is not directly comparable to the normal-distribution conditions because distributional overlap and detection probability depend on shape, not only on the first two moments. Note that Conditions E and F have a smaller effect size than Condition G by design, because the variance ratio is part of the experimental manipulation. Power comparisons are made between the *t*-test and Welch's test within each condition, not across conditions with different effect sizes.

For each sample size and condition, the proportion of *p*-values smaller than 0.05 was calculated and depicted visually. For Conditions A–D, the population means were equal; thus, a *p*-value lower than 0.05 indicates a false positive. For Conditions E–H, the population means were unequal; thus, a *p*-value lower than 0.05 indicates a true positive.

2.2 Simulations with varied skewness levels

The above simulation used only a single lognormal distribution (Condition D) alongside normal distributions (Conditions A–C), but did not show how Welch's false positive rate varies as a function of skewness. Furthermore, because sample skewness systematically underestimates population skewness (especially in small samples), a researcher inspecting descriptive statistics may underestimate the risk. To address both questions, i.e., how the false positive rate scales with skewness, and how observed sample skewness relates to the underlying population skewness driving the inflation, we conducted a further simulation. Using Fleishman's (1978) power method (see also Rasch et al. 2011), we generated data from populations with controlled skewness ranging from 0 to 3. We compared a large sample ($n_1 = 50$) against a smaller one ($n_2 = 10$), with both drawn from the same skewed population (i.e., the null hypothesis was true). Because both samples were drawn from the same population, the variances were equal at the population level ($\sigma_1 = \sigma_2$); any inflation of the false positive rate observed in this simulation can therefore be attributed to skewness rather than to heteroscedasticity. For each level of population skewness, excess kurtosis was fixed at $(5/3) \cdot (\text{skewness})^2$. We selected this relationship because it characterizes the Inverse Gaussian (Wald) distribution, a standard theoretical model for response times, while also representing an approximation of the skewness-kurtosis scaling generally observed in human subject data (Schwarz 2001). The skewness simulation used 10,000,000 replications because the false positive rate had to be estimated precisely across 31 skewness levels.

2.3 Empirical sampling study

Simulation outcomes may not be representative of actual (psychological) data. To ground our simulation findings in real-world data, we conducted an empirical study using two large psychological datasets. This approach allows for an examination of how the *t*-test and Welch's test perform with naturally occurring distributional properties often found in psychological research.

In the introduction of their article, Delacre et al. referred to the *variability hypothesis*, i.e., the proposition that males exhibit greater standard deviations than females in intellectual tasks. Inspired by this notion, we used the Armed Services Vocational Aptitude Battery (ASVAB) test battery (Bureau of Labor Statistics, 2023, as also used by De Winter et al. 2016). We selected the test with the largest male–female *SD* ratio: Auto & Shop information (see Appendix A). Additionally, we used the Driver Behaviour Questionnaire (DBQ) – Aggressive Violations scale, a dimension of the DBQ with a skewed distribution (Wells et al. 2008; De Winter et al. 2016). Table 1 provides an overview of the gender differences of the two selected variables.

A similar procedure was followed as in the simulations. A sample size of 50 was set for males, while the sample size for females was varied between 2 and 50. This was then repeated by setting the sample size for females to 50 and varying the sample size for males between 2 and 50. Samples were randomly drawn with replacement from the entire population of males and females in the dataset. For each sample size, 1,000,000 sample pairs were drawn. We assessed direction-consistent power by computing the proportion of samples where the *t*-test or Welch's test comparing males and females was statistically significant ($p<0.05$) and the observed sample mean difference was in the expected direction (i.e., males > females).

In a separate subsampling study, we determined the number of false positives. This was done by fixing the sample size for males at 50 and comparing it with a second sample of males, in which the sample size varied between 2 and 50. Since males were compared with males drawn from the same dataset, a statistically significant finding ($p<0.05$) in this context represents a false positive. This process was repeated for females.

2.4 Additional simulations: comparison with alternative tests

The analyses above focus on the classic question of differences between two population means. Groups can indeed have identical means but different variances. For example, measuring a single chemical sample with a precise benchtop spectrometer and a less reliable

Table 1 Population-level descriptive statistics of the two datasets, by gender

	ASVAB—Auto & shop information		DBQ—Aggressive violations	
	Males	Females	Males	Females
N	5975	5939	3323	5754
Minimum	0	0	6	6
Maximum	25	25	36	29
Mean	14.99	10.21	8.79	7.81
SD	5.88	3.94	3.05	2.30
Skewness	−0.18	0.25	2.09	2.08
Excess kurtosis	−0.98	−0.19	7.92	6.38

handheld device would produce the same true mean but different levels of variance due to noise.

Such a situation is, however, seldom seen in the behavioural sciences. As Sawilowsky and Blair (1992) noted: "*We have spent many years examining large data sets but have never encountered a treatment or other naturally occurring condition that produces heterogeneous variances while leaving population means exactly equal*" (p. 358). This raises the question of whether testing a null hypothesis of equal means in the presence of unequal *SD*s, the primary scenario where Welch's test is advocated, is meaningful or realistic at all. As demonstrated in the ASVAB – Auto & Shop information dataset, strongly unequal standard deviations are typically accompanied by unequal means (see Fig. 8 in Appendix A). This pattern is evident not only in cognitive ability data but also in reaction time research, where the mean and standard deviation tend to covary (e.g., Luce 1986; Wagenmakers and Brown 2007; Wagenmakers et al. 2005; Zhang et al. 2019).

Given that neither classical test performed satisfactorily across all conditions, we extended the comparison to seven additional two-sample tests that relax the normality assumption in different ways. Briefly,

(1) the Yuen-Welch test (Wilcox 1994; Yuen 1974) applies Welch's *t*-test to the trimmed means of each sample (here, after removing the highest and lowest 10% of observations from each group), reducing the influence of extreme values.
(2) The *t*-test on ranks (Zimmerman and Zumbo 1993) replaces all observations with their ranks in the combined sample and applies the standard *t*-test to those ranks.
(3) Welch's *t*-test on ranks (Ruxton and Neuhäuser 2018; Zimmerman and Zumbo 1993) follows the same rank transformation but applies Welch's *t*-test (rather than Student's) to the ranks, so that the variance of the ranks is allowed to differ between groups.
(4) The Brunner-Munzel test (Brunner and Munzel 2000; Karch 2021) tests the null hypothesis of stochastic equality. Unlike the *t*-test on ranks, which pools rank variance across groups, it estimates separate rank variances for each group via a Welch-type correction on ranks. We computed *p*-values from the small-sample *t*-approximation of Brunner and Munzel (2000). This statistic is undefined when the two samples are perfectly separated, because the placement-variance denominator collapses to zero and the Satterthwaite-style degrees of freedom reduce to 0/0. In that case we treated the outcome as significant ($p = 0$), since perfect separation constitutes strong evidence against stochastic equality. This convention matches the behaviour of the *brunnermunzel* package in *R*, which returns a vanishingly small *p*-value for non-overlapping samples.
(5) The permutation-based Welch's test (Noguchi et al. 2021) computes Welch's *t*-statistic in the usual way but obtains its *p*-value from the empirical distribution of *t*-statistics generated by randomly reshuffling the group labels (here, $k = 1001$ reshuffles), so that significance does not rely on the *t*-distribution assumption.

In some situations, researchers may not be interested in equality of means but rather in whether the overall distributions differ. This can be tested with the Kolmogorov–Smirnov test (KST; Stephens 1970) or the Anderson–Darling test (ADT; Pettitt 1976; Scholz and Stephens 1987). Both tests evaluate whether two independent samples are drawn from the same underlying distribution by directly comparing their empirical cumulative distribution functions (ECDFs).

(6) The KST quantifies the largest absolute vertical difference between the two ECDFs, while
(7) the ADT calculates a weighted integral of the squared differences, assigning greater weight to discrepancies in the tails.

The simulations conducted above for Welch's test and the *t*-test were repeated for the KST and ADT, with 1,000,000 replications per condition and sample size. For computational efficiency, KST *p*-values were calculated using the asymptotic Kolmogorov distribution with Stephens' correction; ADT *p*-values were obtained via a permutation test with 1000 Monte Carlo replications to ensure robustness at small sample sizes.

Additionally, a 3×3 factorial design was used, crossing three sample-size configurations ($n_1 = 10$, $n_2 = 50$; $n_1 = 50$, $n_2 = 10$; $n_1 = 30$, $n_2 = 30$) with three levels of population skewness (0.3, 2.0, and 4.0), yielding nine simulation conditions. In each condition, samples were drawn from two populations generated using Fleishman's (1978) power method; both populations were assigned the skewness specified by the condition, with excess kurtosis again fixed at $(5/3)\cdot(\text{skewness})^2$. Population 1 had a mean of 1.0 and a standard deviation of 0.4 throughout. The mean of Population 2 was varied across 51 equally spaced values from 1.0 to 2.0, and its standard deviation was scaled proportionally (from 0.4 to 0.8) so that the coefficient of variation remained 0.4 in both populations. This design reflects the empirical regularity that the standard deviation tends to scale with the mean (e.g., in the DBQ data of Table 1, male and female aggressive-violation scores have nearly identical sample skewness of 2.09 and 2.08, despite differing in mean and standard deviation). The value 0.4 sits just above the empirical coefficients of variation in Table 1, which ranged from 0.29 to 0.39 across males and females in each of the ASVAB and DBQ datasets. At a mean shift of zero, the two populations were identical, so any rejection is a false positive; as the mean shift grows, the populations separate in mean and standard deviation while retaining the same shape. The simulation used 100,000 replications per increment, with all nine tests applied to the same drawn sample pairs.

3 Results

3.1 Simulations with equal and unequal variances

Figure 1 shows the results of the simulations for the eight conditions. Subfigures A–D show the false positive rate for Conditions A–D, while subfigures E–H show the statistical power for Conditions E–H.

As expected and consistent with prior literature (e.g., Delacre et al. 2017), when sample sizes are unequal in combination with unequal *SD*s of the populations, the *t*-test is either conservative (Condition A) or shows an excessive number of false positives (Condition B). Regarding the skewed (lognormal) distribution (Condition D), Welch's test results in a high number of false positives when n_2 is small. For example, for $n_1 = 50$ and $n_2 = 10$, this is 9.4% instead of the nominal 5%. The *t*-test yields a false positive rate close to the nominal value of 5%, regardless of n_2. Note that the high false positive rate for Welch's test occurs espe-

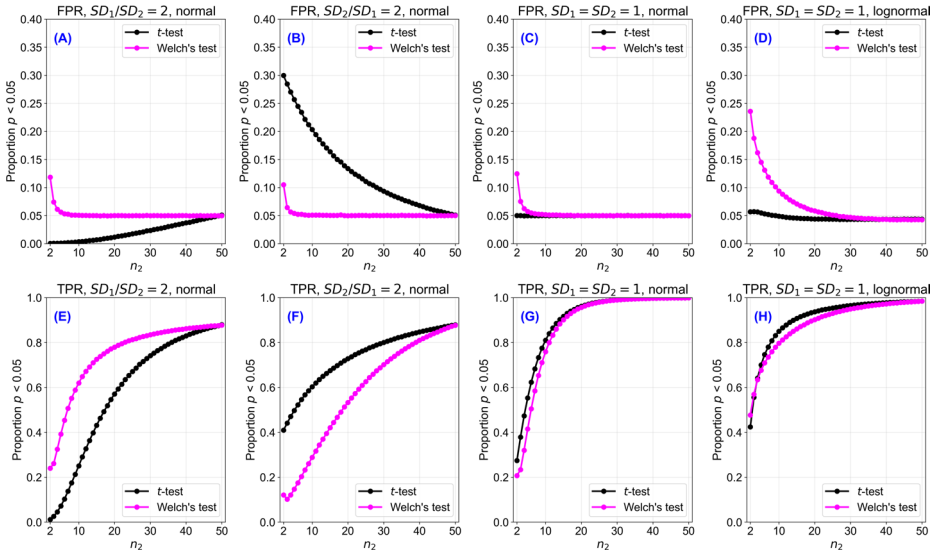

Fig. 1 Simulation results for the *t*-test and Welch's test, for eight conditions. Each point represents the proportion of statistically significant *p*-values ($p < 0.05$) out of 1,000,000 replications for the given condition and sample size ($n_1 = 50$, n_2 shown on the *x*-axis). FPR, False positive rate; TPR, True positive rate

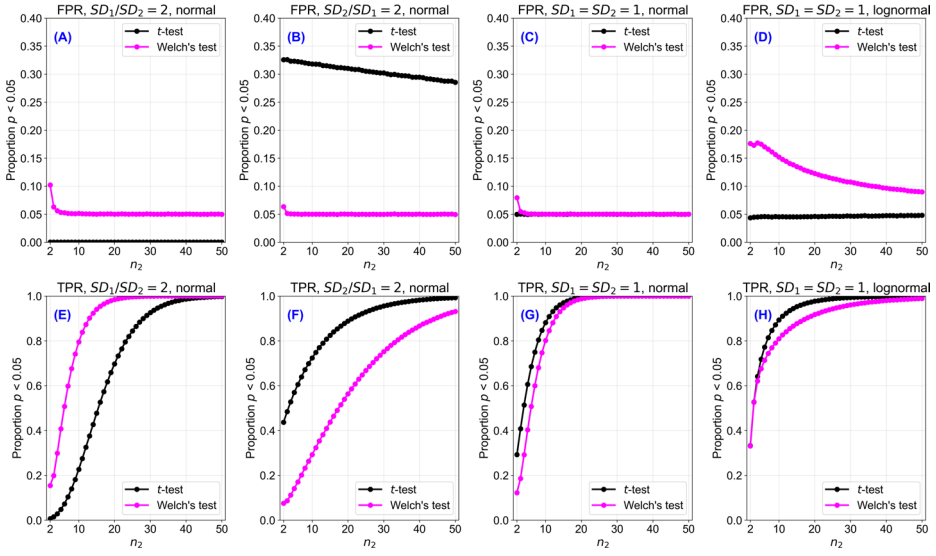

Fig. 2 Simulation results for eight conditions with a larger fixed sample size. Each point represents the proportion of statistically significant *p*-values ($p < 0.05$) out of 1,000,000 replications for the given condition and sample size ($n_1 = 1000$, n_2 shown on the *x*-axis). FPR, False positive rate; TPR, True positive rate. This figure is similar to Fig. 1 except that $n_1 = 1000$ instead of $n_1 = 50$

cially with a larger fixed sample. For example, when using $n_1 = 1000$ and $n_2 = 50$, Welch's test yields 9.0% false positives compared to 4.8% for the *t*-test (see Condition D in Fig. 2).

Consistent with Conditions A and B, the *t*-test is less powerful than Welch's test in Condition E but more powerful in Condition F. The higher power of the *t*-test in Condition F could be seen as a theoretical artifact of miscalibration (since fair power comparisons require maintained Type I error rates). In Condition G, where the *SD*s in the population are equal and there is only a mean shift, the *t*-test has higher power than Welch's test. For example, with $n_1 = 50$ and $n_2 = 10$, the power is 81.0% for the *t*-test and 75.9% for Welch's test. In the case of the lognormal distribution, the *t*-test generally features higher power than Welch's test (Condition H).

3.2 Simulations with varied skewness levels

The mean sample skewness values observed in the small group ($n_2 = 10$) as well as in the larger group ($n_1 = 50$) are considerable underestimates of the true population skewness. This difference is visualized in Fig. 3 by plotting the false positive rates against the three different measures of skewness. The lines for sample skewness (red and blue) are horizontally compressed compared to the line for population skewness (black), yet they all map to the same inflated false positive rates on the *y*-axis. This demonstrates that a seemingly moderate level of observed sample skewness can correspond to a high, and easily underestimated, risk of Type I errors when using Welch's test with unequal sample sizes.

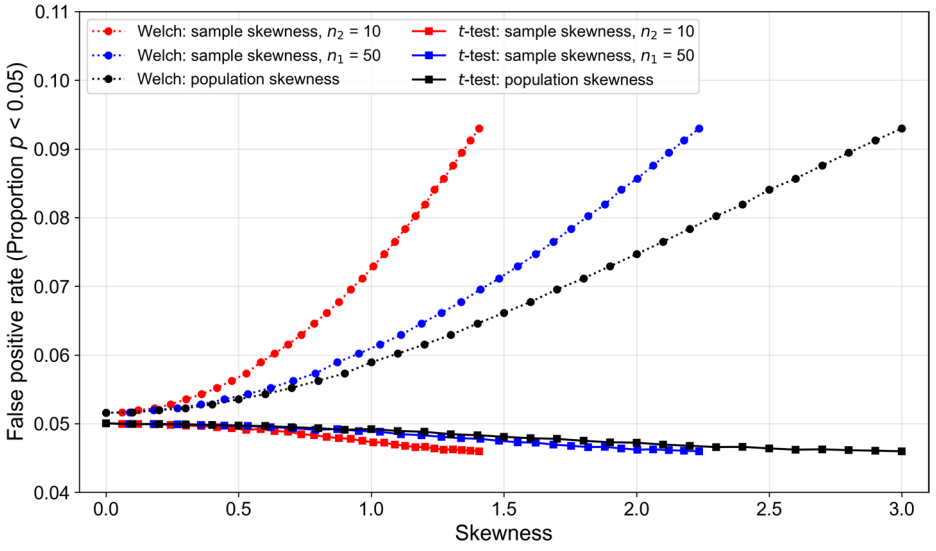

Fig. 3 False positive rate of Welch's test and the *t*-test as a function of skewness, for unequal sample sizes ($n_1 = 50$, $n_2 = 10$). The same false positive rate is plotted against three different measures of skewness: the true population skewness (black lines), the mean observed sample skewness in the small group (red lines, $n_2 = 10$), and the mean observed sample skewness in the large group (blue lines, $n_1 = 50$)

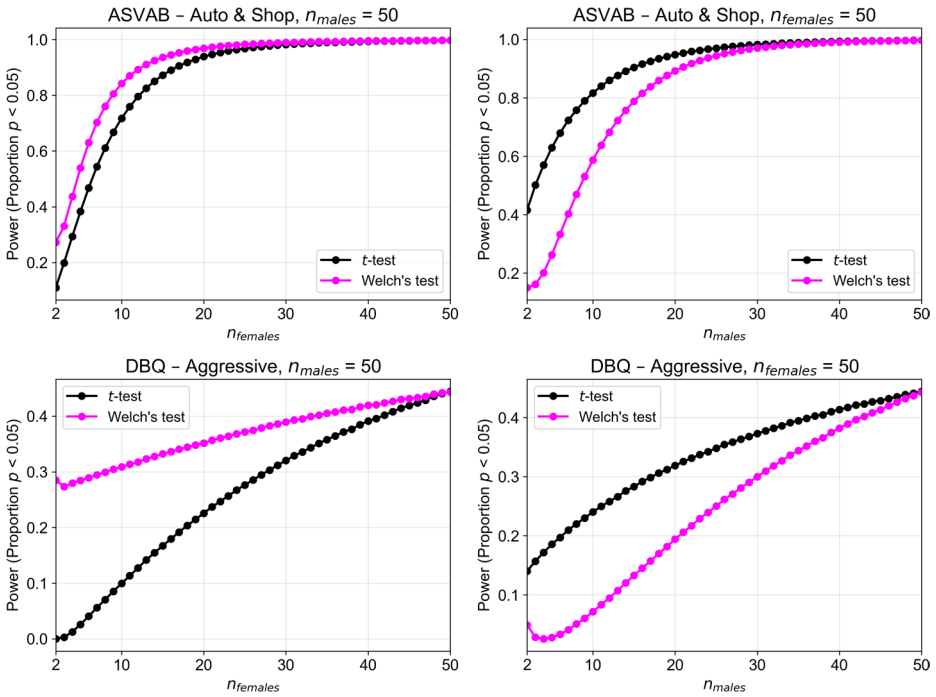

Fig. 4 Left figures: Direction-consistent power when comparing a fixed sample size of 50 males against a variable sample size of females (shown on the *x*-axis). Right figures: Direction-consistent power when comparing a fixed sample size of 50 females against a variable sample size of males (shown on the *x*-axis). Direction-consistent power is the proportion of samples in which the test result is significant ($p<0.05$) and the obtained mean for males is higher than that of females. A distinction is made between two variables (top: ASVAB—Auto & Shop information; bottom: DBQ—Aggressive Violations)

3.3 Empirical sampling study

The results of the empirical sampling study in Fig. 4 correspond with those from the simulation study results in Fig. 1. The *t*-test has higher power when the group with greater variability (in this case: males) has a smaller sample size (Fig. 4, top right), while Welch's test has higher power when the group with lower variability (in this case: females) has a smaller sample size (Fig. 4, top left).

When averaged across all sample sizes, the direction-consistent power for the *t*-test is 88.4% and for Welch's test is 86.3% for the ASVAB; for the DBQ, the *t*-test achieves 29.7% power compared to 30.6% for Welch's test. Thus, overall, the power of the *t*-test and Welch's test is similar and varies depending on which group has the smaller sample size.

One might object that, in the normal-distribution simulations, comparing power between the *t*-test and Welch's test is unfair: the *t*-test's higher power in Condition F (and lower power in Condition E) could simply reflect its miscalibrated false positive rate under unequal variances (Conditions A and B). For the skewed empirical data, however, this objection does not hold. Here, it is Welch's test, not the *t*-test, that exhibits an inflated false positive rate

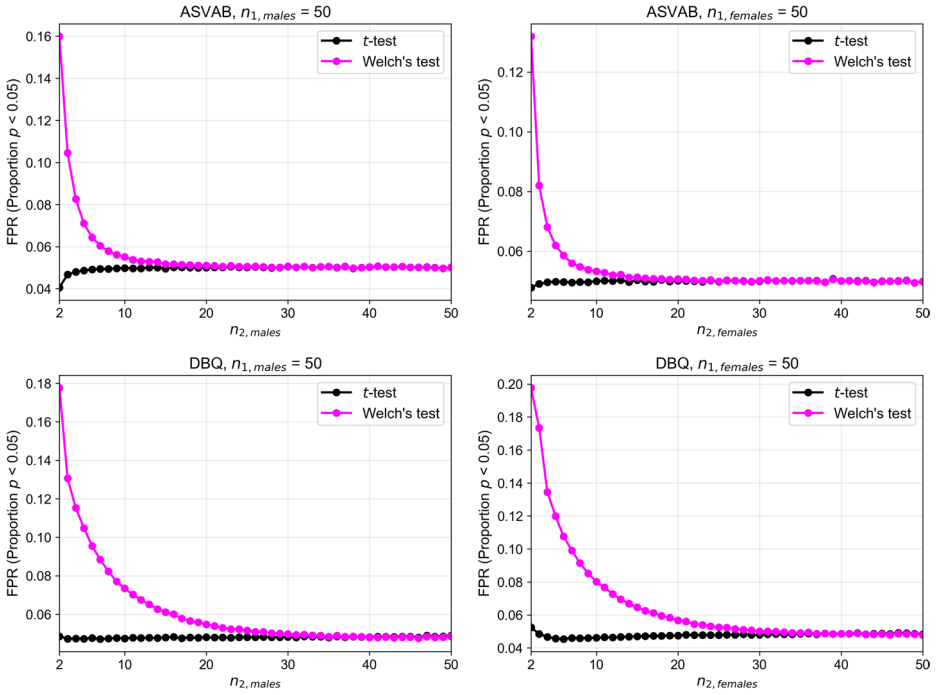

Fig. 5 Left figures: False positive rate when comparing a fixed sample size of 50 males against a second sample of males with size n_2 (shown on the *x*-axis). Right figures: Comparison of a fixed sample size of 50 females against a second sample of females with size n_2 (shown on the *x*-axis). The false positive rate is the proportion of samples in which there is a significant difference ($p<0.05$). A distinction is made between two variables (top: ASVAB—Auto & Shop information; bottom: DBQ—Aggressive Violations)

(Fig. 5). Specifically, in the case of the DBQ, which exhibits strong skewness (Table 1), we observe results qualitatively similar to Condition D (lognormal), with an inflation of false positives. The ASVAB dataset, which displayed near-normal skewness (Table 1), yielded inflated false positive rates at the smallest sample sizes only.

3.4 Additional simulations: comparison with alternative tests

Figure 6 shows that when distributions are identical (Conditions C and D), the false positive rates of the Kolmogorov–Smirnov test (KST) and the Anderson–Darling test (ADT) are close to the nominal 5%, indicating that neither test suffers from inflated Type I error. For normal distributions with unequal means and unequal standard deviations (Conditions E and F), statistical power is high, as it is for the lognormal distribution with a mean shift (Condition H). An interesting contrast with the t-test and Welch's test occurs under Conditions A and B, where population means are equal but the standard deviations differ. In this case, the KST, and especially the more powerful ADT, successfully detect that the distributions are different, particularly at larger sample sizes. The results for the individual tests are depicted in Appendix B (Figs. 9, 10, 11, 12, 13, 14, 15).

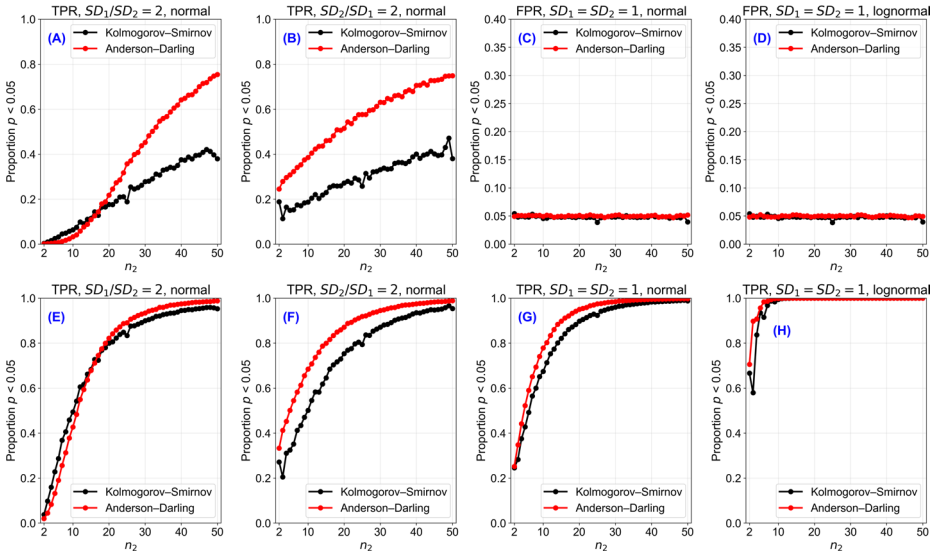

Fig. 6 Simulation results for the Kolmogorov–Smirnov test (KST) and the Anderson–Darling test (ADT), for eight conditions. Each point represents the proportion of statistically significant *p*-values ($p < 0.05$) out of 1,000,000 replications for the given condition and sample size ($n_1 = 50$, n_2 shown on the *x*-axis). FPR, False positive rate; TPR, True positive rate

Figure 7 shows the rejection rate of nine tests as a function of the mean shift between the two populations, for the nine conditions of the 3 × 3 factorial design (sample-size configuration × population skewness). At a mean shift of zero, the two populations are identical, so any rejection is a false positive. In both unequal-sample-size conditions ($n_1 = 10$, $n_2 = 50$ and $n_1 = 50$, $n_2 = 10$), the rejection rate of Welch's test scaled monotonically with population skewness: approximately 5% at a skewness of 0.3, 7.5% at a skewness of 2.0, and 10% at a skewness of 4.0. This pattern reproduced the inflation observed in Condition D of Fig. 1 and demonstrates that the inflation is not specific to the lognormal distribution. The *t*-test maintained a false positive rate at or near the nominal 5% across all conditions. The *t*-test on ranks, Welch's *t*-test on ranks, and the permutation-based Welch's test maintained the nominal false positive rate across all nine conditions, with deviations from the nominal 5% within about 0.6 percentage points (and within about 0.25 percentage points for the *t*-test on ranks and permutation-based Welch's test). The Yuen-Welch test, despite trimming, also exhibited some inflation at strong skewness with unequal sample sizes (7.6% at a skewness of 4.0), although less severely than Welch's test. The KST was slightly conservative across all conditions (3.4–4.6%). The ADT maintained the nominal level under most conditions but showed mild inflation in some high-skewness situations, reaching 6.5% at a skewness of 4.0 with balanced sample sizes and 5.8–6.1% in some unequal-sample-size conditions at skewness values of 2.0 and above. Finally, the Brunner-Munzel test was close to the nominal level in the balanced cells (about 5.1–5.4%) and only slightly liberal under sample-size imbalance (about 5.6–5.8%). In the balanced sample-size conditions ($n_1 = n_2 = 30$), all para-

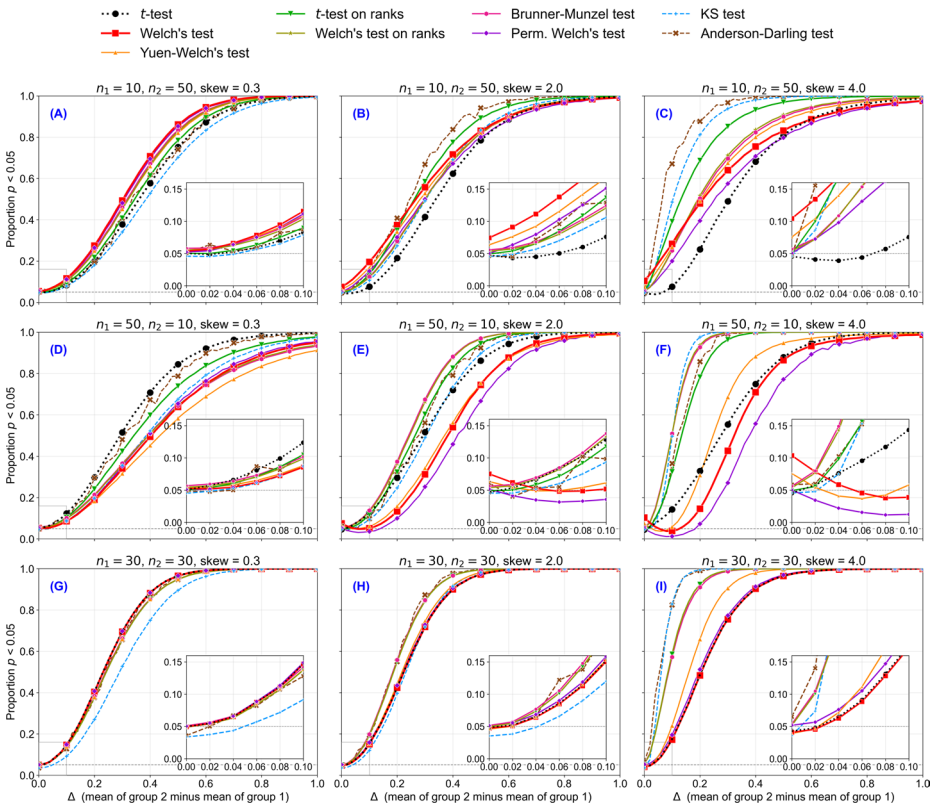

Fig. 7 Rejection rate of nine tests (the *t*-test, Welch's test, the Yuen-Welch test, the *t*-test on ranks, Welch's *t*-test on ranks, a permutation-based Welch's test, the Brunner-Munzel test, the KST, and the ADT) as a function of the mean shift between two populations, for the nine conditions of the 3 × 3 factorial design. Rows correspond to sample-size configurations (top: $n_1 = 10$, $n_2 = 50$; middle: $n_1 = 50$, $n_2 = 10$; bottom: $n_1 = n_2 = 30$); columns correspond to population skewness (left: 0.3; middle: 2.0; right: 4.0). Both populations are generated with the same Fleishman skewness and a constant coefficient of variation of 0.4, so the standard deviation scales with the mean. The inset of each panel shows the small mean-shift region where the false positive rate is most clearly visible, with a dashed line at $\alpha = 0.05$. Each point represents the proportion of statistically significant *p*-values out of 100,000 replications

metric tests perform near the nominal level even at strong skewness, reinforcing the broader observation that imbalance is the main driver of Welch's test failure under skewness.

When the populations differ in mean and variability, the distributional tests (the Kolmogorov–Smirnov and Anderson–Darling tests) achieved substantially higher power than the *t*-test and Welch's test under strong skewness with unequal sample sizes; for example, with $n_1 = 10$, $n_2 = 50$, and a population skewness of 4.0, the ADT detects a mean shift of 0.5 in 99.8% of replications, compared to 83.2% for Welch's test. Both rank-based *t*-tests showed a similar advantage; for example, at the same sample sizes and skewness with a mean shift of 0.5, the *t*-test on ranks attained 97% power and Welch's *t*-test on ranks 91%, compared with 81% for the *t*-test. The Brunner-Munzel test attained comparable power, with rejection rates near 0.99 at the largest mean shifts in the skewness = 4 panels.

4 Discussion

This study compared the *t*-test to Welch's test using simulations and a subsampling study on empirical datasets. The simulation results in Fig. 1 showed that Welch's test is advantageous when it comes to adhering to a nominal false positive rate (0.05) under unequal sample sizes combined with unequal *SD*s in a normally distributed population (Conditions A, B). In contrast, the *t*-test featured higher power in the case of equal variances in combination with unequal sample sizes (Conditions G and H), and controlled Type I error much better under skew (Condition D). Each classical test thus has a regime in which it fails to maintain the nominal false positive rate under unequal sample sizes: the *t*-test under heteroscedasticity (becoming highly inflated when the smaller sample has the larger variance, and conservative in the opposite case), and Welch's test under skewness (with the inflation scaling monotonically with population skewness, as documented in Fig. 3); with balanced sample sizes both tests hold the nominal level. More importantly, neither classical test is satisfactory under skewness with unequal sample sizes: Welch's test inflates Type I error, while the *t*-test, despite controlling Type I error there, loses substantial power to robust alternatives (Fig. 7). The high false positive rates of Welch's test under skewness were confirmed both in a simulation precisely controlling the level of skewness (Fig. 3) and in an empirical sampling study of psychological data (Figs. 4 and 5). The convergence between controlled simulations and resampling from real datasets indicates that the inflation is not an artifact of the lognormal generator used in Condition D but arises in distributional shapes routinely encountered in psychological research.

Our detailed simulation study (Fig. 3) showed that as population skewness increased, the false positive rate of Welch's test became substantially inflated, an issue the *t*-test did not share. This finding helps to contextualize previous work, such as that by Rasch et al. (2011), who noted Type I error inflation for Welch's test but "*only in the extreme case*" (p. 231) of a skewness of 3. The issue is that Rasch et al.'s focus on population-level skewness values underestimates the practical risk. The observed sample skewness, particularly in the smaller of two groups, is an underestimate of the true population skewness that actually drives the error rate. For example, as shown in Fig. 3, a population skewness of 2.4, which already produces a false positive rate of 8.2% for Welch's test, exhibits a mean sample skewness of only about 1.2 in the small group ($n = 10$). This implies that a researcher, observing what appears to be only moderate skewness in their sample, could unknowingly be operating in a condition where Welch's test is prone to a high false positive rate. Standard diagnostic advice to 'check for skewness' before choosing a test may therefore be insufficient, because the very sample sizes that make Welch's test vulnerable are the ones that most severely attenuate the observed skewness statistic.

The extent to which skewed distributions occur in reality depends on various factors such as the type of measurement instrument used, whether there are floor or ceiling effects, and whether the researcher has transformed the data or removed outliers. This study used an alpha value of 0.05. Recent guidelines recommend a more stringent alpha value, such as 0.005, for claims of new discoveries in order to reduce the rate of false positive findings in the literature (Benjamin et al. 2018). The use of lower alpha values implies that differences in false positive rates between the *t*-test and Welch's test may only increase

under skew. To illustrate, we ran an additional simulation for Condition D with $n_1 = 50$ and $n_2 = 20$ (100,000,000 replications): the false positive rate for Welch's test at alpha = 0.001 was 0.00176, i.e., 76% above the nominal level. Beyond this classical comparison, a more fundamental issue concerns the scenario in which Welch's test is most often advocated. The test addresses the null hypothesis of equal means while allowing variances to differ; yet the configuration of equal means with unequal variances is rarely observed in the behavioural data examined here, since the same processes that shift means also tend to alter variances. As we showed, heterogeneity of variance frequently co-occurs with distributions where the mean and variance are inherently linked (see Fig. 8; see also e.g., Sawilowsky and Blair 1992; Wagenmakers and Brown 2007). Distributional equality tests achieved substantially higher power than the *t*-test and Welch's test when populations differ in shape. Thus, contrary to Ruxton's (2006) assertion that Welch's test is an underused alternative compared to the *t*-test, robust tests such as the *t*-test on ranks or Welch's test on ranks, as well as distributional tests such as the KST and ADT, potentially represent underused alternatives relative to both the *t*-test and Welch's test. However, distributional tests have limitations of their own. Because they are sensitive to any distributional difference, whether in means, variances, or shape, a significant result does not indicate which aspect of the distributions differs. This makes them less suitable when the research question specifically concerns a difference in central tendency, or when the goal is to rule out group differences on a particular parameter before proceeding with a more complex analysis. In such cases, the *t*-test or Welch's test remains the more targeted tool, and distributional tests are best viewed as complements rather than replacements. For skewed data, several alternatives have been advocated as preferable to both the *t*-test and Welch's test: rank-based tests such as Wilcoxon-Mann–Whitney or tests on ranks (Zimmerman and Zumbo 1993), trimmed-mean tests such as Yuen-Welch (Wilcox 1994; Yuen 1974), resampling and bootstrapping methods (Karch 2021; Noguchi et al. 2021), and logistic regression for binary outcomes (Curtis 2024). Following Anscombe's (1960) perspective, the preference for these alternative tests to protect against the effects of outliers or other disturbances can be seen as paying a small insurance premium (i.e., a slight loss of efficiency under ideal conditions) to gain protection when deviating from a normal distribution. The Brunner-Munzel test, which Karch (2021) recommended as a replacement for the Mann–Whitney test when testing stochastic equality, showed a false positive rate close to nominal in the balanced cells of Fig. 7 (about 5.1–5.4%) and only mildly liberal under sample-size imbalance (about 5.6–5.8%). However, it became increasingly liberal when the smaller sample was very small, as can be seen in Fig. 12.

At a mean shift of zero, the two rank-based *t*-tests (Student's and Welch's) and the permutation-based Welch's test maintained the nominal Type I error rate across the nine conditions of the factorial design (Fig. 7). The classical tests (Student's *t*-test and Welch's *t*-test) retained a small power advantage under normality with equal SDs, but this advantage reversed under skewness: as Fig. 7 shows, at equal sample sizes the rank-based *t*-tests attained substantially higher power than the *t*-test under skewness (e.g., 99% versus 76% at population skewness 4 with a moderate mean shift), and at unequal sample sizes the gap remained substantial. These findings have implications for which test should serve as a default in statistical software and in routine applied research. A common position is

that researchers should inspect their data and select a test based on its features. However, the robust-statistics tradition takes a different view: a default is chosen to perform acceptably across the broadest set of plausible conditions, without requiring prior data inspection (Wilcox and Keselman 2003). Many users are likely to rely on software defaults; one that requires judging skewness and sample-size imbalance therefore places a heavy demand on them. An inspect-then-select approach additionally introduces researcher degrees of freedom that an unconditional robust choice avoids, which is particularly relevant for preregistered or confirmatory analyses.

On these grounds, a robust alternative such as the rank-based *t*-tests or a permutation-based Welch's test deserves consideration as a routine choice. Three caveats apply. First, the *t*-test on ranks pools variance across groups, so under heteroscedasticity with equal means, it exhibits the same Behrens-Fisher pattern as the classical *t*-test. Welch's *t*-test on ranks largely mitigates this asymmetry but is slightly liberal even under equal variances. Second, because rank-based tests evaluate stochastic equality rather than strict mean equality (Conover and Iman 1981), rejections under pure heteroscedasticity often reflect distributional differences rather than formal Type I errors. Furthermore, because variance heterogeneity in behavioural data typically accompanies true mean differences (Sawilowsky and Blair 1992), the strict equal-means-unequal-variances configuration that inflates the rank-based false positive rate is rarely encountered in practice. Third, the permutation-based Welch's test preserves the original measurement scale and thus permits direct reporting of mean differences, whereas rank-based *t*-tests sacrifice this interpretability for power. Robustness is not free, however: trimming discards usable observations when the distribution is well behaved, so the Yuen-Welch test loses power under normality (in Condition G, its power of 71.3% trailed both Welch's test at 75.9% and the *t*-test at 81.0% for $n_1 = 50$, $n_2 = 10$).

In conclusion, neither Student's *t*-test nor Welch's *t*-test is a safe default under the common combination of skewness and unequal sample sizes: Welch's test inflates Type I error under skewness, while the *t*-test, although it controls Type I error there, loses substantial power to robust alternatives. Among the nine tests examined, the permutation-based Welch's test held the nominal Type I error at a mean shift of zero across the factorial design (Fig. 7) while preserving the original measurement scale, making it a defensible default when the research question concerns equality of means. The Anderson–Darling test is preferable when the question concerns whether two distributions differ, since it attained the highest power when the two skewed populations differed simultaneously in mean and variability, though it shows mild Type I error inflation at strong skewness (reaching 6.5% at population skewness 4.0 with balanced sample sizes).

Appendix A

ASVAB-related statistics

See Table 2 and Fig. 8.

Table 2 Gender differences in the ASVAB (n = 11,914)

	Males Mean	Females Mean	Males *SD*	Females *SD*	*SD* ratio
1. General science (0–25)	**14.99**	13.65	**5.56**	4.83	1.15
2. Arithmetic reasoning (0–30)	**16.53**	15.11	**7.49**	6.87	1.09
3. Word knowledge (0–35)	23.40	**23.71**	**8.77**	8.27	1.06
4. Paragraph comprehension (0–15)	9.59	**10.30**	**3.81**	3.56	1.07
5. Numerical operations (0–50)	30.44	**33.03**	**11.55**	11.35	1.02
6. Coding speed (0–84)	38.73	**45.67**	16.11	**16.69**	0.96
7. Auto & Shop information (0–25)	**14.99**	10.21	**5.88**	3.94	1.49
8. Mathematics knowledge (0–25)	**12.14**	11.86	**6.34**	6.00	1.06
9. Mechanical comprehension (0–25)	**14.26**	11.00	**5.67**	4.32	1.31
10. Electronics information (0–20)	**11.44**	8.92	**4.62**	3.69	1.25

Note: The higher number (between males and females) is shown in boldface for each pair of male/female means and male/female standard deviations

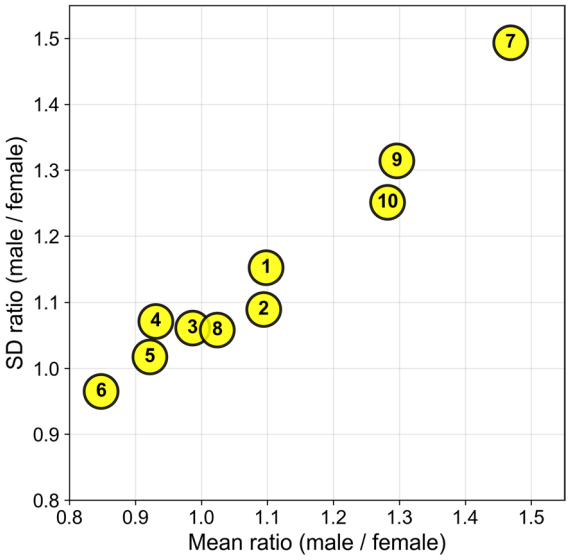

Fig. 8 Scatter plot of the standard deviation (*SD*) ratio (male *SD* / female *SD*) against the mean ratio (male mean / female mean) for the 10 subtests of the Armed Services Vocational Aptitude Battery (ASVAB). Test numbers correspond to those listed in Table 2. Higher mean ratios indicate higher male average scores relative to females, and higher *SD* ratios indicate greater male variability relative to females

Appendix B

Simulation results for extra tests (Yuen-Welch test, *t*-test on ranks, Welch's *t*-test on ranks, Brunner-Munzel test, permutation test based on Welch's *t*-statistic, Kolmogorov-Smirnov test, Anderson-Darling test)

See Figs. 9, 10, 11, 12, 13, 14 and 15.

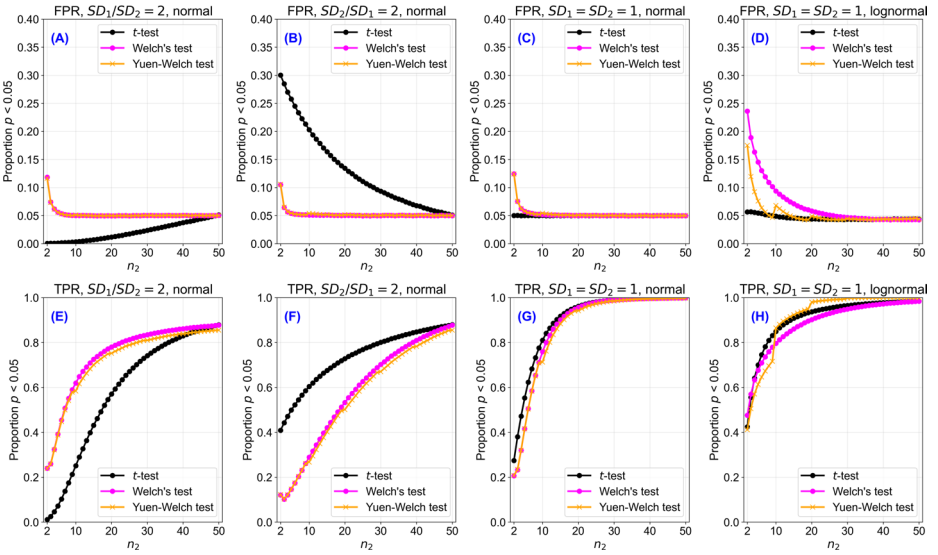

Fig. 9 Simulation results for eight conditions. Each point represents the proportion of statistically significant *p*-values ($p < 0.05$) out of 1,000,000 replications for the given condition and sample size ($n_1 = 50$, n_2 shown on the *x*-axis). FPR, False positive rate; TPR, True positive rate. This figure is similar to Fig. 1, except it now includes results for the **Yuen-Welch test**, using a trimming proportion (gamma) of 10% ($\gamma = 0.1$)

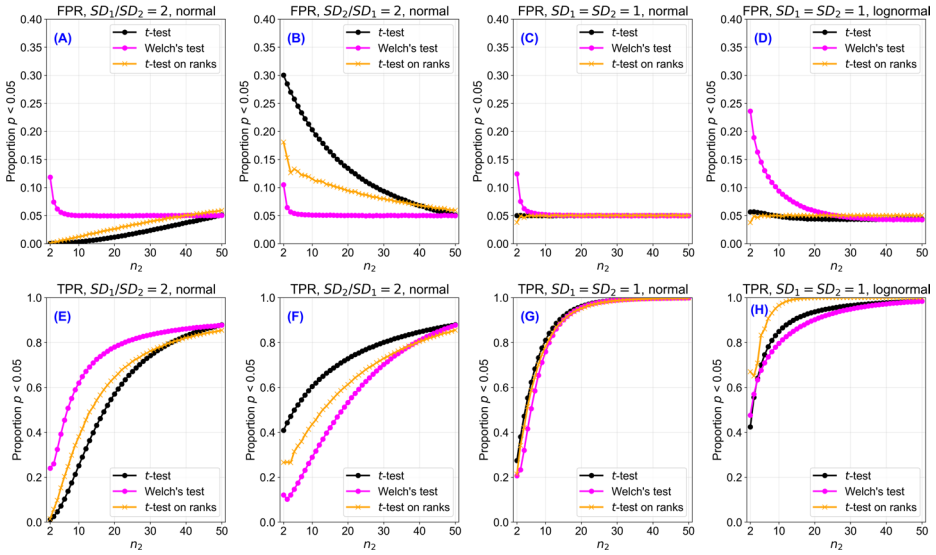

Fig. 10 Simulation results for eight conditions. Each point represents the proportion of statistically significant *p*-values ($p < 0.05$) out of 1,000,000 replications for the given condition and sample size ($n_1 = 50$, n_2 shown on the *x*-axis). FPR, False positive rate; TPR, True positive rate. This figure is similar to Fig. 1, except it now includes results for the ***t*-test on ranks**. For this test, the original data were ranked across combined groups, and a standard *t*-test was then performed on these ranks

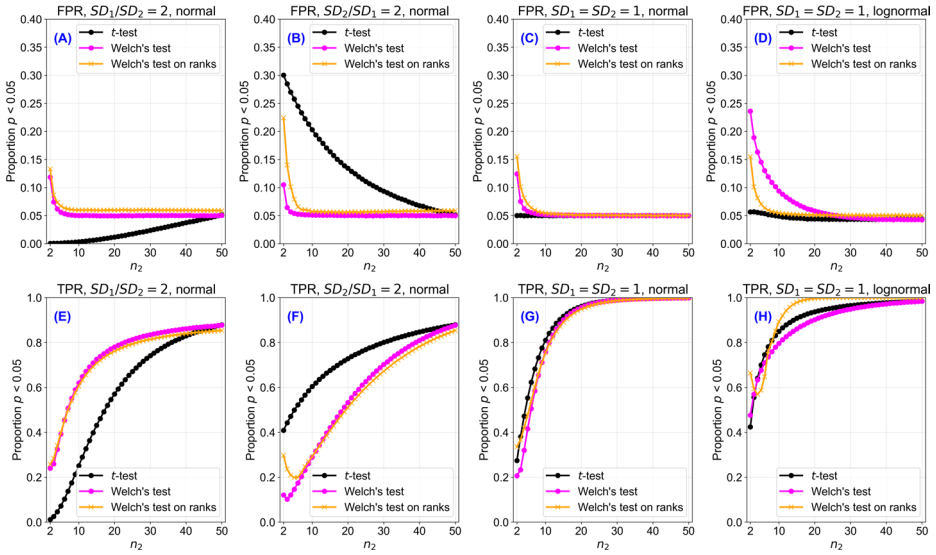

Fig. 11 Simulation results for eight conditions. Each point represents the proportion of statistically significant p-values ($p < 0.05$) out of 1,000,000 replications for the given condition and sample size ($n_1 = 50$, n_2 shown on the x-axis). FPR, False positive rate; TPR, True positive rate. This figure is similar to Fig. 1, except it now includes results for **Welch's *t*-test on ranks** (Ruxton and Neuhäuser 2018; Zimmerman and Zumbo 1993). For this test, the original data were ranked across the combined groups, and Welch's *t*-test (rather than Student's *t*-test) was then performed on these ranks

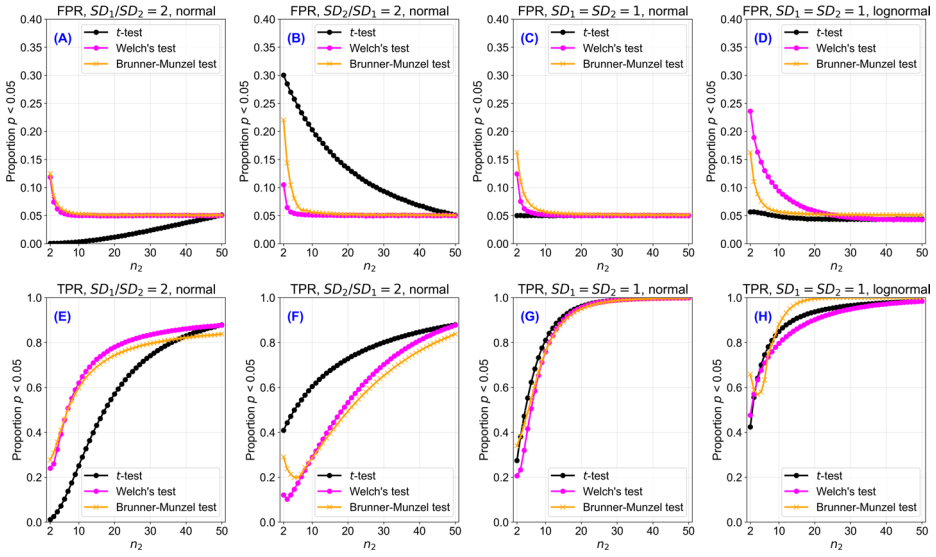

Fig. 12 Simulation results for eight conditions. Each point represents the proportion of statistically significant p-values ($p < 0.05$) out of 1,000,000 replications for the given condition and sample size ($n_1 = 50$, n_2 shown on the x-axis). FPR, False positive rate; TPR, True positive rate. This figure is similar to Fig. 1, except it now includes results for the **Brunner-Munzel test**

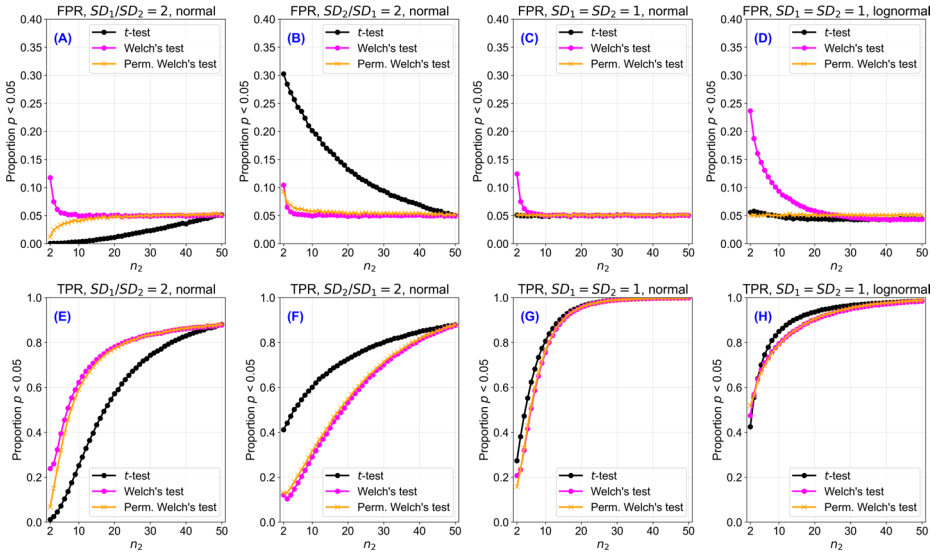

Fig. 13 Simulation results for eight conditions. Each point represents the proportion of statistically significant *p*-values ($p < 0.05$) out of 100,000 replications (reduced from 1,000,000 due to the computational cost of permutation testing) for the given condition and sample size ($n_1 = 50$, n_2 shown on the *x*-axis). FPR: False positive rate; TPR: True positive rate. This figure is similar to Fig. 1, except it now includes results for the **permutation test based on Welch's *t*-statistic**. The permutation test determines significance by comparing the observed Welch's *t*-statistic to a distribution of *t*-statistics generated by repeatedly ($k = 1001$) shuffling the data between the two groups and recalculating the statistic

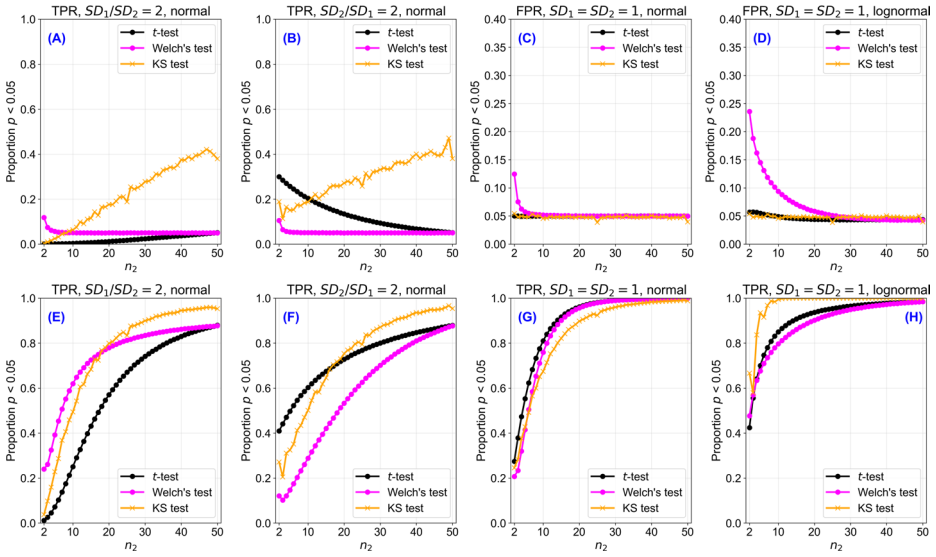

Fig. 14 Simulation results for eight conditions. Each point represents the proportion of statistically significant p-values ($p < 0.05$) out of 1,000,000 replications for the given condition and sample size ($n_1 = 50$, n_2 shown on the x-axis). FPR, False positive rate; TPR, True positive rate. In Conditions A and B (equal means, unequal SDs), the **Kolmogorov–Smirnov test (KST)** is correctly detecting that the two distributions differ, so the rejection rate is reported as a true positive rate rather than a false positive rate. This figure is similar to Fig. 1, except it now includes results for the KST

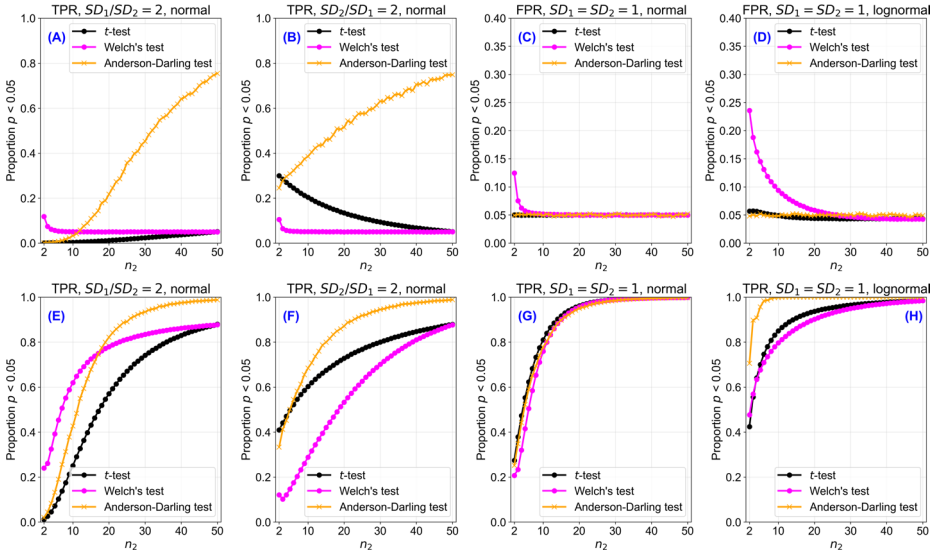

Fig. 15 Simulation results for eight conditions. Each point represents the proportion of statistically significant p-values ($p < 0.05$) out of 1,000,000 replications for the given condition and sample size ($n_1 = 50$, n_2 shown on the x-axis). FPR, False positive rate; TPR, True positive rate. As with Fig. 14, the rejection rates in Conditions A and B are reported as true positive rates because the **Anderson–Darling test (ADT)** is correctly detecting the distributional difference. This figure is similar to Fig. 1, except it now includes results for the ADT

Acknowledgements The author declares that no funds, grants, or other support were received during the preparation of this manuscript. I thank Julian Karch for several valuable suggestions and comments: for recommending that the permutation-based Welch's test be evaluated, for his comments on the handling of non-overlapping samples, and for the observation that higher power under unequal variances can reflect miscalibration of its Type I error rate rather than a genuine advantage. I also thank Mitch Earleywine for his comments on an earlier version of this manuscript.

Author contribution J.C.F. de Winter: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization.

Funding The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability To ensure reproducibility, the full simulation code is available as a supplement at: https://doi.org/10.4121/e8e6861a-7ab0-4b6d-bd67-5f95029322c5.

Declarations

Conflict of interest The author declares that he has no conflict of interest.

References

Algina, J., Oshima, T.C., Lin, W.-Y.: Type I error rates for Welch's test and James's second-order test under nonnormality and inequality of variance when there are two groups. J. Educ. Behav. Stat. **19**(3), 275–291 (1994). https://doi.org/10.3102/10769986019003275

Anscombe, F.J.: Rejection of outliers. Technometrics **2**, 123–146 (1960). https://doi.org/10.1080/00401706.1960.10489888

Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.-J., Berk, R., Bollen, K.A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C.D., Clyde, M., Cook, T.D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A.P., Forster, M., George, E.I., Gonzalez, R., Goodman, S., Green, E., Green, D.P., Greenwald, A.G., Hadfield, J.D., Hedges, L.V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D.J., Imai, K., Imbens, G., Ioannidis, J.P.A., Jeon, M., Jones, J.H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S.E., McCarthy, M., Moore, D.A., Morgan, S.L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T.H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F.D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D.J., Winship, C., Wolpert, R.L., Xie, Y., Young, C., Zinman, J., Johnson, V.E.: Redefine statistical significance. Nat. Hum. Behav. 2(1), 6–10 (2018). https://doi.org/10.1038/s41562-017-0189-z

Boneau, C.A.: The effects of violations of assumptions underlying the t test. Psychol. Bull. **57**, 49–64 (1960). https://doi.org/10.1037/h0041412

Brunner, E., Munzel, U.: The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. Biom. J. **42**(1), 17–25 (2000). https://doi.org/10.1002/(SICI)1521-4036(200001)42:1%3C17::AID-BIMJ17%3E3.0.CO;2-U

Bureau of Labor Statistics, U.S. Department of Labor: National Longitudinal Survey of Youth 1979 cohort, 1979–2020. Center for Human Resource Research, Ohio State University (2023)

Conover, W.J., Iman, R.L.: Rank transformations as a bridge between parametric and nonparametric statistics. Am. Stat. **35**(3), 124–129 (1981). https://doi.org/10.1080/00031305.1981.10479327

Curtis, D.: Welch's t test is more sensitive to real world violations of distributional assumptions than Student's t test but logistic regression is more robust than either. Stat. Pap. **65**, 3981–3989 (2024). https://doi.org/10.1007/s00362-024-01531-7

De Winter, J.C.F.: Using the Student's *t*-test with extremely small sample sizes. Pract. Assess. Res. Eval. **18**, 10 (2013). https://doi.org/10.7275/e4r6-dj05

De Winter, J.C.F.: Comments on the blog post "Always use Welch's t-test instead of Student's t-test". The 20% Statistician. https://daniellakens.blogspot.com/2015/01/always-use-welchs-t-test-instead-of.html?showComment=1422328885403#c1081063886480631670 (2015)

De Winter, J.C.F., Gosling, S.D., Potter, J.: Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: a tutorial using simulations and empirical data. Psychol. Methods **21**, 273–290 (2016). https://doi.org/10.1037/met0000079

Delacre, M., Lakens, D., Leys, C.: Why psychologists should by default use Welch's *t*-test instead of Student's *t*-test. Int. Rev. Soc. Psychol. **30**, 92–101 (2017). https://doi.org/10.5334/irsp.82

Delacre, M., Lakens, D., Leys, C.: Correction: Why psychologists should by default use Welch's t-test instead of Student's t-test. Int. Rev. Soc. Psychol. **35**(1), 21 (2022). https://doi.org/10.5334/irsp.661

Fleishman, A.I.: A method for simulating non-normal distributions. Psychometrika **43**, 521–532 (1978). https://doi.org/10.1007/BF02293811

Karch, J.D.: Psychologists should use Brunner-Munzel's instead of Mann-Whitney's U test as the default nonparametric procedure. Adv. Methods Pract. Psychol. Sci. **4**(2), 1–14 (2021). https://doi.org/10.1177/2515245921999602

Lakens, D.: Always use Welch's t-test instead of Student's t-test. The 20% Statistician. https://daniellakens.blogspot.com/2015/01/always-use-welchs-t-test-instead-of.html (2015)

Luce, R.D.: Response Times: Their Role in Inferring Elementary Mental Organization. Oxford University Press (1986)

Noguchi, K., Konietschke, F., Marmolejo-Ramos, F., Pauly, M.: Permutation tests are robust and powerful at 0.5% and 5% significance levels. Behav. Res. Methods **53**, 2712–2724 (2021). https://doi.org/10.3758/s13428-021-01595-5

Pettitt, A.N.: A two-sample Anderson-Darling rank statistic. Biometrika **63**, 161–168 (1976). https://doi.org/10.1093/biomet/63.1.161

Rasch, D., Kubinger, K.D., Moder, K.: The two-sample t test: pre-testing its assumptions does not pay off. Stat. Pap. **52**, 219–231 (2011). https://doi.org/10.1007/s00362-009-0224-x

Ruxton, G.D.: The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. Behav. Ecol. **17**, 688–690 (2006). https://doi.org/10.1093/beheco/ark016

Ruxton, G.D., Neuhäuser, M.: Striving for simple but effective advice for comparing the central tendency of two populations. J. Mod. Appl. Stat. Methods **17**(2), eP2567 (2018). https://doi.org/10.22237/jmasm/1551908612

Sakai, T.: Two sample t-tests for IR evaluation: Student or Welch? Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, 1045–1048, Pisa, Italy (2016). https://doi.org/10.1145/2911451.2914684

Sawilowsky, S.S., Blair, R.C.: A more realistic look at the robustness and Type II error properties of the t test to departures from population normality. Psychol. Bull. **111**, 352–360 (1992). https://doi.org/10.1037/0033-2909.111.2.352

Scholz, F.W., Stephens, M.A.: K-sample Anderson-Darling tests. J. Am. Stat. Assoc. **82**, 918–924 (1987). https://doi.org/10.1080/01621459.1987.10478517

Schwarz, W.: The ex-Wald distribution as a descriptive model of response times. Behav. Res. Methods Instrum. Comput. **33**, 457–469 (2001). https://doi.org/10.3758/BF03195403

Stephens, M.A.: Use of the Kolmogorov–Smirnov, Cramér–Von Mises and Related statistics without extensive tables. J. R. Stat. Soc. Ser. B Stat Methodol. **32**, 115–122 (1970). https://doi.org/10.1111/j.2517-6161.1970.tb00821.x

Student: The probable error of a mean. Biometrika **6**(1), 1–25 (1908). https://doi.org/10.1093/biomet/6.1.1

Wagenmakers, E.-J., Brown, S.: On the linear relation between the mean and the standard deviation of a response time distribution. Psychol. Rev. **114**, 830–841 (2007). https://doi.org/10.1037/0033-295X.114.3.830

Wagenmakers, E.-J., Grasman, R.P.P.P., Molenaar, P.C.M.: On the relation between the mean and the variance of a diffusion model response time distribution. J. Math. Psychol. **49**(3), 195–204 (2005). https://doi.org/10.1016/j.jmp.2005.02.003

Welch, B.L.: The generalization of 'student's' problem when several different population variances are involved. Biometrika 34(1–2), 28–35 (1947). https://doi.org/10.1093/biomet/34.1-2.28

Wells, P., Tong, S., Sexton, B., Grayson, G., Jones, E.: Cohort II: a study of learner and new drivers, vol. 1. Main Report (Report no. 81). Department for Transport, London, UK (2008)

Wilcox, R.R.: Some results on the Tukey-McLaughlin and Yuen methods for trimmed means when distributions are skewed. Biom. J. **36**, 259–273 (1994). https://doi.org/10.1002/bimj.4710360302

Wilcox, R.R., Keselman, H.J.: Modern robust data analysis methods: measures of central tendency. Psychol. Methods **8**(3), 254–274 (2003). https://doi.org/10.1037/1082-989X.8.3.254

Yuen, K.K.: The two-sample trimmed t for unequal population variances. Biometrika **61**, 165–170 (1974). https://doi.org/10.1093/biomet/61.1.165

Zhang, B., De Winter, J., Varotto, S., Happee, R., Martens, M.: Determinants of take-over time from automated driving: a meta-analysis of 129 studies. Transport. Res. Part F: Traffic Psychol. Behav. **64**, 285–307 (2019). https://doi.org/10.1016/j.trf.2019.04.020

Zimmerman, D.W., Zumbo, B.D.: Rank transformations and the power of the Student t test and Welch t' test for non-normal populations with unequal variances. Can. J. Exp. Psychol. **47**(3), 523–539 (1993). https://doi.org/10.1037/h0078850