INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY

Delacre, M., et al (2017). Why Psychologists Should by Default Use Welch's *t*-test Instead of Student's *t*-test. *International Review of Social Psychology*, 30(1), 92–101, DOI: https://doi.org/10.5334/irsp.82

RESEARCH ARTICLE

Why Psychologists Should by Default Use Welch's *t*-test Instead of Student's *t*-test

Marie Delacre*, Daniël Lakens† and Christophe Leys*

When comparing two independent groups, psychology researchers commonly use Student's *t*-tests. Assumptions of normality and homogeneity of variance underlie this test. More often than not, when these conditions are not met, Student's *t*-test can be severely biased and lead to invalid statistical inferences. Moreover, we argue that the assumption of equal variances will seldom hold in psychological research, and choosing between Student's *t*-test and Welch's *t*-test based on the outcomes of a test of the equality of variances often fails to provide an appropriate answer. We show that the Welch's *t*-test provides a better control of Type 1 error rates when the assumption of homogeneity of variance is not met, and it loses little robustness compared to Student's *t*-test when the assumptions are met. We argue that Welch's *t*-test should be used as a default strategy.

Keywords: Welch's *t*-test; Student's *t*-test; homogeneity of variance; Levene's test; Homoscedasticity; statistical power; type 1 error; type 2 error

Independent sample *t*-tests are commonly used in the psychological literature to statistically test differences between means. There are different types of *t*-tests, such as Student's *t*-test, Welch's *t*-test, Yuen's *t*-test, and a bootstrapped *t*-test. These variations differ in the underlying assumptions about whether data is normally distributed and whether variances in both groups are equal (see, e.g., Rasch, Kubinger, & Moder, 2011; Yuen, 1974). Student's *t*-test is the default method to compare two groups in psychology. The alternatives that are available are considerably less often reported. This is surprising, since Welch's *t*-test is often the preferred choice and is available in practically all statistical software packages.

In this article, we will review the differences between Welch's *t*-test, Student's *t*-test, and Yuen's *t*-test, and we suggest that Welch's *t*-test is a better default for the social sciences than Student's and Yuen's *t*-tests. We do not include the bootstrapped *t*-test because it is known to fail in specific situations, such as when there are unequal sample sizes and standard deviations differ moderately (Hayes & Cai, 2007).

When performing a *t*-test, several software packages (i.e., R and Minitab) present Welch's *t*-test by default. Users can request Student's *t*-test, but only after explicitly stating that the assumption of equal variances is met. Student's *t*-test is a parametric test, which means it relies on assumptions about the data that are analyzed. Parametric tests are believed to be more powerful than non-parametric tests (i.e., tests that do not require assumptions about the population parameters; Sheskin, 2003). However, Student's *t*-test is generally only more powerful when the data are normally distributed (the assumption of normality) and the variances are equal in both groups (homoscedasticity; the assumption of homogeneity of variance; Carroll & Schneider, 1985; Erceg-Hurn & Mirosevich, 2008).

When sample sizes are equal between groups, Student's *t*-test is robust to violations of the assumption of equal variances as long as sample sizes are big enough to allow correct estimates of both means and standard deviations (i.e., $n \geq 5$),[1] except when distributions underlying the data have very high skewness and kurtosis, such as a chi-square distribution with 2 degrees of freedom. However, if variances are *not* equal across groups and the sample sizes differ across independent groups, Student's *t*-test can be severely biased and lead to invalid statistical inferences (Erceg-Hurn & Mirosevich, 2008).[2,3] Here, we argue that there are no strong reasons to assume equal variances in the psychological literature by default nor substantial costs in abandoning this assumption.

In this article, we will first discuss why we need a default test and why a two-step procedure where researchers decide whether or not to use Welch's *t*-test based on a check of the assumption of normality and equal variances is undesirable. Then, we will discuss whether the assumption of equal variances is plausible in psychology and point out research areas where this assumption is implausible.

* Université Libre de Bruxelles, Service of Analysis of the Data (SAD), Bruxelles, BE

† Eindhoven University of Technology, Human Technology Interaction Group, Eindhoven, NL

Corresponding author: Marie Delacre (marie.delacre@ulb.ac.be)

We will then review differences between Student's *t*-test, Welch's *t*-test, and Yuen's *t*-test and show through simulations that the bias in Type 1 error rates when Yuen's *t*-test is used is often severely inflated (above 0.075, which is "critical inflation", following Bradley, 1978) and that the bias in Type 1 error rates when Student's *t*-test is used has a larger impact on statistical inferences than the rather modest impact on the Type 2 error rate of always using Welch's *t*-test by default. Given our analysis and the availability of Welch's *t*-test in all statistical software, we recommend a procedure where Welch's *t*-test is used by default when sample sizes are unequal.

Limitations of Two-Step Procedures

Readers may have learned that the assumptions of normality and of equal variances (or the homoscedasticity assumption) must be examined using assumption checks prior to performing any *t*-test. When data are not normally distributed, with small sample sizes, alternatives should be used. Classic nonparametric statistics are well-known, such as the Mann-Whitney U-test and Kruskal-Wallis. However, unlike a *t*-test, tests based on rank assume that the distributions are the same between groups. Any departure to this assumption, such as unequal variances, will therefore lead to the rejection of the assumption of equal distributions (Zimmerman, 2000). Alternatives exist, known as the "modern robust statistics" (Wilcox, Granger, & Clark, 2013). For example, data sets with low kurtosis (i.e., a distribution flatter than the normal distribution) should be analyzed with the two-sample trimmed *t*-test for unequal population variances, also called Yuen's *t*-test (Luh & Guo, 2007; Yuen, 1974). However, analyses in a later section will show that the normality assumption is not very important for Welch's *t*-test and that there are good reasons to, in general, prefer Welch's *t*-test over Yuen's *t*-test.

With respect to the assumption of homogeneity of variance, if the test of the equality of variance is non-significant and the assumption of equal variances cannot be rejected, homoscedastic methods such as the Student's *t*-test should be used (Wilcox et al., 2013). If the test of the equality of variances is significant, Welch's *t*-test should be used instead of Student's *t*-test because the assumption of equal variances is violated. However, testing the equality of variances before deciding which *t*-test is performed is problematic for several reasons, which will be explained after having described some of the most widely used tests of equality of variances.

Different Ways to Test for Equal Variances

Researchers have proposed several tests for the assumption of equal variances. Levene's test and the F-ratio test are the most likely to be used by researchers because they are available in popular statistical software (Hayes & Cai, 2007). Levene's test is the default option in SPSS. Levene's test is the One-Way ANOVA computed on the terms $|x_{ij}-\hat{\theta}_j|$, where x_{ij} is the ith observation in the jth group, and $\hat{\theta}_j$ is the "center" of the distribution for the jth group (Carroll & Schneider, 1985). In R, the "center" is by default the median, which is also called "Brown Forsythe test for equal variances". In SPSS, the "center" is by default the mean (which is the most powerful choice when the underlying data are symmetrical).[4] The F-ratio statistic is obtained by computing SD2/SD1 (standard deviation ratio, SDR). A generalization of the F-ratio test, to be used when there are more than two groups to compare, is known as the Bartlett's test.

The F-ratio test and the Bartlett test are powerful, but they are only valid under the assumption of normality and collapse as soon as one deviates even slightly from the normal distribution. They are therefore not recommended (Rakotomalala, 2008).

Levene's test is more robust than Bartlett's test and the F-ratio test, but there are three arguments against the use of Levene's test. First, there are several ways to compute Levene's test (i.e., using the median or mean as center), and the best version of the test for equal variances depends on how symmetrically the data is distributed, which is itself difficult to statistically quantify.

Second, performing two tests (Levene's test followed by a *t*-test) on the same data makes the alpha level and power of the *t*-test dependent upon the outcome of Levene's test. When we perform Student's or Welch's *t*-test conditionally on a significant Levene's test, the long-run Type 1 and Type 2 error rates will depend on the power of Levene's test. When the power of Levene's test is low, the error rates of the conditional choice will be very close to Student's error rates (because the probability of choosing Student's *t*-test is very high). On the other hand, when the power of Levene's test is very high, the error rates of the conditional choice will be very close to Welch's error rate (because the probability of choosing Welch's *t*-test is very high; see Rasch, Kubinger, & Moder, 2011). When the power of Levene's test is medium, the error rates of the conditional choice will be somewhere between Student's and Welch's error rates (see, e.g., Zimmerman, 2004). This is problematic when the test most often performed actually has incorrect error rates.

Third, and relatedly, Levene's test can have very low power, which leads to Type 2 errors when sample sizes are small and unequal (Nordstokke & Zumbo, 2007). As an illustration, to estimate the power of Levene's test, we simulated 1,000,000 simulations with balanced designs of different sample sizes (ranging from 10 to 80 in each condition, with a step of 5) under three SDR where the true variances are unequal, respectively, 1.1, 1.5, and 2, yielding 45,000,000 simulations in total. When SDR = 1, the equal variances assumption is true when SDR > 1 the standard deviation of the second sample is bigger than the standard deviation of the first sample and when SDR < 1 the standard deviation of the second sample is smaller than the standard deviation of the first sample. We ran Levene's test centered around the mean and Levene's test centered around the median and estimated the power (in %) to detect unequal variances with equal sample sizes (giving the best achievable power for a given total N; see **Figure 1**).[5]

As we can see in the graph, the further SDR is from 1, the smaller the sample size needed to detect a statistically significant difference in the SDR. Furthermore, for each SDR, power curves of the Levene's test based on the mean

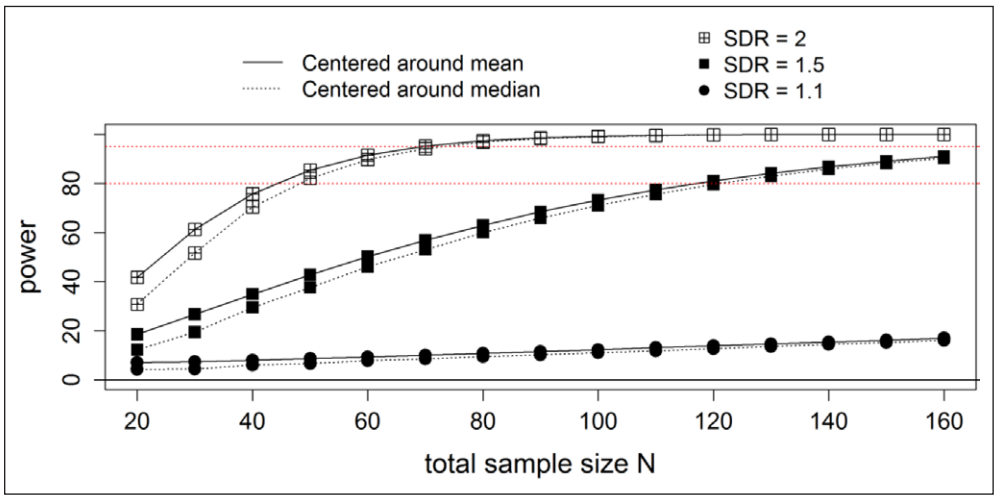

Figure 1: Estimated power of Levene's test as a function of sample size, SDR and centering parameter.

are slightly above power curves of the Levene's test based on the median, meaning that it leads to slightly higher power than Levene's test based on the median. This can be due to the fact that data is extracted from normal distributions. With asymmetric data, the median would perform better. When SDR = 2, approximately 50 subjects are needed to have 80 percent power to detect differences, while approximately 70 subjects are needed to have 95 percent power to detect differences (for both versions of Levene's test). To detect an SDR of 1.5 with Levene's test, approximately 120 subjects are needed to reach a power of 0.80 and about 160 to reach a power of 0.95. Since such an SDR is already very problematic in terms of the type 1 error rate for the Student's *t*-test (Bradley, 1978), needing such a large sample size to detect it is a serious hurdle. This issue becomes even worse for lower SDR, since an SDR as small as 1.1 already calls for the use of Welch's *t*-test (See table A3.1 to A3.9 in the additional file). Detecting such a small SDR calls for a huge sample size (a sample size of 160 provides a power rate of 0.16).

Since Welch's *t*-test has practically the same power as Student's *t*-test, even when SDR = 1, as explained using simulations later, we should seriously consider using Welch's *t*-test by default.

The problems in using a two-step procedure (first testing for equality of variances, then deciding upon which test to use) have already been discussed in the field of statistics (see e.g., Rasch, Kubinger, & Moder, 2011; Ruxton, 2006; Wilcox, Granger, & Clark, 2013; Zimmerman, 2004), but these insights have not changed the current practices in psychology, as of yet. More importantly, researchers do not even seem to take the assumptions of Student's *t*-test into consideration before performing the test, or at least rarely discuss assumption checks.

We surveyed statistical tests reported in the journal *SPPS* (*Social Psychological and Personality Science*) between April 2015 and April 2016. From the total of 282 studies, 97 used a *t*-test (34.4%), and the homogeneity of variance was explicitly discussed in only 2 of them. Moreover, based on the reported degrees of freedom in the results section, it seems that Student's *t*-test is used most often and that alternatives are considerably less popular. For 7 studies, there were decimals in the values of the degrees of freedom, which suggests Welch's *t*-test might have been used, although the use of Welch's *t*-test might be higher but not identifiable because some statisticians recommend rounding the degrees of freedom to round numbers.

To explain this lack of attention to assumption checks, some authors have argued that researchers might have a lack of knowledge (or a misunderstanding) of the parametric assumptions and consequences of their violations or that they might not know how to check assumptions or what to do when assumptions are violated (Hoekstra, Kiers, & Johnson, 2012).[6] Finally, many researchers don't even know there are options other than the Student's *t*-test for comparing two groups (Erceg-Hurn & Mirosevich, 2008). How problematic this is depends on how plausible the assumption of equal variances is in psychological research. We will discuss circumstances under which the equality of variances assumption is especially improbable and provide real-life examples where the assumption of equal variances is violated.

Homogeneity of Variance Assumptions

The homogeneity of variances assumption is rarely true in real life and cannot be taken for granted when performing a statistical test (Erceg-Hurn & Mirosevich, 2008; Zumbo & Coulombe, 1997). Many authors have examined real data and noted that SDR is often different from the 1:1 ratio (see, e.g., Grissom, 2000; Erceg-Hurn & Mirosevich, 2008). This shows that the presence of unequal variances is a realistic assumption in psychological research.[7] We will discuss three different origins of unequal standard deviations across two groups of observations.

A first reason for unequal variances across groups is that psychologists often use *measured variables* (such as age, gender, educational level, ethnic origin, depression level, etc.) instead of random assignment to condition. In their review of comparing psychological findings from all fields of the behavioral sciences across cultures, Henrich, Heine, and Norenzayan (2010) suggest that parameters vary largely from one population to another. In other words, variance is not systematically the same in every pre-existing group. For example, Feingold (1992) has

shown that intellectual abilities of males were more variable than intellectual abilities of females when looking at several standardized test batteries measuring general knowledge, mechanical reasoning, spatial visualization, quantitative ability, and spelling. Indeed, the variability hypothesis (that men demonstrate greater variability than women) is more than a century old (for a review, see Shields, 1975). In many research domains, such as mathematics performance, there are strong indicators that variances ratios differ between 1.1 and 1.2, although variances ratios do not differ in all countries, and the causes for these differences are not yet clear. Nevertheless, it is an empirical fact that variances ratios can differ among pre-existing groups.

Furthermore, some pre-existing groups have different variability by definition. An example from the field of education is the comparison of selective school systems (where students are accepted on the basis of selection criterions) versus comprehensive school systems (where all students are accepted, whatever their aptitudes; see, e.g., Hanushek & Wößmann, 2006). At the moment that a school accepts its students, variability in terms of aptitude will be greater in a comprehensive school than in a selective school, by definition.

Finally, a quasi-experimental treatment can have a different impact on variances between groups. Hanushek and Wößmann (2006) suggest that there is an impact of the educational system on variability in achievement. Even if variability, in terms of aptitude, is greater in a comprehensive school than in a selective school at first, a selective school system at primary school increases inequality (and then variability) in achievement in secondary school. Another example is variability in moods. Cowdry, Gardner, O'Leary, Leibenluft, & Rubinow (1991) noted than intra-individual variability is larger in patients suffering from premenstrual syndrome (PMS) than in normal patients and larger in normal patients than in depressive patients. Researchers studying the impact of an experimental treatment on mood changes can expect a bigger variability of mood changes in patients with PMS than in normal or depressive patients and thus a higher standard deviation in mood measurements.

A second reason for unequal variances across groups is that while variances of two groups are the same when group assignment is completely randomized, deviation from equality of variances can occur later, as a *consequence of an experimental treatment* (Cumming, 2013; Erceg-Hurn & Mirosevich, 2008; Keppel, 1991). For example, psychotherapy for depression can increase the variability in depressive symptoms, in comparison with a control group, because the effectiveness of the therapy will depend on individual differences (Bryk & Raudenbush, 1988; Erceg-Hurn & Mirosevich, 2008). Similarly, Kester (1969) compared the IQs of students from a control group with the IQs of students when high expectancies about students were induced in the teacher. While no effect of teacher expectancy on IQ was found, the variance was bigger in the treatment group than in the control group (56.52 vs. 32.59, that is, SDR ≈ 1.32). As proposed by Bryk and Raudenbush (1988), this can result from the interaction between the treatment and the students' reactions: students can react differently to the induced expectations. More generally, whenever a manipulation has individual moderators, variability should increase compared to a control condition.

Knowing whether standard deviations differ across conditions is important information, but in many fields, we have no accurate estimates of the standard deviation in the population. Whereas we collect population effect sizes in meta-analyses, these meta-analyses often do not include the standard deviations from the literature. As a consequence, we regrettably do not have easy access to aggregated information about standard deviations across research areas, despite the importance of this information. It would be useful if meta-analysts start to code information about standard deviations when performing meta-analyses (Lakens, Hilgard, & Staaks, 2016), such that we can accurately quantify whether standard deviations differ between groups, and how large the SDR is.

The Mathematical Differences Between Student's *t*-test, Welch's *t*-test, and Yuen's *t*-test

So far, we have simply mentioned that Welch's *t*-test differs from Student's *t*-test in that it does not rely on the equality of variances assumption. In this section, we will explain why this is the case. The Student's *t* statistic is calculated by dividing the mean difference between group $\bar{x}_1 - \bar{x}_2$ by a pooled error term, where s_1^2 and s_2^2 are variance estimates from each independent group, and where n_1 and n_2 are the respective sample sizes for each independent group (Student, 1908):

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\right) * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (1)$$

The degrees of freedom are computed as follows (Student, 1908):

$$df = n_1 + n_2 - 2 \quad (2)$$

Student's *t*-test is calculated based on a *pooled* error term, which implies that both samples' variances are estimates of a common population variance. Whenever the variances of the two normal distributions are not similar and the sample sizes in each group are not equal, Student's *t*-test results are biased (Zimmerman, 1996). The more unbalanced the distribution of participants across both independent groups, the more Student's *t*-test is based on the incorrect standard error (Wilcox et al., 2013) and, consequently, the less accurate the computation of the *p*-value will be.

When the larger variance is associated with the *larger* sample size, there is a decrease in the nominal Type 1 error rate (Nimon, 2012; Overall, Atlas, & Gibson, 1995). The reason for this is that the error term increases, and,

as a consequence, the Student's *t*-value decreases, leading to fewer significant findings than expected with a specific alpha level. When the larger variance is associated with the *smaller* sample size, the Type 1 error rate is inflated (Nimon, 2012; Overall, Atlas, & Gibson, 1995). This inflation is caused by the under-evaluation of the error term, which increases Student's *t* value and thus leads to more significant results than are expected based on the alpha level.

As discussed earlier, Student's *t*-test is robust to unequal variances as long as the sample sizes of each group are similar (Nimon, 2012; Ruxton, 2006; Wallenstein, Zucker, & Fleiss, 1980), but, in practice, researchers often have different sample sizes in each of the independent groups (Ruxton, 2006). Unequal sample sizes are particularly common when examining measured variables, where it is not always possible to determine *a priori* how many of the collected subjects will fall in each category (e.g., sex, nationality, or marital status). However, even with complete randomized assignment to conditions, where the same number of subjects are assigned to each condition, unequal sample sizes can emerge when participants have to be removed from the data analysis due to being outliers because the experimental protocol was not followed when collecting the data (Shaw & Mitchell-Olds, 1993) or due to missing values (Wang et al., 2012).

Previous work by many researchers has shown that Student's *t*-test performs surprisingly poorly when variances are unequal and sample sizes are unequal (Glass, Peckham, & Sanders, 1972; Overall, Atlas, & Gibson, 1995; Zimmerman, 1996), especially with small sample sizes and low alpha levels (e.g., alpha = 1%; Zimmerman, 1996). The poor performance of Student's *t*-test when variances are unequal becomes visible when we look at the error rates of the test and the influence of both Type 1 errors and Type 2 errors. An increase in the Type 1 error rate leads to an inflation of the number of false positives in the literature, while an increase in the Type 2 error rate leads to a loss of statistical power (Banerjee et al., 2009).

To address these limitations of Student's *t*-test, Welch (1947) proposed a separate-variances *t*-test computed by dividing the mean difference between group $\bar{x}_1 - \bar{x}_2$ by an unpooled error term, where s_1^2 and s_2^2 are variance estimates from each independent group, and where n_1 and n_2 are the respective sample sizes for each independent group:[8]

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3}$$

The degrees of freedom are computed as follows:

$$\mathrm{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{4}$$

When both variances and sample sizes are the same in each independent group, the *t*-values, degrees of freedom, and the *p*-values in Student's *t*-test and Welch's *t*-test are the same (see **Table 1**). When the variance is the same in both independent groups but the sample sizes differ, the *t*-value remains identical, but the degrees of freedom differ (and, as a consequence, the *p*-value differs). Similarly, when the variances differ between independent groups but the sample sizes in each group are the same, the *t*-value is identical in both tests, but the degrees of freedom differ (and, thus, the *p*-value differs). The most important difference between Student's *t*-test and Welch's *t*-test, and indeed the main reason Welch's *t*-test was developed, is when both the variances and the sample sizes differ between groups, the *t*-value, degrees of freedom, and *p*-value all differ between Student's *t*-test and Welch's *t*-test. Note that, in practice, samples practically never show exactly the same pattern of variance as populations, especially with small sample sizes (Baguley, 2012; also see table A2 in the additional file).

Yuen's *t*-test, also called "20 percent trimmed means test", is an extension of Welch's *t*-test and is allegedly more robust in case of non-normal distributions (Wilcox & Keselman, 2003). Yuen's *t*-test consists of removing the lowest and highest 20 percent of the data and applying Welch's *t*-test on the remaining values. The procedure is explained and well-illustrated in a paper by Erceg-Hurn and Mirosevich (2008).

Simulations: Error Rates for Student's *t*-test versus Welch's *t*-test

When we are working with a balanced design, the statistical power (the probability of finding a significant effect, when there is a true effect in the population, or 1 minus the Type 2 error rate) is very similar for Student's *t*-test

	Equal variances	**Unequal variances**
Balanced design	$t_{Welch} = t_{Student}$ $df_{Welch} = df_{Student}$ $p_{Welch} = p_{Student}$	$t_{Welch} = t_{Student}$ $df_{Welch} \neq df_{Student}$ $p_{Welch} \neq p_{Student}$
Unbalanced design	$t_{Welch} = t_{Student}$ $df_{Welch} \neq df_{Student}$ $p_{Welch} \neq p_{Student}$	$t_{Welch} \neq t_{Student}$ $df_{Welch} \neq df_{Student}$ $p_{Welch} \neq p_{Student}$

Table 1: Comparison of *t*-value and Degrees of Freedom of Welch's and Student's *t*-test.

and Welch's *t*-test. Even with extremely large SDR (respectively, 0.01, 0.1, 10, and 100) and small sample sizes (10 subjects per group), the biggest increase in power of Student's *t*-test compared to Welch's-*t* test is approximately 5 percent when the test is applied on two normal skewed distributions with unequal shapes. In all other cases, the difference in power between both tests is smaller (See table A1.1 to A1.9 in the additional file).

Considering the cases where sample sizes are unequal and SDR = 1, Student's-*t* test is sometimes better than Welch's *t*-test, and sometimes the reverse is true. The difference is small, except in three scenarios (See table A5.2, A5.5, and A5.6 in the additional file). However, because there is no correct test to perform that assures SDR = 1, and because variances are likely not to be equal in certain research areas, our recommendation is to always use Welch's *t*-test instead of Student's *t*-test.

To illustrate the differences in Type 1 error rates between Student's *t*-test and Welch's *t*-test, we simulated 1,000,000 studies under the null hypothesis (no difference between the means in each group) under four scenarios. We chose a small sample ratio (n_1 = 40 vs. n_2 = 60) to show that when the equal variances assumption was not met and SDR = 2, biased error rates are observed in Student's *t*-test. We compared Scenario 1, where the variance is the same in each group (SDR = 1; homoscedasticity assumption met) and sample sizes are unequal (See **v 2a**), with Scenario 2, where the variance differs between groups (SDR = 2) but sample sizes are equal ($n_1 = n_2 = 50$; see **Figure 2b**). Furthermore, we simulated Scenario 3, where both sample sizes and variances were unequal between groups and the larger variance is associated with the larger sample size (SDR = 2; see **Figure 2c**), and a similar Scenario 4, where the larger variance is associated with the smaller sample size (SDR = 0.5; see **Figure 2d**). *P*-value distributions for both Student's and Welch's *t*-tests were then plotted. When there is no true effect, *p*-values are distributed uniformly.

As long as the variances are equal between groups or sample sizes are equal, the distribution of Student's *p*-values is uniform, as expected (see **Figures 2a** and **2b**),

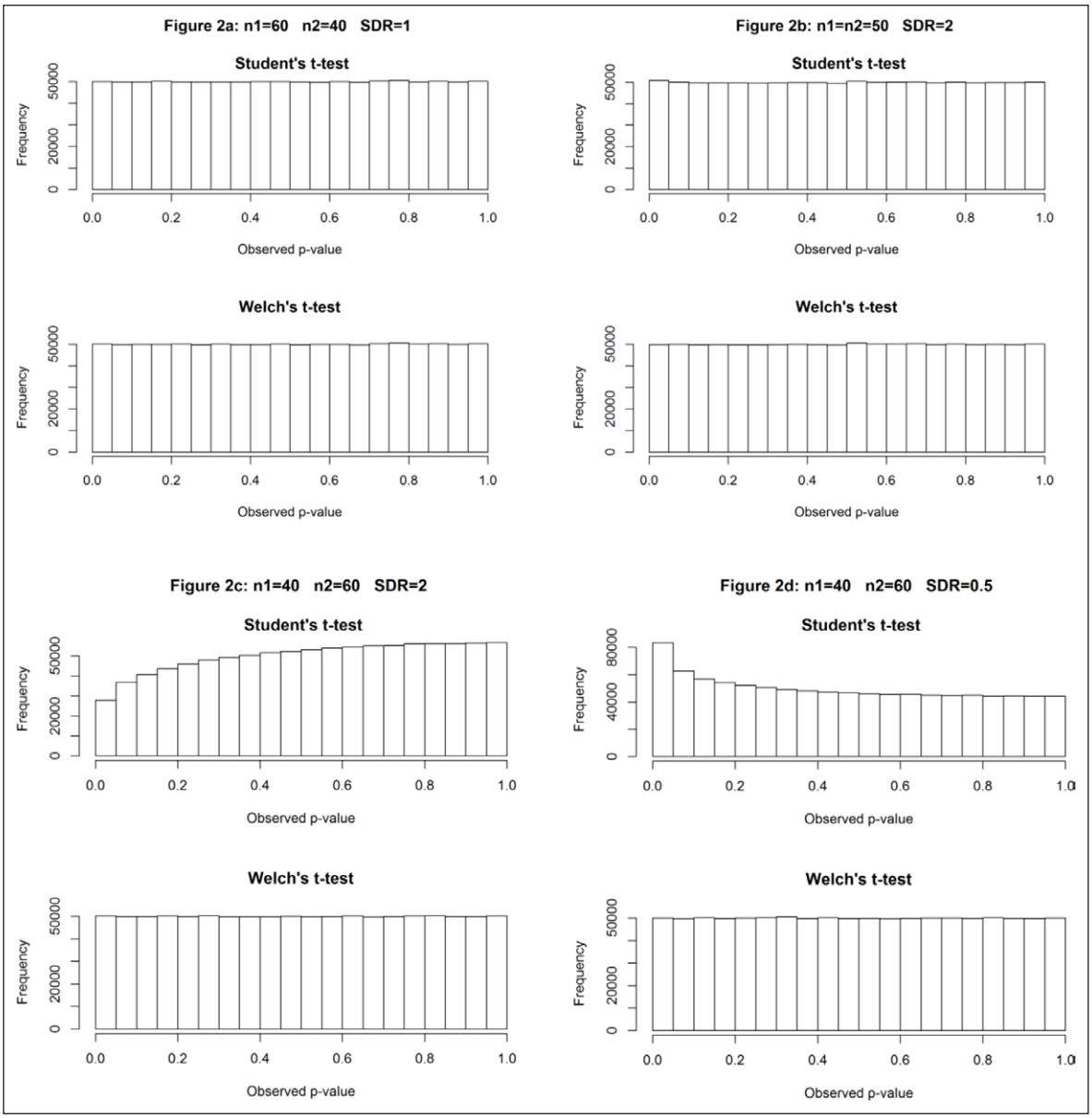

Figure 2: *P*-value distributions for Student's and Welch's *t*-test under the null as a function of SDR, and sample size.

which implies that the probability of rejecting a true null hypothesis equals the alpha level for any value of alpha. On the other hand, when the larger variance is associated with the larger sample size, the frequency of *p*-values less than 5 percent decreases to 0.028 (see **Figure 2c**), and when the larger variance is associated with the smaller sample size, the frequency of *p*-values less than 5 percent increases to 0.083 (see **Figure 2d**). Welch's *t*-test has a more stable Type 1 error rate (see Keselman et al., 1998; Keselman, Othman, Wilcox, & Fradette, 2004; Moser & Stevens, 1992; Zimmerman, 2004) Additional simulations, presented in the additional file, show that these scenarios are similar for several shapes of distributions (see tables A3.1 to A3.9 and table A4 in the additional file).

Moreover, as discussed previously, with very small SDRs, Welch's *t*-test still has a better control of Type 1 error rates than Student's *t*-test, even if neither of them give critical values (i.e., values under 0.025 or above 0.075, according to the definition of Bradley, 1975). With SDR = 1.1, when the larger variance is associated with the larger sample size, the frequency of Student's *p*-value being less than 5 percent decreases to 0.046, and when the larger variance is associated with the smaller sample size, the frequency of Student's *p*-value being less than 5 percent increases to 0.054. On the other side, the frequency of Welch's *p*-values being below 0.05 is exactly 5 percent in both cases.

Yuen's *t*-test is not a good unconditional alternative because we observe an unacceptable departure from the nominal alpha risk of 5 percent for several shapes of distributions (see tables A3.1, A3.4, A3.7, A3.8, and A3.9 in the additional file), particularly when we are studying asymmetric distributions of unequal shapes (see tables A3.8 and A3.9 in the additional file). Moreover, even when Yuen's Type 1 error does not show a critical departure from the nominal alpha risk (i.e., values above 0.075), Welch's *t*-test more accurately controls the Type 1 error rate (see tables A3.2, A3.3, A3.5, and A3.6 in the additional file). The Type 1 error rate of Welch's *t*-test remains closer to the nominal size (i.e., 5%) in all the previously discussed cases and also performs better with very extreme SDRs and unbalanced designs, as long as there are at least 10 subjects per groups (See table A4 in the additional file).

In **Figure 3**, *p*-values from Welch's *t*-test and Student's *t*-test tests, shown separately in **Figure 2** (through histograms), are now plotted against each other. **Figure 3a** shows Student's *p*-values plotted against Welch's *p*-values of Scenario 1, where the variance is the same in each group (SDR = 1) and sample sizes are unequal. **Figure 3b** displays Student's *p*-values plotted against Welch's *p*-values of Scenario 2, where the

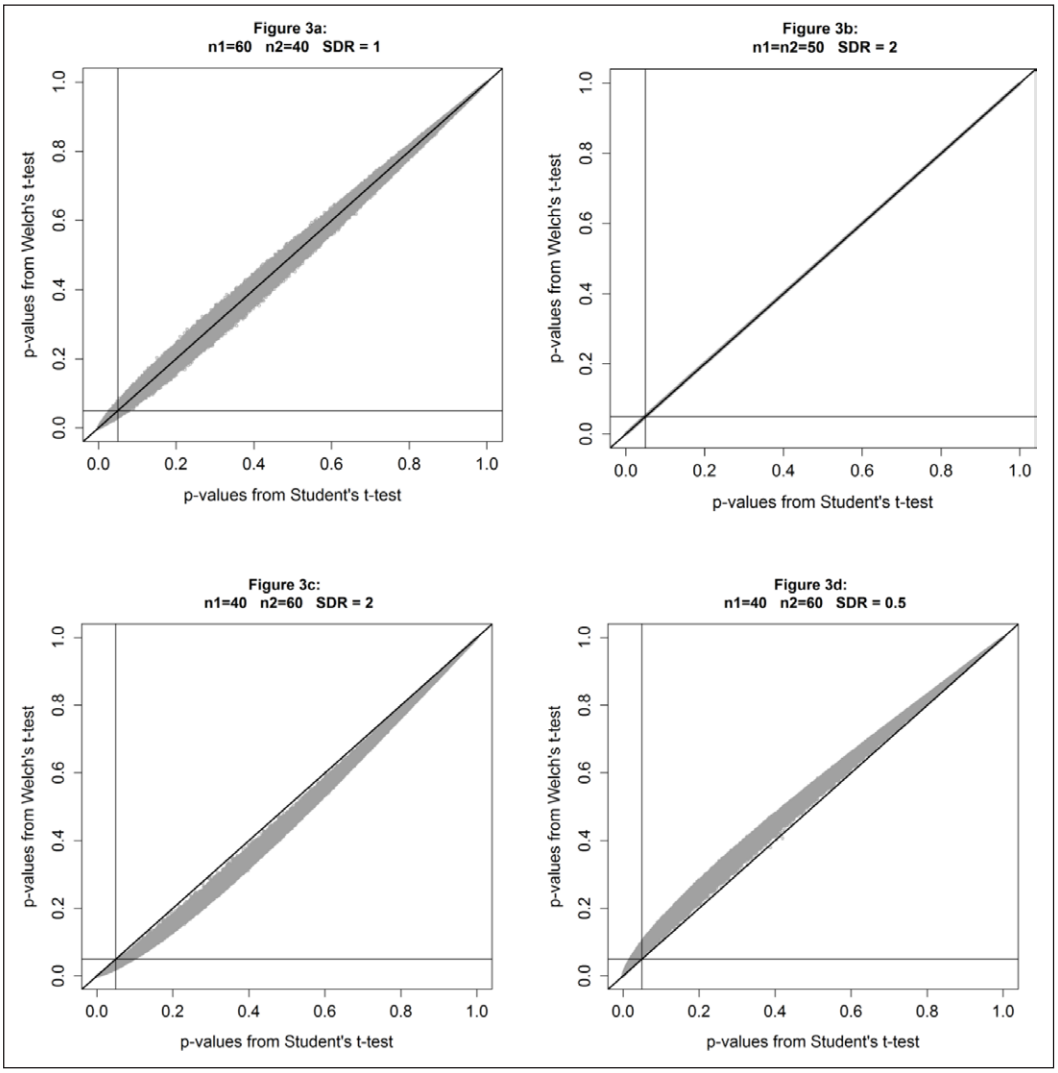

Figure 3: *P*-values from Student's *t*-test against *p*-values from Welch's *t*-test under the null.

variance differs between groups (SDR = 2) but sample sizes are equal ($n_1 = n_2 = 50$). **Figure 3c** shows Student's *p*-values plotted against Welch's *p*-values of Scenario 3, where both sample sizes and variances are unequal between groups and the larger variance is associated with the larger sample size (SDR = 2). And, finally, **figure 3d** plots Student's *p*-values against Welch's *p*-values of Scenario 4, where the greater variance is associated with the smaller sample size (SDR = 0.5).

Dots are marked on the black diagonal line when both tests return the same *p*-value. The top left quadrant contains all *p*-values less than 0.05 according to a Student's *t*-test, but greater than 0.05 according to Welch's *t*-test. The bottom right quadrant reports all *p*-values less than 0.05 according to Welch's *t*-test, but greater than 0.05 according to Student's *t*-test. The larger the standard deviations ratio and the greater the sample sizes ratio, the larger the difference between *p*-values from Welch's *t*-test and Student's *t*-test.

Conclusion

When the assumption of equal variances is not met, Student's *t*-test yields unreliable results, while Welch's *t*-test controls Type 1 error rates as expected. The widely recommended two-step approach, where the assumption of equal variances is tested using Levene's test and, based on the outcome of this test, a choice of Student's *t*-test or Welch's *t*-test is made, should not be used. Because the statistical power for this test is often low, researchers will inappropriately choose Student's *t*-test instead of more robust alternatives. Furthermore, as we have argued, it is reasonable to assume that variances are unequal in many studies in psychology, either because measured variables are used (e.g., age, culture, gender) or because, after random assignment to conditions, variance is increased in the experimental condition compared to the control condition due to the experimental manipulation. As it is explained in the additional file, Yuen's *t*-test is not a better test than Welch's *t*-test, since it often suffers high departure from the alpha risk of 5 percent. Therefore, we argue that Welch's *t*-test should always be used instead of Student's *t*-test.

When using Welch's *t*-test, a very small loss in statistical power can occur, depending on the shape of the distributions. However, the Type 1 error rate is more stable when using Welch's *t*-test compared to Student's *t*-test, and Welch's *t*-test is less dependent on assumptions that cannot be easily tested. Welch's *t*-test is available in practically all statistical software packages (and already the default in R and Minitab) and is easy to use and report. We recommend that researchers make clear which test they use by specifying the analysis approach in the result section.

Convention is a weak justification for the current practice of using Student's *t*-test by default. Psychologists should pay more attention to the assumptions underlying the tests they perform. The default use of Welch's *t*-test is a straightforward way to improve statistical practice.

Notes

[1] There is a Type 1 error rate inflation in a few cases where sample sizes are extremely small and SDR is big (e.g., when $n_1 = n_2 = 3$ are sampled from uniform distributions and SDR = 2, the Type 1 error rate = 0.083; or when $n_1 = 3$ is sampled from a uniform distribution and $n_2 = 3$ is sampled from a double exponential distribution). However, with extremely small sample sizes ($N \leq 5$), the estimate of means and standard deviations is extremely inaccurate anyway. As we mentioned in table A2 (see the additional file), the smaller the sample size, the further the average standard deviation is from the population standard deviation, and the larger the dispersion around this average.

[2] This is called the Behren-Fisher problem (Hayes & Cai, 2007).

[3] In a simulation that explored Type 1 error rates, we varied the size of the first sample from 10 to 40 in steps of 10 and the sample sizes ratio and the standard deviation ratio from 0.5 to 2 in steps of 0.5, resulting in 64 simulations designs. Each design was tested 1,000,000 times. Considering these parameter values, we found that the alpha level can be inflated up to 0.11 or deflated down to 0.02 (see the additional file).

[4] Other variants have been proposed, such as the percent trimmed mean (Lim & Loh, 1996).

[5] Because sample sizes are equal for each pair of samples, which sample has the bigger standard deviation is not applicable. In this way, SDR = X will return the same answer in terms of percent power of Levene's test as SDR = 1/X. For example, SDR = 2 will return the same answer as SDR = ½ = 0.5.

[6] For example, many statistical users believe that the Mann-Whitney non-parametric test can cope with both normality and homosedasticity issues (Ruxton, 2006). This assumption is false, since the Mann-Whitney test remains sensitive to heterosedasticity (Grissom, 2000; Nachar, 2008; Neuhäuser & Ruxton, 2009).

[7] Like Bryk and Raudenbush (1988), we note that unequal variances between groups does not systematically mean that population variances are different: standard deviation ratios are more or less biased estimates of population variance (see table A2 in the additional file). Differences can be a consequence of bias in measurement, such as response styles (Baumgartner & Steenkamp, 2001). However, there is no way to determine what part of the variability is due to error rather than the true population value.

[8] Also known as the Satterwaite's test, the Smith/Welch/Satterwaite test, the Aspin-Welch test, or the unequal variances *t*-test.

Competing Interests

The authors have no competing interests to declare.

Additional File

The additional file for this article can be found as follows:

- DOI: https://doi.org/10.5334/irsp.82.s1

Author's Note

All code needed to recreate the simulations resulting in the figures and appendices is available at https://osf.io/bver8/files/, as are as the .txt files containing the results of all simulations.

References

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences.* Palgrave Macmillan. Retrieved from https://books.google.fr/books?hl=fr&lr=&id=ObUcBQAAQBAJ&oi=fnd&pg=PP1&dq=baguley+2012&ots=-eiUlHiCYs&sig=YUUKZ7jiGF33wdo3WVO-8l-OUu8.

Banerjee, A., Chitnis, U., Jadhav, S., Bhawalkar, J. & **Chaudhury, S.** (2009). Hypothesis testing, type I and type II errors. *Industrial Psychiatry Journal*, 18(2), 127. DOI: https://doi.org/10.4103/0972-6748.62274

Baumgartner, H. & **Steenkamp, J.-B. E.** (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38(2), 143–156. DOI: https://doi.org/10.1509/jmkr.38.2.143.18840

Bryk, A. S. & **Raudenbush, S. W.** (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, 104(3), 396. DOI: https://doi.org/10.1037/0033-2909.104.3.396

Carroll, R. J. & **Schneider, H.** (1985). A note on Levene's tests for equality of variances. *Statistics & Probability Letters*, 3(4), 191–194. DOI: https://doi.org/10.1016/0167-7152(85)90016-1

Cowdry, R. W., Gardner, D. L., O'Leary, K. M., Leibenluft, E. & **Rubinow, D. R.** (1991). Mood variability: A study of four groups. *American Journal of Psychiatry*, 148(11), 1505–1511. DOI: https://doi.org/10.1176/ajp.148.11.1505

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* Routledge. Retrieved from https://books.google.fr/books?hl=fr&lr=&id=1W6laNc7Xt8C&oi=fnd&pg=PR1&dq=understanding+the+new+statistics:+effect+sizes,+confidence+intervals,+and+meta-analysis&ots=PuJZVHb03Q&sig=lhSjkfzp4o5OXAKhZ_zYzP9nsr8.

Erceg-Hurn, D. M. & **Mirosevich, V. M.** (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63(7), 591. DOI: https://doi.org/10.1037/0003-066X.63.7.591

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research*, 62(1), 61–84. DOI: https://doi.org/10.3102/00346543062001061

Glass, G. V., Peckham, P. D. & **Sanders, J. R.** (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42(3), 237–288. DOI: https://doi.org/10.3102/00346543042003237

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, 68(1), 155. DOI: https://doi.org/10.1037/0022-006X.68.1.155

Hanushek, E. A. & **Wößmann, L.** (2006). Does educational tracking affect performance and inequality? Differences-in-differencesevidenceacrosscountries*. *Economic Journal*, 116(510), C63–C76. DOI: https://doi.org/10.1111/j.1468-0297.2006.01076.x

Hayes, A. F. & **Cai, L.** (2007). Further evaluating the conditional decision rule for comparing two independent means. *British Journal of Mathematical and Statistical Psychology*, 60(2), 217–244. DOI: https://doi.org/10.1348/000711005X62576

Henrich, J., Heine, S. J. & **Norenzayan, A.** (2010). Most people are not WEIRD. *Nature*, 466(7302), 29–29. DOI: https://doi.org/10.1038/466029a

Hoekstra, R., Kiers, H. & **Johnson, A.** (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, 3, 137. DOI: https://doi.org/10.3389/fpsyg.2012.00137

Keppel, G. (1991). Design and analysis: A researcher's handbook. Prentice-Hall, Inc. Retrieved from http://psycnet.apa.org/psycinfo/1991-98751-000.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Levin, J. R., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68(3), 350–386. DOI:https://doi.org/10.3102/00346543068003350

Keselman, H. J., Othman, A. R., Wilcox, R. R. & **Fradette, K.** (2004). The new and improved two-sample *t* test. *Psychological Science*, 15(1), 47–51. DOI: https://doi.org/10.1111/j.0963-7214.2004.01501008.x

Kester, S. W. (1969). The communication of teacher expectations and their effects on the achievement and attitudes of secondary school pupils. University of Oklahoma. Retrieved from https://shareok.org/handle/11244/2570.

Lakens, D., Hilgard, J. & **Staaks, J.** (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, 4(1), 1. DOI: https://doi.org/10.1186/s40359-016-0126-3

Lim, T.-S. & **Loh, W.-Y.** (1996). A comparison of tests of equality of variances. *Computational Statistics & Data Analysis*, 22(3), 287–301. DOI: https://doi.org/10.1016/0167-9473(95)00054-2

Luh, W.-M. & **Guo, J.-H.** (2007). Approximate sample size formulas for the two-sample trimmed mean test with unequal variances. *British Journal of Mathematical and Statistical Psychology*, 60(1), 137–146. DOI: https://doi.org/10.1348/000711006X100491

Moser, B. K. & **Stevens, G. R.** (1992). Homogeneity of variance in the two-sample means test. *American Statistician*, 46(1), 19–21. DOI: https://doi.org/10.1080/00031305.1992.10475839

Nachar, N. (2008). The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, 4(1), 13–20. DOI: https://doi.org/10.20982/tqmp.04.1.p013

Neuhäuser, M. & **Ruxton, G. D.** (2009). Distribution-free two-sample comparisons in the case of heterogeneous variances. *Behavioral Ecology and Sociobiology*, 63(4), 617–623. DOI: https://doi.org/10.1007/s00265-008-0683-4

Nimon, K. F. (2012). Statistical assumptions of substantive analyses across the general linear model: A mini-review. *Frontiers in Psychology*, 3, 322. DOI: https://doi.org/10.3389/fpsyg.2012.00322

Nordstokke, D. W. & **Zumbo, B. D.** (2007). A Cautionary Tale about Levene's Tests for Equal Variances. *Journal of Educational Research & Policy Studies*, 7(1), 1–14.

Overall, J. E., Atlas, R. S. & **Gibson, J. M.** (1995). Tests that are robust against variance heterogeneity in k × 2 designs with unequal cell frequencies. *Psychological Reports*, 76(3), 1011–1017. DOI: https://doi.org/10.2466/pr0.1995.76.3.1011

Rakotomalala, R. (2008). Comparaison de populations. *Tests Non Paramétriques*, Université Lumiere Lyon, 2. Retrieved from http://www.academia.edu/download/44989200/Comp_Pop_Tests_Nonparametriques.pdf.

Rasch, D., Kubinger, K. D. & **Moder, K.** (2011). The two-sample *t* test: Pre-testing its assumptions does not pay off. *Statistical Papers*, 52(1), 219–231. DO: https://doi.org/10.1007/s00362-009-0224-x

Ruxton, G. D. (2006). The unequal variance *t*-test is an underused alternative to Student's *t*-test and the Mann–Whitney U test. *Behavioral Ecology*, 17(4), 688–690. DOI: https://doi.org/10.1093/beheco/ark016

Shaw, R. G. & **Mitchell-Olds, T.** (1993). ANOVA for unbalanced data: An overview. *Ecology*, 74(6), 1638–1645. DOI: https://doi.org/10.2307/1939922

Sheskin, D. J. (2003). *Handbook of parametric and nonparametric statistical procedures* (3rd ed.). Boca Raton, Florida: CRC Press. DOI: https://doi.org/10.1201/9781420036268

Shields, S. (1975). Functionalism, Darwinism, and the psychology of women. *American Psychologist*, 30(7), 739. DOI: https://doi.org/10.1037/h0076948

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

Wallenstein, S., Zucker, C. L. & **Fleiss, J. L.** (1980). Some statistical methods useful in circulation research. *Circulation Research*, 47(1), 1–9. DOI: https://doi.org/10.1161/01.RES.47.1.1

Wang, H., Smith, K. P., Combs, E., Blake, T., Horsley, R. D. & **Muehlbauer, G. J.** (2012). Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. *Theoretical and Applied Genetics*, 124(1), 111–124. DOI: https://doi.org/10.1007/s00122-011-1691-8

Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika*, 34(1/2), 28–35. DOI: https://doi.org/10.2307/2332510

Wilcox, R. R., Granger, D. A. & **Clark, F.** (2013). Modern robust statistical methods: Basics with illustrations using psychobiological data. *Universal Journal of Psychology*, 1(2), 21–31.

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

Yuen, K. K. (1974). The two-sample trimmed *t* for unequal population variances. *Biometrika*, 61(1), 165–170. DOI: https://doi.org/10.1093/biomet/61.1.165

Zimmerman, D. W. (1996). Some properties of preliminary tests of equality of variances in the two-sample location problem. *Journal of General Psychology*, 123(3), 217–231. DOI: https://doi.org/10.1080/00221309.1996.9921274

Zimmerman, D. W. (2000). Statistical significance levels of nonparametric tests biased by heterogeneous variances of treatment groups. *Journal of General Psychology*, 127(4), 354–364. DOI: https://doi.org/10.1080/00221300009598589

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57(1), 173–181. DOI: https://doi.org/10.1348/000711004849222

Zumbo, B. D. & **Coulombe, D.** (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 51(2), 139. DOI: https://doi.org/10.1037/1196-1961.51.2.139

How to cite this article: Delacre, M., Lakens, D. and Leys, C. (2017). Why Psychologists Should by Default Use Welch's *t*-test Instead of Student's *t*-test. *International Review of Social Psychology*, 30(1), 92–101, DOI: https://doi.org/10.5334/irsp.82

Published: 05 April 2017

]u[*International Review of Social Psychology* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS

INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY

Delacre, M., et al. (2019). Taking Parametric Assumptions Seriously: Arguments for the Use of Welch's *F*-test instead of the Classical *F*-test in One-Way ANOVA. *International Review of Social Psychology*, 32(1): 13, 1–12. DOI: https://doi.org/10.5334/irsp.198

RESEARCH ARTICLE

Taking Parametric Assumptions Seriously: Arguments for the Use of Welch's *F*-test instead of the Classical *F*-test in One-Way ANOVA

Marie Delacre*, Christophe Leys*, Youri L. Mora* and Daniël Lakens†

Student's *t*-test and classical *F*-test ANOVA rely on the assumptions that two or more samples are independent, and that independent and identically distributed residuals are normal and have equal variances between groups. We focus on the assumptions of normality and equality of variances, and argue that these assumptions are often unrealistic in the field of psychology. We underline the current lack of attention to these assumptions through an analysis of researchers' practices. Through Monte Carlo simulations, we illustrate the consequences of performing the classic parametric *F*-test for ANOVA when the test assumptions are not met on the Type I error rate and statistical power. Under realistic deviations from the assumption of equal variances, the classic *F*-test can yield severely biased results and lead to invalid statistical inferences. We examine two common alternatives to the *F*-test, namely the Welch's ANOVA (*W*-test) and the Brown-Forsythe test (*F**-test). Our simulations show that under a range of realistic scenarios, the *W*-test is a better alternative and we therefore recommend using the *W*-test by default when comparing means. We provide a detailed example explaining how to perform the *W*-test in SPSS and R. We summarize our conclusions in practical recommendations that researchers can use to improve their statistical practices.

Keywords: ANOVA; Welch test; parametric test; parametric assumptions; replicability crisis

When comparing independent groups researchers often analyze the means by performing a Student's *t*-test or classical Analysis of Variance (ANOVA) *F*-test (Erceg-Hurn & Mirosevich, 2008; Keselman et al., 1998; Tomarken & Serlin, 1986). Both tests rely on the assumptions that independent and identically distributed residuals (1) are sampled from a normal distribution and (2) have equal variances between groups (or homoscedasticity; see Lix, Keselman, & Keselman, 1996). While a deviation from the normality assumption generally does not strongly affect either the Type I error rates (Glass, Peckham, & Sanders, 1972; Harwell, Rubinstein, Hayes, & Olds, 1992; Tiku, 1971) or the power of the *F*-test (David & Johnson, 1951; Harwell et al., 1992; Srivastava, 1959; Tiku, 1971), the *F*-test is not robust against unequal variances (Grissom, 2000). Unequal variances can alter both the Type I error rate (David & Johnson, 1951; Harwell et al., 1992) and statistical power (Nimon, 2012; Overall, Atlas, & Gibson, 1995) of the *F*-test.

* Université Libre de Bruxelles, Service of Analysis of the Data (SAD), Bruxelles, BE

† Eindhoven University of Technology, Human Technology Interaction Group, Eindhoven, NL

Corresponding author: Marie Delacre (marie.delacre@ulb.ac.be)

Although it is important to make sure test assumptions are met before a statistical test is performed, researchers rarely provide information about test assumptions when they report an *F*-test. We examined statistical tests reported in 116 articles in the *Journal of Personality and Social Psychology* published in 2016. Fourteen percent of these articles reported a one-way *F*-test, but only one article indicated that the homogeneity of variances assumption was taken into account. They reported corrected degrees of freedom for unequal variances, which could signal the use of the *W*-test instead of the classical *F*-test. A similar investigation (Hoekstra, Kiers & Johnson, 2012) yielded conclusions about the lack of attention to both the homoscedasticity and the normality assumptions. Despite the fact that the *F*-test is currently used by default, better alternatives exist, such as the Welch's *W* ANOVA (*W*-test), the Alexander-Govern test, James' second order test, and the Brown-Forsythe ANOVA (*F**-test). Although not the focus of the current article, additional tests exist that allow researchers to compare groups either based on other estimators of central tendency than the mean (see for example Erceg-Hurn & Mirosevich, 2008; Wilcox, 1998), or based on other relevant parameters of distribution than the central tendency, such as standard deviations and the shape of the distribution (Grissom, 2000; Tomarken & Serlin, 1986). However, since most

researchers currently generate hypotheses about differences between means (Erceg-Hurn & Mirosevich, 2008; Keselman et al., 1998), we think that a first realistic first step towards progress would be to get researchers to correctly test the hypothesis they are used to.

Although the debate surrounding the assumptions of the *F*-test has been widely explored (see for example the meta-analysis of Harwell et al., 1992), applied researchers still largely ignore the consequences of assumption violations. Non-mathematical pedagogical papers summarizing the arguments seem to be lacking from the literature, and the current paper aims to fill this gap. We will discuss the pertinence of the assumptions of the *F*-test, and focus on the question of heteroscedasticity (that, as we will see, can have major consequences on error rates). We will provide a non-mathematical explanation of how alternatives to the classical *F*-test cope with heteroscedasticity violations. We conducted simulations in which we compare the *F*-test with the most promising alternatives. We argue that when variances are equal between groups, the *W*-test has nearly the same empirical Type I error rate and power as the *F*-test, but when variances are unequal, it provides empirical Type I and Type II error rates that are closer to the expected levels compared to the *F*-test. Since the *W*-test is available in practically all statistical software packages, researchers can immediately improve their statistical inferences by replacing the *F*-test by the *W*-test.

Normality and Homogeneity of Variances under Ecological Conditions

For several reasons, assumptions of homogeneity of variances and normality are always more or less violated (Glass et al., 1972). In this section we will summarize the specificity of the methods used in our discipline that can account for this situation.

Normality Assumption

It has been argued that there are many fields in psychology where the assumption of normality does not hold (Cain, Zhang & Yuan, 2017; Micceri, 1989; Yuan, Bentler & Chan, 2004). As argued by Micceri (1989), there are several factors that could explain departures from the normality assumption, and we will focus on three of them: treatment effects, the presence of subpopulations, and the bounded measures underlying residuals.

First, although the mean can be influenced by the treatment effects, experimental treatment can also change the shape of a distribution, either by influencing the *skewness*, quantifying the asymmetry of the shape of the distribution, and *kurtosis*, a measure of the tendency to produce extreme values. A distribution with positive kurtosis will have heavier tails than the normal distribution, which means that extreme values will be more likely, while a distribution with negative kurtosis will have lighter tails than the normal distribution, meaning that extreme values will be less likely (Westfall, 2014; Wilcox, 2005). For example, a training aiming at reducing a bias perception of threat when being exposed to ambiguous words will not uniformly impact the perception of all participants, depending on their level of anxiety (Grey & Mathews, 2000). This could influence the kurtosis of the distribution of bias score.

Second, prior to any experimental treatment, the presence of several subpopulations may lead to departures from the normality assumptions. A subgroup might exist that is unequal on some characteristics relevant to the measurements, that are not controlled within the studied group, which results in mixed distributions. This unavoidable lack of control is inherent of our field given its complexity. As an illustration, Wilcox (2005) writes that pooling two normally-distributed populations that have the same mean but different variances (e.g. normally distributed scores for schizophrenic and not schizophrenic participants) could result in distributions that are very similar to the normal curve, but with thicker tails. As another example, when assessing a wellness score for the general population, data may be sampled from a left-skewed distribution, because most people are probably not depressed (see Heun et al., 1999). In this case, people who suffer from depression and people who do not suffer from depression are part of the same population, which can leads to asymmetry in the distribution.

Third, bounded measures can also explain non-normal distributions. For example, response time can be very large, but never below zero, which results in right-skewed distributions. In sum, there are many common situations in which normally distributed data is an unlikely assumption.

Homogeneity of Variances Assumption

Homogeneity of variances (or homoscedasticity) is a mathematical requirement that is also ecologically unlikely (Erceg-Hurn & Mirosevich, 2008; Grissom, 2000). In a previous paper (Delacre, Lakens & Leys, 2017), we identified three different causes of heteroscedasticity: the variability inherent to the use of measured variables, the variability induced by quasi-experimental treatments on measured variables, and the variability induced by different experimental treatments on randomly assigned subjects. One additional source of variability is the presence of unidentified moderators (Cohen et al., 2013).

First, psychologists, as many scholars from various fields in human sciences, often use measured variables (e.g. age, gender, educational level, ethnic origin, depression level, etc.) instead of random assignment to conditions. Prior to any treatment, parameters of pre-existing groups can vary largely from one population to another, as suggested by Henrich, Heine, and Norenzayan (2010). For example, Green, Deschamps, and Páez (2005) have shown that the scores of competitiveness, self-reliance and interdependence are more variable in some ethnic groups than in others. This stands true for many pre-existing groups such as gender, cultures, or religions and for various outcomes (see for example Adams et al., 2014; Beilmann et al., 2014; Church et al., 2012; Cohen & Hill, 2007; Haar et al., 2014; Montoya & Briggs, 2013). Moreover, groups are sometimes defined with the intention to have different variabilities. For example, as soon as a selective school admits its students based on the results of aptitude tests,

the variability will be smaller compared to a school that accepts all students.

Second, a quasi-experimental treatment can have different impacts on variances between pre-existing groups, that can even be of theoretical interest. For example, in the field of linguistics and social psychology, Wasserman and Weseley (2009) investigated the impact of language gender structure on sexist attitudes of women and men. They tested differences between sexist attitude scores of subjects who read a text in English (i.e. a language without grammatical gender) or in Spanish (i.e. a language with grammatical gender). The results showed that (for a reason not explained by the authors), the women's score on the sexism dimension was more variable when the text was read in Spanish than in English ($SD_{spanish} = .80 > SD_{english} = .50$). For men, the reverse was true ($SD_{spanish} = .97 < SD_{english} = 1.33$).[1]

Third, even when the variances of groups are the same before treatment (due to a complete succesful randomization in group assignment), unequal variances can emerge later, as a consequence of an experimental treatment (Box, 1954; Bryk & Raudenbush, 1988; Cumming, 2005; Erceg-Hurn & Mirosevich, 2008; Keppel & Wickens, 2004). For example, Koeser and Sczesny (2014) have compared arguments advocating either masculine generic or gender-fair language with control messages in order to test the impact of these conditions on the use of gender-fair wording (measured as a frequency). They report that the standard deviations increase after treatment in all experimental conditions.

Consequences of Assumption Violations

Assumptions violations would not be a matter per se, if the *F*-test was perfectly robust against departures from them (Glass et al., 1972). When performing a test, two types of errors can be made: Type I errors and Type II errors. A Type I error consists of falsely rejecting the null hypothesis in favour of an alternative hypothesis, and the Type I error rate (α) is the proportion of tests that, when sampling many times from the same population, reject the null hypothesis when there is no true effect in the population. A Type II error consists of failing to reject the null hypothesis, and the Type II error rate (β) is the proportion of tests, when sampling many times from the same population, that fail to reject the null hypothesis when there is a true effect. Finally, the statistical power ($1 - \beta$) is the proportion of tests, when sampling many times from the same population, that correctly reject the null hypothesis when there is a true effect in the population.

Violation of the Normality Assumption

Regarding the Type I error rate, the shape of the distribution has very little impact on the *F*-test (Harwell et al., 1992). When departures are very small (i.e. a kurtosis between 1.2 and 3 or a skewness between –0.4 and 0.4), the Type I error rate of the *F*-test is very close to expectations, even with sample sizes as small as 11 subjects per group (Hsu & Feldt, 1969).

Regarding the Type II error rate, many authors underlined that departures from normality do not seriously affect the power (Boneau, 1960; David & Johnson, 1951; Glass et al., 1972; Harwell et al., 1992; Srivastava, 1959; Tiku, 1971). However, we can conclude from Srivastava (1959) and Boneau (1960) that kurtosis has a slightly larger impact on the power than skewness. The effect of non-normality on power increases when sample sizes are unequal between groups (Glass et al., 1972). Lastly the effect of non-normality decreases when sample sizes increase (Srivastava, 1959).

Violation of Homogeneity of Variances Assumption

Regarding the Type I error rate, the *F*-test is sensitive to unequal variances (Harwell et al., 1992). More specifically, the more unequal the *SD* of the population's samples are extracted from, the higher the impact. When there are only two groups, the impact is smaller than when there are more than two groups (Harwell et al., 1992). When there are more than two groups, the *F*-test becomes more liberal, meaning that the Type I error rate is larger than the nominal alpha level, even when sample sizes are equal across groups (Tomarken & Serlin, 1986). Moreover, when sample sizes are unequal, there is a strong effect of the sample size and variance pairing. In case of a positive pairing (i.e. the group with the larger sample size also has the larger variance), the test is too conservative, meaning that the Type I error rate of the test is lower than the nominal alpha level, whereas in case of a negative pairing (i.e. the group with the larger sample size has the smaller variance), the test is too liberal (Glass et al., 1972; Nimon, 2012; Overall et al., 1995; Tomarken & Serlin, 1986).

Regarding the Type II error rate, there is a small impact of unequal variances when sample sizes are equal (Harwell et al., 1992), but there is a strong effect of the sample size and variance pairing (Nimon, 2012; Overall et al., 1995). In case of a positive pairing, the Type II error rate increases (i.e. the power decreases), and in case of a negative pairing, the Type II error decreases (i.e. the power increases).

Cumulative Violation of Normality and Homogeneity of Variance

Regarding both Type I and Type II error rates, following Harwell et al. (1992), there is no interaction between normality violations and unequal variances. Indeed, the effect of heteroscedasticity is relatively constant regardless of the shape of the distribution.

Based on mathematical explanations and Monteo Carlo simulations we chose to compare the *F*-test with the *W*-test and *F**-test and to exclude the James' second-order and Alexander-Govern's test because the latter two yield very similar results to the *W*-test, but are less readily available in statistical software packages. Tomarken and Serlin (1986) have shown that from the available alternatives, the *F**-test and the *W*-test perform best, and both tests are available in SPSS, which is widely used software in the psychological sciences (Hoekstra et al., 2012). For a more extended description of the James' second-order and Alexander-Govern's test, see Schneider and Penfield (1997).

The Mathematical Differences Between the *F*-test, *W*-test, and *F**-test

The mathematical differences between the *F*-test, *W*-test and *F**-test can be explained by focusing on how standard deviations are pooled across groups. As shown in (1) the *F* statistic is calculated by dividing the inter-group variance by a pooled error term, where s_j^2 and n_j are respectively the variance estimates and the sample sizes from each independent group, and where k is the number of independent groups:

$$F = \frac{\frac{1}{k-1}\sum_{j=1}^{k}\left[n_j(\bar{X}_j - \bar{X}_{..})^2\right]}{\frac{1}{N-k}\sum_{j=1}^{k}(n_j-1)s_j^2} \quad (1)$$

The degrees of freedom in the numerator (2) and in the denominator (3) of the *F*-test are computed as follows:

$$df_n = k-1 \quad (2)$$

$$df_d = N-k, \quad (3)$$

With $N = \sum_{j=1}^{k} n_j$. As a generalization of the Student's *t*-test, the *F*-test is calculated based on a pooled error term. This implies that all samples are considered as issued from a common population variance (hence the assumption of homoscedasticity). When there is heteroscedasticity, and if the larger variance is associated with the larger sample size, the error term, which is the denominator in (1), is overestimated. The *F*-value is therefore smaller, leading to fewer significant findings than expected, and the *F*-test is too conservative. When the larger variance is associated with the smaller sample size the denominator in (1) is underestimated. The *F*-value is then inflated, which yields more significant results than expected.

The *F** statistic proposed by Brown and Forsythe (1974) is computed as follows:

$$F^* = \frac{\sum_{j=1}^{k}\left[n_j\left(\bar{X}_j - \bar{X}_{..}\right)^2\right]}{\sum_{j=1}^{k}\left[\left(1-\frac{n_j}{N}\right)s_j^2\right]} \quad (4)$$

Where x_j and s_j^2 are respectively the group mean and the group variance, and $\bar{X}_{..}$ is the overall mean. As it can be seen in (4) the numerator of the *F** statistic is equal to the sum of squares between groups (which is equal to the numerator of the *F* statistic when one compares two groups). In the denominator, the variance of each group is weighted by 1 minus the relative frequency of each group. This adjustment implies that the variance associated with the group with the smallest sample size is given more weight compared to the *F*-test. As a result, when the larger variance is associated with the larger sample size, *F** is larger than *F*, because the denominator decreases, leading to more significant findings compared to the *F*-test. On the other hand, when the larger variance is associated with the smaller sample size, *F** is smaller than *F*, because the denominator increases, leading to fewer significant findings compared to the *F*-test. The degrees of freedom in the numerator and in the denominator of *F**-test are computed as follows (with the same principle as the denominator computation of the *F** statistic):

$$df_n = k-1 \quad (5)$$

$$df_d = \frac{1}{\sum_{j=1}^{k}\left[\frac{\left(\frac{\left(1-\frac{n_j}{N}\right)s_j^2}{\sum_{j=1}^{k}\left[\left(1-\frac{n_j}{N}\right)s_j^2\right]}\right)^2}{n_j-1}\right]} \quad (6)$$

Formula (7) provides the computation of the *W*-test, or Welch's *F*-test. In the numerator of the *W*-test the squared deviation between group means and the general mean are weighted by $\frac{n_j}{s_j^2}$ instead of n_j (Brown & Forsythe, 1974). As a consequence, for equal sample sizes, the group with the highest variance will have smaller weight (Liu, 2015).

$$W = \frac{\frac{1}{k-1}\sum_{j=1}^{k}\left[w_j\left(\bar{X}_j - \bar{X}'\right)^2\right]}{1+\frac{2(k-2)}{k^2-1}\sum_{j=1}^{k}\left[\left(\frac{1}{n_j-1}\right)\left(1-\frac{w_j}{w}\right)^2\right]} \quad (7)$$

where:

$$w_j = \frac{n_j}{s_j^2}$$

$$w = \sum_{j=1}^{k}\left(\frac{n_j}{s_j^2}\right)$$

$$\bar{X}' = \frac{\sum_{j=1}^{k}\left(w_j\bar{X}_j\right)}{w}$$

The degrees of freedom of the *W*-test are approximated as follows:

$$df_n = k-1 \quad (8)$$

$$df_d = \frac{k^2-1}{3\sum_{j=1}^{k}\left[\frac{(1-\frac{w_j}{w})^2}{n_j-1}\right]} \quad (9)$$

When there are only two groups to compare, the *F**-test and *W*-test are identical (i.e., they have exactly the same statistical value, degrees of freedom and significance). However, when there are more than two groups to compare, the tests differ. In the appendix we illustrate the calculation of all three statistics in detail for a fictional three-group design for educational purposes.

Monte Carlo simulations: *F*-test versus *W*-test versus *F**-test

We performed Monte Carlo simulations using R (version 3.5.0) to assess the Type I and Type II error rates for the three tests. One million datasets were generated for 3840 scenarios that address the arguments present in the literature. In 2560 scenarios, means were equal across all groups (i.e. the null hypothesis is true), in order to assess the Type I error rate of the tests. In 1280 scenarios, there were differences between means (i.e. the alternative hypothesis is true) in order to assess the power of the tests. In all scenarios, when using more than 2 samples, all samples but one was generated from the same population, and only one group had a different population mean.

Population parameter values were chosen in order to illustrate the consequences of factors known to play a key role on both the Type I error rate and the statistical power when performing an ANOVA. Based on the literature review presented above, we manipulated the number of groups, the sample sizes, the sample size ratio (n-ratio = $\frac{n_k}{n_j}$), the SD-ratio (SD-ratio = $\frac{\sigma_k}{\sigma_j}$), and the sample size and variance pairing. In our scenarios, the number of compared groups (k) varied from 2 to 5. Sample sizes of k-1 groups (n_j) were 20, 30, 40, 50, or 100. The sample size of the last group was a function of the n-ratio, ranging from 0.5 to 2, in steps of 0.5. The simulations for which the n-ratio equals 1 are known as a balanced design (i.e. sample sizes are equal across all groups). The SD of the population from which was extracted last group was a function of the SD-ratio, with values of 0.5, 1, 2 or 4. The simulations for which the SD-ratio equals 1 are the particular case of homoscedasticity (i.e. equal variances across groups).

All possible combinations of n-ratio and SD-ratio were performed in order to distinguish positive pairings (the group with the largest sample size is extracted from the population with the largest SD), negative pairings (the group with the smallest sample size is extracted from the population with the smallest SD), and no pairing (sample sizes and/or population SD are equal across all groups). All of those conditions were tested with normal and non-normal distributions. When two groups are compared, conclusions for the three ANOVA tests (F, F^*, W) should yield identical error rates when compared to their equivalent t-tests (the F-test is equivalent to Student's t-test, and the F^*-test and W-test are equivalent to Welch's t-test; Delacre et al., 2017). When there are more than three groups, the F-test becomes increasingly liberal as soon as the variances of the distributions in each group are not similar, even when sample sizes are equal between groups (Harwell et al., 1992; Quensel, 1947).

For didactic reasons, we will report only the results where we compared three groups (k = 3). Increasing the number of groups increases how liberal all tests are. For interested readers, all figures for cases where we compare more than three groups are available here: https://osf.io/h4ks8/. Overall, the larger the sample sizes, the less the distributions of the population underlying the samples impact the robustness of the tests (Srivastava, 1959). However, increasing the sample sizes does not improve the robustness of the test when there is heteroscedasticity. Interested reader can see all details in the following Excel spreadsheet, available on github: « Type I error rate.xlsx ».

In sum, the simulations grouped over different sample sizes yield 9 conditions based on the n-ratio, SD-ratio, and sample size and variance pairing, as summarized in **Table 1**.

In all Figures presented below, averaged results for each sub-condition are presented under seven different configurations of distributions, using the following legend.

Type I Error Rate of the *F*-test, *W*-test, and *F**-test

As previously mentioned, the Type I error rate (α) is the long-run frequency of observing significant results when the null-hypothesis is true. When means are equal across all groups the Type I error rate of all test should be equal to the nominal alpha level. We assessed the Type I error rate of the F-test, W-test and F^*-test under 2560 scenarios using a nominal alpha level of 5%.

When there is no difference between means, the nine cells of **Table 1** simplify into five sub-conditions:

- Equal n and SD across groups (a)
- Unequal n but equal SD across groups (b and c)
- Unequal SD but equal n across groups (d and g)
- Unequal n and SD across groups, with positive correlation between n and SD (e and i)
- Unequal n and SD across groups, with negative correlation between n and SD (f and h)

Table 1: 9 conditions based on the n-ratio, SD-ratio, and sample size and variance pairing.

		***n*-ratio**		
		1	**>1**	**<1**
SD-ratio	1	a	b	c
	>1	d	e	f
	<1	g	h	i

Note: The n-ratio is the sample size of the last group divided by the sample size of the first group. When all sample sizes are equal across groups, the n-ratio equals 1. When the sample size of the last group is higher than the sample size of the first group, n-ratio >1, and when the sample size of the last group is smaller than the sample size of the first group, n-ratio <1. SD-ratio is the population SD of the last group divided by the population SD of the first group. When all samples are extracted from populations with the same SD, the SD-ratio equals 1. When the last group is extracted from a population with a larger SD than all other groups, the SD-ratio >1. When the last group is extracted from a population with a smaller SD than all other groups, the SD-ratio <1.

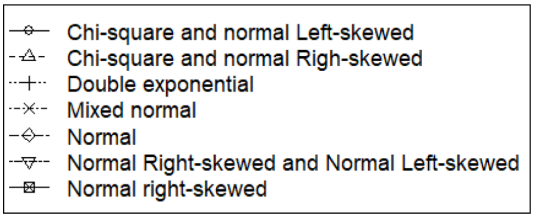

Figure 1: Legend.

In **Figures 2** to **6** (see **Figure 1** for the legend), we computed the average Type I error rate of the three tests under these five subcategories. The light grey area corresponds to the liberal criterion from Bradley (1978), who regards a departure from the nominal alpha level as acceptable whenever the Type I error rate falls within the interval [0.5 × α; 1.5 × α]. The dark grey area corresponds to the more conservative criterion from which departures from the nominal alpha is considered negligible as long as the Type I error rate falls within the interval [0.9 × α; 1.1 × α].

In **Figures 2** and **3** (cells a, b, and c in **Table 1**), the population variance is equal between all groups, so the homoscedasticity assumption is met. The *F*-test and *F**-test only marginally deviate from the nominal 5%, regardless of the underlying distribution and the *SD*-ratio. The *W*-test also only marginally deviates from the nominal 5%, except under asymmetry (the tests becomes a little more liberal) or extremely heavy tails (the test becomes a bit more conservative), consistently with observations in Harwell et al. (1992). However, deviations don't exceed the liberal criterion of Bradley (1978).

In **Figures 4, 5** and **6** (cells d to i, **Table 1**) the population variance is unequal between groups, so that the homoscedasticity assumption is not met. When sample sizes are equal across groups (**Figure 4**) and when there is a positive correlation between sample sizes and *SDs* (**Figure 5**), the Type I error rate of the *W*-test is closer to the nominal 5% than the Type I error rate of the *F**-test and the *F*-test, the latter which is consistently at the lower limit of the liberal interval suggested by Bradley, in line with Harwell et al. (1992), Glass et al. (1972), Nimon (2012) and Overall et al. (1995). Heteroscedasticity does not impact the Type I error rate of the *W*-test, regardless of the distribution (the order of the distribution shape remains the same in all conditions).

When there is a negative correlation between sample sizes and *SDs* (**Figure 6**), the Type I error rate of the *F**-test is slightly closer of the nominal 5% than the Type I error rate of the *W*-test, for which the distributions (more specifically, the skewness) has a larger impact on the Type I error rate than when there is homoscedasticity. This is consistent with conclusions of Lix et al. (1996) about

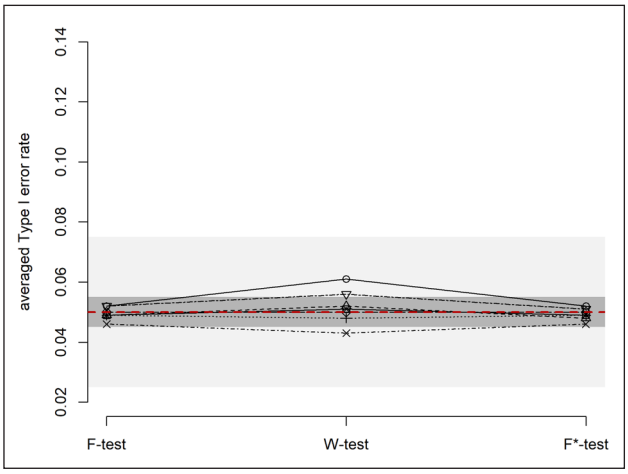

Figure 2: Type I error rate of the F-test, W-test and F*-test when there are equal SDs across groups and equal sample sizes (cell a in Table 1).

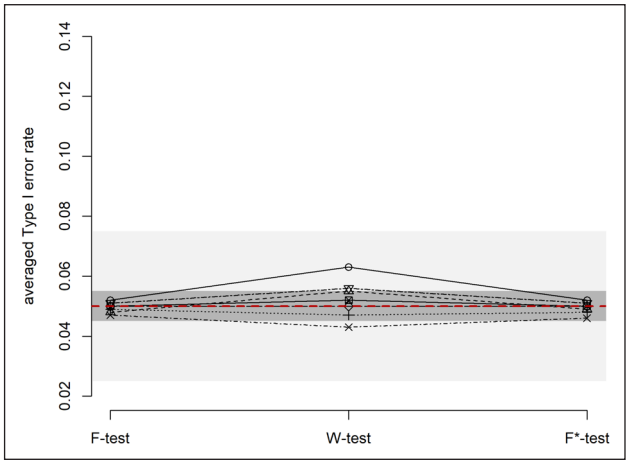

Figure 3: Type I error rate of the F-test, W-test and F*-test when there are equal SDs across groups and unequal sample sizes (cells b and c in Table 1).

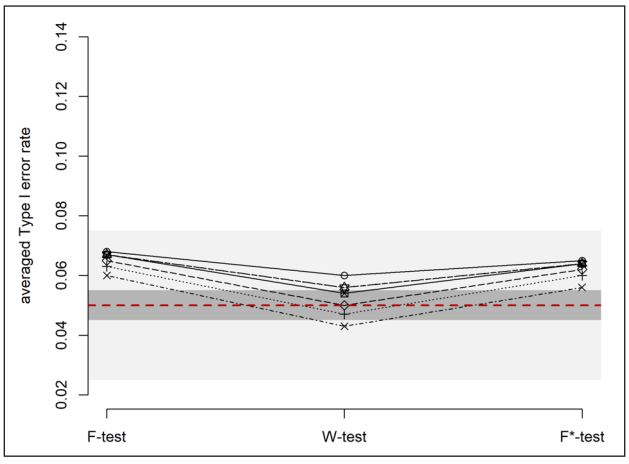

Figure 4: Type I error rate of the F-test, W-test and F*-test when there are unequal SDs across groups and equal sample sizes (cells d and g in Table 1).

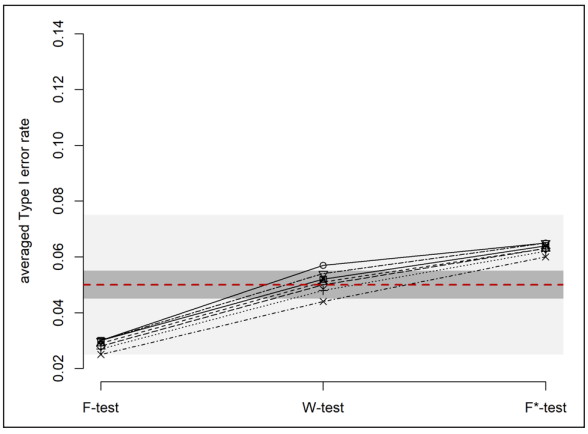

Figure 5: Type I error rate of the F-test, W-test and F*-test when there are unequal SDs across groups, and positive correlation between sample sizes and SDs (cells e and i in Table 1).

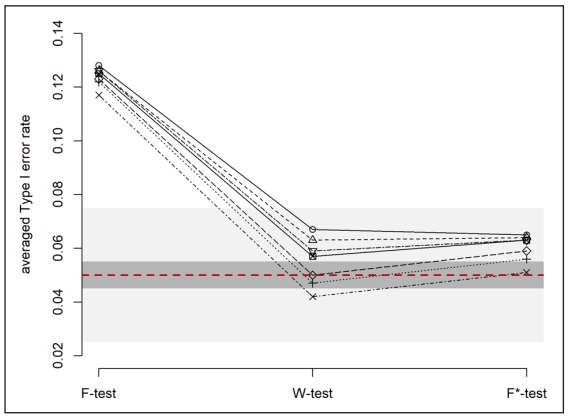

Figure 6: Type I error rate of the F-test, W-test and F*-test when there are unequal SDs across groups, and negative correlation between sample sizes and SDs (cells f and g in Table 1).

the Alexander-Govern and the James' second order tests (which return very similar results as the *W*-test, as we already mentioned). However, both tests still perform relatively well, contrary to the *F*-test that is much too liberal, in line with observations by Harwell et al. (1992), Glass et al. (1972), Nimon (2012) and Overall et al. (1995).

Conclusions

We can draw the following conclusions for the Type I error rate:

1) When all assumptions are met, all tests perform adequately.
2) When variances are equal between groups and distributions are not normal, the *W*-test is a little less efficient than both the *F*-test and the *F**-test, but departures from the nominal 5% Type I error rate never exceed the liberal criterion of Bradley (1978).
3) When the assumption of equal variances is violated, the *W*-test clearly outperforms both the *F**-test (which is more liberal) and the *F*-test (which is either more liberal or more conservative, depending on the *SDs* and *SD* pairing).
4) The last conclusion generally remains true when both the assumptions of equal variances and normality are not met.

Statistical power for the *F*-test, *W*-test, and *F**-test

As previously mentioned, the statistical power $(1 - \beta)$ of a test is the long-run probability of observing a statistically significant result when there is a true effect in the population. We assessed the power of the *F*-test, *W*-test and *F**-test under 1280 scenarios, while using the nominal alpha level of 5%. In all scenarios, the last group was extracted from a population that had a higher mean than the population from where were extracted all other groups ($\mu_k = \mu_j + 1$). Because of that, in some scenarios there is a positive correlation between the *SD* and the mean (i.e. the last group has the largest *SD* and the largest mean) and in other scenarios, there is a negative correlation between *SD* and the mean (i.e. the last group has the smallest *SD* and the largest mean). As we know that the correlation between the *SD* and the mean matters for the *W*-test (see Liu, 2015), the 9 sub-conditions in **Table 1** were analyzed separately.

We computed two main outcomes: the consistency and the power. The consistency refers to the relative difference between the observed power and the nominal power, divided by the expected power:

$$Consistency = \frac{0 - E}{E} \tag{10}$$

When consistency equals zero, the observed power is consistent with the nominal power (under the parametric assumptions of normality and homoscedasticity); a negative consistency shows that the observed power is lower than the expected power; and a positive consistency shows that the observed power is higher than the expected power.

In **Figures 7, 8** and **9** (cells a, b, and c in **Table 1** see **Figure 1** for the legend), the population variance is equal between all groups, meaning that the homoscedasticity assumption is met. When distributions are normal,

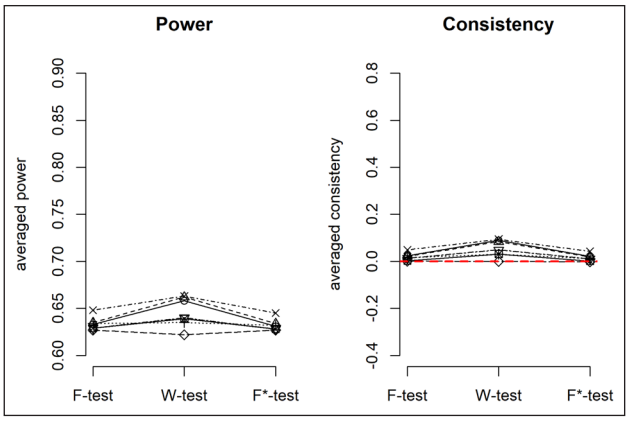

Figure 7: Power and consistency of the F-test, W-test and F*-test when there are equal SDs across groups and equal sample sizes (cell a in Table 1).

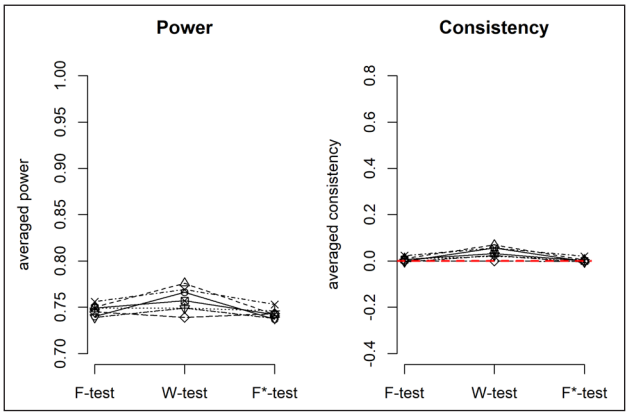

Figure 8: Power and consistency of the F-test, W-test and F*-test when there are equal SDs across groups, and positive correlation between sample sizes and means (cell b in Table 1).

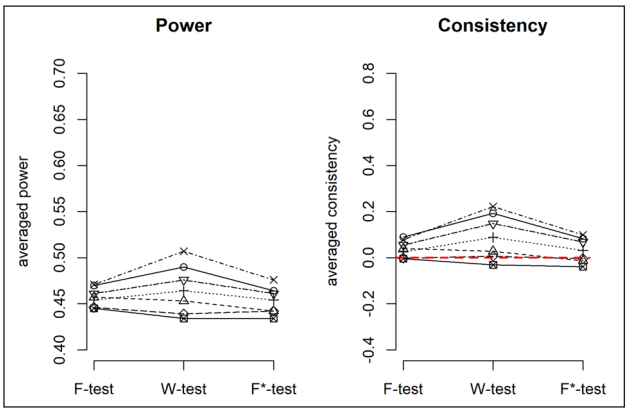

Figure 9: Power and consistency of the F-test, W-test and F*-test when there are equal SDs across groups, and negative correlation between sample sizes and means (cell c in Table 1).

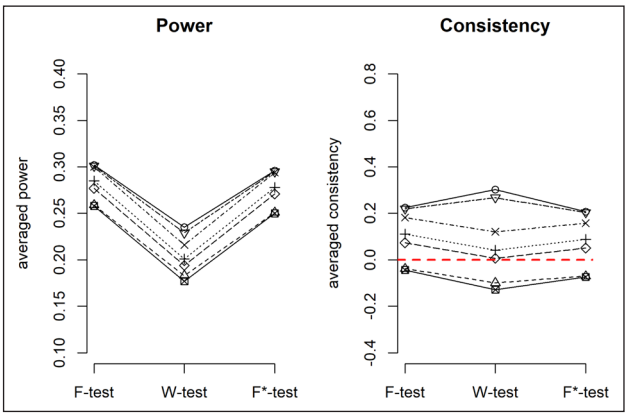

Figure 10: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, positive correlation between SDs and means, and equal sample sizes across groups (cell d in Table 1).

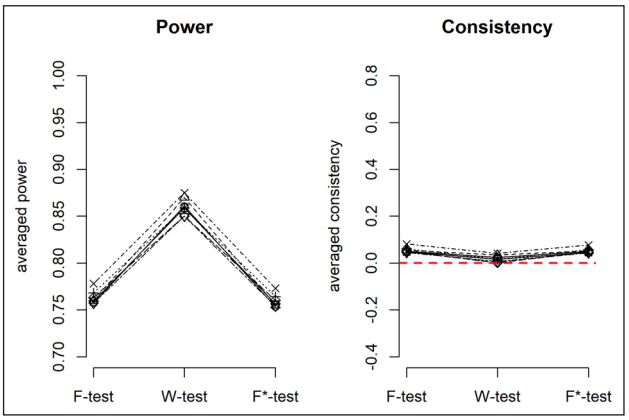

Figure 11: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, negative correlation between SDs and means, and equal sample sizes across groups (cell g in Table 1).

the *W*-test is slightly less powerful than the *F*-test and *F**-test, even though differences are very small. With all other distributions, the *W*-test is generally more powerful than the *F**-test and *F*-test, even with heavy-tailed distributions, which is in contrast with previous findings (Wilcox, 1998). Wilcox (1998) concluded that there is a loss of power when means from heavy-tailed distributions (e.g. double exponential or a mixed normal distribution) are compared to means from normal distributions. This finding is based on the argument that heavy-tailed distributions are associated with bigger standard deviations than normal distributions, and that the effect size for such distributions is therefore smaller (Wilcox, 2011). However, this conclusion is based on a common conflation of kurtosis and the standard deviation, which are completely independent (DeCarlo, 1997). One can find distributions that have similar *SD* but different kurtosis (see Appendix 2). However, while the *W*-test is more powerful than the *F*-test and the *F**-test in many situations, it is a bit less consistent with theoretical expectations than both other tests in the sense that the *W*-test is generally more powerful than expected (especially with high kurtosis, or when asymmetries go in opposite directions). This is due to the fact that the *W*-test is more impacted by the distribution shape, in line with observations by Harwell et al. (1992). Note that differences between *W*-test and other tests, in terms of consistency, are very small.

In **Figures 10** to **15** (cells d to i in **Table 1** see **Figure 1** for the legend), the population variance is unequal between groups, meaning that the homoscedasticity assumption is not met. When sample sizes are equal across groups (**Figures 10** and **11**), the *F*-test and the *F**-tests are equally powerful, and have the same consistency, whatever the correlation between the *SD* and the mean. On the other hand, the power of the *W*-test depends on the correlation between the *SD* and the mean (in line with Liu, 2015). When the group with the largest mean has the largest variance (**Figure 10**), the largest deviation between group means and the general mean is given less weight, and as a consequence the *W*-test is less powerful than both other tests. At the same time, the test is slightly less consistent than both other tests. When the group with the largest mean has the smallest variance (**Figure 11**), the largest deviation between group means and the general mean is given more weight, and therefore the *W*-test is more powerful than both other tests. The test is also slightly more consistent than both other tests.

When sample sizes are unequal across groups, the power of the *F**-test and the *F*-test are a function of the correlation between sample sizes and *SDs*. When there is a negative correlation between sample sizes and *SDs* (**Figures 12** and **13**), the *F*-test is always more powerful than the *F**-test. Indeed, as was explained in the previous mathematical section, the *F*-test gives more weight to the smallest variance (the statistic is therefore increased) while the *F**-test gives more weight to the largest variance (the statistic is therefore decreased). Conversely, when there is a positive correlation between sample sizes and *SDs* (**Figures 14** and **15**), the *F*-test is always more conservative than the *F**-test, because the *F*-test gives more weight to the largest variance while the *F**-test gives more weight to the smallest variance.

The power of the *W*-test is not a function of the correlation between sample sizes and *SDs*, but rather a function of the correlation between *SDs* and means. The test is more powerful when there is a negative correlation between *SDs* and means, and less powerful when there is a positive

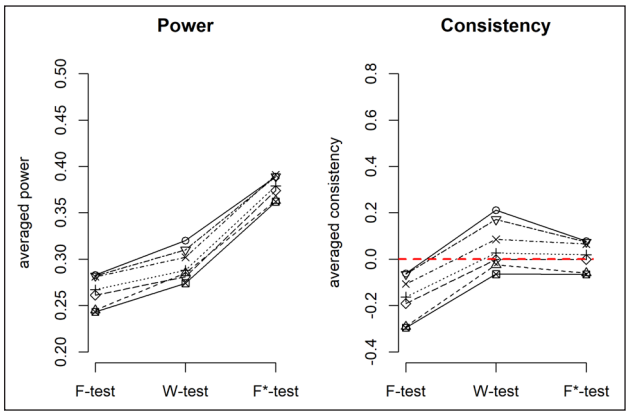

Figure 12: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, negative correlation betwen sample sizes and SDs, and positive correlation between SDs and means (cell f in Table 1).

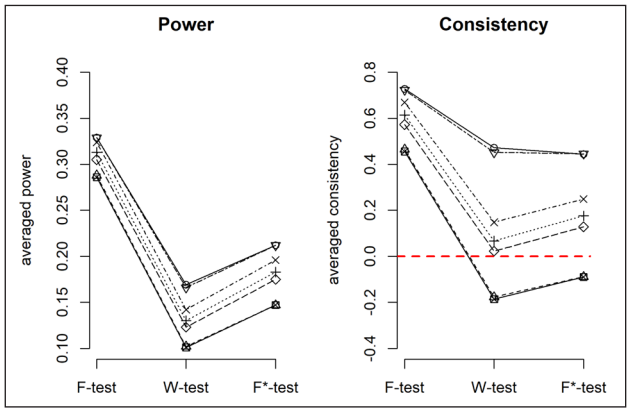

Figure 13: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, negative correlation betwen sample sizes and SDs, and negative correlation between SDs and means (cell h in Table 1).

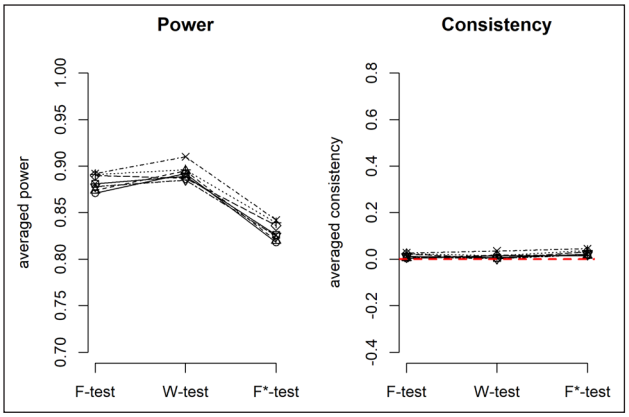

Figure 14: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, positive correlation betwen sample sizes and SDs, and positive correlation between SDs and means (cell e in Table 1).

correlation between *SDs* and means. Note that for all tests, the effect of heteroscedasticity is approximately the same regardless of the shape of the distribution. Moreover,

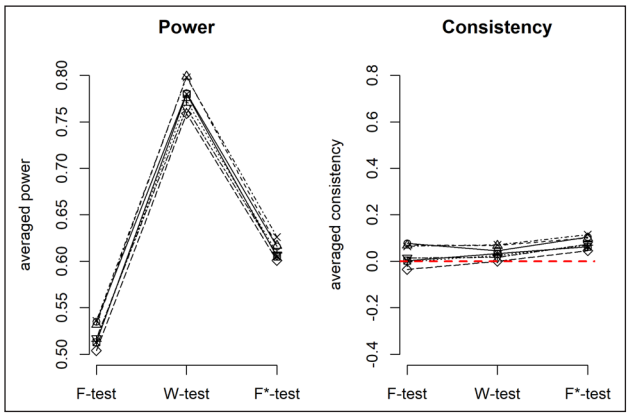

Figure 15: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, positive correlation betwen sample sizes and SDs, and negative correlation between SDs and means (cell i in Table 1).

there is one constant observation in our simulations: whatever the configuration of the *n*-ratio, the consistency of the three tests is closer to zero when there is a negative correlation between the *SD* and the mean (meaning that the group with the highest mean has the lower variance).

We can draw the following conclusions about the statistical power of the three tests:

1) When all assumptions are met, the *W*-test falls slightly behind the *F*-test and the *F**-test, both in terms of power and consistency.
2) When variances are equal between groups and distributions are not normal, the *W*-test is slightly more powerful than both the *F*-test and the *F**-test, even with heavy-tailed distributions.
3) When the assumption of equal variances is violated, the *F*-test is either too liberal or too conservative, depending on the correlation between sample sizes and *SDs*. On the other side, the *W*-test is not influenced by the sample sizes and *SDs* pairing. However, it is influenced by the *SD* and means pairing.
4) The last conclusion generally remains true when both assumptions of equal variances and normality are not met.

Recommendations

Taking both the effects of the assumption violations on the alpha risk and on the power, we recommend using the *W*-test instead of the *F*-test to compare groups means. The *F*-test and *F**-test should be avoided, because a) the equal variances assumption is often unrealistic, b) tests of the equal variances assumption will often fail to detect differences when these are present, c) the loss of power when using the *W*-test is very small (and often even negligible), and d) the gain in Type I error control is considerable under a wide range of realistic conditions. Also, we recommend the use of balanced designs (i.e. same sample sizes in each group) whenever possible. When using the *W*-test, the Type I error rate is a function of criteria such as the skewness of the distributions, and whether skewness is combined with unequal variances and unequal samples sizes between groups. Our simulations show that the Type

I error rate control is in general slightly better with balanced designs.

Note that the *W*-test suffers from limitations and cannot be used in all situations. First, as previously mentioned, *W*-test, as all tests based on means, does not allow researchers to compare other relevant parameters of a distribution than the mean. For these reason, we recommend to never neglect the descriptive analysis of the data. A complete description of the shape and characteristics of the data (e.g. histograms and boxplots) is important. When at least one statistical parameter relating to the shape of the distribution (e.g. variance, skewness, kurtosis) seems to vary between groups, comparing results of the *W*-test with results of a nonparametric procedure is useful in order to better understand the data. Second, with small sample sizes (i.e. less than 50 observations per group when comparing at most four groups, 100 observations when comparing more than four groups), the *W*-test will not control Type I error rate when skewness is present and detecting departures for normality is therefore especially important in small samples. Unless you have good reasons to believe that distributions underlying the data have small kurtosis and skewness, we recommend to avoid alternative tests that are based on means comparison, in favour of alternatives such as the trimmed means test (Erceg-Hurn & Mirosevich, 2008)[2] or nonparametric tests. For more information about robust alternatives that are based on other parameters than the mean, see Erceg-Hurn and Mirosevich (2008).

Notes

[1] Note that this is a didactic example, the differences have not been tested and might not differ statistically.

[2] The null hypothesis of the trimmed means test assumes that trimmed means are the same between groups. A trimmed mean is a mean computed on data after removing the lowest and highest values of the distribution. Trimmed means and means are equal when data are symmetric. On the other hand, when data are asymmetric, trimmed means and means differ.

Additional File

The additional file for this article can be found as follows:

- **Supplemental Materials.** A numerical example of the mathematical development of the F-test, W-test, and F*-test (Appendix 1) and justification for the choice of distributions in simulation (Appendix 2). DOI: https://doi.org/10.5334/irsp.198.s1

Competing Interests

The authors have no competing interests to declare.

Author Contribution

The first author performed simulations. The first, second and fourth authors contributed to the design. All authors contributed to the writing and the review of the literature. The Supplemental Material, including the full R code for the simulations and plots can be obtained from https://github.com/mdelacre/W-ANOVA. This work was supported by the Netherlands Organization for Scientific Research (NWO) VIDI grant 452-17-013. The authors declare that they have no conflicts of interest with respect to the authorship or the publication of this article.

References

Adams, B. G., Van de Vijver, F. J., de Bruin, G. P., & **Bueno Torres, C.** (2014). Identity in descriptions of others across ethnic groups in South Africa. *Journal of Cross-Cultural Psychology*, *45*(9), 1411–1433. DOI: https://doi.org/10.1177/0022022114542466

Beilmann, M., Mayer, B., Kasearu, K., & **Realo, A.** (2014). The relationship between adolescents' social capital and individualism-collectivism in Estonia, Germany, and Russia. *Child Indicators Research*, *7*(3), 589–611. DOI: https://doi.org/10.1007/s12187-014-9232-z

Boneau, C. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin*, *57*(1), 49–64. DOI: https://doi.org/10.1037/h0041412

Box, G. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, i. Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, *25*(2), 290–302. DOI: https://doi.org/10.1214/aoms/1177728786

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. DOI: https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Brown, M. B., & **Forsythe, A. B.** (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(346), 364–367. DOI: https://doi.org/10.2307/2285659

Bryk, A. S., & **Raudenbush, S. W.** (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, *104*(3), 396–404. DOI: https://doi.org/10.1037/0033-2909.104.3.396

Cain, M. K., Zhang, Z., & **Yuan, K.-H.** (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*(5), 1716–1735. DOI: https://doi.org/10.3758/s13428-016-0814-1

Church, A. T., Willmore, S. L., Anderson, A. T., Ochiai, M., Porter, N., Mateo, N. J., Ortiz, F. A., et al. (2012). Cultural differences in implicit theories and self-perceptions of traitedness: Replication and extension with alternative measurement formats and cultural dimensions. *Journal of Cross-Cultural Psychology*, *43*(8), 1268–1296. DOI: https://doi.org/10.1177/0022022111428514

Cohen, A. B., & **Hill, P. C.** (2007). Religion as culture: Religious individualism and collectivism among American Catholics, Jews, and Protestants. *Journal of Personality*, *75*(4), 709–742. DOI: https://doi.org/10.1111/j.1467-6494.2007.00454.x

Cohen, J., Cohen, P., West, S. G., & **Aiken, L. S.** (2013). *Applied multiple regression/correlation analysis for the behavioural sciences*. Mahwah, NJ: Erlbaum. DOI: https://doi.org/10.4324/9780203774441

Cumming, G. (2005). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

David, F. N., & **Johnson, N. L.** (1951). The effect of non-normality on the power function of the f-test in the analysis of variance. *Biometrika*, *38*(1–2), 43–57. DOI: https://doi.org/10.2307/2332316

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods*, *2*(3), 292–307. DOI: https://doi.org/10.1037//1082-989X.2.3.292

Delacre, M., Lakens, D., & **Leys, C.** (2017). Why psychologists should by default use Welch's t-test instead of student's t-test. *International Review of Social Psychology*, *30*(1), 92–101. DOI: https://doi.org/10.5334/irsp.82

Erceg-Hurn, D. M., & **Mirosevich, V. M.** (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*(7), 591–601. DOI: https://doi.org/10.1037/0003-066X.63.7.591

Glass, G. V., Peckham, P. D., & **Sanders, J. R.** (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237–288. DOI: https://doi.org/10.3102/00346543042003237

Green, E. G., Deschamps, J.-C., & **Páez, D.** (2005). Variation of individualism and collectivism within and between 20 countries: A typological analysis. *Journal of Cross-Cultural Psychology*, *36*(3), 321–339. DOI: https://doi.org/10.1177/0022022104273654

Grey, S., & **Mathews, A.** (2000). Effects of training on interpretation of emotional ambiguity. *The Quarterly Journal of Experimental Psychology*, *53*(4), 1143–1162. DOI: https://doi.org/10.1080/713755937

Grissom, R. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, *68*(1), 155–165. DOI: https://doi.org/10.1037//0022-006X.68.1.155

Haar, J. M., Russo, M., Suñe, A., & **Ollier-Malaterre, A.** (2014). Outcomes of work-life balance on job satisfaction, life satisfaction and mental health: A study across seven cultures. *Journal of Vocational Behavior*, *85*(3), 361–373. DOI: https://doi.org/10.1016/j.jvb.2014.08.010

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & **Olds, C. C.** (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects anova cases. *Journal of Educational Statistics*, *17*(4), 315–339. DOI: https://doi.org/10.3102/10769986017004315

Henrich, J., Heine, S. J., & **Norenzayan, A.** (2010). Most people are not weird. *Nature*, *466*, 29–29. DOI: https://doi.org/10.1038/466029a

Heun, R., Burkart, M., Maier, W., & **Bech, P.** (1999). Internal and external validity of the who well-being scale in the elderly general population. *Acta Psychiatrica Scandinavica*, *99*(3), 171–178. DOI: https://doi.org/10.1111/j.1600-0447.1999.tb00973.x

Hoekstra, R., Kiers, H. A., & **Johnson, A.** (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, *3*(137), 1–9. DOI: https://doi.org/10.3389/fpsyg.2012.00137

Hsu, T.-C., & **Feldt, L. S.** (1969). The effect of limitations on the number of criterion score values on the significance level of the f-test. *American Educational Research Journal*, *6*(4), 515–527. DOI: https://doi.org/10.3102/00028312006004515

Keppel, G., & **Wickens, T. D.** (2004). *Design and analysis: A researcher's handbook*. Upper Saddle River, New Jersey: Prentice Hall.

Keselman, H., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Levin, J. R., et al. (1998). Statistical practices of educational researchers: An analysis of their anova, manova, and ancova analy. *Review of Educational Research*, *68*(3), 350–386. DOI: https://doi.org/10.3102/00346543068003350

Koeser, S., & **Sczesny, S**. (2014). Promoting gender-fair language: The impact of arguments on language use, attitudes, and cognitions. *Journal of Language and Social Psychology*, *33*(5), 548–560. DOI: https://doi.org/10.1177/0261927X14541280

Liu, H. (2015). *Comparing welch anova, a kruskal-wallis test, and traditional anova in case of heterogeneity of variance* (PhD thesis). Virginia Commonwealth University.

Lix, L. M., Keselman, J. C., & **Keselman, H.** (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance *f* test. *Review of Educational Research*, *66*(4), 579–619. DOI: https://doi.org/10.3102/00346543066004579

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. DOI: https://doi.org/10.1037/0033-2909.105.1.156

Montoya, D. Y., & **Briggs, E.** (2013). Shared ethnicity effects on service encounters: A study across three us subcultures. *Journal of Business Research*, *66*(3), 314–320. DOI: https://doi.org/10.1016/j.jbusres.2011.08.011

Nimon, K. F. (2012). Statistical assumptions of substantive analyses across the general linear model: A mini-review. *Frontiers in Psychology*, *3*(322), 1–5. DOI: https://doi.org/10.3389/fpsyg.2012.00322

Overall, J. E., Atlas, R. S., & **Gibson, J. M.** (1995). Tests that are robust against variance heterogeneity in k × 2 designs with unequal cell frequencies. *Psychological Reports*, *76*(3), 1011–1017. DOI: https://doi.org/10.2466/pr0.1995.76.3.1011

Quensel, C.-E. (1947). The validity of the *z*-criterion when the variates are taken from different normal populations. *Scandinavian Actuarial Journal*, *30*(1), 44–55. DOI: https://doi.org/10.1080/03461238.1947.10419648

Schneider, P. J., & **Penfield, D. A.** (1997). Alexander and Govern's approximations: Providing an alternative to anova under variance heterogeneity. *The Journal of Experimental Education*, *65*(3), 271–286. DOI: https://doi.org/10.1080/00220973.1997.9943459

Srivastava, A. B. L. (1959). Effects of non-normality on the power of the analysis of variance test. *Biometrika*, *46*(1–2), 114–122. DOI: https://doi.org/10.2307/2332813

Tiku, M. (1971). Power function of the f-test under non-normal situations. *Journal of the American Statistical Association*, *66*, 913–916. DOI: https://doi.org/10.1080/01621459.1971.10482371

Tomarken, A. J., & **Serlin, R. C.** (1986). Comparison of anova alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*(1), 90–99. DOI: https://doi.org/10.1037//0033-2909.99.1.90

Wasserman, B. D., & **Weseley, A. J.** (2009). ?'Qué? Quoi? Do languages with grammatical gender promote sexist attitudes? *Sex Roles*, *61*, 634–643. DOI: https://doi.org/10.1007/s11199-009-9696-3

Westfall, P. H. (2014). Kurtosis as peakedness, 1905–2014. R.I.P. *The American Statistician*, *68*(3), 191–195. DOI: https://doi.org/10.1080/00031305.2014.917055

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, *53*(3), 300–314. DOI: https://doi.org/10.1037/0003-066X.53.3.300

Wilcox, R. R. (2005). Comparing medians: An overview plus new results on dealing with heavy-tailed distributions. *The Journal of Experimental Education*, *73*(3), 249–263. DOI: https://doi.org/10.3200/JEXE.73.3.249-263

Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing*. Cambridge, Massachusetts, US: Academic Press. DOI: https://doi.org/10.1016/B978-0-12-386983-8.00010-X

Yuan, K.-H., Bentler, P. M., & **Chan, W.** (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika*, *69*(3), 421–436. DOI: https://doi.org/10.1007/BF02295644

How to cite this article: Delacre, M., Leys, C., Mora, Y. L., & Lakens, D. (2019). Taking Parametric Assumptions Seriously: Arguments for the Use of Welch's *F*-test instead of the Classical *F*-test in One-Way ANOVA. *International Review of Social Psychology*, 32(1): 13, 1–12. DOI: https://doi.org/10.5334/irsp.198

Submitted: 05 June 2018 **Accepted:** 20 May 2019 **Published:** 01 August 2019

]u[*International Review of Social Psychology* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS