Why Researchers Should Always Prefer the *W*-test to the *F*-Test in One-Way ANOVA Designs.

Marie Delacre*, Christophe Leys, and Youri Mora

Université Libre de Bruxelles

Daniël Lakens

Eindhoven University

Author Note

Marie Delacre, Service of analysis of the data, Université Libre de Bruxelles, Belgium; Daniël Lakens, School of Innovation Sciences, Eindhoven University of Technology, The Netherlands; Mora Youri, Service of analysis of the data, Université Libre de Bruxelles, Belgium; Leys Christophe, Service of analysis of the data, Université Libre de Bruxelles, Belgium;

The Supplemental Material, including the full R code for the simulations and plots can be obtained from https://github.com/mdelacre/Welch-ANOVA

Correspondence concerning this article should be addressed to Marie Delacre, Service of analysis of the data, Université Libre de Bruxelles, Bruxelles. Email: marie.delacre@ulb.ac.be

Word Count: 5229 (including 232 words in boxes)

Abstract

When comparing independent groups, researchers in psychology commonly use Analysis of Variance (ANOVA), which compare groups based on their means and assumes data is normally distributed, and variances are equal across conditions. When these assumptions are not met, the classical ANOVA (*F*-test) can be severely biased, which leads to invalid statistical inferences. However, despite their importance, test assumptions are rarely explicitly considered in scientific articles. We discuss why the assumptions of normality and homogeneity of variances will often not hold in psychological research. We explain when and why this is problematic, especially for the assumption of homogeneity of variances. Our simulations show that Welch's ANOVA (*W*-test) controls the Type 1 error rate better than the *F*-test when the assumption of homogeneity of variance is not met, and loses little robustness compared to the *F*-test when the assumptions are met. Because assumption tests for the equality of variances often fail to provide an informative answer, we argue that the *W*-test should be a preferred choice than the *F*-test when comparing means.

When comparing groups, researchers often use the mean to summarize them (Troendle, 2008). The classical Analysis of Variance (*F*-test), Welch's *W* ANOVA (*W-test*), the Alexander-Govern test, James' test and the Brown-Forsythe ANOVA (*F**-test) are the main tests to compare means. They rely on different assumptions about whether data are sampled from a normal distribution or not, and whether the distributions across groups have equal variances or not (Lix, Keselman, & Keselman, 1996). However, in psychological research and possibly in other fields, the *F*-test is the default method to compare different groups (Erceg-Hurn & Mirosevich, 2008). Alternatives tests are considerably less often reported in the literature. Moreover, researchers rarely provide information about the homogeneity of variances assumption[1]. Despite the fact that the *F*-test is currently used by default, the *W*-test is often a better choice, as long as one can consider that the mean correctly reflects the whole distribution. As we argue in this article, the test has nearly the same statistical power, but provides better Type 1 error control than the *F*-test when variances are unequal (Liu, 2015). Moreover, the *W*-test is available in practically all statistical software packages. R and Minitab present the *W*-test by default: Users can request the *F*-test, but only after explicitly stating that the assumption of equal variances is met (see the box "Conducting the *W*-test in R or SPSS").

In this paper, we review the differences between the *F*-test, *W*-test and *F**-test. Based on extensive simulations that compare the Type 1 error rate and the statistical power of these three tests, we suggest that the *W*-test is a better option for psychological science than the *F*-test and *F**-test when comparing the means of groups.

Moreover, there are additional options to compare groups distributions, such as the trimmed means test and nonparametric tests (e.g. the Kruskal-Wallis test and the Mann-Whitney U test)[2]. These options can be useful to compare groups based on other parameters than their means (e.g. in situations where means are not a good or complete representation of a group; Hayes & Cai, 2007). However, since most hypotheses in psychology are based on a comparison of means, we limit our analysis to these tests in order to highlight two main points. First, there are situations where tests comparing means will yield invalid results. Second, when comparing means provide valid results, researchers can improve their statistical inferences by replacing the *F*-test with the *W*-test.[3]

All tests comparing means rely on assumptions about the data. While the *F*-test relies on the normality and homogeneity of variances, some alternatives (e.g. *W*-test) only rely on the normality assumption. When both normality and homogeneity of variances assumptions are met, the *F*-test is slightly more powerful than alternatives. When groups are extracted from populations that have unequal variances, the *F*-test can be severely biased and lead to invalid statistical inferences[4] (i.e., incorrect Type 1 error rates and deviations from the desired power). When comparing only two groups, the problem of unequal variances can be dealt with through experimental design (i.e., collecting the same number of participants in each group). When comparing more than two groups with unequal variances, however, the *F*-test is too liberal even when sample sizes are equal across groups (Box, 1954).

We first explain why the assumptions of normality and equal variances are not always plausible in psychology and provide examples of research areas where unequal variances should be expected. We will then review differences between the *F*-test, *W*-test and *F**-test and show through simulations that unequal variances between groups have a larger impact on the Type 1 error rate and statistical power than violations of the normality assumption. We will argue that the Type 1 error inflation observed with the *F*-test or *F**-test when variances are unequal is much more problematic than the possible small loss of statistical power when the *W*-test is used when variances are equal. Finally, we will point out cases where the *W*-test is not recommended. As we will show, the test is not robust against departures from the normality assumption, when sample sizes are small (i.e., n < 50). We provide approaches to detect such violations, and recommendations to deal with these situations.

Why you Should Think about the Assumptions Underlying Parametric Tests

When the assumptions of parametric tests (i.e. tests having assumptions about the distribution underlying the data) are not met, the conclusions based on parametric tests can be severely biased (Lix et al., 1996), both in terms of Type 1 error rate and power. Assumption checks are rarely reported in the literature, but when researchers do check for assumptions, they often follow a two-step procedure that is recommended in many textbooks (Field, 2013; Howell, 2012). As a first step, researchers are recommended to statistically and/or visually examine the assumptions of normality and equal variances (or the assumption of equal variances), before in the second step choosing the best statistical test (Delacre et al., 2017). However, this two-step procedure is not recommended. Several authors have shown the limitations of conducting such a procedure when comparing two groups (Rasch, Kubinger, & Moder, 2011; Ruxton, 2006; Zimmerman, 2004), and the limitations of this two-step procedure also hold for the *F*-test or regression (Wilcox, Granger, & Clark., 2013). Assumption checks for normality can have low statistical power to actually detect deviations from normality. For example, while the Kolmogorov-Smirnov test is very often used, it will often fail to detect differences between the normal distribution and other distributions (such as the normal skewed distribution, see Supplemental Material 1, and Ghasemi & Zahediasl, 2012; Thode, 2002; Wilcox, 2005), and the Kolmogorov-Smirnov test is therefore not recommended. The Shapiro-Wilk test (available in SPSS)[5] is is a better choice because it is more powerful (Ghasemi & Zahediasl, 2012; Supplemental Material 1), and will almost always detect highly skewed distributions, even when sample sizes are very small. However, there are still two main limitations of using the Shapiro-Wil test in a two-step procedure when examining the normality assumption. First, when sample sizes are small (i.e. $n < 50$; see

Supplemental Material 1), except when distributions are highly skewed, all tests have too low power to detect departures from the normality assumption. This is problematic since the normality assumption is especially crucial for small sample sizes (Supplemental Material 2 and 3). Second, with more than 50 subjects per group, the Shapiro-Wilk test will detect departures from the normal distribution, even when those departures have no negative consequences for the Type 1 error rate or statistical power[6]. Because of the limitations of testing for normality, it is often advised to combine the Shapiro-Wilk test with graphical methods (Ghasemi & Zahediasl, 2012; Öztuna, Elhan, & Tüccar, 2006).

The same arguments for tests for normality apply to tests for the homogeneity of variances assumption (such as Levene's test). These tests will often fail to reject the null hypothesis (i.e. lack of power) when differences in variances and sample sizes are small. This is problematic given that even small differences can inflate error rates in the *F*-test and Student's *t*-tests (Delacre et al., 2017). To conclude, assumption tests are at best a very limited approach to deciding whether or not a statistical test should be performed that relies on the normality and homogeneity assumption, or not. At the same time, as we will argue in the next section, the normality and equal variances assumptions are often unrealistic.

Is the Normality Assumption Realistic?

It has been argued that there are many fields in psychology where the assumption of normality does not hold (Cain, Zhang, & Yuan, 2016). For example, Micceri (1989) reviewed 440 large datasets (i.e. $n \geq 400$) published in a wide variety of journals between 1982 and 1984[7], which contained psychometric and/or abilities measures. He found that the skewness[8] and kurtosis[9] of observed distributions seemed to be closer to an exponential curve than to a normal distribution (Figure 1).

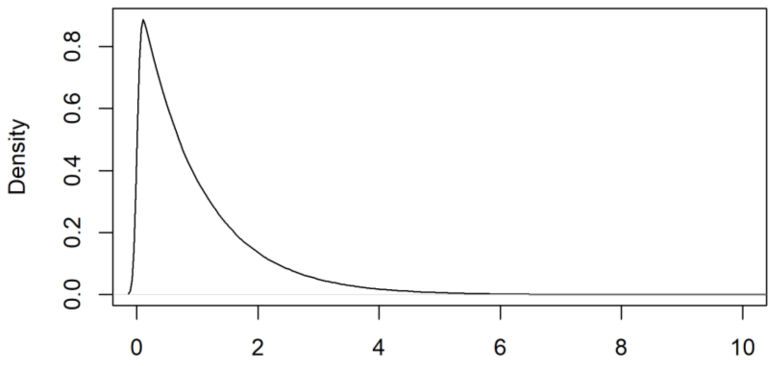

Figure 1. Simulated standard exponential curve.

There is convincing data indicating that in social and behavioral science data are often heavy-tailed (Yuan, Bentler, & Chan, 2004). According to Wilcox (2005), it commonly happens that distributions are very similar to a normal curve but with “thicker” tails than a normal distribution (Figure 2).

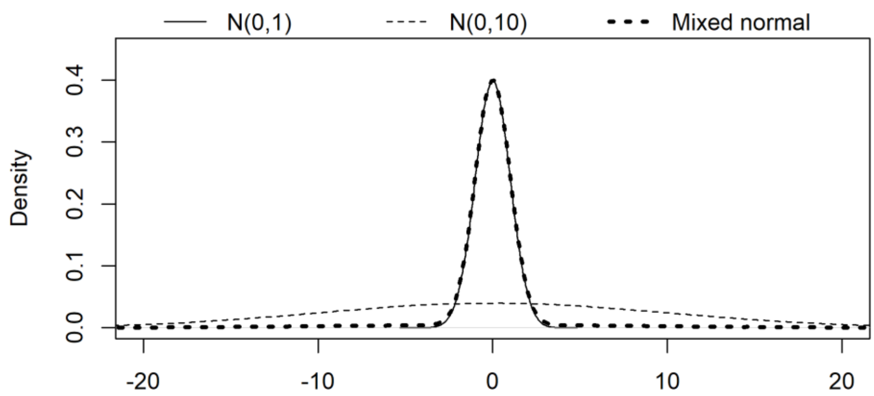

Figure 2. Mixed normal distribution where P(X~N(0,1))= .9 and where P(X~N(0,10))= .1, vs. N(0,1) and N(0,10). The solid line in very close to the bold dashed line, however, the bold dashed line represents a distribution that has a higher kurtosis (≈25.3 vs. 3).

High kurtosis and skewness can also both be present in a distribution. For example, when assessing a wellness score for the general population, data may be sampled from an asymmetric distribution with negative skewness, because most people are probably not depressed (Figure 3, left). An example is provided by the study of Heun, Burkart, Maier and Bech (1999), who evaluated the validity of the WHO Well-Being Scale (WBS) in elderly, and found that the three versions of the WBS yielded highly skewed data. Moreover, when studying reaction times, data are often sampled from asymmetric distributions with positive skewness because it is uncommon to observe much longer response time (Cain et al., 2016; Palmer, Horowitz, Torralba, & Wolfe, 2011; Van Zandt, 2000). Thus, there are many common situations in which perfectly normally distributed data is an unlikely assumption.

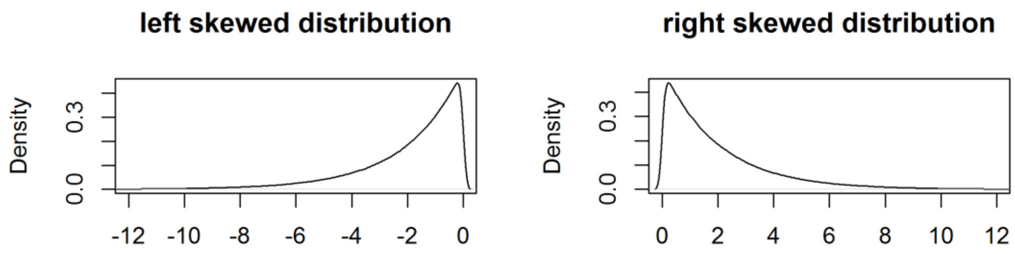

Figure 3. Example of fictive distributions, where skewness is negative (left) or positive (right)

Is the Homogeneity of Variance Assumption Unrealistic?

Discussions in the literature have pointed out that homogeneity of variances assumption is problematic in psychological research (Erceg-Hurn & Mirosevich, 2008; Grissom, 2000). In a previous paper (Delacre et al., 2017), we identified three different causes of unequal standard deviations across groups of observations: 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.

First, psychologists 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 Paez (2005) have shown that the scores of competitiveness, self-reliance and interdependence are more variable in some ethnic groups than in others[10]. Many other examples could be cited where constructs have different variances when pre-existing groups from different gender, cultures, religions, or ethnicity are compared (see for example Adams, Van de Vijver, De Bruin, & Bueno Torres, 2014; Beilmann, Mayer, Kasearu, & Realo, 2014; Church et al., 2012; Cohen & Hill, 2007; Haar, Russo, Suñe, & Ollier-Malaterre, 2014; Montoya & Briggs, 2013). Differences in variability between groups are also often plausible in other fields, such as when different school systems are compared in educational psychology (Delacre et al., 2017). In this last example, groups are sometimes defined in order to have different variability: 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 a different impact on variances between pre-existing groups. 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; Wasserman & Weseley, 2009).

Third, even when the variances of groups are the same before treatment (due to a complete randomization in the group assignment), unequal variances can emerge later, as a consequence of an experimental treatment (Bryk & Raudenbush, 1988; Cumming, 2013; Erceg-Hurn & Mirosevich, 2008; Keppel & Wickens, 2004). For example, Koeser & 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. Thus, there are many common situations in which the homogeneity of variances assumption is an unlikely to be true.

Disclosures

Conflicts of interest : The author(s) declare that they have no conflicts of interest with respect to the authorship or the publication of this article.

Author Contributions : First author performed simulations. First, second and fourth authors contributed to the design. All authors contributed to the writing and the review of the literature.

Simulations Comparing the *F*-test vs. *W*-test vs. *F**-test

We performed simulations to examine the Type 1 error rate and statistical power for different underlying distributions for the *F*-test, *W*-test and *F**-test. The differences between the three tests are mathematically explained in the appendix, which mainly concern the way standard deviations are pooled across groups.

Type 1 Error Rate of the *F*-test vs. *W*-test vs. *F**-test

Simulating error rates when the normality assumption is met. To examine the differences in Type 1 error rate between the *F*-test, *W*-test and *F**-test, we simulated

1,000,000 studies under the null hypothesis (where there are no differences between the means in each group) for four scenarios. For each scenario, we examine the *p*-value distribution. When 5% of the *p*-values fall below 0.05, the Type 1 error rate is controlled as intended. Each scenario was repeated twice, once for an ANOVA with two groups, and once for an ANOVA with three groups. As explained in the appendix, when comparing two groups, *W*-test and *F**-test are mathematically identical and should yield identical error rates. The Type 1 error rate of the three tests under all scenarios are summarized in Table 1. In scenario 1, the variances are the same in each group (SD-ratio = 1; assumption of equal variances met) and sample sizes are unequal (n=20 in the last group; n=40 in all other groups). Table 1 shows that the Type 1 error rate is controlled as intended for all three ANOVA tests, when comparing 2 and 3 groups. In Scenario 2, the variances differ between groups (SD-ratio = the ratio between the biggest standard deviation and the smallest standard deviation = 4) but sample sizes are equal (n = 40 in all groups). Table 1 shows that only *W*-test controls the Type 1 error rate as intended when comparing three groups. In Scenario 3, both sample sizes and variances were unequal between groups and the larger variance is associated with the larger sample size (SD-ratio =4; n=80 in the last group; n=40 in all other groups). Table 1 again shows the *W*-test controls better the Type 1 error rate than the *F*-test. Finally, Scenario 4 is the same as Scenario 3, but the larger variance is associated with the smaller sample size (SD-ratio = 4; n=20 in the last group; n=40 in all other groups, with the same results as Scenario 3).

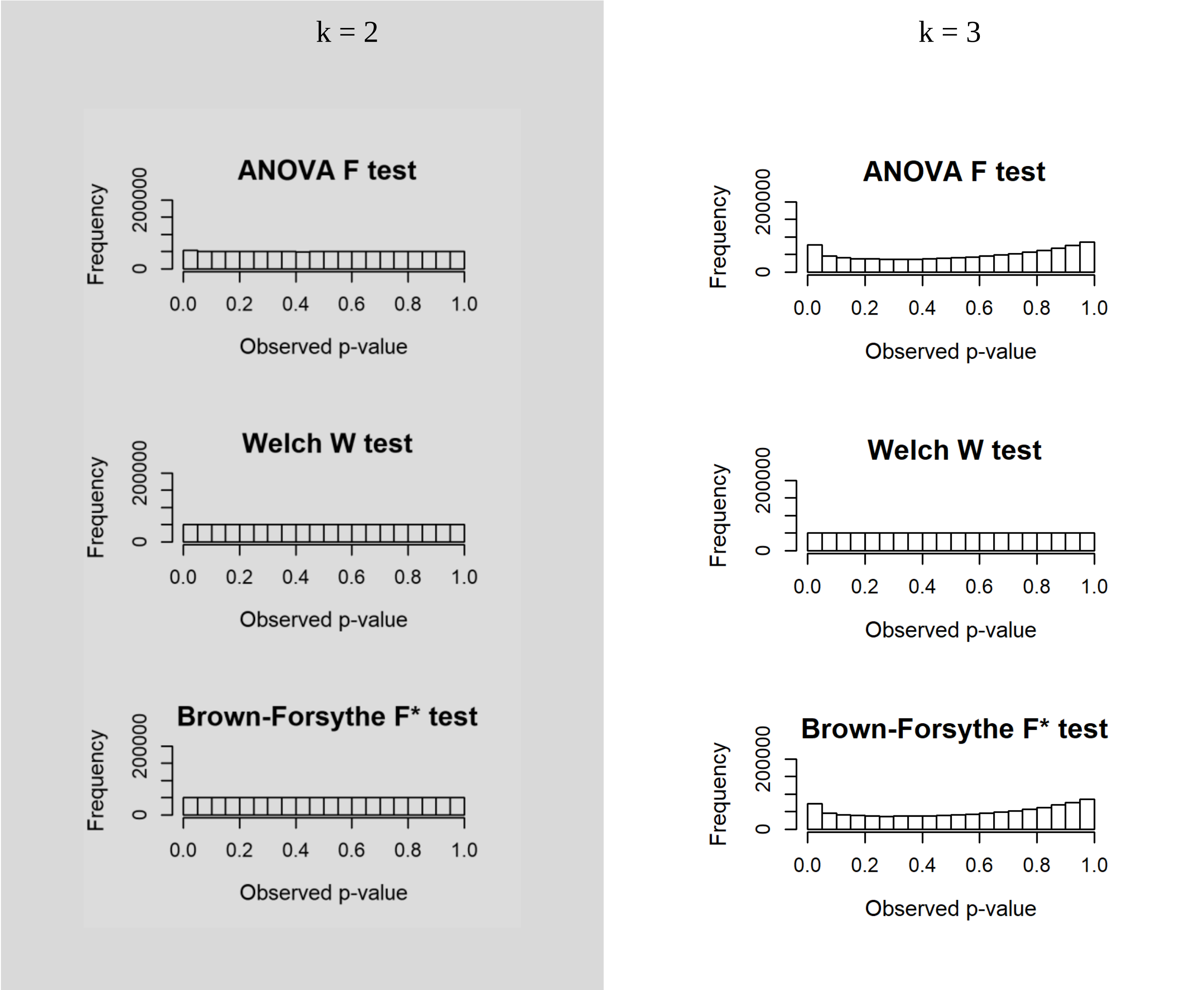

Figure 4. *P*-value distributions for the *F*-test, *W*-test and *F**-test under the null hypothesis when variances are unequal between groups (SD-ratio =4) and sample sizes are equal between groups (n= 40 in all groups), as a function of the number of groups to compare.

As shown in Table 1, when there are only two groups to compare, and as long as the variances are equal between groups, the *p*-value distribution of the *F*-test is uniform, as expected. When sample sizes are equal between groups, the impact of unequal variances is very small and the *p*-value distribution is very close to a uniform distribution. However, when there is a positive (or negative) correlation between sample sizes and standard deviations (i.e. the larger variance is associated with the larger – or smaller – sample size), the percentage of *p*-values smaller than the alpha of 0.05 decreases (or increases).

When there are three groups to compare, the *p*-value distribution of the *F*-test is uniform only when variances are equal. When variances are unequal, the percentage of *p*-values smaller than 0.05 (i.e. the Type 1 error rate) differs from the nominal 5%, even when sample sizes are equal between groups (as shown in Figure 4). In this latter case, the *F*-test becomes more liberal.

This tendency can be generalized: when the number of group increases, the test becomes increasingly liberal. The Type 1 error rate is too low when there is a positive correlation between sample sizes and standard deviations, but too high when there is either a negative correlation between sample sizes and standard deviations or heteroscedasticity with balanced designs[11]. The *F**-test is robust against unequal variances when there are two groups to compare (Table 1). When there are three groups to compare, the test is less affected by violations of the assumption of equal variances than the *F*-test, but the Type 1 error rate still increases when there are unequal variances between groups. Additional simulations, presented in the Supplemental Material, show that the test gets more liberal as the sample size is smaller, and as the SD-ratio and the number of groups to compare increases. Finally, the *W*-test yields a more stable Type 1 error rate, regardless the number of groups that is compared, and regardless of the SD-ratio.

Table 1.

Comparison of Type 1 error rate of the F-test, W-test and F-test, as a function of the number of groups*

	Two groups			Three groups		
Scenario	F	F*	W	F	F*	W
1	0.050	0.050	0.050	0.050	0.050	0.050
2	0.053	0.050	0.050	0.078	0.072	0.050
3	0.009	0.050	0.050	0.016	0.072	0.050
4	0.155	0.050	0.050	0.192	0.071	0.050

Note. Type 1 error rates for the *F*-test, *W*-test and *F**-test are compared when variances are equal (SD-ratio=1) and sample sizes are unequal between groups (n=20 in one group; n = 40 in all other groups; Scenario 1), when variances are unequal between groups (SD-ratio=4) and sample sizes are equal (n = 40 in all groups; Scenario 2), positively correlated with the variance (SD-ratio=4, n=80 in one group, n = 40 in all other groups; Scenario 3), or negatively correlated with the variance (SD-ratio=4, n=20 in one group; n=40 in all other groups; Scenario 4).

Simulating error rates when the normality assumption is not met. While the *W*-test is more robust than both the *F*-test and *F**-test when there are unequal variances, it is less robust than the two other tests when the normality assumption is not met (Supplemental Material 2). The *W*-test is more affected by heavy-tailed and skewed distributions than the *F*-test, becoming more conservative with heavy-tailed distributions (Table A2.2 and A2.3), and more liberal with skewed distributions (Table A2.4, A2.5, A2.6 and A2.7). Furthermore, the *W*-test becomes very liberal when highly skewed distributions are combined with unequal variances and sample sizes between groups[12].

When the data is not normally distributed, and variances are unequal, the *F*-test requires 20 subjects per group to control the Type 1 error rate within an interval of .025 to .075 (i.e. a deviation from the Type 1 error rate deemed acceptable in the literature; Bradley,1978). However, regardless of the sample size, the Type 1 error rate will commonly be out of this interval when variances are unequal (the same holds for the *F**-test). When distributions look symmetric or are moderately skewed (see Supplemental Material 2) *W*-test can be used with only 20 subjects per group. With highly skewed distributions, at least 50 subjects per group are required (when comparing a maximum of four groups), and with even more groups, a larger sample size per group is required. Nevertheless, because highly skewed distributions are easier to detect with a Shapiro-Wilk test than unequal variances with a test of homogeneity of variances (Delacre et al., 2017), the *W*-test is still preferable to the *F*-test.

Power for the *F*-test, *W*-test, and *F-test**

In addition to the Type 1 error rate, the power (1 – the Type 2 error rate) is an important aspect of statistical tests. In order to examine the power of the *F*-test, *W*-test and *F**-test, we performed simulations in which we introduced a true effect (the mean = 1 in the last group, mean = 0 in all other groups). In the same way as when examining the Type 1 error rate in the simulations reported earlier, we manipulated the distribution and variances across groups.

First, it is often believed that the *W*-test and *F**-test are less powerful than the *F*-test when the assumption of the *F*-test are met[13]. When both assumptions are met, our simulations show that the relative loss of power is never larger than 3% when performing a *W*-test or a *F**-test[14]. This relative loss is very small in comparison with the deviation in the Type 1 error rate from the nominal 5% when performing a *F*-test with groups of unequal variances. Moreover, the relative differences in power between *F*-test and both *W*-test and *F**-test tend towards zero when the number of subjects per group increases[15]. When data are extracted from skewed distributions, the relative loss of power can reach up to 23%, meaning that the *W*-test and *F**-test can be almost a quarter less powerful than the *F*-test. However, it considerably decreases when sample sizes increase: with at least 50 subjects per group, the relative loss of power is less than 3% when performing a *W*-test instead of a *F*-test, and almost null when performing a F*_test instead of the F-test. The larger the sample size, the smaller the relative loss of power. Remember that with less than 50 subjects per group, the *W*-test should be avoided when distributions are highly skewed (as indicated by the Shapiro-Wilk test)[16].

When examining the power of both tests under different distributions, it is important not to confuse kurtosis and the standard deviation. In previous work Wilcox (1998) concluded that there is a loss of power when comparing means from heavy-tailed distributions (e.g. double exponential or some mixed normal distribution; Figure 2). 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). DeCarlo (1997) explains that kurtosis and SD are totally independent, meaning that one can find distributions that have similar SD but different kurtosis. When heavy-tailed distributions have equal standard deviations and SD-ratios as normal distributions, there are no substantial differences in power as a function of the kurtosis of the underlying distribution (see Supplemental Material 3).

Conducting Shapiro-Wilk test in R or SPSS

In R, the Shapiro-Wilk test for each compared groups can be run by the function "shapiro.test", using the following syntax:

shapiro.test(*data.name$dv.name[data.name$iv.name*= =*x]*)[1], where x corresponds to one level of the iv.

In SPSS, the Shapiro-Wilk test can be run using the following syntax:

EXAMINE VARIABLES=DV BY IV

/PLOT NPPLOT

Figure 5 shows the output, obtained in SPSS, when performing a Shapiro-Wilk test on data summarized in Table A1.

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	IV	Statistic	df	Sig.	Statistic	df	Sig.
DV	1	,081	41	,200*	,954	41	,094
	2	,134	21	,200*	,932	21	,150
	3	,115	31	,200*	,968	31	,462

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Figure 5. Output in SPSS

Conducting the *W*-test in R or SPSS

In R, the *W*-test can be run by the function “oneway.test”, using the following syntax: **oneway.test(*dv.name* ~ *iv.name*, data=*data.name*, var.equal=FALSE)** [1].

The last argument is used to specify that the *W*-test should be used instead of the *F*-test (which assumes the assumption of equal variances is true). This argument is optional, and when the var.equal is not specified, the *W*-test is reported by default.

In SPSS, the *W*-test can be run using the following syntax:

ONEWAY *dv.name* BY *iv.name*

/STATISTICS WELCH

Figure 6 shows the output, obtained in SPSS, when performing a *W*-test on data summarized in Table A1. As one can see, the degrees of freedom in the numerator of the *W*-test and *F*-test are the same. However, the degrees of freedom in the denominator differ, and in the *W*-test the degrees of freedom has decimal numbers (which should be reported, not rounded).

Robust Tests of Equality of Means

DV

	Statistic[a]	df1	df2	Sig.
Welch	4,606	2	59,320	,014

a. Asymptotically F distributed.

Figure 6. Output in SPSS

Recommendations

In sum, we provide five recommendations:

1) Use the *W*-test instead of the *F*-test to compare the group means. The *F*-test and *F**-test should be avoided, because the equal variances assumption is often unrealistic, tests of the equal variances assumption will often fail to detect differences when these are present, the loss of power is very small (and often even negligible), and the gain in Type 1 error control is considerable under a wide range of realistic conditions.

2) Do not 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 distributions have unequal shape, compare results of the *W*-test with results of a nonparametric procedure is useful in order to better understand the data, since these tests are testing the null hypothesis of the equality of distributions.

3) Use the Shapiro-Wilk test to detect departures from normality (combined with graphical methods). Contrary to the Kolmogorov-Smirnov test, the Shapiro-Wilk test will almost always detect distributions with high skewness, even with very small sample sizes. With small sample sizes, the *W*-test will not control Type 1 error rate when skewness is present, and detecting departures for normality is therefore especially important in small samples. When comparing at most four groups, the *W*-test should be avoided if the Shapiro-Wilk test reject the normality assumption, with less than 50 observations per groups. When comparing more than four groups, the *W*-test should be avoided if the Shapiro-Wilk test reject the normality assumption, with less than 100 subjects per groups. When normality cannot be assumed because of high kurtosis or high skewness, we recommend the use of alternative tests that are not based on means comparison, such as the trimmed means test or nonparametric tests. For more information, see Erceg-Hurn and Mirosevich (2008).

4) Perform a-priori power-analysis. Fifty subjects per groups are generally enough to control the Type 1 error, but power analyses are important in order to determine the required sample sizes to achieve sufficient power to detect a statistically significant difference (see Albers & Lakens, 2018).

5) Use balanced designs (i.e. same sample size in each group) whenever possible. When using the *W*-test, the Type 1 error rate is a function of criteria such as the skewness of the distributions, and whether skewness is combined with unequal variances and unequal sample sizes between groups. Our simulations show that the Type 1 error rate control is in general slightly better for balanced designs.

---

[1] Hoekstra, Kiers, and Johnson (2012) have shown that from 50 randomly selected publications in *Psychological Science* that reported at least one ANOVA, *t*-test, or regression, only three articles discussed the normality and heterogeneity of variances assumption. To generalize this result, we surveyed statistical tests reported in 116 articles in the *Journal of Personality and Social Psychology* published in the year 2016. In 14% of these articles a One-Way ANOVA was reported, but none of the articles explicitly reported examining the homogeneity of variances, and only one article implicitly reported taking into account the homogeneity assumption by reporting corrected degrees of freedom as used in the *W*-test.

[2] Non parametric tests such as the Mann-Whitney *U* test and Kruskal-Wallis are very appropriate in psychological field because they only require ordinal scale (Gibbons & Chakraborti, 2011). Their null hypothesis assumes 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 (Grissom, 2000; Nachar, 2008; Neuhäuser & Ruxton, 2009; Tomarken & Serlin, 1986; Zimmerman, 2000). Other alternatives, such as trimmed means test, exist (Wilcox et al., 2013). The null hypothesis of the trimmed means test assumes that trimmed means are the same between groups. A trimmed means is a mean

computed on data after removing the lowest and highest values of the distribution (Erceg-Hurn & Mirosevich, 2008). Trimmed means and means are equal when data are symmetric. On the other hand, when data are asymmetric, trimmed means and means differ.

[3] We do not include the Alexander-Govern and James' tests (i.e. two alternatives that are robust against unequal variances between groups) because these tests are not available in all statistical software (such as SPSS). However, additional simulations indicated that these tests give results very close to the *W*-test, which means that they have very similar strengths and limitations. When data are symmetrically distributed, the biggest difference we found between the *W*-test and the Alexander-Govern test and James test in Type 1 error rates is .008. When data is skewed, with unequal skewness between groups, the difference in Type 1 error rates can increase to maximum .01.

[4] Every time we mention «unequal variances», we refer to variances in the populations that the data are sampled from, and not the sample variances.

[5] The Shapiro-Wilk test is based on the correlation between the observed data and their corresponding normal score (i.e. the vector or quantile of the observed data, and the vector or quantile that should be obtained if data were normally distributed; Ghasemi & Zahediasl, 2012; Öztuna, Elhan, & Tüccar, 2006).

[6] For example, in a previous paper, we have shown that when comparing two groups where data are uniformly distributed (i.e. kurtosis = 1.8) tests comparing means are still valid, both in terms of the Type 1 error rate and the power (Delacre, Lakens, & Leys, 2017).

[7] *Applied Psychology*, *Journal of Research in Personality*, *Journal of Personality*, *Journal of Personality Assessment*, *Multivariate Behavioral Research*, *Perceptual and Motor Skills*, *Applied Psychological Measurement*, *Journal of Experimental Education*, *Journal of Educational Psychology*, *Journal of Educational Research*, and *Personnel Psychology*.

[8] Skewness is a measure symmetry, used to describe the shape of the distributions underlying the data (Joanes & Gill, 1998). A distribution with positive skewness will be right-skewed. A distribution with negative skewness will be left-skewed (see Figure 3). A distribution with a null skewness will be symmetric.

[9] Kurtosis is a measure used to describe the shape of the distributions underlying the data (Joanes & Gill, 1998). A distribution with positive kurtosis will be more peaked and have heavier tails than the normal distribution. On the other hand, a distribution with negative kurtosis will be flatter and have lighter tails than the normal distribution. Finally, a distribution with a null kurtosis will have the same tails and peakedness than the normal distribution (DeCarlo, 1997).

[10] Among others, the score of competitiveness in Switzerland (M = 1.9; SD = .57) and Spain (M = 1.65; SD = .56) are less variable than the score of competitiveness in Italy (M = 2.12; SD = .79) or France (M = 2.28; SD = .75).

[11] To yield a robust test, the Type 1 error rate has to be sufficiently close to the nominal 5% level. In order to assess the robustness of the three tests in our simulations, we follow Bradley (1978) and consider the Type 1 error rate as 'close enough' to the nominal 5% if it falls in the interval [0.025; 0.075].

[12] It is illustrated in simulations presented in the Supplemental Material 2. We performed the W-test in order to compare samples of equal means with the following conditions : one sample (40 subjects) was extracted from a normal left-skewed distribution and all other samples (20 subjects per sample) were extracted from a chi-square distribution. The SD-ratio was 0.5. As a result, the type 1 error rate was .90 when we compared three groups, .100 when we compared four groups and .108 when we compared five groups.

[13] We will only compare results of the *W*-test and the *F*-test when the assumption of equality of variances is met because when variances between groups are unequal, results of the *F*-test

are not valid. When there is a negative correlation between sample sizes and standard deviation, or when there are unequal standard deviations between groups, the power of the *F*-test is overestimated ($\text{Power}_{F\text{-test}} > \text{Power}_{W\text{-test}}$); when there is a positive correlation between sample sizes and standard deviations, the power of the *F*-test is underestimated ($\text{Power}_{F\text{-test}} < \text{Power}_{W\text{-test}}$).

[14] The relative loss of power is computed as follows : ($\text{power}_{\text{student}}$-$\text{power}_{\text{welch}}$)/$\text{power}_{\text{student}}$. For example, if one achieves a power of .487 when performing an *F*-test, and a power of .472 when performing a *W*-test, the relative loss of power is (.487-.472)/.487 = .031

[15] For example, with at least 50 subjects per group the relative loss of power is approximately 1% when performing a *W*-test or a *F**-test, in comparison with a F-test. With at least 100 subjects per group, the relative loss in null.

[16] Note that 50 subjects per group is enough to achieve robustness in terms of Type 1 error rate, however, it is also important to have a good power. The required number of subjects to achieve a sufficient power will be a function of parameters such as the effect size, the sample sizes ratio, and the power criterion. In general, power is deemed acceptable at .80 (Cohen, 1988)

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. http://dx.doi.org/10.1177/0022022114542466

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. Journal of Experimental Social Psychology, 74, 187–195. http://dx.doi.org/10.1016/j.jesp.2017.09.004

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.http://dx.doi.org/10.1007/s12187-014-9232-z

Box, G. E. (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. http://dx.doi.org/10.1214/aoms/1177728786

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. http://dx.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. http://dx.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. http://dx.doi.org/10.1037/0033-2909.104.3.396

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2016). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior research methods*, *49*(5), 1716-1735. http://dx.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. (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. http://dx.doi.org/10.1177/0022022111428514

Cohen, J. (1988). Statistical power analysis for the behavioral sciences . Hilsdale. NJ: Lawrence Earlbaum Associates, 2.

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. http://dx.doi.org/10.1111/j.1467-6494.2007.00454.x

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=cumming+understanding+the+new&ots=PvDWTKb44N&sig=fQA-XUVLIkWXx1iNznrHdZgxlUA

DeCarlo, L. (1997). On the meaning and use of kurtosis. Psychological Methods, 2(3), 292-307. http://dx.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 (in press for the International Review of Social Psychology). Retrieved from https://osf.io/preprints/sbp6k/

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. https://doi.org/10.1037/0003-066X.63.7.591

Fan, W., & Hancock, G. R. (2012). Robust means modeling: An alternative for hypothesis testing of independent means under variance heterogeneity and nonnormality. *Journal of Educational and Behavioral Statistics*, *37*(1), 137–156. http://dx.doi.org/10.3102/1076998610396897

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage. Retrieved from https://books.google.fr/books?hl=fr&lr=&id=c0Wk9IuBmAoC&oi=fnd&pg=PP2&dq=discovering+statistics+using+spss&ots=LbzoNGWvZI&sig=TfzzF_g05GtfHPF3iL_CbH8aIbk

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-Statisticians. *International Journal of Endocrinology and Metabolism*, *10*(2), 486-489. https://doi.org/10.5812/ijem.3505

Gibbons, J. D., & Chakraborti, S. (2011). *Nonparametric statistical inference*. Boca Raton, FL: Chapman and Hall/CRC.

Green, E. G., Deschamps, J.-C., & Paez, D. (2005). Variation of individualism and collectivism within and between 20 countries: A typological analysis. *Journal of cross-cultural psychology*, *36*(3), 321–339. http://dx.doi.org/10.1177/0022022104273654

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of consulting and clinical psychology*, *68*(1), 155–165. http://dx.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. http://dx.doi.org/10.1016/j.jvb.2014.08.010

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior research methods*, *39*(4), 709–722. http://dx.doi.org/10.3758/bf03192961

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29-29. 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. http://dx.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*. http://dx.doi.org/10.3389/fpsyg.2012.00137

Howell, D. (2013). *Statistical methods for psychology*. Australia: Thomson/Wadsworth.

Joanes, D. N., & Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *47*(1), 183–189. http://dx.doi.org/10.1111/1467-9884.00122

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

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., … Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of educational research*, *68*(3), 350–386. http://dx.doi.org/10.2307/1170601

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. http://dx.doi.org/10.1177/0261927x14541280

Liu, H. (2015). Comparing Welch ANOVA, a Kruskal-Wallis test, and traditional ANOVA in case of heterogeneity of variance. Virginia Commonwealth University.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (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.

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.

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. http://dx.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. http://dx.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*. http://dx.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 x 2 designs with unequal cell frequencies. *Psychological reports*, *76*(3), 1011–1017. http://dx.doi.org/10.2466/pr0.1995.76.3.1011

Öztuna, D., Elhan, A. H., & Tüccar, E. (2006). Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. *Turkish Journal of Medical Sciences*, *36*(3), 171–176.

Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 58. http://dx.doi.org/10.1037/a0020747.supp

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. http://dx.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. http://dx.doi.org/10.1093/beheco/ark016

Thode, H. C. (2002). *Testing for normality*. New York : Marcel Dekker.

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

Troendle, J. F. (2008). Testing for Group Effect in a 2 x k Heteroscedastic ANOVA Model. *Biometrical Journal*, *50*(4), 571–583.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic bulletin & review*, *7*(3), 424–465. http://dx.doi.org/10.3758/bf03214357

Wasserman, B. D., & Weseley, A. J. (2009). ¿Qué? Quoi? Do languages with grammatical gender promote sexist attitudes? *Sex Roles*, *61*(9-10), 634–643. http://dx.doi.org/10.1007/s11199-009-9696-3

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*(3/4), 330–336. http://dx.doi.org/10.2307/2332579

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

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 Rand, R., & Wilcox Rand, R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*. Elsevier Academic Press.

Yuan, K.-H., Bentler, P. M., & Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika*, *69*(3), 421–436. http://dx.doi.org/10.1007/bf02295644

Zimmerman, D. W. (2000). Statistical significance levels of nonparametric tests biased by heterogeneous variances of treatment groups. *The Journal of general psychology*, *127*(4), 354–364. http://dx.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. http://dx.doi.org/10.1348/000711004849222

Appendix

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

In this section, we will explain the mathematical differences in how the *F*-test, *W*-test and *F**-test are computed, with a focus on the differences in how standard deviations are pooled across groups.

As shown in formula 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:

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

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

$$Df_n = k-1 \tag{2}$$

$$Df_d = N-k, \text{ where } N = \sum_{j=1}^{k} n_j \tag{3}$$

As a generalization of the Student's *t*-test, the *F*-test is calculated based on a pooled error term, which implies that all samples are estimates of a common population variance. The *F*-test suffers from the same limitations as the *t*-test when sample sizes are unequal between groups, in that the Type 1 error rate is no longer controlled at the desired level when variances are unequal between groups. When the larger variance is associated with the larger sample size, there is a decrease in the Type 1 error rate (Nimon, 2012; Overall, Atlas, & Gibson, 1995), because the error term increases, and therefore, the *F*-value decreases, leading to fewer significant findings than expected with a specific type 1 error level. When the larger variance is associated with the smaller sample size, the Type 1 error rate is inflated (Nimon, 2012; Overall et al., 1995). This inflation is caused by the under evaluation of the error term, which increases the *F*-value, and thus leads to more significant results than expected based on the nominal Type 1 error level. Moreover, when the number of groups increases, 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.

To address the problems with error control in the *F*-test when variances are unequal, several authors have proposed alternative approaches to statistical tests on more than two means, which do not rely on the homogeneity of variances assumption (e.g., Welch, 1951). Tomarken and Serlin (1986) have shown that from the available alternatives, *F**-test and *W*-test are the best choice. Both tests are available in SPSS, which is a widely used software in psychological science (Hoekstra et al., 2012). The *F** statistic proposed by Brown and Forsythe (1974) is computed as follows:

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

Where x_j and s_j^2 are respectively the group mean and the group variance, and $\overline{x_{..}}$ is the overall mean.

As can be seen in formula 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 of the statistic, the variance of each group is weighted by 1 minus the relative frequency of each group, so that the variance associated with the group with the smallest sample size is given more weight. 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 with 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 than expected with the *F*-test.

The degrees of freedom in the numerator and in the denominator of *F**-test are computed as follow:

$$\mathrm{Df_n} = k-1 \qquad (5)$$

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

As shown in our simulations, the *F**-test appears to be more robust than the *F*-test in many situations where there are unequal variances between groups, when looking at the Type 1 error rate, but in many circumstances, it is too liberal. Our simulations also show that the *W*-test has better Type 1 error control than both *F*-test and *F**-test when there are unequal variances between groups. As can be seen in formula 7, the squared deviation between groups means and the general mean are weighted by $\frac{n_j}{s_j^2}$ instead of n_j in the numerator of the *W*-test (Brown & Forsythe, 1974).

$$W = \frac{\frac{1}{k-1}\sum_{j=1}^{k}\left[w_j(\overline{X_j}-\overline{X'})^2\right]}{1+\frac{2(k-2)}{k^2-1}\sum_{j=1}^{k}\left[\left(\frac{1}{n_j-1}\right)\left(1-\frac{w_j}{w}\right)^2\right]}, \qquad (7)$$

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

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

$$\overline{X'} = \frac{\sum_{j=1}^{k}(w_j\ \overline{x_j})}{w}$$

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

$$\mathrm{Df_n} = k-1 \qquad (8)$$

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

When there are only two groups to compare, the *F**-test and *W*-test 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.

To better understand how to compute all statistics, a set of fictional raw data simulate the example of a three-groups design. A summary is presented in Table A1. The complete example is available on Github. The DV is a score that can vary from 0 to 40. The IV is a three-level factor A (levels = A_1, A_2 and A_3).

Table A1. *Summary of the data of the fictive case*

	A1	A2	A3
n_i	41.00	21.00	31.00
$\bar{X}$	24	23	27
s^2	81.75	10.075	38.40

The global mean (i.e. the mean of the global dataset) is a weighted mean of the group means:

$$\frac{(41 \times 24) + (21 \times 23) + (31 \times 20.5)}{41 + 21 + 31} = \frac{2304}{93} \approx 24.77$$

The *F*-test statistic and degrees of freedom are computed by applying formulas 1, 2 and 3:

$$\mathrm{F} = \frac{\frac{1}{3-1}\left[41\times(24-\frac{2304}{93})^2+21\times(23-\frac{2304}{93})^2+31\times(27-\frac{2304}{93})^2\right]}{\frac{1}{93-3}[(41-1)\times 81.75+(21-1)\times 10.075+(31-1)\times 38.4]} \approx 2.377$$

$$\mathrm{df_n} = 3\text{-}1 = 2$$

$$\mathrm{df_d} = 93\text{-}3 = 90$$

The F^*-test and his degrees of freedom are computed by applying formulas 4, 5 and 6.

$$F^* = \frac{41\times(24-\frac{2304}{93})^2+21\times(23-\frac{2304}{93})^2+31\times(27-\frac{2304}{93})^2}{\left(1-\frac{41}{93}\right)\times 81.75+\left(1-\frac{21}{93}\right)\times 10.075+\left(1-\frac{31}{93}\right)\times 38.4} \approx 3.088$$

$$df_n = 3\text{-}1 = 2$$

$$df_d = \frac{1}{\frac{\left(\frac{\left(1-\frac{41}{93}\right)\times 81.75}{\sum_{j=1}^{k}\left(1-\frac{n_j}{N}\right)s_j^2}\right)^2}{41-1}+\frac{\left(\frac{\left(1-\frac{21}{93}\right)\times 10.075}{\sum_{j=1}^{k}\left(1-\frac{n_j}{N}\right)s_j^2}\right)^2}{21-1}+\frac{\left(\frac{\left(1-\frac{31}{93}\right)\times 38.4}{\sum_{j=1}^{k}\left(1-\frac{n_j}{N}\right)s_j^2}\right)^2}{31-1}} \approx 81{,}149$$

where $\sum_{j=1}^{k}\left(1-\frac{n_j}{N}\right)s_j^2 \approx 79{,}11$

Finally, the *W*-test and his degrees of freedom are computed in applying formulas 7, 8 and 9:

$$W = \frac{\frac{1}{3-1}\left[\frac{41}{81{,}75}\left(24-\overline{X'}\right)^2+\frac{21}{10.075}\left(23-\overline{X'}\right)^2+\frac{31}{38.4}\left(27-\overline{X'}\right)^2\right]}{\frac{2(3-2)}{3^2-1}\times\left[\left(\frac{1}{41-1}\right)\left(1-\frac{41/81{,}75}{w}\right)^2+\left(\frac{1}{21-1}\right)\left(1-\frac{21/10.075}{w}\right)^2+\left(\frac{1}{31-1}\right)\left(1-\frac{31/38.4}{w}\right)^2\right]+1} \approx 4.606$$

$$w = \sum_{j=1}^{k} w_j \approx 3{,}39$$

Where $$\overline{X'} = \frac{\sum_{j=1}^{k}(w_j\overline{x_j})}{w} \approx 24{,}10$$

$$df_n = 3\text{-}1 = 2$$

$$df_d = = \frac{3^2-1}{3\left[\frac{\left(1-\frac{w_j}{w}\right)^2}{41-1}+\frac{\left(1-\frac{w_j}{w}\right)^2}{21-1}+\frac{\left(1-\frac{w_j}{w}\right)^2}{31-1}\right]} \approx 59{,}32$$

One should notice that in this example, the biggest sample size has the biggest variance. As previously mentioned, it means that the *F*-test will be too conservative, because the *F* value decreases. The F^* ANOVA will also be a little too conservative, even if the test is less affected than the *F*-test. As a consequence: $W > F^* > F$.