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[1], Christophe Leys[1], Youri L. Mora[1], & Daniël Lakens[2]

[1] Université Libre de Bruxelles, Service of Analysis of the Data (SAD), Bruxelles, Belgium

[2] Eindhoven University of Technology, Human Technology Interaction Group, Eindhoven, the Netherlands

Author Note

All materials required to reproduce the analyses reported in this article are available at https://github.com/mdelacre/W-ANOVA

This work was supported by the Netherlands Organization for Scientific Research (NWO) VIDI grant 452-17-013.

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. The Supplemental Material, including the full R code for the simulations and plots can be obtained from https://github.com/mdelacre/W-ANOVA. The authors declare that they have no conflicts of interest with respect to the authorship or the publication of this article.

Correspondence concerning this article should be addressed to Marie Delacre, CP191, avenue F.D. Roosevelt 50, 1050 Bruxelles. E-mail: marie.delacre@ulb.ac.be

Abstract

Student's t-test and classical F-test ANOVA rely on the assumptions that two or more samples are independent and that independant 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 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 five practical recommendations that researchers can use to improve their statistical practices.

Keywords: W-test; ANOVA; homogeneity of variance; statistical power; type I error, parametric assumptions

Word count: X

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

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; Tomarken & Serlin, 1986). Both tests rely on the assumptions that iid[1] 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) or the power of the F-test (**???**, **???**, **???**, **???**), the F-test is not robust against unequal variances (Grissom, 2000), that can alter both Type I error rate (**???**, **???**) and power (Nimon, 2012; Overall, Atlas, & Gibson, 1995). Type I and type II error rates and power will be formally defined in the section called "Stakes Underlying the F-test Assumptions of Normality and Homogeneity of variances".

Yet, researchers rarely provide information about these 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 the year 2016. Fourteen percent of these articles reported a One-Way F-test, but only one article indicated taking the homogeneity of variances assumption 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, alternatives exist that are often a better choice, such as the Welch's W ANOVA (W-test), the Alexander-Govern test, James second order test and the Brown-Forsythe ANOVA (F*-test). To be clear, we focus exclusively on tests that

[1] iid means independent and identically distributed

compare groups based on their means. Our goal, throughout this paper, is not to give methodological considerations about the hypotheses. We agree that they exist many tests that allow to consider not only the mean but also other relevant parameters of the distributions (such as standard deviations and the shape of the distribution), in changing the main hypothesis (see for example Erceg-Hurn & Mirosevich, 2008; Wilcox, 1998), however researchers still rely mostly on the mean differences (**???**; Erceg-Hurn & Mirosevich, 2008). We think that a first realistic first step towards progress would be to get researchers to correctly test the hypothesis they are used to. Hence, although the debate surrounding the F-tests assumptions is widely explored (see for example the meta-analysis of **???**), applied researchers still seem to ignore the consequences of their violations and some non-mathematical pedagogical paper summarizing the arguments seems to be lacking. This paper aims at filling 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, is of major consequences). We will explain, with as few mathematical demonstrations as possible, how the alternative tests cope with heteroscedasticity in order to overcome its impact on the results.We conducted simulations comparing F-test with the most adequate alternatives. We argue that when variances are equal between groups, the W-test has nearly the same empirital Type I error rate and power than the F-test but when variances are unequal, it provides empirical Type 1 error rate and power that are closer to the expected level than the F-test.Since the W-test is available in practically all statistical software packages, researchers can improve their statistical inferences by replacing the F-test by the W-test.

Normality and Homogeneity of variances in 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; K.-H. Yuan, Bentler, & Chan, 2004). As argued by Micceri (1989), there are many factors that could explain departures from the normality assumption. Micceri (1989) identified many factors, and we will focus on three of them: the treatment effects, the presence of subpopulations within the studied one and the bounded measures underlying residuals.

First, while it is obvious that the mean can be changed by the treatment effects, experimental treatment could also change the shape of a distribution, either on one or both indicators of skewness, a measure of asymmetry of the shape of the distribution, and kurtosis, a measure of the tendency to produce extreme values [2](**???**, **???**). For example, (**???**) show that practicing juggling during an amount of time can have various effect of juggle learning efficiency between subjects. This could influence the kurtosis of the distribution.

Second, prior any experimental treatments, departures from the normality assumptions may be due to the presence of several subpopulations that present unequal characteristics and that are not controlled within the studied group (resulting in mixed distributions). This unavoidable lack of control is inherent of our field given its complexity. As an illustration, (**???**) 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 is 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, Burkart, Maier, & Bech, 1999). In this case, people suffering from

[2] a distribution with positive kurtosis will have heavier tails than the normal distribution meaning 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.

depression and people that do not are included within the same population group, explaining the asymmetry.

Third, bounded measure can also explain the non-normal distributions. Such examples can be found in the field of neurosciences such as reaction times, that can be very large but never below zero (resulting in right-skewed distributions; see (**???**) for a discussion on the reaction time distribution shape). 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 demand that is ecologically very 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 (J. Cohen, Cohen, West, & Aiken, 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, Van de Vijver, de Bruin, & Bueno Torres, 2014; Beilmann, Mayer, Kasearu, & Realo, 2014; Church et al., 2012; A. B. Cohen & Hill, 2007; Haar, Russo, Suñe, & Ollier-Malaterre, 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. In this example, the goal is not to alter the variability but is an inherent statistical implication of such theoretical positions.

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 ($SD_{spanish} = .97 < SD_{english} = 1.33$) [3]

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, 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

[3] Note that this example is for didactic reasons, the differences have not been tested and might not differ significantly

treatment in all experimental conditions.

Fourth, more often than not, psychological processes are captured in situations where many variables are unidentified and/or left uncontrolled (J. Cohen et al., 2013). Since some of these variables can act as moderators, they can generate heteroscedasticity.Indeed, by definition, a moderator is a variable that will interact with factors, which implies that the effect of the moderator will be different in one condition of the factor than in another condition of the same factor.

To conclude, there are many common situations in which the homogeneity of variances assumption is unlikely to be true.

Stakes Underlying the *F*-test Assumptions of Normality and Homogeneity of variances

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 error can be made: Type I error and Type II error. Type I error consists in falsely reject the main hypothesis in favour of an alternative hypothesis, and Type I error rate (α) is the proportion of tests, when sampling many times from the same population, that wrongly reject the main hypothesis in favour of an alternative hypothesis (**???**). Type II error consists in wrongly reject the alternative hypothesis and type II error rate (β) is the proportion of tests, when sampling many times from the same population, that wrongly reject the alternative hypothesis (**???**). Finally, the statistical power (1-β) is the proportion of tests, when sampling many times from the same population, that correctly reject the main hypothesis in favour of an alternative hypothesis.

Violation of the Normality Assumption

Regarding the type I error rate, the distribution shape hasvery little impact on the *F*-test (**???**). When departures are very small (i.e. a kurtosis between 1.2 and 3 or a skewness between -.4 and .4), the Type I error rate of the *F*-test is very close of expectations even with very small samples of 11 subjects per group (**???**).

Regarding the type II error rate, many author underlined that departures from normality does not seriously affect the power (**???**, **???**, **???**, **???**, **???**; Glass et al., 1972). However, following (**???**) and (**???**), kurtosis impacts slightly more 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 (**???**).

Violation of Homogeneity of Variances Assumption

Regarding the type I error, the *F*-test is sensitive to unequal variances (**???**). More specifically, the higher the *SD*-ratio, the higher the impact. When there are only two groups, the impact is smaller than when there are more than two groups (**???**). When there are more than two groups, the *F*-test becomes more liberal, 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 positive pairing (i.e. the group with the larger sample size also has the larger variance), the test is too conservative whereas in case of 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, there is a small impact of unequal variances when sample sizes are equal (**???**), however, there is a strong effect of the sample size and variance pairing

(Nimon, 2012; Overall et al., 1995). In case of positive pairing, the type II error rate increases (i.e. the power decreases) and in case of 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, following (**???**), 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.

Throughout mathematical explanations and Monteo Carlo simulations, among the five previously mentioned alternatives, we chose to examine the F-test, W-test and F*-test and to exclude the James second-order and Alexander-Govern's test. Tomarken and Serlin (1986) have shown that from the available alternatives, the F*-test and the W-test are the best choices. Both tests are available in SPSS, which is a widely used software in psychological sciences (Hoekstra et al., 2012). The two later tests were not included, because they yield very similar results to the W-test but are less readily available in statistical software packages. For a more extended description of these two alternatives, see Schneider and Penfield (1997).

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 way standard deviations are pooled across groups to stress the implications of heteroscedasticity (see appendix for a numerical example). We suggest readers to focus their attention on the terms involving the variances.

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:

$$F = \frac{\frac{1}{k-1}\sum_{j=1}^{k}[n_j(\bar{x_j} - \bar{x_{..}})^2]}{\frac{1}{N-k}\sum_{j=1}^{k}(n_j - 1)s_j^2} \tag{1}$$

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

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

$$df_d = N - k, \tag{3}$$

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, if the larger variance is associated with the larger sample size, the error term, denominator in (1), is overestimated. F-value is therefore decreased, leading to fewer significant findings than expected with a specific Type I error level. It explains why the F-test is too conservative in this configuration.

On the other side, when the larger variance is associated with the smaller sample size, the error term, denominator in (1), is underestimated. F-value is therefore inflated, yielding more significant results than expected under the nominal Type I error level.

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

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

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 adjustement implies that the variance associated with the group with the smallest sample size is given more weight than in 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 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 follows (with the same principle than the denominator computation):

$$df_n = k - 1 \tag{5}$$

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

Formula (7) provides the W-test computation. 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 (**???**).

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

where:

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

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

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

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

$$df_n = k - 1 \tag{8}$$

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

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. To better understand how to compute all statistics, a set of fictional raw data simulating the example of a three-group design is available in the appendix. The following section will present Monte Carlo simulations assessing these three tests on both Types I and II error rates.

Monte Carlo simulations: *F*-test vs. *W*-test vs. *F**-test

Simulations methods (and justification)

Monte Carlo simulation studies are performed using R-gui (version 3.5.0). 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 main hypothesis is true), in order to assess the Type I error rate of the tests. In 1280 scenarios, there were means differences (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. Only the remaining sample was different.

Population parameters values were chosen in order to illustrate the consequences of factors known to play a key role on both Type I error rate and power, when performing ANOVA, consistently with the literature review presented above: number of groups, sample sizes, sample sizes ratio, *SD*-ratio, sample size and variance pairing.

Number of groups to compare. In our scenarios, the number of compared groups (k) varied from 2 to 5.

Sample sizes. Sample sizes of k-1 groups (n_j) were 20,30,40,50 or 100.

Sample sizes ratio. The sample size of the last group was a function of the sample size ratio (n-ratio $= \frac{n_k}{n_j}$), ranging from 0.5 to 2, in steps of 0.5. The simulations for which the n-ratio equals 1 are the particular case of balanced design (i.e. sample sizes are equal across all groups).

***SD*-ratio.** The *SD* of the population from which was extracted last group was a function of the *SD*-ratio (*SD*-ratio $= \frac{\sigma_k}{\sigma_j}$) that was 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).

Sample size and variance pairing. All possible combinations of *n*-ratio and *SD*-ratio were performed in order to distinguish positive pairing (the group with the largest sample size is extracted from the population with the largest *SD*), negative pairing (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 these conditions were tested considering normal and non-normal distributions.

When two groups are compared, conclusions for the three ANOVA tests (F, F^*, W) should yield identical error rates than their equivalent t-test (F-test is equivalent to the Student's t-test as well as F^*-test and W-test are equivalent to the Welch's t-test; Delacre et al., 2017). On the other side, 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 (**???**, **???**).

For didactic reasons, we will report only the results when we compare three groups (k=3). Increasing the number of groups have as main effect to increase the liberality of all test. For interested readers, all figures for cases when we compare more than three groups are available here: https://github.com/mdelacre/W-ANOVA/tree/revisionbranch/Figures%2C%20Type%20I%20error%20rate%20and%20power%20for%20all%20k%20between%202%20and%205.

Similarly, we merged all sample sizes. Unsurprisingly, the higher the sample sizes, the less the distributions of the population underlying the samples impact the robustness of the tests (**???**). On the other side, increasing the sample sizes does not improve the robustness of the test when there is heteroscedasticity. Interested reader could see all details in the following Excell spreadsheet, available on github : « Type I error rate.xlsx ».

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

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

Note. 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, 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 finally, 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*, *SD*-ratio equals 1. When the last group is extracted from a population with a larger *SD* than all other groups, *SD*-ratio > 1 and finally, when the last group is extracted from a population with a smaller *SD* than all other groups, *SD*-ratio < 1.

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

As previously mentioned, the Type I error rate (α) is the proportion of tests, when sampling many times from the same population (e.g. 1,000,000 times), that wrongly reject the main hypothesis in favour of an alternative hypothesis (**???**). 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, considering the nominal alpha level of 5%.

When there is no difference mean, the 9 cells of Table 1 boils down into 5 subconditions:

- 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)

In Figure 1, we computed the average Type I error rate of the three tests under the 5 prementioned subcategories. The light grey area corresponds to the liberal criterion from Bradley (1978), from which a departure from the nominal alpha is acceptable as long as the type I error rate falls within the interval [.5 × α; 1.5 × α]. The dark grey area corresponds to the more conservative criterion from which departures from the nominal alpha is negligible as long as the Type I error rate falls within the interval [.9 × α; 1.1 × α].

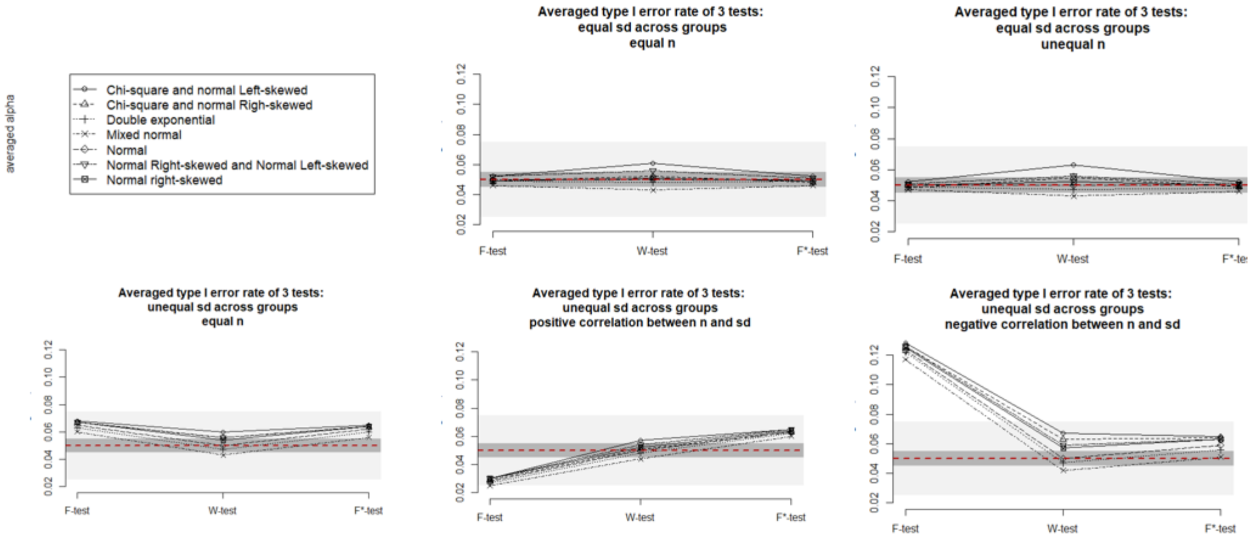


Figure 1. Type I error rate of the F-test, W-test and F*-test under 5 subcategories. Top left: cell a in Table 1; top right: cells b and c; bottom left: cells d and g; bottom middle: cells e and i; bottom right: cells f and h.

Simulating datasets equal variances between groups. In the two top plots in Figure 1 (see cells a, b, and c in Table 1), the population variance is equal between all groups, meaning that the homoscedasticity assumption is met. One observes that the F-test and F^*-test only marginally deviates from the nominal 5%, regardless the underlying distribution and the SD-ratio. The W-tests also only marginally deviates from the nominal

5%, except under asymmetry (the tests becomes a little more sensitive) or extremely heavy tails (the test becomes more conservative), consistently with (**???**). However, deviations don't exceed the liberal criterion of Bradley (1978).

Simulating datasets with unequal variances between groups. In the three bottom plots in Figure 1 (cells d to i, Table 1), the population variance is unequal between groups, meaning that the homoscedasticity assumption is not met.

When sample sizes are equal across groups (cells d and g, bottom left in Figure 1) and when there is a positive correlation between sample sizes and *SDs* (cells e and i, bottom center in Figure 1), the Type I error rate of the W-test is closer of the nominal 5% than the Type I error rate of the F*-test and even more than the F-test that is persistently at the lower limit of the liberal Bradley's interval, consistently with (**???**), Glass et al. (1972), Nimon (2012) and Overall et al. (1995). Moreover, the 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* (cells h and f, bottom right in Figure 1), 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) impacts a little more the type I error rate than when there is homoscedasticity. This is consistent with the statement of Lix et al. (1996) about the Alexander-Govern and the James second order tests (which return very similar results than the W-test, as we already mentioned). However, both tests still perform well considering liberal Bradley's interval, contrary to the F-test that is too liberal, consistently with (**???**), Glass et al. (1972), Nimon (2012) and Overall et al. (1995).

In summary. 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, while the *W*-test is a little less efficient than both *F*-test and *F**-test, departures from the nominal 5% never exceed the liberal criterion of Bradley (1978).
3) When the assumption of equal variances is violated, the *W*-test highly outperforms both the *F**-test (more liberal) and the *F*-test (either more liberal or more conservative, depending on the *SDs* and *SD* pairing).
4) The last conclusion generally remains true when both 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-β) of a test is the proportion of tests, when sampling many times from the same population, that rightly reject the main hypothesis (i.e. the hypothesis that all means are equal across groups; **???**). We assessed the power of the *F*-test, *W*-test and *F**-test, under 1280 scenarios, considering the nominal alpha level of 5%. In all scenarios, the last group was extracted from a population of mean of one unit more than the means of the other groups. Because of it, in some scenarios, there is a positive correlation between the *SD* and the mean (i.e. the last group has the largest *SD* and 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 **???**), the 9 subconditions in Table 1 were considered separately.

We computed two main outcomes: the consistency (Figure 2) and the power (Figure 3). The consistency refers to the relative difference between the observed power and the expected theoretical power (i.e. the power theoretically computed, based on the power curve of each test) divided by the expected power:

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

In other words, a consistency equal to zero shows that the observed power is consistent with theoretical 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.

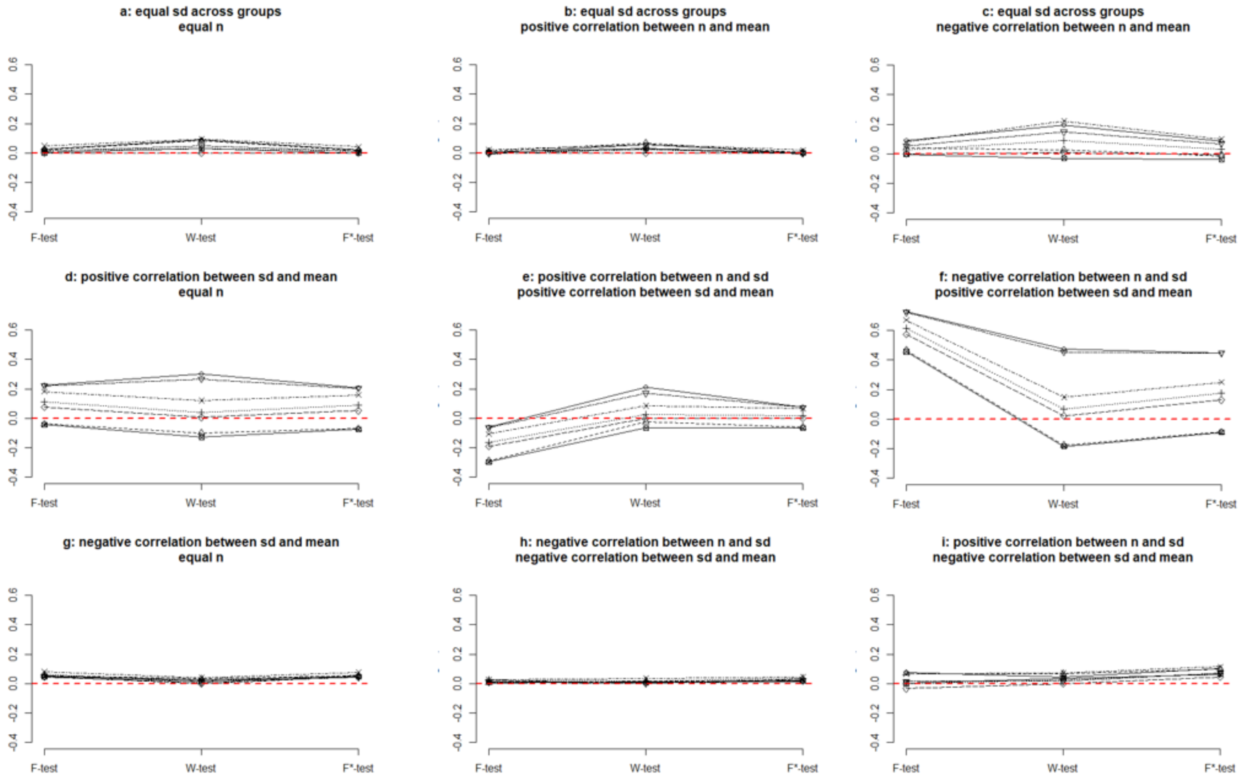


Figure 2. Consistency of the F-test, W-test and F*-test under 9 subcategories

Simulating datasets with equal variances between groups. In the three top plots in Figures 2 and 3 (plots a, b and c), the population variance is equal between all groups, meaning that the homoscedasticity assumption is met.

When distributions are normal, the *W*-test is slightly less powerful than the *F*-test and

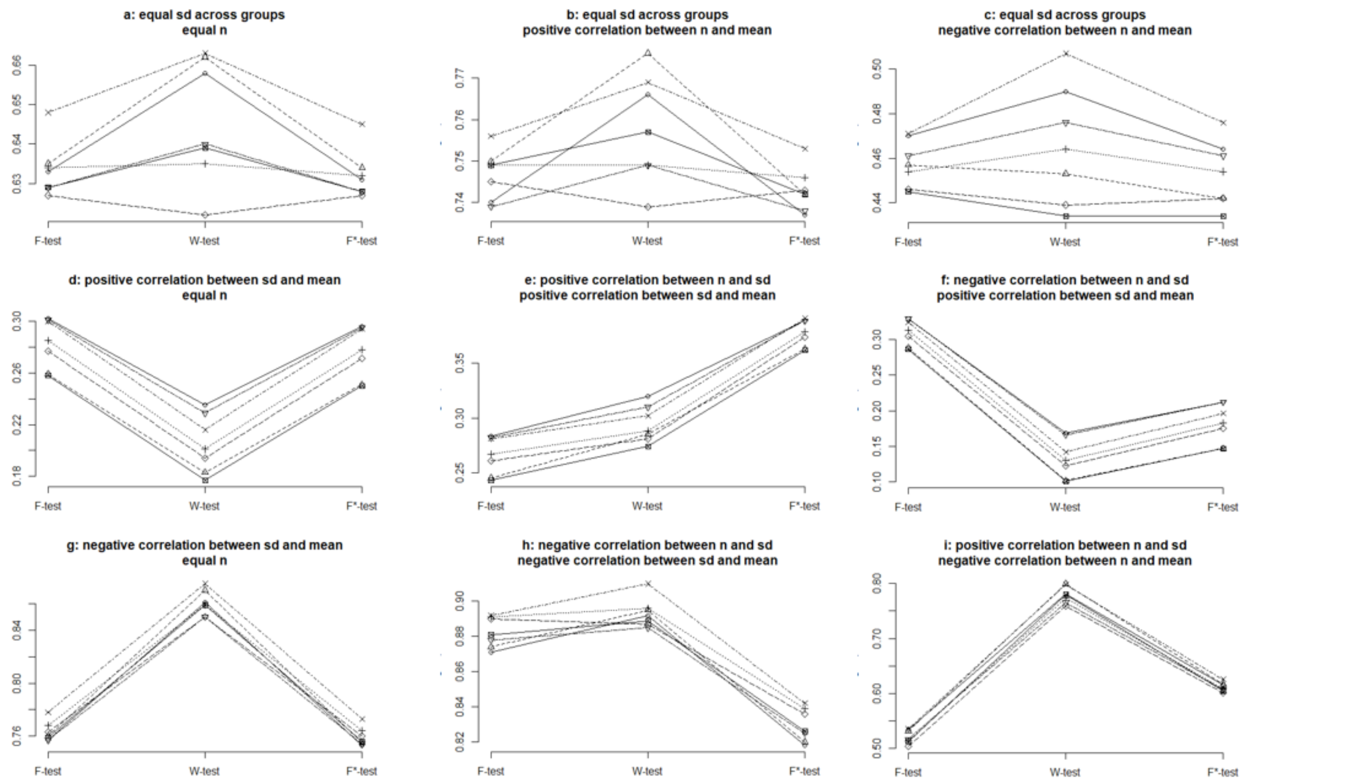


Figure 3. Power of the F-test, W-test and F*-test under 9 subcategories

*F**-test, even if differences are very small (always smaller than 3%). 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 comparing means from heavy-tailed distributions (e.g. double exponential or a mixed normal distribution) when compared to 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 is due to a common confusion between kurtosis and the standard deviation, while they are completely independent (DeCarlo, 1997), meaning that one can find distributions that have similar *SD* but different kurtosis (see Supplemental Material). Note that our observation on the power of the *W*-test is temperate by the fact that the consistency is a little smaller than the consistency of both other tests. Indeed, the *W*-test is more impacted by the distribution shape, consistently with (**???**). The W-test becomes generally more liberal when distributions underlying the data are not normal (especially with high kurtosis, or when asymmetries go in opposite directions). Notice that differences between W-test and other tests, in terms of consistency, are very small!

Simulating datasets with unequal variances between groups. In all but the three top plots in Figure 2 and 3 (plots d to i), the population variance is unequal between groups, meaning that the homoscedasticity assumption is not met.

When sample sizes are equal across groups (plots d and g in Figure 2 and 3), F-test and F*-tests are equally powerful, and have the same consistency, whatever the correlation between the *SD* and the mean. On the other side, the power of the W-test depends on correlation between the *SD* and the mean (consistently with **???**): when the group with the largest mean has the largest variance (plot d in Figure 2 and 3), 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 (plot g in Figure 2 and 3), 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* (plots f and h in Tables 2 and 3), the F-test is always more powerful than the F*-test. Indeed, as it 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* (plots e and i), the F-test is always more conservative than the F*-test. Indeed, as it was explained in the previous mathematical section, the *F*-test gives more weight to the largest variance (the statistic is therefore decreased) while the *F**-test gives more weight to the smallest variance (the statistic is therefore increased).

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, as previously mentioned: the test is more powerful when there is a negative correlation between *SDs* and means, and less powerful when there is a positive correlation between *SDs* and means.

Note that for all tests, the effect of heteroscedasticity is about the same regardless of the shape of the distribution. Moreover, there is one persistent observation in our simulations: whatever the configuration of *n*-ratio, the consistency of the three tests is closer of zero when there is a negative correlation between *SD* and mean (meaning that the group with the higest mean has the lower variance).

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

1) When all assumptions are met, the *W*-test slighly falls 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 *F*-test and *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 provide five recommendations:

1. Use the W-test instead of the F-test to compare groups 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 when using the W-test is very small (and often even negligible), and the gain in Type I 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 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.

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 I 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 group. When comparing more than four groups, the W-test should be avoided if the Shapiro-Wilk test rejects the normality assumption, with less than 100 subjects per group 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 (Erceg-Hurn & Mirosevich, 2008) [4] or nonparametric tests. For more information, see Erceg-Hurn and Mirosevich (2008).

4. Perform a-priori power-analyses. Fifty subjects per group are generally enough to control the Type I error rate, but power analyses are still 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 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 sample sizes between groups. Our simulations show that the Type I error rate control is in general slightly better with balanced designs.

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: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. doi: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.

[4] 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.

Child Indicators Research, *7*(3), 589–611. doi:10.1007/s12187-014-9232-z

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. doi: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: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: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: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. doi: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: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* (Erlbaum.). Mahwah, NJ.

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

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods*, *2*(3), 292–307. doi: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: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: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: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:10.1177/0022022104273654

Grissom, R. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, *68*(1), 155–165. doi: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:10.1016/j.jvb.2014.08.010

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not weird. *Nature*, *466*, 29–29. doi: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: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:10.3389/fpsyg.2012.00137

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

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:10.1177/0261927x14541280

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:10.3102/00346543066004579

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. doi: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: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: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. doi:10.2466/pr0.1995.76.3.1011

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:10.1080/00220973.1997.9943459

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: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:10.1007/s11199-009-9696-3

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

Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing* (Academic Press.). Cambridge, Massachusetts, US.

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

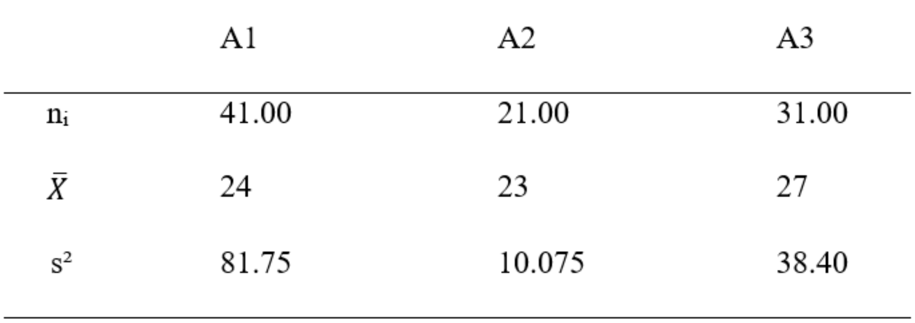
		***n*-ratio**		
		1	**>1**	**<1**
***SD*-ratio**	**1**	a	b	c
	>1	d	e	f
	<1	g	h	i

Appendix

The Mathematical Development of the *F*-test, *W*-test, and *F**-test: Numerical Example

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:

$$\frac{(41 * 24) + (21 * 23) + (31 * 27)}{41 + 21 + 31} = \frac{2304}{93} \approx 24.77$$

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

$$F = \frac{\frac{1}{3-1}[41 * (24 - \frac{2304}{93})^2 + 21 * (23 - \frac{2304}{93})^2 + 31 * (27 - \frac{2304}{93})^2]}{\frac{1}{93-3}[(41-1) * 81.75 + (21-1) * 10.07 + (31-1) * 38.40]} \approx 2.38$$

$$df_n = 3 - 1 = 2$$

$$df_d = 93 - 3 = 90$$

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

$$F^* = \frac{41 * (24 - \frac{2304}{93})^2 + 21 * (23 - \frac{2304}{93})^2 + 31 * (27 - \frac{2304}{93})^2}{(1 - \frac{41}{93}) * 81.75 + (1 - \frac{21}{93}) * 10.07 + (1 - \frac{31}{93}) * 38.40} \approx 3.09$$

$$df_n = 3 - 1 = 2$$

$$df_d = \frac{1}{\frac{(\frac{(1-\frac{41}{93})*81.75}{\sum_{j=1}^{k}(1-\frac{n_j}{N})s_j^2})^2}{41-1} + \frac{(\frac{(1-\frac{21}{93})*10.07}{\sum_{j=1}^{k}(1-\frac{n_j}{N})s_j^2})^2}{21-1} + \frac{(\frac{(1-\frac{31}{93})*38.40}{\sum_{j=1}^{k}(1-\frac{n_j}{N})s_j^2})^2}{31-1}} \approx 81.15$$

$$Where \sum_{j=1}^{k}(1 - \frac{n_j}{N}) * s_j^2 \approx 79.11$$

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

616 and 9:

$$W = \frac{\frac{1}{3-1}[\frac{41}{81.75}(24 - \bar{X}')^2 + \frac{21}{10.07}(23 - \bar{X}')^2 + \frac{31}{38.40}(27 - \bar{X}')^2]}{\frac{2(3-2)}{3^2-1}[(\frac{1}{41-1})(1 - \frac{\frac{41}{81.75}}{w})^2 + (\frac{1}{21-1})(1 - \frac{\frac{21}{10.07}}{w})^2 + (\frac{1}{31-1})(1 - \frac{\frac{31}{38.40}}{w})^2] + 1} \approx 4.61$$

617 Where:

618 $w = \sum_{j=1}^{k} w_j \approx 3.39$

$$\bar{X}\prime = \frac{\sum_{j=1}^{k}(w_j \bar{x_j})}{w} \approx 24.10$$

$$df_n = 3 - 1$$

$$df_d = \frac{3^2 - 1}{3[\frac{(1-\frac{w_j}{w})^2}{41-1} + \frac{(1-\frac{w_j}{w})^2}{21-1} + \frac{(1-\frac{w_j}{w})^2}{31-1}]} \approx 59.32$$

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*-test will also be a little too conservative, even if the test is less affected than the F-test. As a consequence: $W > F^* > F$.