- Taking Parametric Assumptions Seriously: Arguments for the Use of Welch's F-test instead of the Classical F-test in One-way ANOVA
- Marie Delacre¹, Christophe Leys¹, Youri L. Mora¹, & Daniël Lakens² 3
- 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 6

Author Note

- 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 10
- obtained from https://github.com/mdelacre/W-ANOVA. This work was supported by the 11
- Netherlands Organization for Scientific Research (NWO) VIDI grant 452-17-013. The
- authors declare that they have no conflicts of interest with respect to the authorship or the 13
- publication of this article.

2

- Correspondence concerning this article should be addressed to Marie Delacre, CP191, 15
- 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 18 samples are independent, and that independent and identically distributed residuals are 19 normal and have equal variances between groups. We focus on the assumptions of 20 normality and equality of variances, and argue that these assumptions are often unrealistic 21 in the field of psychology. We underline the current lack of attention to these assumptions 22 through an analysis of researchers' practices. Through Monte Carlo simulations we illustrate the consequences of performing the classic parametric F-test for ANOVA when the test assumptions are not met on the Type I error rate and statistical power. Under realistic deviations from the assumption of equal variances the classic F-test can yield severely biased results and lead to invalid statistical inferences. We examine two common alternatives to the F-test, namely the Welch's ANOVA (W-test) and the Brown-Forsythe test (F^* -test). Our simulations show that under a range of realistic scenariosthe W-test is 29 a better alternative and we therefore recommend using the W-test by default when 30 comparing means. We provide a detailed example explaining how to perform the W-test in 31 SPSS and R. We summarize our conclusions in practical recommendations that researchers 32 can use to improve their statistical practices. 33

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

36 Word count: 5870

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 40 (Erceg-Hurn & Mirosevich, 2008; Keselman et al., 1998; Tomarken & Serlin, 1986). Both 41 tests rely on the assumptions that independent and identically distributed residuals (1) are sampled from a normal distribution and (2) have equal variances between groups (or homoscedasticity; see Lix, Keselman, & Keselman, 1996). While a deviation from the normality assumption generally does not strongly affect either the Type I error rates (Glass, Peckham, & Sanders, 1972; Harwell, Rubinstein, Hayes, & Olds, 1992; Tiku, 1971) or the power of the F-test (David & Johnson, 1951; Harwell et al., 1992; Srivastava, 1959; Tiku, 1971), the F-test is not robust against unequal variances (Grissom, 2000). Unequal variances can alter both the Type I error rate (David & Johnson, 1951; Harwell et al., 1992) and statistical power (Nimon, 2012; Overall, Atlas, & Gibson, 1995) of the F-test. Although it important to make sure test assumptions are met before a statistical test 51 is performed, researchers rarely provide information about test assumptions when they 52 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 that the homogeneity of variances assumption was taken into account. They reported corrected degrees of freedom for unequal variances, which could signal the use of the W-test instead of the classical F-test. A similar investigation (Hoekstra, Kiers, & Johnson, 2012) yielded conclusions about the lack of attention to both the homoscedasticity and the normality assumptions. Despite the fact that the F-test is currently used by default, alternatives that are often a better choice exist, such as the Welch's W ANOVA (W-test), the Alexander-Govern test, 61 James' second order test, and the Brown-Forsythe ANOVA (F^* -test). Although not the

focus of the current article, additional tests exist that allow researchers to examine
hypotheses about other relevant parameters of a distribution than the mean (such as
standard deviations and the shape of the distribution (see for example Erceg-Hurn &
Mirosevich, 2008; Wilcox, 1998). However, since most researchers currently generate
hypotheses about differences between means (Erceg-Hurn & Mirosevich, 2008; Keselman et
al., 1998), we think that a first realistic first step towards progress would be to get
researchers to correctly test the hypothesis they are used to.

Although the debate surrounding the assumptions of the F-test has been widely 70 explored (see for example the meta-analysis of Harwell et al., 1992), applied researchers still largely ignore the consequences of assumption violations. Non-mathematical pedagogical papers summarizing the arguments seem to be lacking from the literature, and the current paper aims to fill this gap. We will discuss the pertinence of the assumptions of the F-test, and focus on the question of heteroscedasticity (that, as we will see, can have major consequences on error rates). We will provide a non-mathematical explanation of 76 how alternatives to the classical F-test cope with heteroscedasticity violations. We 77 conducted simulations in which we compare the F-test with the most promising alternatives. We argue that when variances are equal between groups, the W-test has nearly the same empirical Type I error rate and power as the F-test, but when variances are unequal, it provides empirical Type I and Type II error rates that are closer to the 81 expected levels compared to the F-test. Since the W-test is available in practically all statistical software packages, researchers can immediately improve their statistical inferences by replacing the F-test by the W-test.

Normality and Homogeneity of variances under Ecological Conditions

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

89 Normality Assumption

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

First, while it is obvious that the mean can be influenced by the treatment effects, 95 experimental treatment could also change the shape of a distribution, either by influencing the skewness, quantifying the asymmetry of the shape of the distribution, and kurtosis, a measure of the tendency to produce extreme values. A distribution with positive kurtosis will have heavier tails than the normal distribution, which means that extreme values will be more likely, while a distribution with negative kurtosis will have lighter tails than the 100 normal distribution, meaning that extreme values will be less likely (Westfall, 2014; 101 Wilcox, 2005). For example, a training aiming at reducing a bias perception of threat when 102 being exposed to ambiguous words will not uniformly impact the perception of all 103 participants, depending on their level of anxiety (Grey & Mathews, 2000). This could 104 influence the kurtosis of the distribution of bias score. 105

Second, prior to any experimental treatment, the presence of several subpopulations may lead to departures from the normality assumptions. Subgroup might exist that are unequal on some characteristics relevant to the measurements, that are not controlled within the studied group, which results in mixed distributions. This unavoidable lack of control is inherent of our field given its complexity. As an illustration, Wilcox (2005) writes that pooling two normally-distributed populations that have the same mean but different variances (e.g. normally distributed scores for schizophrenic and not schizophrenic participants) could result in distributions that are very similar to the normal curve, but with thicker tails. As another example, when assessing a wellness score for the general

population, data may be sampled from a left-skewed distribution, because most people are probably not depressed (see Heun, Burkart, Maier, & Bech, 1999). In this case, people who suffer from depression and people who do not suffer from depression are part of the same population, which can leads to asymmetry in the distribution.

Third, bounded measures can also explain non-normal distributions. Examples can
be found in the fields that analyze reaction time data, where measurements can be very
large, but never below zero, which results in right-skewed distributions; see Ratcliff (1979)
for a discussion on the shape of reaction time distributions. In sum, there are many
common situations in which normally distributed data is an unlikely assumption.

²⁴ Homogeneity of Variances Assumption

Homogeneity of variances (or homoscedasticity) is a mathematical requirement that 125 is also ecologically unlikely (Erceg-Hurn & Mirosevich, 2008; Grissom, 2000). In a previous 126 paper (Delacre, Lakens, & Leys, 2017), we identified three different causes of 127 heteroscedasticity: the variability inherent to the use of measured variables, the variability 128 induced by quasi-experimental treatments on measured variables, and the variability 129 induced by different experimental treatments on randomly assigned subjects. One 130 additional source of variability is the presence of unidentified moderators (Cohen, Cohen, 131 West, & Aiken, 2013). 132

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; 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.

Second, a quasi-experimental treatment can have different impacts on variances 147 between pre-existing groups, that can even be of theoretical interest. For example, in the 148 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 150 differences between sexist attitude scores of subjects who read a text in English (i.e. a 151 language without grammatical gender) or in Spanish (i.e. a language with grammatical 152 gender). The results showed that (for a reason not explained by the authors), the women's 153 score on the sexism dimension was more variable when the text was read in Spanish than 154 in English $(SD_{spanish} = .80 > SD_{english} = .50)$. For men, the reverse was true 155 $(SD_{spanish} = .97 < SD_{english} = 1.33)^{-1}$ 156

Third, even when the variances of groups are the same before treatment (due to a 157 complete successful randomization in group assignment), unequal variances can emerge 158 later, as a consequence of an experimental treatment (Box, 1954; Bryk & Raudenbush, 159 1988; Cumming, 2005; Erceg-Hurn & Mirosevich, 2008; Keppel & Wickens, 2004). For 160 example, Koeser and Sczesny (2014) have compared arguments advocating either masculine 161 generic or gender-fair language with control messages in order to test the impact of these 162 conditions on the use of gender-fair wording (measured as a frequency). They report that 163 the standard deviations increase after treatment in all experimental conditions. # 164

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

Consequences of Assumption Violations. Assumptions violations would not be a matter 165 per se, if the F-test was perfectly robust against departures from them (Glass et al., 1972). 166 When performing a test, two types of errors can be made: Type I errors and Type II 167 errors. A Type I error consists of falsely rejecting the null hypothesis in favour of an 168 alternative hypothesis, and the Type I error rate (α) is the proportion of tests that, when 169 sampling many times from the same population, reject the null hypothesis when there is no 170 true effect in the population. A Type II error consists of failing to reject the null 171 hypothesis, and the Type II error rate (β) is the proportion of tests, when sampling many 172 times from the same population, that fail to reject the null hypothesis when there is a true 173 effect. Finally, the statistical power $(1-\beta)$ is the proportion of tests, when sampling many 174 times from the same population, that correctly reject the null hypothesis when there is a 175 true effect in the population.

177 Violation of the Normality Assumption

Regarding the Type I error rate, the shape of the distribution has very little impact on the F-test (Harwell et al., 1992). When departures are very small (i.e. a kurtosis between 1.2 and 3 or a skewness between -.4 and .4), the Type I error rate of the F-test is very close to expectations, even with sample sizes as small as 11 subjects per group (Hsu & Feldt, 1969).

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

Violation of Homogeneity of Variances Assumption

Regarding the Type I error rate, the F-test is sensitive to unequal variances (Harwell 191 et al., 1992). More specifically, the more unequal the SD of the populations samples are 192 extracted from, the higher the impact. When there are only two groups, the impact is 193 smaller than when there are more than two groups (Harwell et al., 1992). When there are 194 more than two groups, the F-test becomes more liberal, meaning that the Type I error rate 195 is larger than the nominal alpha level, even when sample sizes are equal across groups 196 (Tomarken & Serlin, 1986). Moreover, when sample sizes are unequal, there is a strong 197 effect of the sample size and variance pairing. In case of a positive pairing (i.e. the group 198 with the larger sample size also has the larger variance), the test is too conservative, 199 meaning that the Type I error rate of the test is lower than the nominal alpha level, 200 whereas in case of a negative pairing (i.e. the group with the larger sample size has the 201 smaller variance), the test is too liberal (Glass et al., 1972; Nimon, 2012; Overall et al., 202 1995; Tomarken & Serlin, 1986). 203

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

comulative Violation of Normality and Homogeneity of Variance

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

Based on mathematical explanations and Monteo Carlo simulations we chose to compare the F-test with the W-test and F*-test and to exclude the James' second-order

229

and Alexander-Govern's test because the latter two yield very similar results to the

W-test, but are less readily available in statistical software packages. Tomarken and Serlin

(1986) have shown that from the available alternatives, the F*-test and the W-test

perform best, and both tests are available in SPSS, which is widely used software in the

psychological sciences (Hoekstra et al., 2012). For a more extended description of the

James' second-order and Alexander-Govern's test, see Schneider and Penfield (1997).

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

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

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

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

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

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

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

253

denominator in (1) is underestimated. The F-value is then inflated, which yields more significant results than expected.

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

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

Where x_j and s_j^2 are respectively the group mean and the group variance, and $\bar{x}_{..}$ is the 240 overall mean. As it can be seen in (4) the numerator of the F^* statistic is equal to the sum 241 of squares between groups (which is equal to the numerator of the F statistic when one 242 compares two groups). In the denominator, the variance of each group is weighted by 1 243 minus the relative frequency of each group. This adjustment implies that the variance 244 associated with the group with the smallest sample size is given more weight compared to 245 the F-test. As a result, when the larger variance is associated with the larger sample size, 246 F^* is larger than F, because the denominator decreases, leading to more significant 247 findings compared to the F-test. On the other hand, when the larger variance is associated 248 with the smaller sample size, F^* is smaller than F, because the denominator increases, 240 leading to fewer significant findings compared to the F-test. The degrees of freedom in the numerator and in the denominator of F^* -test are computed as follows (with the same 251 principle as the denominator computation of the F^* statistic):

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

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

$$(6)$$

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

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

258 where:

259

260

261

262

268

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

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

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

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

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

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

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

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

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

Population parameter values were chosen in order to illustrate the consequences of factors known to play a key role on both the Type I error rate and the statistical power

when performing an ANOVA. Based on the literature review presented above, we 279 manipulated the number of groups, the sample sizes, the sample size ratio, the SD-ratio 280 $(SD\text{-ratio} = \frac{\sigma_k}{\sigma_j})$, and the sample size and variance pairing. In our scenarios, the number of 281 compared groups (k) varied from 2 to 5. Sample sizes of k-1 groups (n_i) were 20, 30, 40, 282 50, or 100. The sample size of the last group was a function of the sample size ratio 283 $(n\text{-ratio} = \frac{n_k}{n_i})$, ranging from 0.5 to 2, in steps of 0.5. The simulations for which the *n*-ratio 284 equals 1 are known as a balanced design (i.e. sample sizes are equal across all groups). The 285 SD of the population from which was extracted last group was a function of the SD-ratio, 286 with values of 0.5, 1, 2 or 4. The simulations for which the SD-ratio equals 1 are the 287 particular case of homoscedasticity (i.e. equal variances across groups). 288

All possible combinations of n-ratio and SD-ratio were performed in order to 289 distinguish positive pairings (the group with the largest sample size is extracted from the 290 population with the largest SD), negative pairings (the group with the smallest sample size 291 is extracted from the population with the smallest SD), and no pairing (sample sizes 292 and/or population SD are equal across all groups). All these conditions were tested with 293 normal and non-normal distributions. When two groups are compared, conclusions for the 294 three ANOVA tests (F, F^*, W) should yield identical error rates when compared to their equivalent t-tests (the F-test is equivalent to Student's t-test, and the F^* -test and W-test are equivalent to Welch's t-test; Delacre et al., 2017). When there are more than three 297 groups, the F-test becomes increasingly liberal as soon as the variances of the distributions 298 in each group are not similar, even when sample sizes are equal between groups (Harwell et 299 al., 1992; Quensel, 1947). 300

For didactic reasons, we will report only the results where we compare three groups (k=3). Increasing the number of groups increases how liberal all tests are. For interested readers, all figures for cases where we compare more than three groups are available here: https://osf.io/h4ks8/. Overall, the larger the sample sizes, the less the distributions of the population underlying the samples impact the robustness of the tests (Srivastava, 1959).

However, increasing the sample sizes does not improve the robustness of the test when
there is heteroscedasticity. Interested reader can see all details in the following Excell
spreadsheet, available on github: « Type I error rate.xlsx ».

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

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

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

Note. The *n*-ratio is the sample size of the last group divided by the sample size of 313 the first group. When all sample sizes are equal across groups, the n-ratio equals 1. When 314 the sample size of the last group is higher than the sample size of the first group, n-ratio > 315 1, and when the sample size of the last group is smaller than the sample size of the first 316 group, n-ratio < 1. SD-ratio is the population SD of the last group divided by the 317 population SD of the first group. When all samples are extracted from populations with 318 the same SD, the SD-ratio equals 1. When the last group is extracted from a population 319 with a larger SD than all other groups, the SD-ratio > 1. When the last group is extracted from a population with a smaller SD than all other groups, the SD-ratio < 1.

In all Figures presented below, averaged results for each subcondition are presented

under 7 different configurations of distributions, using the following legend:

- -- Chi-square and normal Left-skewed
- △- Chi-square and normal Righ-skewed
- —— Double exponential
- -⇔- Normal
- Normal right-skewed

Figure 1. Legend

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

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

When there is no difference between means, the 9 cells of Table 1 simplify into 5 subconditions:

• Equal n and SD across groups (a)

332

333

• 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 2 to 6, we computed the average Type I error rate of the three tests under these 5 subcategories. The light grey area corresponds to the liberal criterion from Bradley (1978), who regards a departure from the nominal alpha level as acceptable whenever the Type I error rate falls within the interval $[.5 \times \alpha; 1.5 \times \alpha]$. The dark grey area corresponds to the more conservative criterion from which departures from the nominal alpha is considered negligible as long as the Type I error rate falls within the interval $[.9 \times \alpha; 1.1 \times \alpha]$.

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

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

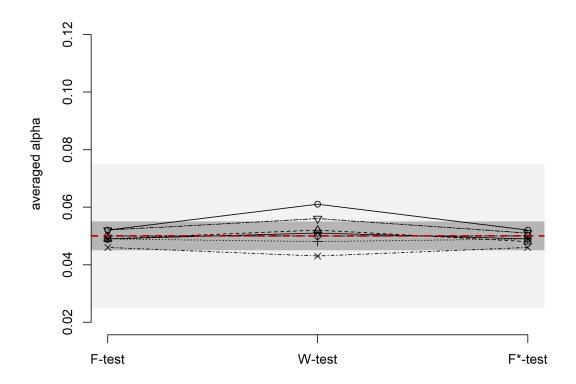


Figure 2. Type I error rate of the F-test, W-test and F*-test when there are equal SDs across groups and equal sample sizes (cell a in Table 1)

does not impact the Type I error rate of the W-test, regardless of the distribution (the order of the distribution shape remains the same in all conditions).

When there is a negative correlation between sample sizes and SDs (Figure 6), the Type I error rate of the F^* -test is slightly closer of the nominal 5% than the Type I error rate of the W-test, for which the distributions (more specifically, the skewness) has a larger impact on the Type I error rate than when there is homoscedasticity. This is consistent with conclusions by Lix et al. (1996) about the Alexander-Govern and the James' second order tests (which return very similar results as the W-test, as we already mentioned). However, both tests still perform relatively well, contrary to the F-test that is much too

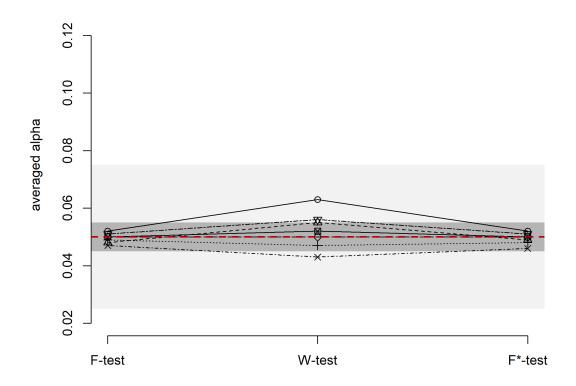


Figure 3. Type I error rate of the F-test, W-test and F*-test when there are equal SDs across groups and unequal sample sizes (cells b and c in Table 1)

liberal, in line with observations by Harwell et al. (1992), Glass et al. (1972), Nimon (2012) and Overall et al. (1995).

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

1) When all assumptions are met, all tests perform adequately.

371

372

377

- When variances are equal between groups and distributions are not normal, the W-test is a little less efficient than both the F-test and the F^* -test, but departures
 from the nominal 5% Type I error rate never exceed the liberal criterion of Bradley
 (1978).
 - 3) When the assumption of equal variances is violated, the W-test clearly outperforms

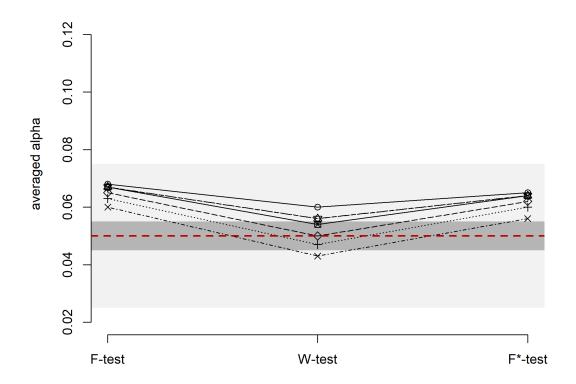


Figure 4. Type I error rate of the F-test, W-test and F*-test when there are unequal SDs across groups and equal sample sizes (cells d and g in Table 1)

- both the F^* -test (which is more liberal) and the F-test (which is either more liberal or more conservative, depending on the SDs and SD pairing).
- 4) The last conclusion generally remains true when both the assumptions of equal variances and normality are not met.

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

378

380

381

As previously mentioned, the statistical power $(1-\beta)$ of a test is the long-run probability of observing a statistically significant result when there is a true effect in the population. We assessed the power of the F-test, W-test and F*-test under 1280 scenarios, while using the nominal alpha level of 5%. In all scenarios, the last group was extracted

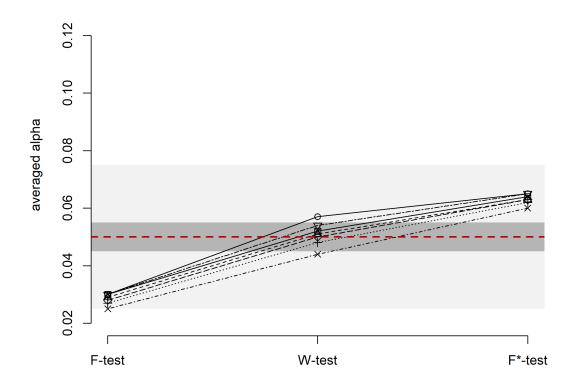


Figure 5. Type I error rate of the F-test, W-test and F*-test when there are unequal SDs across groups, and positive correlation between sample sizes and SDs (cells e and i in Table 1)

from a population that has a higher mean than the population from where were extracted all other groups ($\mu_k = \mu_j + 1$). Because of this, in some scenarios there is a positive correlation between the SD and the mean (i.e. the last group has the largest SD and the largest mean) and in other scenarios, there is a negative correlation between SD and the mean (i.e. the last group has the smallest SD and the largest mean). As we know that the correlation between the SD and the mean matters for the W-test (see Liu, 2015), the 9 subconditions in Table 1 were analyzed separately.

We computed two main outcomes: the consistency and the power. The consistency

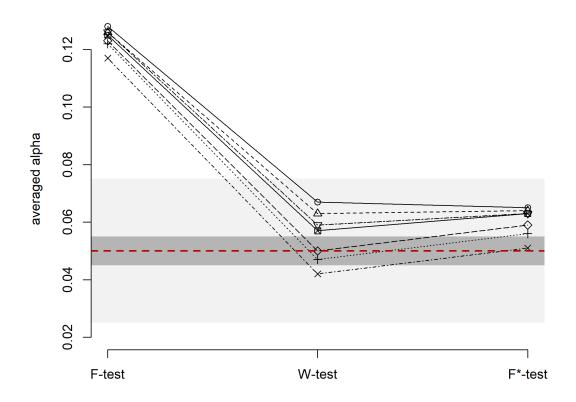


Figure 6. Type I error rate of the F-test, W-test and F*-test when there are unequal SDs across groups, and negative correlation between sample sizes and SDs (cells f and g in Table 1)

refers to the relative difference between the observed power and the nominal power, divided by the expected power:

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

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

In Figures 7, 8 and 9 (cells a, b, and c in Table 1), the population variance is equal

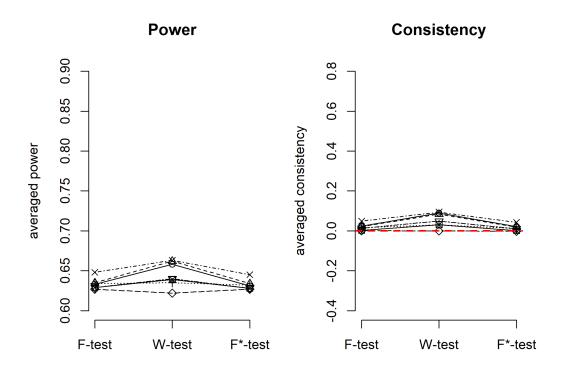


Figure 7. Power and consistency of the F-test, W-test and F*-test when there are equal SDs across groups and equal sample sizes (cell a in Table 1)

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 F^* -test, 403 even through differences are very small. With all other distributions, the W-test is 404 generally more powerful than the F^* -test and F-test, even with heavy-tailed distributions, 405 which is in contrast with previous findings (Wilcox, 1998). Wilcox (1998) concluded that there is a loss of power when means from heavy-tailed distributions (e.g. double exponential or a mixed normal distribution) are compared to means from normal distributions. This finding is based on the argument that heavy-tailed distributions are 409 associated with bigger standard deviations than normal distributions, and that the effect 410 size for such distributions is therefore smaller (Wilcox, 2011). However, this conclusion is 411

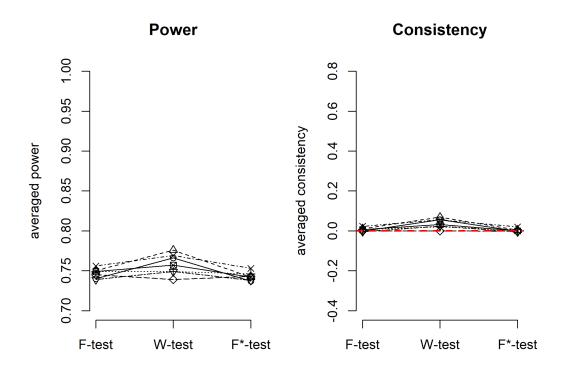


Figure 8. Power and consistency of the F-test, W-test and F*-test when there are equal SDs across groups, and positive correlation between sample sizes and means (cell b in Table 1)

based on a common conflation of kurtosis and the standard deviation, which are completely 412 independent (DeCarlo, 1997). One can find distributions that have similar SD but different 413 kurtosis (see Appendix 2). However, while the W-test is more powerful than the F-test 414 and the F^* -test in many situations, it is a bit less consistent with theoretical expectations than both other tests in the sense that the W-test is generally more powerful than 416 expected (especially with high kurtosis, or when asymmetries go in opposite directions). 417 This is due to the fact that the W-test is more impacted by the distribution shape, in line 418 with observations by Harwell et al. (1992). Note that differences between W-test and other 419 tests, in terms of consistency, are very small. 420

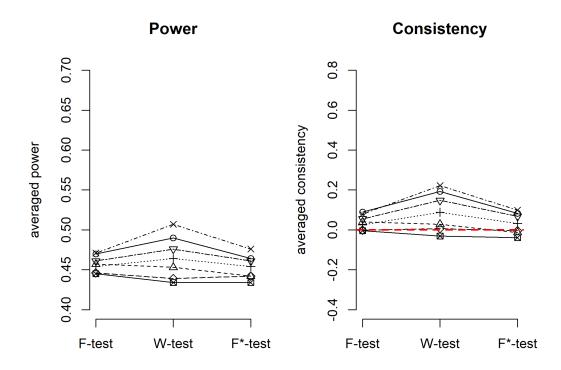


Figure 9. Power and consistency of the F-test, W-test and F*-test when there are equal SDs across groups, and negative correlation between sample sizes and means (cell c in Table 1)

In Figures 10 to 15 (cells d to i in Table 1), the population variance is unequal 421 between groups, meaning that the homoscedasticity assumption is not met. When sample 422 sizes are equal across groups (Figures 10 and 11), the F-test and the F^* -tests are equally 423 powerful, and have the same consistency, whatever the correlation between the SD and the 424 mean. On the other hand, the power of the W-test depends on the correlation between the 425 SD and the mean (in line with Liu, 2015). When the group with the largest mean has the 426 largest variance (Figure 10), the largest deviation between group means and the general 427 mean is given less weight, and as a consequence the W-test is less powerful than both other 428 tests. At the same time, the test is slightly less consistent than both other tests. When the 429 group with the largest mean has the smallest variance (Figure 11), the largest deviation 430

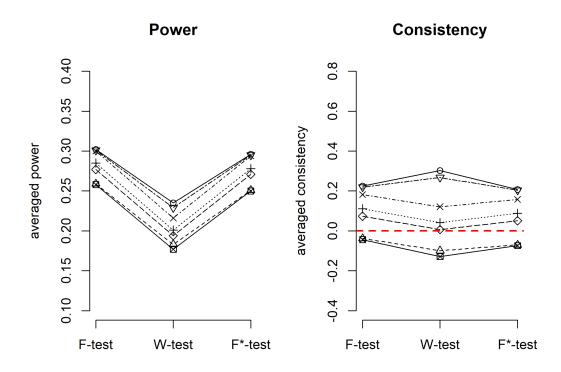


Figure 10. Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, positive correlation between SDs and means, and equal sample sizes across groups (cell d in Table 1)

between group means and the general mean is given more weight, and therefore the W-test is more powerful than both other tests. The test is also slightly more consistent than both other tests.

When sample sizes are unequal across groups, the power of the F^* -test and the F-test are a function of the correlation between sample sizes and SDs. When there is a negative correlation between sample sizes and SDs (Figures 12 and 13), the F-test is always more powerful than the F^* -test. Indeed, as was explained in the previous mathematical section, the F-test gives more weight to the smallest variance (the statistic is

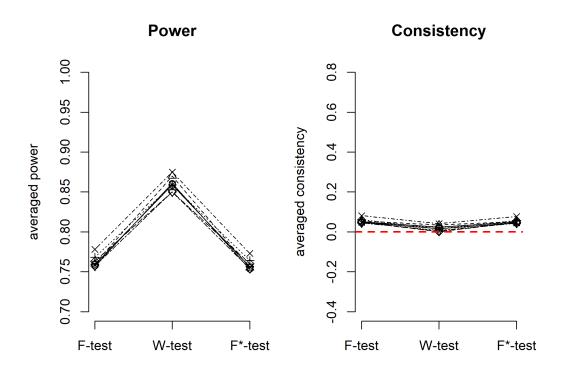


Figure 11. Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, negative correlation between SDs and means, and equal sample sizes across groups (cell g in Table 1)

therefore increased) while the F^* -test gives more weight to the largest variance (the statistic is therefore decreased). Conversely, when there is a positive correlation between sample sizes and SDs (Figures 14 and 15), the F-test is always more conservative than the F^* -test, because the F-test gives more weight to the largest variance while the F^* -test gives more weight to the smallest variance.

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

454

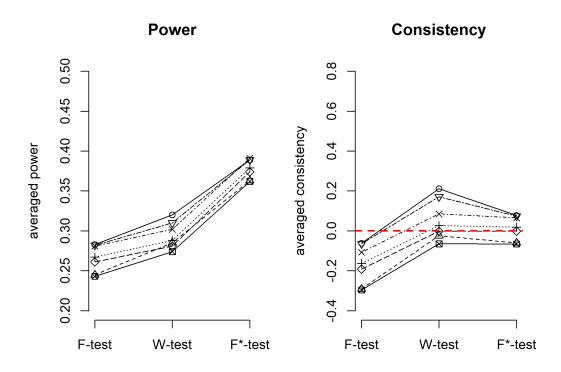


Figure 12. Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, negative correlation between sample sizes and SDs, and positive correlation between SDs and means (cell f in Table 1)

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

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

1) When all assumptions are met, the W-test falls slightly behind the F-test and the

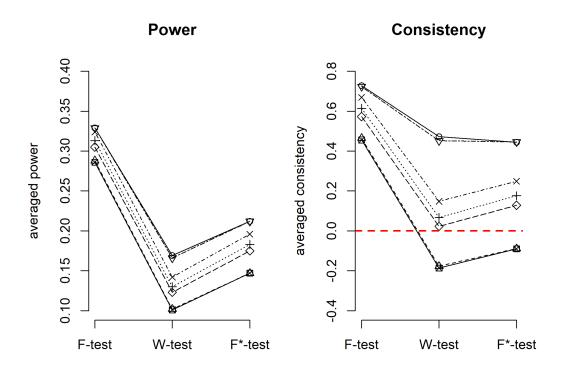


Figure 13. Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, negative correlation between sample sizes and SDs, and negative correlation between SDs and means (cell h in Table 1)

 F^* -test, both in terms of power and consistency.

455

456

457

458

459

460

461

462

463

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

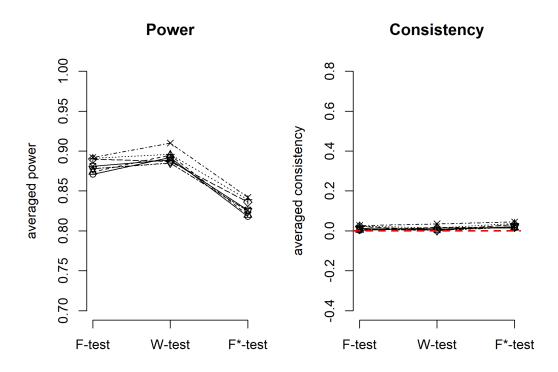


Figure 14. Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, positive correlation between sample sizes and SDs, and positive correlation between SDs and means (cell e in Table 1)

and normality are not met.

465

Recommendations

Taking both the effects of the assumption violations on the alpha risk and on the power, we recommend to use the W-test instead of the F-test to compare groups means. The F-test and F*-test should be avoided, because a) the equal variances assumption is often unrealistic, b) tests of the equal variances assumption will often fail to detect differences when these are present, c) the loss of power when using the W-test is very small (and often even negligible), and d) the gain in Type I error control is considerable under a

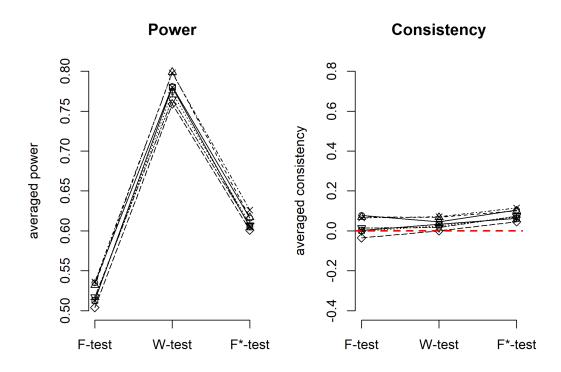


Figure 15. Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, positive correlation between sample sizes and SDs, and negative correlation between SDs and means (cell i in Table 1)

wide range of realistic conditions. Also, we recommend the use of balanced designs
(i.e. same sample sizes in each group) whenever possible. When using the W-test, the
Type I error rate is a function of criteria such as the skewness of the distributions, and
whether skewness is combined with unequal variances and unequal samples sizes between
groups. Our simulations show that the Type I error rate control is in general slightly better
with balanced designs.

Note that the W-test suffers from limitations and cannot be used in all situations. First, as previously mentioned, W-test, as all tests based on means, does not allow

researchers to compare other relevant parameters of a distribution than the mean. For 480 these reason, we recommend to never neglect the descriptive analysis of the data. A 481 complete description of the shape and characteristics of the data (e.g. histograms and 482 boxplots) is important. When at least one statistical parameter relating to the shape of the 483 distribution (e.g. variance, skewness, kurtosis) seems to vary between groups, comparing 484 results of the W-test with results of a nonparametric procedure is useful in order to better 485 understand the data. Second, with small sample sizes (i.e. less than 50 observations per 486 group when comparing at most four groups, 100 observations when comparing more than 487 four groups), the W-test will not control Type I error rate when skewness is present and 488 detecting departures for normality is therefore especially important in small samples. 489 Unless you have good reasons to believe that distributions underlying the data have small 490 kurtosis and skewness, we recommend to avoid alternative tests that are not based on 491 means comparison, in favour of alternatives such as the trimmed means test (Erceg-Hurn & Mirosevich, 2008) ² or nonparametric tests. For more information, see Erceg-Hurn and Mirosevich (2008).

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

495 References

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

Beilmann, M., Mayer, B., Kasearu, K., & Realo, A. (2014). The relationship between adolescents' social capital and individualism-collectivism in estonia, germany, and

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

- Boneau, C. (1960). The effects of violations of assumptions underlying the t test.

 Psychological Bulletin, 57(1), 49–64.
- Box, G. (1954). Some theorems on quadratic forms applied in the study of analysis of
 variance problems, i. Effect of inequality of variance in the one-way classification.

 The Annals of Mathematical Statistics, 25(2), 290–302.
- 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.
- David, F. N., & Johnson, N. L. (1951). The effect of nonnormality on the power function of the f-test in the analysis of variance. *Biometrika*, 38(1-2), 43–57. doi:10.2307/2332316
- 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
- Grey, S., & Mathews, A. (2000). Effects of training on interpretation of emotional
 ambiguity. The Quarterly Journal of Experimental Psychology, 53(4), 1143–1162.
 doi:10.1080/713755937
- 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
- Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing
 monte carlo results in methodological research: The one- and two-factor fixed effects
 anova cases. Journal of Educational Statistics, 17(4), 315–339.
- 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
- Hsu, T.-C., & Feldt, L. S. (1969). The effect of limitations on the number of criterion score
 values on the significance level of the f-test. *American Educational Research*Journal, 6(4), 515–527. doi:10.3102/00028312006004515

- Keppel, G., & Wickens, T. D. (2004). Design and analysis: A researcher's handbook (Prentice Hall.). Upper Saddle River, New Jersey.
- Keselman, H., 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 analy. Review of Educational Research, 68(3),
- 350-386.
- 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
- Liu, H. (2015). Comparing welch anova, a kruskal-wallis test, and traditional anova in case
 of heterogeneity of variance (PhD thesis). Virginia Commonwealth University.
- Lix, L. M., Keselman, J. C., & Keselman, H. (1996). Consequences of assumption
- violations revisited: A quantitative review of alternatives to the one-way analysis of
- variance *f* test. Review of Educational Research, 66(4), 579–619.
- doi: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
- Quensel, C.-E. (1947). The validity of the *z*-criterion when the variates are taken from
- different normal populations. Scandinavian Actuarial Journal, 30(1), 44–55.
- doi:10.1080/03461238.1947.10419648
- 8599 Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution
- statistics. Psychological Bulletin, 86(3), 446–461. doi:10.1037/0033-2909.86.3.446
- Schneider, P. J., & Penfield, D. A. (1997). Alexander and govern's approximations:
- Providing an alternative to anova under variance heterogeneity. The Journal of
- $Experimental\ Education,\ 65(3),\ 271-286.\ doi:10.1080/00220973.1997.9943459$
- 604 Srivastava, A. B. L. (1959). Effects of non-normality on the power of the analysis of
- variance test. *Biometrika*, 46(1-2), 114–122. doi:10.2307/2332813
- Tiku, M. (1971). Power function of the f-test under non-normal situations. Journal of the
- 607 American Statistical Association, 66, 913–916. doi:10.1080/01621459.1971.10482371
- Tomarken, A. J., & Serlin, R. C. (1986). Comparison of anova alternatives under variance
- heterogeneity and specific noncentrality structures. Psychological Bulletin, 99(1),
- 90–99. doi: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
- Westfall, P. H. (2014). Kurtosis as peakedness, 1905-2014. R.I.P. The American
- Statistician, 68(3), 191-195. doi:10.1080/00031305.2014.917055
- Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical
- methods? American Psychologist, 53(3), 300–314. doi:10.1037/0003-066x.53.3.300
- Wilcox, R. R. (2005). Comparing medians: An overview plus new results on dealing with
- heavy-tailed distributions. The Journal of Experimental Education, 73(3), 249–263.

doi:10.3200/JEXE.73.3.249-263

- 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

Appendix

Appendix 1: The Mathematical Development of the F-test, W-test, and

 F^* -test: Numerical Example

629

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: 631

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

The F-test statistic and degrees of freedom are computed by applying formulas (1), (2) and (3): 633

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

$$df_n = 3 - 1 = 2$$

$$df_d = 93 - 3 = 90$$

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

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

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

637 Where:

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

639
$$\bar{X'} = \frac{\sum_{j=1}^{k} (w_j \bar{x_j})}{w} \approx 24.10$$

$$df_n = 3 - 1$$

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

651

652

653

654

655

656

657

658

- One should notice that in this example, the biggest sample size has the biggest variance. 640
- As previously mentioned, it means that the F-test will be too conservative, because the F641
- value decreases. The F^* -test will also be a little too conservative, even if the test is less 642
- affected than the F-test. As a consequence: $W > F^* > F$. 643

Appendix 2: Justification for the choice of distributions in simulations

- The set of simulations described in the article was repeated for 7 distributions. We used R 645 commands to generate data from different distributions: 646
- k normal distributions (Figure A1): in order to assess the Type I error rate and power of the different tests under the assumption of normality, data were generated 648 by means of the function "rnorm" (from the package "stats"; "R: The Normal 649 Distribution," 2016). 650
 - k double exponential distributions (Figure A2): In order to assess the impact of high kurtosis on the Type I error rate and power of all tests, data were generated by means of the function "rdoublex" (from the package "smoothmest"; "R: The double exponential (Laplace) distribution," 2012).
 - k mixed normal distributions (Figure A3): In order to assess the impact of extremely high kurtosis on the Type I error rate and power of all tests, regardless of variance, data were generated by means of the function "rmixnorm" (from the package "bda"; Wang & Wang, 2015).
- k normal right skewed distributions (Figure A4): In order to assess the impact of 659 moderate skewness on the Type I error rate and power, data were generated by 660 means of the function "rsnorm" (from the package "fGarch"; "R: Skew Normal 661 Distribution," 2017). The normal skewed distribution was chosen because it is the 662

666

667

668

669

only skewed distribution where the standard deviation ratio can vary without having an impact on skewness.

- k-1 normal left skewed distributions (Figure A5) and 1 normal right skewed distribution (Figure A2.4): In order to assess the impact of unequal shapes, in terms of skewness, on the Type I error rate and power, when data have moderate skewness, data were generated by means of the functions "rsnorm" (from the package "fGarch"; "R: Skew Normal Distribution," 2017).
- k-1 chi-squared distributions with two degrees of freedom (See Figure A6), and one 670 normal rigt skewed distribution (Figure A2.4): In order to assess the impact of high 671 asymetry on the Type I error rate an power, k-1 distributions were generated by 672 means of the functions "rchisq" ("R: The (non-central) Chi-squared Distribution," 673 2016). The last distribution was generated by means of "rsnorm" in order to follow a 674 normal right skewed distribution with a mean of 2 (from the package "fGarch"; "R: 675 Skew Normal Distribution," 2017). Because the chi-squared is non-negative, it is not 676 possible to generate chi-squared where population SD= 1, 4 or 8 and population 677 mean is the same than the chi-squared with two degrees of freedom. However, we 678 wanted to assess the impact of different SD-ratio on Type I error rate. For these 679 reasons, the last distribution was generated by means of "rsnorm" in order to follow a 680 normal skewed distribution with positive skewness of +0.99 and mean = 2 (from the 681 package "fGarch"; "R: Skew Normal Distribution," 2017).
- k-1 chi-squared distributions with two degrees of freedom (See Figure A6), and one
 normal left skewed distribution (Figure A5): In order to assess the impact of unequal
 shapes, in terms of skewness, on Type I error rate and power when distributions have
 extreme skewness, k-1 distributions were generated by means of the functions "rchisq"

 ("R: The (non-central) Chi-squared Distribution," 2016). The last distribution was
 generated by means of "rsnorm" in order to follow a normal right skewed distribution

with a mean of 2 (from the package "fGarch"; "R: Skew Normal Distribution," 2017)

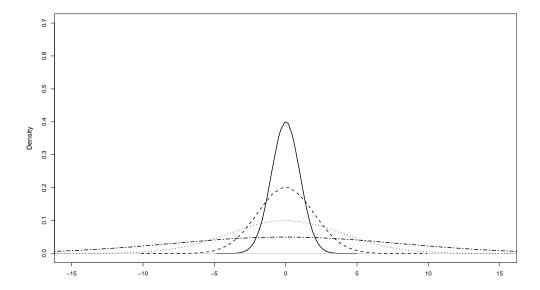


Figure A1. centered normal probability density function, as a function of the population SD

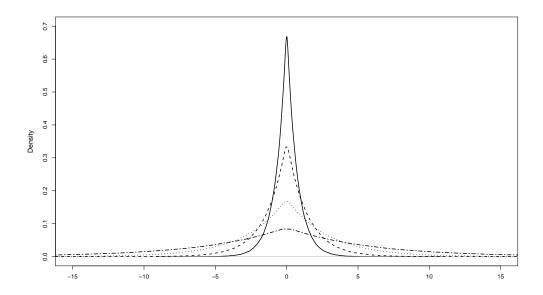


Figure A2. centered double exponential probability density function, as a function of the population SD

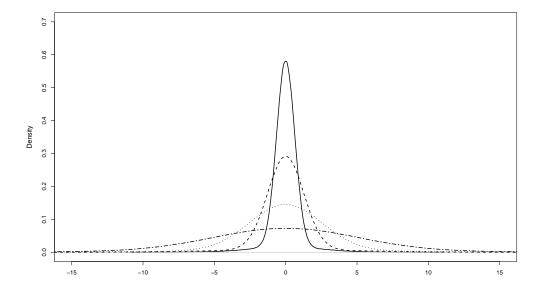


Figure A3. centered mixed normal probability density function, as a function of the population SD

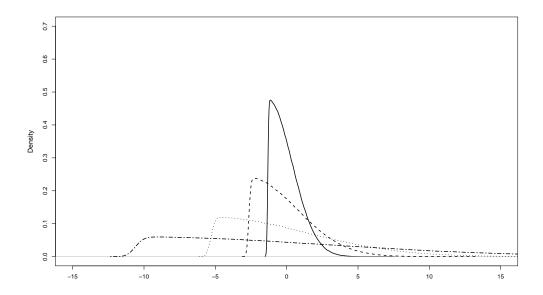


Figure A4. centered normal right skewed robability density function, as a function of the population SD

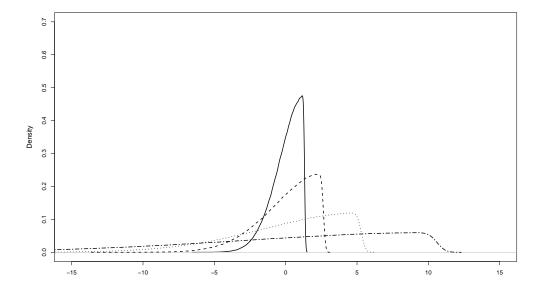
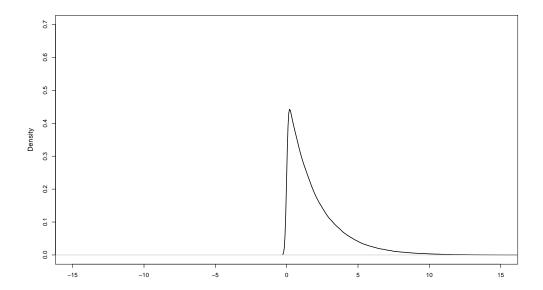


Figure A5. centered normal left skewed robability density function, as a function of the population SD



 $\label{eq:Figure A6} \emph{E}. \ \emph{Chi-squared with 2 degrees of freedom probability density function}, \ \emph{as a function}$ of the population SD