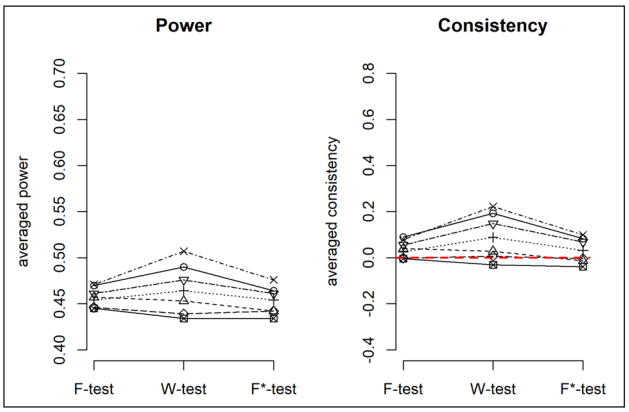

Figure 9: Power and consistency of the F-test, W-test and F*-test when there are equal SDs across groups, and negative correlation between sample sizes and means (cell c in Table 1).

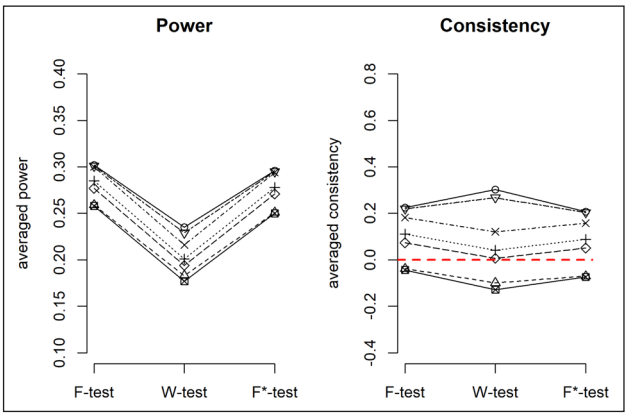

Figure 10: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, positive correlation between SDs and means, and equal sample sizes across groups (cell d in Table 1).

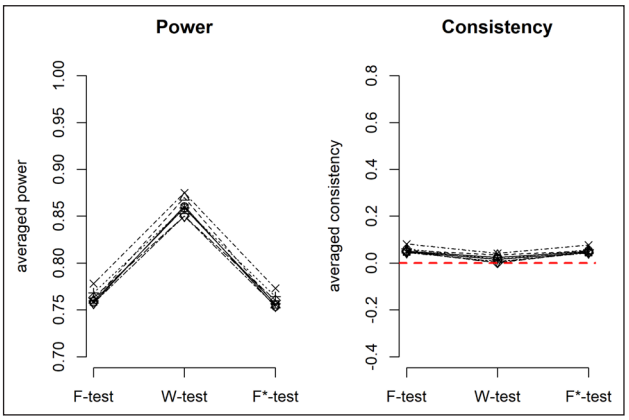

Figure 11: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, negative correlation between SDs and means, and equal sample sizes across groups (cell g in Table 1).

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

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

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

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