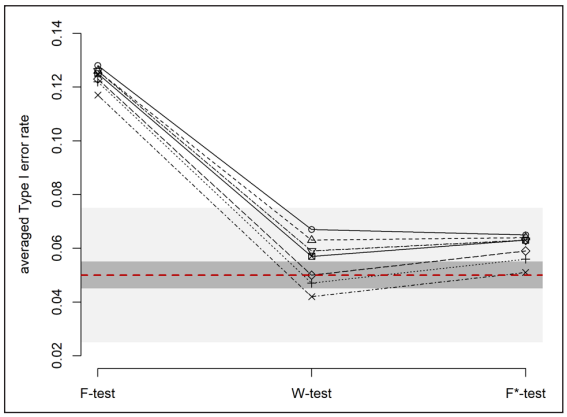

Figure 6: Type I error rate of the F-test, W-test and F*-test when there are unequal SDs across groups, and negative correlation between sample sizes and SDs (cells f and g in Table 1).

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

Conclusions

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

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

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

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

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

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

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

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

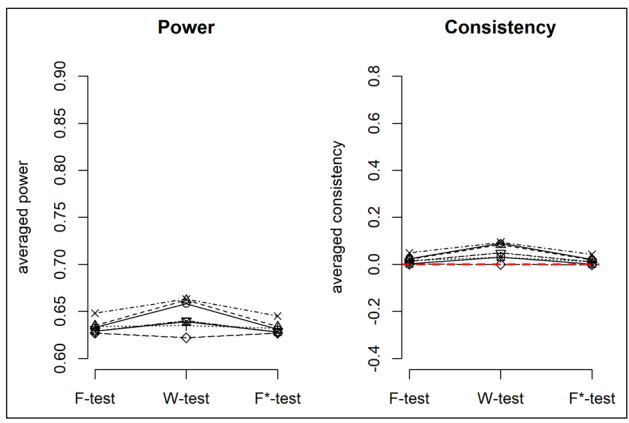

Figure 7: Power and consistency of the F-test, W-test and F*-test when there are equal SDs across groups and equal sample sizes (cell a in Table 1).

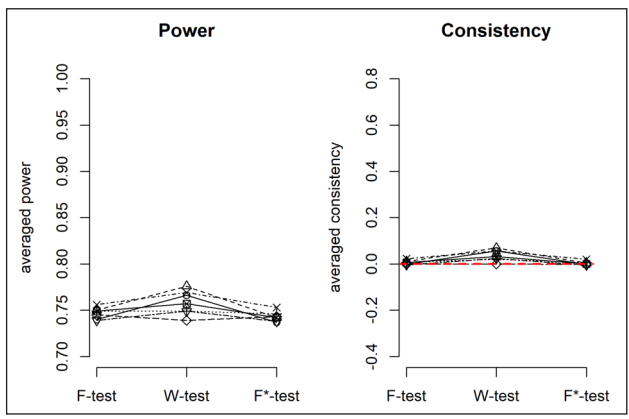

Figure 8: Power and consistency of the F-test, W-test and F*-test when there are equal SDs across groups, and positive correlation between sample sizes and means (cell b in Table 1).