the variability will be smaller compared to a school that accepts all students.

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 ($SD_{spanish} = .97 < SD_{english} = 1.33$).[1]

Third, even when the variances of groups are the same before treatment (due to a complete succesful randomization in group assignment), unequal variances can emerge later, as a consequence of an experimental treatment (Box, 1954; 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 treatment in all experimental conditions.

Consequences of Assumption Violations

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 errors can be made: Type I errors and Type II errors. A Type I error consists of falsely rejecting the null hypothesis in favour of an alternative hypothesis, and the Type I error rate (α) is the proportion of tests that, when sampling many times from the same population, reject the null hypothesis when there is no true effect in the population. A Type II error consists of failing to reject the null hypothesis, and the Type II error rate (β) is the proportion of tests, when sampling many times from the same population, that fail to reject the null hypothesis when there is a true effect. Finally, the statistical power ($1 - \beta$) is the proportion of tests, when sampling many times from the same population, that correctly reject the null hypothesis when there is a true effect in the population.

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 –0.4 and 0.4), 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 impact 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 et al., 1992). More specifically, the more unequal the *SD* of the population's samples are extracted from, the higher the impact. When there are only two groups, the impact is smaller than when there are more than two groups (Harwell et al., 1992). When there are more than two groups, the *F*-test becomes more liberal, meaning that the Type I error rate is larger than the nominal alpha level, 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 a positive pairing (i.e. the group with the larger sample size also has the larger variance), the test is too conservative, meaning that the Type I error rate of the test is lower than the nominal alpha level, whereas in case of a 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 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).

Cumulative 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 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).