We will then review differences between Student's *t*-test, Welch's *t*-test, and Yuen's *t*-test and show through simulations that the bias in Type 1 error rates when Yuen's *t*-test is used is often severely inflated (above 0.075, which is "critical inflation", following Bradley, 1978) and that the bias in Type 1 error rates when Student's *t*-test is used has a larger impact on statistical inferences than the rather modest impact on the Type 2 error rate of always using Welch's *t*-test by default. Given our analysis and the availability of Welch's *t*-test in all statistical software, we recommend a procedure where Welch's *t*-test is used by default when sample sizes are unequal.

Limitations of Two-Step Procedures

Readers may have learned that the assumptions of normality and of equal variances (or the homoscedasticity assumption) must be examined using assumption checks prior to performing any *t*-test. When data are not normally distributed, with small sample sizes, alternatives should be used. Classic nonparametric statistics are well-known, such as the Mann-Whitney U-test and Kruskal-Wallis. However, unlike a *t*-test, tests based on rank assume that the distributions are the same between groups. Any departure to this assumption, such as unequal variances, will therefore lead to the rejection of the assumption of equal distributions (Zimmerman, 2000). Alternatives exist, known as the "modern robust statistics" (Wilcox, Granger, & Clark, 2013). For example, data sets with low kurtosis (i.e., a distribution flatter than the normal distribution) should be analyzed with the two-sample trimmed *t*-test for unequal population variances, also called Yuen's *t*-test (Luh & Guo, 2007; Yuen, 1974). However, analyses in a later section will show that the normality assumption is not very important for Welch's *t*-test and that there are good reasons to, in general, prefer Welch's *t*-test over Yuen's *t*-test.

With respect to the assumption of homogeneity of variance, if the test of the equality of variance is non-significant and the assumption of equal variances cannot be rejected, homoscedastic methods such as the Student's *t*-test should be used (Wilcox et al., 2013). If the test of the equality of variances is significant, Welch's *t*-test should be used instead of Student's *t*-test because the assumption of equal variances is violated. However, testing the equality of variances before deciding which *t*-test is performed is problematic for several reasons, which will be explained after having described some of the most widely used tests of equality of variances.

Different Ways to Test for Equal Variances

Researchers have proposed several tests for the assumption of equal variances. Levene's test and the F-ratio test are the most likely to be used by researchers because they are available in popular statistical software (Hayes & Cai, 2007). Levene's test is the default option in SPSS. Levene's test is the One-Way ANOVA computed on the terms $|x_{ij}-\hat{\theta}_j|$, where x_{ij} is the ith observation in the jth group, and $\hat{\theta}_j$ is the "center" of the distribution for the jth group (Carroll & Schneider, 1985). In R, the "center" is by default the median, which is also called "Brown Forsythe test for equal variances". In SPSS, the "center" is by default the mean (which is the most powerful choice when the underlying data are symmetrical).[4] The F-ratio statistic is obtained by computing SD2/SD1 (standard deviation ratio, SDR). A generalization of the F-ratio test, to be used when there are more than two groups to compare, is known as the Bartlett's test.

The F-ratio test and the Bartlett test are powerful, but they are only valid under the assumption of normality and collapse as soon as one deviates even slightly from the normal distribution. They are therefore not recommended (Rakotomalala, 2008).

Levene's test is more robust than Bartlett's test and the F-ratio test, but there are three arguments against the use of Levene's test. First, there are several ways to compute Levene's test (i.e., using the median or mean as center), and the best version of the test for equal variances depends on how symmetrically the data is distributed, which is itself difficult to statistically quantify.

Second, performing two tests (Levene's test followed by a *t*-test) on the same data makes the alpha level and power of the *t*-test dependent upon the outcome of Levene's test. When we perform Student's or Welch's *t*-test conditionally on a significant Levene's test, the long-run Type 1 and Type 2 error rates will depend on the power of Levene's test. When the power of Levene's test is low, the error rates of the conditional choice will be very close to Student's error rates (because the probability of choosing Student's *t*-test is very high). On the other hand, when the power of Levene's test is very high, the error rates of the conditional choice will be very close to Welch's error rate (because the probability of choosing Welch's *t*-test is very high; see Rasch, Kubinger, & Moder, 2011). When the power of Levene's test is medium, the error rates of the conditional choice will be somewhere between Student's and Welch's error rates (see, e.g., Zimmerman, 2004). This is problematic when the test most often performed actually has incorrect error rates.

Third, and relatedly, Levene's test can have very low power, which leads to Type 2 errors when sample sizes are small and unequal (Nordstokke & Zumbo, 2007). As an illustration, to estimate the power of Levene's test, we simulated 1,000,000 simulations with balanced designs of different sample sizes (ranging from 10 to 80 in each condition, with a step of 5) under three SDR where the true variances are unequal, respectively, 1.1, 1.5, and 2, yielding 45,000,000 simulations in total. When SDR = 1, the equal variances assumption is true when SDR > 1 the standard deviation of the second sample is bigger than the standard deviation of the first sample and when SDR < 1 the standard deviation of the second sample is smaller than the standard deviation of the first sample. We ran Levene's test centered around the mean and Levene's test centered around the median and estimated the power (in %) to detect unequal variances with equal sample sizes (giving the best achievable power for a given total N; see **Figure 1**).[5]

As we can see in the graph, the further SDR is from 1, the smaller the sample size needed to detect a statistically significant difference in the SDR. Furthermore, for each SDR, power curves of the Levene's test based on the mean