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

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

Additional File
The additional file for this article can be found as follows:

· **Supplemental Materials.** A numerical example of the mathematical development of the F-test, W-test, and F*-test (Appendix 1) and justification for the choice of distributions in simulation (Appendix 2). DOI: https://doi.org/10.5334/irsp.198.s1

Competing Interests
The authors have no competing interests to declare.

Author Contribution
The first author performed simulations. The 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

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: https://doi.org/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: https://doi.org/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. DOI: https://doi.org/10.1037/h0041412

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. DOI: https://doi.org/10.1214/aoms/1177728786

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144–152. DOI: https://doi.org/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: https://doi.org/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: https://doi.org/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: https://doi.org/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., et al. (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: https://doi.org/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: https://doi.org/10.1111/j.1467-6494.2007.00454.x