Permutation tests of scale using deviances

By: Scott J. Richter & Melinda H. McCann

Richter, S. J. and McCann M. H. (2017). Permutation tests of scale using deviances. *Communications in Statistics: Simulation and Computation, 46*(7), 5553-5565. DOI: 10.1080/03610918.2016.1165844

Abstract:

Robust tests for comparing scale parameters, based on deviances—absolute deviations from the median—are examined. Higgins (2004) proposed a permutation test for comparing two treatments based on the ratio of deviances, but the performance of this procedure has not been investigated. A simulation study examines the performance of Higgins' test relative to other tests of scale utilizing deviances that have been shown in the literature to have good properties. An extension of Higgins' procedure to three or more treatments is proposed, and a second simulation study compares its performance to other omnibus tests for comparing scale. While no procedure emerged as a preferred choice in every scenario, Higgins' tests are found to perform well overall with respect to Type I error rate and power.

Keyword: Deviance | Medians | Scale | Permutation test | Robust test

Article:

1. Introduction

Comparison of scale is of interest in many areas, including industrial quality control, agricultural production, and experimental education (Marozzi, 2011). However, it is well known that the parametric F-test for comparing two treatments, as well as parametric tests for more than two treatments (e.g., tests due to Bartlett (1937), Cochran (1937), Hartley (1950)) are generally not robust to nonnormality (see Sharma and Kibria (2013)). Consequently, more robust alternatives are of interest.

Levene (1960) proposed an approximate test using the ANOVA F-test on the absolute deviations from the mean. Brown and Forsythe (1974) suggested instead using absolute deviations from the median (referred to as *deviances* in the remainder of this paper), which they referred to as the "*W50*" test. Note that no uniformly best test for scale has been demonstrated in the literature. In fact, without more stringent distributional assumptions, the minimal sufficient statistic would generally be the n-dimensional vector of order statistics. Thus, no single statistic exists that summarizes the information contained in ***X***, and a uniformly best test statistic does not generally exist. In spite of this, the *W50* test has been recommended as a computationally simple test showing good overall performance with respect to power and robustness to nonnormality in several comparative studies: Keselman et al. (1979), Conover et al. (1981) and Balakrishnan and Ma (1990). Most recently, a study by Sharma and Kibria (2013) comparing 25 omnibus tests for homogeneity of variance recommended the *W50*test as "superior." O'Brien (1979) proposed a modification of Levene's (1960) test which was recommended over the *W50* test for light-tailed distributions (Olejnik and Algina, 1988). Marozzi (2011) considered the *W50* and O'Brien tests, as well as permutation versions of these tests, and found that the permutation versions of these tests tended to be more robust and to have higher power. He recommended the *W50* test as a computationally simple robust test, but also the permutation version of O'Brien's test, which had higher power for symmetric and light-tailed skewed distributions. Higgins (2004) also proposed a test for scale utilizing deviances, based on the ratio of the mean deviances. This test will be referred to as the *RMD*test. This test has intuitive appeal, as it can be viewed as a robust analog to the parametric variance ratio test. To our knowledge, the performance of the *RMD* test has not been studied in the literature. Thus, the robustness and power of the *RMD* test will be examined in this study. In addition, an extension of the *RMD* test to the case of three or more treatments is proposed. Finally, since O'Brien's (1979) test has been shown to be more powerful than the *W50* test for lighter-tailed distributions, but less powerful for heavier-tailed distributions, replacing the mean by the median in O'Brien's statistic may improve its performance. Thus, a modification to O'Brien's test, using deviances instead of deviations from the mean will also be considered in this study. This test will be referred to as the *OB50* test. A simulation study will compare the performance of five tests: (1)Levene's test based on deviations from the mean, (2) the Brown and Forsythe *W50* test, (3) the O'Brien *OB* test, (4) the modification of Obrien's test using deviances (*OB50*), (5a) Higgins' *RMD* test for two treatments, and (5b) the extension of Higgins' test for three or more treatments. Permutation distributions will be used to compute *p*-values for all tests.

2. Methods

Consider a one-way layout with t treatments and n_i observations per treatment. We assume a location-scale model, $y_{ij} = \mu_i + \sigma_i \epsilon_{ij}$, $i = 1, \ldots, t$, $j = 1, \ldots, n_i$, where μ_i and σ_i are the location and scale parameters, respectively, of the treatment i, and ε_{ij} are independent and identically distributed with median 0. It is desired to test H_0: $\sigma_1 = \sigma_2 = \cdots = \sigma_t$ versus H_a: $\sigma_i \neq \sigma_j$ for some i and j.

2.1. Levene's (LEV) test

Levene (1960) proposed an approximate test for scale differences using the ANOVA *F*-test on absolute deviations from the mean, utilizing the scores $\bar{z}_{ij} = |y_{ij} - \bar{y}_i|$. Levene suggested that the *p*-value be based on the *F* distribution with $t - 1$and $n - t$ degrees of freedom.

2.2. Brown–Forsythe (W50) test

Brown and Forsythe (1974) suggested instead using deviances in Levene's test, utilizing the scores $\tilde{z}_{ij} = |y_{ij} - \tilde{y}_i|$, where $\tilde{y}$ is the sample median. The ANOVA *F* test is performed on these scores, and as with Levene's test, the *p*-value is based on the *F* distribution with $t - 1$ and $n - t$ degrees of freedom.

2.3. O'Brien's (OB) test

O'Brien (1979) proposed another modification of Levene's test, using the scores $r_{ij(w)} = \left[(w + n_i - 2)n_{ij}(y_{ij} - \bar{y}_i)^2 - ws_i^2(n_i - 1)\right] / [(n_i - 1)(n_i - 2)]$, where $0 \leqslant w \leqslant 1$. At one extreme, when $w = 0$, the statistic reduces to $r_{ij}(0) = n_i(y_{ij} - \bar{y}_i)^2 / (n_i - 1)$, which is a slight modification of Levene's test based on $\bar{z}_{ij}^2 = (y_i; -\bar{y}_i)^2$. At the other extreme, when $w = 1$, $r_{ij}(1) = q_{ij} = \left[n_i(y_{ij} - \bar{y}_i)^2 - s_i^2\right] / (n_i - 2) = n_i s_i^2 - (n_i - 1)s_{i-1}^2$ which O'Brien (1979) referred to as a “jackknife pseudovalue of s^2_i.” The ANOVA *F* test is performed on these scores and, the *p*-value is based on the *F* distribution with $t - 1$ and $n - t$ degrees of freedom. Tests based on $\bar{z}_2$ have been shown to have inflated Type I error rates, while those based on *q* tend to have low power. Since, $r(w)$ is a weighted average of the two tests, it provides a way to balance the drawbacks of the two tests. O'Brien (1979) suggested that a “utility” value of $w = 0.5$ would work satisfactorily for a majority of situations, and this is the value employed in this study.

2.4. O'Brien's test based on medians (OB50)

Marozzi (2011) found that O'Brien's test was more powerful than the Brown–Forsythe test for symmetric and lighter tailed distributions. To try to improve the power for skewed and heavy-tailed distributions, we propose using deviances rather than deviations from the mean in O'Brien's statistic, utilizing the scores $\tilde{r}_{ij}(w) = \left[(w + n_i - 2)n_i(y_{ij} - \tilde{y}_i)^2 - ws_i^2(n_i - 1)\right] / [(n_i - 1)(n_i - 2)]$, where $\tilde{y}_i$ is the sample median of the i^{th} treatment.

2.5. Higgins’ (RMD) test

Higgins (2004) suggested the following statistic for comparing two scale parameters: RDM $= \frac{\max(\bar{\tilde{z}}_1, \bar{\tilde{z}}_2)}{\min(\bar{\tilde{z}}_1, \bar{\tilde{z}}_2)}$, where $\bar{\tilde{z}}_i$ is the mean of the scores $\tilde{z}_{ij}$ for treatment *i*. The deviances $\tilde{z}_{ij} = |y_{ij} - \tilde{y}_i|$ are the same as those used by Brown and Forsythe (1974). Higgins (2004) suggested using the permutation distribution of the *RMD* statistic to calculate a *p*-value.

2.6. Extension of RMD to more than two treatments ($\mathrm{RMD}_{\max}$)

A simple extension to more than two treatments is to compute the maximum and minimum of mean deviances over all treatments, and then form the ratio, that is, $\mathrm{RDM}_{\max} = \frac{\max(\bar{\tilde{z}}_1, \bar{\tilde{z}}_2, \ldots, \bar{\tilde{z}}t)}{\min(\bar{\tilde{z}}_1, \bar{\tilde{z}}_2, \ldots, \bar{\tilde{z}}_t)}$ The permutation distribution of RMD_{max} will be used to compute the *p*-value.

2.7. Permutation tests

All of the tests described in Sections 2.1–2.4 were proposed as approximate tests based on the *F* distribution. However, exact *p*-values can be calculated using permutation distributions.

Marozzi (2011) found for the two-treatment case that the permutation versions tended to be more robust and have greater power than the approximate tests. Thus, we will consider only the permutation versions of these tests. Test statistics will be computed for a large number of random reassignments of observations to treatments, and the *p*-value will be calculated as the proportion of values of the permutation distribution that are at least as extreme as the observed test statistic value.

3. Simulation study

3.1. Procedures studied

A simulation study compared the Type I error rate and power of the methods described in Section 2:

1. Levene's test using absolute deviations from the mean (*LEV*).

2. Brown and Forsyth's *W50* test (*W50*);

3. O'Brien's method using means (*OB*);

4. O'Brien's method using medians (*OB50*);

5. a. Higgins *RMD* procedure (*RMD*);b. Extension of the *RMD* procedure for more than two treatments (RMD_{max}).

3.2. Sample sizes and differences of scales parameters

Both equal and unequal sample size settings were examined. For the two-treatment setting, equal sample sizes of $n_i = 10$ and $n_i = 30$ were examined, as well as unequal sample size settings of $n_1 = 10$, $n_2 = 30$. Scale ratios of σ_1/σ_2 from 1 to 5 were considered, as were the reciprocals of these, in order to examine the effect of whether the larger or smaller scale parameter was associated with the larger sample size.

Settings at both three and five treatments were also examined. For each, equal sample sizes of $n_i = 10$ and $n_i = 20$ were used. For three treatments unequal sample size settings of $n_1 = 5$, $n_2 = 10$, $n_3 = 15$ and $n_1 = 20$, $n_2 = 40$, $n_3 = 60$ were considered, and for five treatments $n_1 = 5$, $n_2 = 5$, $n_3 = 10$, $n_4 = 15$, $n_5 = 15$, $n_1 = 10$, $n_2 = 10$, $n_3 = 15$, $n_4 = 20$, $n_5 = 20$ and $n_1 = 10$, $n_2 = 10$, $n_3 = 20$, $n_4 = 30$, $n_5 = 30$ were utilized. Maximum scale parameter ratios $\sigma_{max}/\sigma_{min}$ from 1 to 5 were examined, with different patterns of smaller ratios present. For three treatments the patterns (σ, 1, 1) and (σ, (σ + 1)/2, 1) were used. The first setting we refer to as the "single extreme scale parameter" setting, while the second setting has an intermediate scale value midway between the minimum and maximum. For five treatments, settings of (σ, 1, 1, 1, 1) and (σ, (σ + 1)/2, 1, 1, 1) were used, as well as an additional setting where each pair of adjacent treatments had the same scale parameter difference, e.g., (5, 4, 3, 2, 1).

3.3. Distributions

Several different *g* and *h* distributions (Hoaglin, 1985) were used to simulate data from distributions with different characteristics. *g* and *h* distributions are monotonic functions of normal distributions, and allow investigation of nonnormal distributions with specific characteristics. The g-and-h random variable is defined as $Y_{g,h}(Z) = \frac{\exp(gZ)-1}{g}\exp(hZ^2 / 2)$ where $Z \sim N(0, 1)$. When $g = h = 0$, $Y_{g,h}(Z) \sim N(0, 1)$. Nonzero values of g increase the skewness and positive values of h increase the elongation (tail heaviness) of the distribution. Changing the values of g and h do not affect the location of the distribution. The following cases were considered:

1. $g = 0$, $h = 0$—Normally distributed (symmetric, light tails);
2. $g = 0$, $h = 0.4$—Symmetric, moderately heavy tails;
3. $g = 0$, $h = 0.8$—Symmetric, very heavy tails;
4. $g = 0.8$, $h = 0$—Skewed, light tails;
5. $g = 0.8$, $h = 0.4$—Skewed, moderately heavy tails.

Type I error rate and power were estimated based on 1,000 randomly selected data sets from each distribution, for each setting of sample sizes and scale parameter patterns. Marozzi (2016) suggested that only 253 random permutations are necessary with 1,000 random data sets if the goal of the simulation is to estimate the power of a test and only a "rough" estimate of the permutation *p*-value is required, while Keller-McNulty and Higgins (1987) recommended a random sample of at least 1,600 permutations to estimate the exact *p*-value for a permutation test. Since precise estimation of the permutation test *p*-values was considered important, a conservative 2,000 random permutations was utilized.

4. Results

In this section, representative simulation results are presented. Additional results are available from the corresponding author upon request.

4.1. Two treatments

4.1.1. Type I error

All tests were robust in the sense that estimated rates of Type I error were close to the nominal level of 0.05, with none exceeding 0.075, and thus there were no serious problems with inflated error rates (Note that in the tables the first row of each scenario represents the equal scale case, and thus the value given is the estimated Type I error rate).

4.1.2. Power

For all of the equal sample size scenarios, *RMD* always was at least as powerful as the other methods, and tended to be substantially more powerful than *OB* for heavier-tailed distributions (See Table 1). The power advantage over *W50* was more modest, and was even less pronounced as sample sizes increased. As expected, *OB* had higher power than *W50* for the two lighter-tailed distributions, while the reverse was true for the heavier-tailed distributions. *OB50* was slightly more powerful than *OB* for n_i = 30, but slightly less powerful when n_i = 10. Even in cases where *OB50* was more powerful than *OB*, it was not competitive with either *RMD* or *W50*.

Table 1. Proportion of rejections at α = 0.05, two treatments, equal sample sizes.

	Method					
Distribution	**σ_1/σ_2**	***RMD***	***W50***	***OB***	***OB50***	***LEV***
n_1 = 10, n_2= 10						
g = 0, h = 0	1	0.044	0.04	0.041	0.048	0.043
	2	0.367	0.278	0.346	0.318	0.271
	4	0.883	0.659	0.742	0.677	0.616
g = 0, h = 0.4	1	0.049	0.045	0.054	0.048	0.045
	2	0.137	0.114	0.098	0.094	0.064
	4	0.406	0.341	0.227	0.238	0.174
g = 0, h = 0.8	1	0.044	0.048	0.059	0.051	0.055
	2	0.088	0.079	0.070	0.066	0.065
	4	0.212	0.187	0.098	0.123	0.097
g = 0.8, h = 0	1	0.050	0.041	0.050	0.047	0.044
	2	0.206	0.187	0.243	0.168	0.154
	4	0.680	0.539	0.565	0.437	0.398
g = 0.8, h = 0.4	1	0.049	0.045	0.054	0.050	0.059
	2	0.125	0.110	0.090	0.088	0.071
	4	0.330	0.284	0.203	0.202	0.139
n_1 = 30, n_2= 30						
g = 0, h = 0	1	0.046	0.052	0.045	0.044	0.043
	2	0.907	0.795	0.883	0.859	0.789
	4	1.000	0.988	1.000	1.000	0.981
g = 0, h = 0.4	1	0.041	0.039	0.046	0.048	0.045
	2	0.257	0.246	0.166	0.169	0.125
	4	0.662	0.657	0.457	0.461	0.324
g = 0, h = 0.8	1	0.042	0.042	0.044	0.047	0.043
	2	0.107	0.102	0.073	0.087	0.058
	4	0.280	0.274	0.156	0.178	0.078
g = 0.8, h = 0	1	0.057	0.058	0.044	0.047	0.047
	2	0.518	0.510	0.516	0.359	0.358
	4	0.977	0.952	0.935	0.805	0.846
g = 0.8, h = 0.4	1	0.048	0.045	0.038	0.030	0.045
	2	0.199	0.192	0.130	0.135	0.090
	4	0.515	0.501	0.316	0.331	0.218

When sample sizes were unequal, relative power tended to depend on which sample was associated with the larger scale parameter. When the larger scale parameter was associated with the larger sample, *RMD* was substantially more powerful for all scenarios than the other methods (Table 2). *W50* tended to be more powerful than *OB*, even for the lighter-tailed distributions, while again *OB50* was generally no more powerful than *OB*. Alternatively, when the larger scale parameter was associated with the smaller sample, *W50* was most powerful for heavier-tailed distributions and *OB* was most powerful for lighter-tailed distributions. For the heavy-tailed distributions, the power of *OB* could drop below the nominal 0.05 level initially as the variance for the smaller sample increased. This is likely due to the fact that the test statistic for *OB* is a function of the sample variance. As the variance for the smaller sample increased further (not shown in the tables) the power did eventually rise above 0.05, but the relative power remained very low. For all scenarios, however, *RMD* maintained respectable power relative to the other methods. Thus, if specific information regarding the expected scale difference is unknown, *RMD* can be recommended as a general test for comparing scale parameters for two treatments.

Table 2. Proportion of rejections at $\alpha = 0.05$, two treatments, unequal sample sizes.

	Method					
Distribution	σ_1/σ_2	***RMD***	***W50***	***OB***	***OB50***	***LEV***
$n_1 = 30$, $n_2 = 10$						
$g = 0$, $h = 0$	1	0.048	0.058	0.043	0.045	0.053
	2	0.672	0.499	0.058	0.091	0.410
	4	0.998	0.892	0.232	0.385	0.847
$g = 0$, $h = 0.4$	1	0.044	0.052	0.047	0.041	0.047
	2	0.246	0.038	0.011	0.011	0.022
	4	0.609	0.135	0.002	0.002	0.009
$g = 0$, $h = 0.8$	1	0.045	0.044	0.048	0.047	0.047
	2	0.174	0.021	0.022	0.022	0.026
	4	0.364	0.012	0.011	0.009	0.017
$g = 0.8$, $h = 0$	1	0.038	0.057	0.047	0.046	0.048
	2	0.392	0.198	0.025	0.025	0.124
	4	0.863	0.544	0.092	0.072	0.294
$g = 0.8$, $h = 0.4$	1	0.041	0.052	0.050	0.046	0.046
	2	0.223	0.029	0.012	0.019	0.023
	4	0.505	0.084	0.002	0.050	0.032
$n_1 = 10$, $n_2 = 30$						
$g = 0$, $h = 0$	1	0.042	0.049	0.058	0.051	0.048
	2	0.447	0.513	0.680	0.651	0.523
	4	0.967	0.898	0.947	0.927	0.867

g = 0, h = 0.4	1	0.047	0.049	0.062	0.055	0.054
	2	0.114	0.222	0.181	0.166	0.150
	4	0.345	0.489	0.370	0.358	0.331
g = 0, h = 0.8	1	0.047	0.054	0.060	0.055	0.054
	2	0.044	0.122	0.101	0.106	0.108
	4	0.117	0.258	0.187	0.175	0.158
g = 0.8, h = 0	1	0.047	0.044	0.038	0.041	0.051
	2	0.196	0.351	0.413	0.293	0.273
	4	0.714	0.767	0.807	0.686	0.671
g = 0.8, h = 0.4	1	0.044	0.051	0.051	0.051	0.048
	2	0.092	0.180	0.155	0.147	0.128
	4	0.256	0.395	0.301	0.273	0.250

4.2. Omnibus test, three or more treatments

Tables 3 through 12 show simulation results for three and five treatment scenarios.

Table 3. Proportion of rejections at α = 0.05, three treatments, n_1 = 10, n_2= 10,n_3 = 10.

	Method					
Distribution	**$\sigma_1, \sigma_2, \sigma_3$**	***RMD_{max}***	***W50***	***OB***	***OB50***	***LEV***
g = 0, h = 0	1,1,1	0.043	0.046	0.038	0.038	0.051
	3,1,1	0.317	0.631	0.709	0.652	0.565
	3,2,1	0.632	0.404	0.455	0.399	0.378
	5,1,1	0.950	0.863	0.892	0.832	0.796
	5,3,1	0.944	0.631	0.625	0.595	0.578
g = 0, h = 0.4	1,1,1	0.036	0.046	0.052	0.053	0.065
	3,1,1	0.230	0.246	0.184	0.174	0.146
	3,2,1	0.230	0.141	0.110	0.101	0.104
	5,1,1	0.453	0.457	0.292	0.285	0.233
	5,3,1	0.432	0.231	0.151	0.154	0.131
g = 0, h = 0.8	1,1,1	0.040	0.046	0.051	0.057	0.057
	3,1,1	0.124	0.114	0.082	0.079	0.069
	3,2,1	0.109	0.084	0.061	0.068	0.062
	5,1,1	0.213	0.214	0.123	0.129	0.094
	5,3,1	0.236	0.113	0.075	0.082	0.071
g = 0.8, h = 0	1,1,1	0.052	0.048	0.050	0.043	0.051
	3,1,1	0.449	0.461	0.500	0.352	0.324
	3,2,1	0.402	0.266	0.279	0.188	0.178
	5,1,1	0.744	0.721	0.728	0.561	0.544
	5,3,1	0.708	0.430	0.391	0.288	0.282

g = 0.8, *h* = 0.4	1,1,1	0.039	0.050	0.051	0.050	0.060
	3,1,1	0.193	0.193	0.150	0.156	0.108
	3,2,1	0.193	0.125	0.100	0.097	0.083
	5,1,1	0.378	0.359	0.252	0.242	0.179
	5,3,1	0.379	0.180	0.124	0.133	0.101

Table 4. Proportion of rejections at α = 0.05, n_1 = 15, n_2= 10,n_3 = 5.

	Method					
Distribution	**$\sigma_1, \sigma_2, \sigma_3$**	***RMD_{max}***	***W50***	***OB***	***OB50***	***LEV***
g = 0, *h* = 0	1,1,1	0.045	0.049	0.052	0.059	0.053
	3,1,1	0.543	0.691	0.297	0.187	0.573
	3,2,1	0.578	0.345	0.058	0.044	0.266
	5,1,1	0.858	0.878	0.456	0.334	0.756
	5,3,1	0.900	0.519	0.121	0.073	0.404
g = 0, *h* = 0.4	1,1,1	0.051	0.054	0.050	0.055	0.051
	3,1,1	0.223	0.103	0.011	0.016	0.055
	3,2,1	0.261	0.038	0.011	0.015	0.036
	5,1,1	0.412	0.210	0.010	0.009	0.071
	5,3,1	0.454	0.064	0.009	0.010	0.037
g = 0, *h* = 0.8	1,1,1	0.057	0.053	0.046	0.055	0.057
	3,1,1	0.144	0.037	0.019	0.033	0.030
	3,2,1	0.183	0.023	0.018	0.030	0.030
	5,1,1	0.356	0.126	0.012	0.009	0.058
	5,3,1	0.291	0.020	0.015	0.018	0.024
g = 0.8, *h* = 0	1,1,1	0.050	0.052	0.050	0.054	0.050
	3,1,1	0.374	0.384	0.069	0.054	0.219
	3,2,1	0.395	0.152	0.013	0.015	0.093
	5,1,1	0.683	0.640	0.143	0.101	0.356
	5,3,1	0.683	0.238	0.022	0.027	0.137
g = 0.8, *h* = 0.4	1,1,1	0.049	0.047	0.044	0.051	0.047
	3,1,1	0.197	0.062	0.016	0.019	0.040
	3,2,1	0.238	0.022	0.014	0.015	0.025
	5,1,1	0.356	0.038	0.014	0.026	0.025
	5,3,1	0.410	0.037	0.009	0.001	0.028

Table 5. Proportion of rejections at α = 0.05, n_1 = 5, n_2= 10,n_3 = 15.

	Method					
Distribution	**$\sigma_1, \sigma_2, \sigma_3$**	***RMD_{max}***	***W50***	***OB***	***OB50***	***LEV***
g = 0, *h* = 0	1,1,1	0.047	0.047	0.049	0.045	0.054
	3,1,1	0.243	0.450	0.643	0.618	0.468
	3,2,1	0.199	0.376	0.551	0.522	0.400
	5,1,1	0.554	0.700	0.783	0.737	0.644
	5,3,1	0.548	0.611	0.691	0.652	.0599
g = 0, *h* = 0.4	1,1,1	0.059	0.051	0.052	0.054	0.055
	3,1,1	0.081	0.222	0.210	0.211	0.161
	3,2,1	0.058	0.183	0.157	0.156	0.142

	5,1,1	0.160	0.395	0.339	0.311	0.264
	5,3,1	0.142	0.304	0.222	0.217	0.199
$g = 0, h = 0.8$	1,1,1	0.055	0.056	0.044	0.052	0.051
	3,1,1	0.038	0.138	0.117	0.119	0.101
	3,2,1	0.038	0.112	0.090	0.096	0.095
	5,1,1	0.058	0.215	0.172	0.168	0.150
	5,3,1	0.059	0.159	0.121	0.120	0.117
$g = 0.8, h = 0$	1,1,1	0.046	0.037	0.048	0.055	0.042
	3,1,1	0.109	0.340	0.466	0.375	0.296
	3,2,1	0.086	0.288	0.352	0.275	0.246
	5,1,1	0.286	0.580	0.654	0.559	0.484
	5,3,1	0.311	0.481	0.468	0.378	0.311
$g = 0.8, h = 0.4$	1,1,1	0.051	0.046	0.045	0.048	0.050
	3,1,1	0.068	0.182	0.196	0.178	0.145
	3,2,1	0.049	0.150	0.142	0.135	0.119
	5,1,1	0.121	0.315	0.276	0.255	0.225
	5,3,1	0.101	0.240	0.191	0.174	0.179

4.2.1. Equal sample sizes

For the “single extreme scale” case, *OB* had highest power for the two lighter-tailed distributions when scale differences were small (scale patterns [3,1,1] and [3,1,1,1,1]), while RMD_{max} and *W50* had higher power than *OB* for heavier-tailed distributions, with little difference in power between the two methods. With larger scale differences (scale patterns [5,1,1] and [5,1,1,1,1]), however, there was little difference in power for the three methods for the lighter-tailed distributions, while RMD_{max} and *W50* continued to have higher power for the heavier-tailed distributions. *OB50* sometimes showed higher power for heavier-tailed distributions, but the difference was never substantial.

When there were intermediate scale magnitudes present, RMD_{max} tended to have substantially higher power than *W50* and *OB*. For example, for three treatments (Table 3) all estimated power values for *W50* and *OB* were lower when the scale pattern was [5,3,1] than for [5,1,1], and for five treatments (Table 8[Table 9, Table 10, Table 11]), power for *W50* and *OB* were highest for the scale pattern [5,1,1,1,1] and lowest for the scale pattern [5,4,3,2,1]. In contrast, the power of RMD_{max} appeared relatively unaffected by the pattern of scale differences. With larger equal sample sizes (e.g., $n_i \geqslant 20$), RMD_{max} was most powerful for nonnormal distributions except for the single extreme scale setting, where its power was still very close to that of the most powerful procedure. Thus, for equal size samples, RMD_{max} is recommended as an overall omnibus test for scale difference.

4.2.2. Unequal sample sizes

When sample sizes were unequal (See Tables 4–7 and 10–12) the power of the tests depended not only on the distribution and magnitudes of the scale differences, but also on the pattern of scale magnitudes relative to the sample sizes. When larger sample sizes tended to be associated with the larger scale magnitudes (Tables 4, 7 and 10), the RMD_{max} test always had highest power, often by a substantial amount. When the pattern of scale magnitudes was reversed (Tables 5, 6, 11 and 12), *W50* tended to have the highest power for heavier-tailed distributions, and *OB* for lighter-tailed distributions. *OB50* was never substantially more powerful than *OB*.

The scale pattern again played a role in relative power of the methods, as did the sample sizes and the magnitude of scale differences among treatments. The power advantages of *W50* and *OB* over RMD_{max} decreased when there were intermediate scale differences. Also, as sample sizes and the magnitude of scale differences increased, RMD_{max} became more powerful than *OB* and had essentially the same power as *W50*. (Tables 6[Table 7], 12). Thus, the choice of best omnibus test for unequal sample sizes was less clear than for the equal sample size case. However, if little is known about anticipated scale differences between treatments, RMD_{max} is recommended, since it tended to be more competitive when it was not the most powerful procedure, and substantially more powerful when it was.

Table 6. Proportion of rejections at $\alpha = 0.05$, $n_1 = 10$, $n_2 = 20$, $n_3 = 30$.

	Method					
Distribution	**$\sigma_1, \sigma_2, \sigma_3$**	***RMD_{max}***	***W50***	***OB***	***OB50***	***LEV***
$g = 0$, $h = 0$	1,1,1	0.058	0.051	0.044	0.044	0.043
	3,1,1	0.862	0.778	0.899	0.890	0.782
	3,2,1	0.810	0.730	0.828	0.831	0.745
	5,1,1	0.989	0.938	0.966	0.946	0.902
	5,3,1	0.993	0.910	0.923	0.921	0.897
$g = 0$, $h = 0.4$	1,1,1	0.044	0.046	0.040	0.048	0.048
	3,1,1	0.208	0.321	0.233	0.241	0.189
	3,2,1	0.146	0.257	0.166	0.173	0.164
	5,1,1	0.418	0.560	0.418	0.420	0.383
	5,3,1	0.355	0.455	0.254	0.252	0.271
$g = 0$, $h = 0.8$	1,1,1	0.042	0.046	0.042	0.048	0.043
	3,1,1	0.059	0.141	0.101	0.108	0.059
	3,2,1	0.042	0.122	0.086	0.092	0.089
	5,1,1	0.107	0.234	0.155	0.170	0.141
	5,3,1	0.083	0.170	0.111	0.117	0.117
$g = 0.8$, $h = 0$	1,1,1	0.034	0.047	0.045	0.044	0.048
	3,1,1	0.515	0.616	0.681	0.577	0.504
	3,2,1	0.416	0.184	0.124	0.139	0.124
	5,1,1	0.831	0.872	0.872	0.829	0.802
	5,3,1	0.820	0.817	0.699	0.593	0.656
$g = 0.8$, $h = 0.4$	1,1,1	0.034	0.045	0.045	0.049	0.045
	3,1,1	0.143	0.246	0.175	0.182	0.146
	3,2,1	0.097	0.184	0.124	0.139	0.097
	5,1,1	0.293	0.432	0.310	0.317	0.245
	5,3,1	0.246	0.325	0.190	0.206	0.170

Table 7. Proportion of rejections at $\alpha = 0.05$, $n_1 = 30$, $n_2 = 20$, $n_3 = 10$.

	Method					
Distribution	**$\sigma_1, \sigma_2, \sigma_3$**	***RMD_{max}***	***W50***	***OB***	***OB50***	***LEV***

g = 0, *h* = 0	1,1,1	0.043	0.054	0.059	0.054	0.043
	3,1,1	0.966	0.948	0.938	0.890	0.926
	3,2,1	0.937	0.691	0.477	0.383	0.618
	5,1,1	0.998	0.993	0.982	0.960	0.983
	5,3,1	0.999	0.881	0.687	0.606	0.839
g = 0, *h* = 0.4	1,1,1	0.054	0.059	0.052	0.056	0.055
	3,1,1	0.351	0.05	0.01	0.011	0.026
	3,2,1	0.338	0.025	0.018	0.022	0.024
	5,1,1	0.656	0.376	0.034	0.025	0.100
	5,3,1	0.592	0.092	0.009	0.006	0.034
g = 0, *h* = 0.8	1,1,1	0.052	0.059	0.045	0.053	0.053
	3,1,1	0.189	0.032	0.028	0.026	0.029
	3,2,1	0.196	0.025	0.018	0.022	0.024
	5,1,1	0.319	0.045	0.019	0.019	0.024
	5,3,1	0.323	0.022	0.011	0.013	0.018
g = 0.8, *h* = 0	1,1,1	0.054	0.055	0.052	0.047	0.047
	3,1,1	0.706	0.727	0.259	0.159	0.450
	3,2,1	0.610	0.317	0.048	0.033	0.143
	5,1,1	0.965	0.934	0.398	0.288	0.670
	5,3,1	0.909	0.528	0.100	0.072	0.271
g = 0.8, *h* = 0.4	1,1,1	0.050	0.055	0.051	0.054	0.050
	3,1,1	0.270	0.088	0.020	0.015	0.035
	3,2,1	0.295	0.030	0.011	0.010	0.015
	5,1,1	0.545	0.205	0.023	0.013	0.058
	5,3,1	0.488	0.044	0.007	0.006	0.015

Table 8. Proportion of rejections at $\alpha = 0.05$, five treatments, $n_i = 10$.

	Method					
Distribution	**$\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5$**	***RMD_{max}***	***W50***	***OB***	***OB50***	***LEV***
g = 0, *h* = 0	1,1,1,1,1	0.047	0.038	0.035	0.046	0.041
	5,1,1,1,1	0.905	0.880	0.916	0.866	0.812
	5,1,3,1,1	0.936	0.827	0.844	0.808	0.772
	5,4,3,2,1	0.927	0.561	0.574	0.557	0.522
g = 0, *h* = 0.4	1,1,1,1,1	0.051	0.047	0.049	0.052	0.051
	5,1,1,1,1	0.286	0.393	0.262	0.246	0.189
	5,1,3,1,1	0.354	0.308	0.181	0.165	0.149
	5,4,3,2,1	0.362	0.153	0.093	0.098	0.081
g = 0, *h* = 0.8	1,1,1,1,1	0.046	0.045	0.056	0.055	0.046
	5,1,1,1,1	0.134	0.139	0.083	0.086	0.074
	5,1,3,1,1	0.180	0.126	0.067	0.068	0.072
	5,4,3,2,1	0.181	0.087	0.068	0.068	0.063
g = 0.8, *h* = 0	1,1,1,1,1	0.038	0.034	0.042	0.041	0.051
	5,1,1,1,1	0.628	0.758	0.766	0.581	0.587
	5,1,3,1,1	0.699	0.671	0.590	0.405	0.444
	5,4,3,2,1	0.649	0.362	0.279	0.196	0.217
g = 0.8, *h* = 0.4	1,1,1,1,1	0.047	0.042	0.057	0.054	0.058

	5,1,1,1,1	0.236	0.298	0.178	0.171	0.133
	5,1,3,1,1	0.309	0.222	0.121	0.121	0.106
	5,4,3,2,1	0.314	0.110	0.085	0.082	0.083

Table 9. Proportion of rejections at $\alpha = 0.05$, five treatments, $n_i = 20$.

	Method					
Distribution	**$\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5$**	***RMD_{max}***	***W50***	***OB***	***OB50***	***LEV***
$g = 0, h = 0$	1,1,1,1,1	0.056	0.047	0.044	0.048	0.041
	5,1,1,1,1	1.000	0.990	0.996	0.994	0.987
	5,1,3,1,1	1.000	0.989	0.995	0.986	0.980
	5,4,3,2,1	1.000	0.925	0.945	0.934	0.924
$g = 0, h = 0.4$	1,1,1,1,1	0.050	0.048	0.046	0.045	0.049
	5,1,1,1,1	0.566	0.607	0.353	0.366	0.312
	5,1,3,1,1	0.652	0.530	0.244	0.251	0.232
	5,4,3,2,1	0.565	0.253	0.117	0.125	0.129
$g = 0, h = 0.8$	1,1,1,1,1	0.052	0.046	0.048	0.047	0.050
	5,1,1,1,1	0.206	0.168	0.084	0.111	0.087
	5,1,3,1,1	0.276	0.146	0.068	0.101	0.085
	5,4,3,2,1	0.243	0.089	0.057	0.065	0.073
$g = 0.8, h = 0$	1,1,1,1,1	0.052	0.046	0.048	0.047	0.044
	5,1,1,1,1	0.974	0.972	0.953	0.904	0.929
	5,1,3,1,1	0.974	0.965	0.868	0.758	0.841
	5,4,3,2,1	0.931	0.756	0.522	0.394	0.510
$g = 0.8, h = 0.4$	1,1,1,1,1	0.056	0.047	0.043	0.038	0.039
	5,1,1,1,1	0.418	0.440	0.246	0.260	0.199
	5,1,3,1,1	0.524	0.357	0.165	0.186	0.162
	5,4,3,2,1	0.445	0.175	0.087	0.107	0.083

Table 10. Proportion of rejections at $\alpha = 0.05$, five treatments, $n_i = 15, 15, 10, 5, 5$.

	Method					
Distribution	**$\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5$**	***RMD_{max}***	***W50***	***OB***	***OB50***	***LEV***
$g = 0, h = 0$	1,1,1,1,1	0.047	0.048	0.044	0.047	0.047
	5,1,1,1,1	0.530	0.935	0.801	0.599	0.845
	5,1,3,1,1	0.666	0.887	0.627	0.445	0.788
	5,4,3,2,1	0.781	0.512	0.099	0.044	0.395
$g = 0, h = 0.4$	1,1,1,1,1	0.055	0.047	0.046	0.049	0.062
	5,1,1,1,1	0.185	0.267	0.051	0.019	0.104
	5,1,3,1,1	0.263	0.167	0.028	0.010	0.072
	5,4,3,2,1	0.324	0.034	0.016	0.013	0.029
$g = 0, h = 0.8$	1,1,1,1,1	0.049	0.041	0.051	0.044	0.054
	5,1,1,1,1	0.128	0.043	0.034	0.027	0.042
	5,1,3,1,1	0.176	0.037	0.123	0.032	0.035
	5,4,3,2,1	0.217	0.018	0.025	0.022	0.031
$g = 0.8, h = 0$	1,1,1,1,1	0.060	0.044	0.056	0.065	0.051

	5,1,1,1,1	0.329	0.758	0.250	0.160	0.460
	5,1,3,1,1	0.466	0.625	0.140	0.093	0.368
	5,4,3,2,1	0.522	0.211	0.016	0.009	0.103
g = 0.8, *h* = 0.4	1,1,1,1,1	0.063	0.045	0.054	0.054	0.057
	5,1,1,1,1	0.164	0.169	0.043	0.024	0.082
	5,1,3,1,1	0.229	0.117	0.032	0.019	0.070
	5,4,3,2,1	0.286	0.025	0.018	0.017	0.033

Table 11. Proportion of rejections at α = 0.05, five treatments, n_i = 5, 5, 10, 15, 15.

	Method					
Distribution	**σ_1, σ_2, σ_3, σ_4, σ_5**	***RMD_{max}***	***W50***	***OB***	***OB50***	***LEV***
g = 0, *h* = 0	1,1,1,1,1	0.065	0.048	0.041	0.043	0.043
	5,1,1,1,1	0.465	0.680	0.768	0.705	0.620
	5,1,3,1,1	0.457	0.722	0.757	0.700	0.667
	5,4,3,2,1	0.382	0.477	0.679	0.661	0.501
g = 0, *h* = 0.4	1,1,1,1,1	0.061	0.047	0.048	0.038	0.038
	5,1,1,1,1	0.123	0.297	0.254	0.210	0.178
	5,1,3,1,1	0.157	0.295	0.213	0.155	0.160
	5,4,3,2,1	0.075	0.207	0.175	0.161	0.154
g = 0, *h* = 0.8	1,1,1,1,1	0.059	0.048	0.048	0.042	0.043
	5,1,1,1,1	0.063	0.140	0.117	0.093	0.093
	5,1,3,1,1	0.090	0.128	0.099	0.076	0.089
	5,4,3,2,1	0.028	0.124	0.105	0.085	0.088
g = 0.8, *h* = 0	1,1,1,1,1	0.055	0.040	0.045	0.043	0.048
	5,1,1,1,1	0.252	0.536	0.603	0.497	0.453
	5,1,3,1,1	0.281	0.563	0.486	0.396	0.430
	5,4,3,2,1	0.167	0.346	0.425	0.320	0.292
g = 0.8, *h* = 0.4	1,1,1,1,1	0.057	0.058	0.048	0.036	0.046
	5,1,1,1,1	0.106	0.237	0.193	0.168	0.169
	5,1,3,1,1	0.137	0.210	0.151	0.123	0.138
	5,4,3,2,1	0.058	0.166	0.142	0.123	0.121

Table 12. Proportion of rejections at α = 0.05, five treatments, n_i = 10, 10, 20, 30, 30.

	Method					
Distribution	**σ_1, σ_2, σ_3, σ_4, σ_5**	***RMD_{max}***	***W50***	***OB***	***OB50***	***LEV***
g = 0, *h* = 0	1,1,1,1,1	0.058	0.057	0.046	0.054	0.055
	5,1,1,1,1	0.983	0.945	0.976	0.962	0.928
	5,1,3,1,1	0.989	0.969	0.980	0.967	0.958
	5,4,3,2,1	0.993	0.890	0.943	0.967	0.958
g = 0, *h* = 0.4	1,1,1,1,1	0.046	0.050	0.047	0.048	0.043
	5,1,1,1,1	0.312	0.443	0.309	0.307	0.292
	5,1,3,1,1	0.344	0.448	0.225	0.231	0.246
	5,4,3,2,1	0.257	0.341	0.212	0.215	0.208
g = 0, *h* = 0.8	1,1,1,1,1	0.048	0.052	0.052	0.049	0.056

	5,1,1,1,1	0.090	0.166	0.107	0.117	0.117
	5,1,3,1,1	0.090	0.166	0.107	0.117	0.117
	5,4,3,2,1	0.053	0.144	0.108	0.112	0.110
g = 0.8, h = 0	1,1,1,1,1	0.051	0.050	0.054	0.059	0.051
	5,1,1,1,1	0.765	0.866	0.853	0.799	0.787
	5,1,3,1,1	0.765	0.866	0.853	0.799	0.787
	5,4,3,2,1	0.732	0.753	0.655	0.553	0.567
g = 0.8, h = 0.4	1,1,1,1,1	0.051	0.048	0.048	0.046	0.051
	5,1,1,1,1	0.218	0.329	0.220	0.243	0.195
	5,1,3,1,1	0.249	0.307	0.170	0.176	0.170
	5,4,3,2,1	0.159	0.254	0.170	0.168	0.165

5. Discussion

The performance of Higgins' *RMD* test for comparing scale parameters of two treatments was investigated using a simulation study, and an extension (RMD_{max}) to more than two treatments was proposed and likewise studied. As with any simulation study, generalization of results requires caution. These results may not extend to situations where the true scales are very different than those studied here, and/or where the data do not come from the distributions studied here. In addition, the conclusions rely on the assumption of a location–scale model, and thus may not be valid if that assumption is not plausible.

For this study, no single procedure emerged as a clear choice for all scenarios. Thus, a future direction may be to investigate combining two or more tests based on theory developed by Pesarin (2001) and Pesarin and Salmaso (2010). Marozzi (2012) proposed a modification of a test of Hall and Padmanabhan (1997), based on the ratio of trimmed sample variances, which could be considered a generalization of the *RMD* test, and found that combined tests can be very effective for detecting scale shifts in different situations.

However, based on the current study, we recommend *RMD* and RMD_{max} as good intuitive robust tests for scale differences. The tests based on *RMD* and RMD_{max} also have intuitive appeal. As the test statistic is a ratio of measures of variation, the *RMD* and RMD_{max} statistics can be viewed as simple, more robust analogs to well-known parametric tests for scale, namely the variance ratio *F*-test for two treatments and Hartley's F_{max} omnibus test for three or more treatments.

While *W50* performed well across a wide range of scenarios, it usually did not have the highest power. *OB* performed well for lighter-tailed distributions, but usually had much lower power than *RMD*/RMD_{max} and *W50* for heavier-tailed distributions. For scenarios when one procedure was clearly more powerful, more often than not that procedure was *RMD*/RMD_{max}. The results of the simulations suggest that when sample sizes are at least moderately large ($n_i \geqslant 20$), *RMD*/RMD_{max} will likely have as much or more power than either *W50* or *OB* over a wide range of situations, and will rarely have much lower relative power. For sample sizes smaller than 20, *W50* may be preferred for heavier-tailed distributions and *OB* preferred for lighter-tailed distributions, but *RMD*/RMD_{max} should still have good relative power.

References

1. Balakrishnan, N., Ma, C. W. (1990). A comparative study of various tests for the equality of two population variances. *Journal of Statistical Computation and Simulation* 35:41–89.
2. Bartlett, M. S. (1937. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London A*: 268–282.
3. Brown, M. B., Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association* 69:364–367.
4. Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society* 4:102–118.
5. Conover, W. J., Johnson, M. E., Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics* 23:351–361.
6. Hall, P., Padmanabhan, R. A. (1997). Adaptive inference for the two-sample scale problem.*Technometrics* 23:351–361.
7. Hartley, H. O. (1950). The use of range in analysis of variance. *Biometrika* 37:271–280.
8. Higgins, J. J. (2004). *Introduction to Modern Nonparametric Statistics*. Pacific Grove, CA: Duxbury.
9. Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distributions. In: Hoaglin, D. C.,Mosteller, F., Tukey, J. W., eds. *Exploring Data Tables, Trends, and Shapes*. Hoboken, NJ: John Wiley & Sons, Inc, pp. 461–511.
10. Keller-McNulty, S., Higgins, J. J. (1987). Effect of tail weight and outliers on power and type-I error of Robust permutation tests for location. *Communications in Statistics—Simulation and Computation*16(1):17–36.
11. Keselman, H. J., Games, P. A., Clinch, J. J. (1979). Tests for homogeneity of variance. *Communications in Statistics—Simulation and Computation* 8:113–119.
12. Levene, H. (1960). *Robust tests for equality of variances*, In: Olkin, I., Hotelling, H., eds. *Contributions to Probability and Statistics*. Palo Alto, CA: Stanford University Press, 278–292.
13. Marozzi, M. (2011). Levene type tests for the ratio of two scales, *Journal of Statistical Computation and Simulation* 81(7):815–826.
14. Marozzi, M. (2012). A modified hall-padmanabhan test for the homogeneity of scales, *Communication in Statistics—Theory and Methods* 41:3068–3078.
15. Marozzi, M. (2016) Multivariate tests based on interpoint distances with application to magnetic resonance imaging, *Statistical Methods in Medical Research* 25(6):2593–2610.

16. O'Brien, R. G. (1979). A general ANOVA method for robust tests of additive models for variances.*Journal of the American Statistical Association* 74:877–880.

17. Olejnik, S. F., Algina, J. (1988). Tests of variance equality when distributions differ in form and location,*Educational and Psychological Measurement* 48:317–329.

18. Pesarin, F. (2001). *Multivariate Permutation Tests with Applications in Biostatistics*. Chichester: Wiley.

19. Pesarin, F., Salmaso, L. (2010). *Permutation Tests for Complex Data*. Chichester: Wiley.

20. Sharma, D., Kibria, B. M. (2013). On some test statistics for testing homogeneity of variances: a comparative study. *Journal of Statistical Computation and Simulation* 83(10):1944–1963.