Using the Tukey–Kramer omnibus test in the Hayter–Fisher procedure

By: Scott J. Richter, Melinda H. McCann

Richter, S. J. and McCann, M. H. (2012). Using the Tukey-Kramer Test as the Omnibus Test in the Hayter-Fisher Procedure. *British Journal of Mathematical and Statistical Psychology*, 65(3), 499-510. doi:10.1111/j.2044-8317.2012.02041.x,

Abstract:

Using Tukey–Kramer versus the ANOVA F-test as the omnibus test of the Hayter–Fisher procedure for comparing all pairs of normally distributed means, when sample sizes are unequal, is investigated. Simulation results suggest that using Tukey–Kramer leads to as much or more any-pairs power compared to using the F-test for certain patterns of mean differences, and equivalent per-pair and all-pairs power for all cases. Furthermore, using Tukey–Kramer results in a consonant test procedure, where there cannot be disagreement between the results of the omnibus test and the subsequent pairwise tests. The results suggest that when sample sizes are unequal, Tukey–Kramer may be preferred over the F-test as the omnibus test for the Hayter–Fisher procedure.

Keywords: statistical methods | omnibus test | Hayter–Fisher procedure | Tukey–Kramer

Article:

Many studies can be analysed within the framework of the one-way, fixed effects analysis of variance (ANOVA) model,

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \ldots, k, \;\; j = 1, \ldots, n_i,$$

where the εij are independent $N(0, \sigma^2)$ random variables, and the μi and σ^2 are unknown parameters. The ANOVA *F*-test can be used to test the hypothesis that all μi are equal. However, the *F*-test does not determine which sample means are statistically different when one is interested in pairwise differences. Thus, multiple pairwise comparisons of the means, especially of all pairwise differences, are a common goal of researchers. Textbooks on statistical methods are virtually unanimous in presenting the Tukey/Tukey–Kramer (Tukey, 1949, 1953;Kramer, 1956) simultaneous pairwise comparison procedure. When confidence intervals are not needed, however, the Hayter–Fisher procedure (Hayter, 1986) is often recommended as a more powerful alternative to Tukey's testing procedure. In fact, although there are other methods that can be slightly more powerful (see, for example, Peritz, 1970; Ramsey, 1978, 1981; Shaffer, 1979, 1986; Welsch, 1977;Westfall, 1997), the relative simplicity of the Tukey–Kramer and Hayter–Fisher methods continues to make them attractive (Ramsey, 2002;Myers & Well, 2003; Ramsey & Ramsey, 2008).

The Tukey/Tukey–Kramer (TK) procedure makes all pairwise comparisons using the statistic

$$q^* = \frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{\frac{1}{2} MSE(1/n_i + 1/n_j)}},$$

which is compared to the quantile, $q_{\alpha,}k_{,\nu}$, of the studentized range distribution for k means with ν error degrees of freedom. When sample sizes are equal ($n_i=n_j=n$) the statistic reduces to

$$q^* = \frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{MSE/n}}.$$

Alternatively, the Hayter–Fisher (HF) method was originally devised as an improvement on Fisher's least significant difference (LSD) procedure with equal sample sizes, and is carried out as follows. First, test the overall null hypothesis, $H_0:\mu_1=\mu_2=\cdots=\mu k$, at level α, using the ANOVA F-test. If the F-test is significant, employ Tukey's procedure for testing all pairwise differences, using $q_{\alpha,}k_{-1,\nu}$, the studentized range distribution quantile for $k-1$ means, instead of $q_{\alpha,}k_{,\nu}$. If the F-test is not significant, make no comparisons and no pairwise differences can be declared significant at familywise significance level α.

A similar, but not equivalent, modification to the Hayter–Fisher procedure is to replace the F-test, in the first step, with the Tukey-Q test as the ‘omnibus’ test. Simulation results have shown that this modification can result in higher power for testing the overall null hypothesis,$H_0:\mu_1=\mu_2=\cdots=\mu k$, for certain configurations of means. David, Lachenbruch, and Brandis (1972) and Seaman, Levin, and Serlin (1991) demonstrated this, for equal sample sizes, for the maximum range configuration – two extreme means with equal means between them. For unequal sample sizes, Ramsey and Ramsey (2008) compared the Hayter–Fisher procedure, using the Tukey–Kramer test procedure with $k-1$ groups after a significant F-test, to the usual Tukey–Kramer procedure. Their simulation results showed that for unequal sample sizes, the Tukey–Kramer procedure can actually have higher power to detect at least the largest studentized pairwise difference (any-pair power) than the Hayter–Fisher procedure for the single extreme mean configuration. These results should not be surprising, since the Tukey–Kramer omnibus test, which uses the studentized range distribution based on maximum pairwise differences, should be more sensitive to pairwise differences. With regard to detecting all pairwise differences (all-pairs power), however, Ramsey and Ramsey (2008) found that the Tukey–Kramer procedure always performed more poorly than the Hayter–Fisher procedure. Thus, although there may sometimes be a slight advantage for Tukey–Kramer in detecting the largest studentized pairwise difference, whenever both procedures are successful at detecting the largest studentized pairwise difference, the Hayter–Fisher procedure will always have at least as much power to further detect other, possibly smaller, pairwise differences, since the Hayter–Fisher procedure employs the Tukey–Kramer critical value based on $k - 1$ groups, rather than k.

Ramsey and Ramsey (2008) did not consider using the Tukey–Kramer test as the omnibus test in the Hayter–Fisher procedure. However, based on their results as well as previous ones, making this modification should result in a procedure superior to both TK and HF for detecting pairwise differences. The focus of this investigation is to compare, for unequal sample sizes, the power of the modified Hayter–Fisher procedure using the Tukey–Kramer test of the largest observed studentized pairwise difference as the omnibus test, to that of the usual Hayter–Fisher procedure using the *F*-test as the omnibus test. To avoid confusion and to emphasize its sequential testing nature, the modified version of the Hayter–Fisher procedure using the Tukey–Kramer test will be referred to as the ‘Tukey–Kramer two-step’ (TK2S), using ‘HF’ to refer to the usual Hayter–Fisher test utilizing the *F*-test as the omnibus test. For greater simplicity in implementation, TK2S might be better thought of as a sequential testing procedure. First, compare the largest observed test statistic, $q*$, to $q_{\alpha,k,\nu}$– if significant, then proceed to the next largest test statistic, comparing it to $q_{\alpha,k-1,\nu}$, and continuing in this fashion, using $q_{\alpha,k-1,\nu}$ for all subsequent comparisons.

2. Simulation

2.1 . Details of the simulation

Consider the one-way fixed-effects ANOVA model described in Section 1:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \ldots, k, \; j = 1, \ldots, n_i,$$

where the εij are independent $N(0, \sigma^2)$ random variables, and the μi and σ^2 are unknown parameters. Three procedures were considered:

1 *TK2S*. Test the pairwise difference with the largest observed test statistic using $q_{\alpha,k,\nu}$– if the test is significant then proceed to test all remaining pairwise differences using $q_{\alpha,k-1,\nu}$.

2 *HF*. Carry out the ANOVA *F*-test – if the test is significant, then proceed to test all pairwise differences using $q_{\alpha,k-1,\nu}$.

3 *TK*. Test all pairwise differences using $q_{\alpha,k,\nu}$.

Following Ramsey and Ramsey (2008), several different mean and sample size configurations were considered (see Table A1 in the Appendix), and the values of μi selected to produce specified values of Cohen's effect size, $f=\sigma m/\sigma$, where $\mu = \sum \mu_i / k$ and $\sigma_m^2 = \sum (\mu_i - \mu)^2 / k$. The *maximum range*configuration is defined by $\mu_1 = -\sigma f \sqrt{k/2},\ \mu_2 = \cdots = \mu_{k-1},\ \mu_k = \sigma f \sqrt{k/2}$, and produces configurations in which the smallest and largest means are as far apart as possible for the specified effect size. Ramsey (1978) showed that this configuration favours tests based on the studentized range distribution compared to those based on the *F*-distribution. The *minimum range* configuration, on the other hand, produces configurations in which the smallest and largest means are as close as possible for the specified effect size. The minimum range configuration will favour tests based on the *F*-distribution, compared to those based on the studentized range distribution (Ramsey, 1978). In addition, the *single extreme mean* configuration, where $\mu_1=\cdots=\mu k_{-1}$ with μk different, was considered. Ramsey and Ramsey (2008) found TK to have higher any-pair power than HF for the single extreme mean configuration. Finally, a configuration where means were equally spaced was also considered.

Three sample size configurations, representing different ranges between the largest and smallest sample sizes, were chosen for each setting of ν, the error degrees of freedom. These are given in Table A1. For each case, sample sizes were randomly assigned to the *k* groups for each simulated data set, to ensure that all possible sample size pairings with groups were equally represented. The power results reported represent the average power, over the three sample size configurations, for each setting of ν.

Familywise Type I error rate (FWER), any-pair, average per-pair (per-pair power, averaged over all non-null pairs), and all-pairs power were estimated, based on 10,000 randomly selected samples, for the TK, HF and TK2S procedures.

2.2 . Simulation results

Tables 1–7 show the estimated power for each procedure, averaged over several patterns of unequal sample sizes (detailed in Table A1), and are presented above. Additional simulation results, including estimated familywise error rates for all procedures, are available from the first author.

Table 1. Estimated any-pair power for the Tukey–Kramer (TK) procedure, and the Hayter–Fisher procedure using either the Tukey–Kramer (TK2S) or ANOVA *F* (HF) omnibus test: single extreme mean configuration, *k* groups, ν degrees of freedom and effect size *f*

k	**ν**	***f***	**Estimated any-pair power**			**Difference**	
			TK2S	**HF**	**TK**	**TK2S–HF**	**TK2S–TK**
4	5	1.60	.7980	.8002	.7980	−.002	.000

		1.31	.6394	.6481	.6393	−.009	.000
	10	1.31	.8608	.8553	.8606	.005	.000
		0.73	.4407	.4502	.4397	−.010	.001
4	20	0.73	.7906	.7883	.7903	.002	.000
		0.44	.3982	.4039	.3967	−.006	.002
4	60	0.44	.7983	.7993	.7979	−.001	.000
		0.36	.6341	.6396	.6328	−.005	.001
6	5	1.60	.7472	.7131	.7472	.034	.000
		1.31	.6228	.5881	.6228	.035	.000
6	10	1.31	.8030	.7649	.8030	.038	.000
		0.73	.3469	.3266	.3458	.020	.001
6	20	0.73	.7671	.7440	.7667	.023	.000
		0.44	.3366	.3135	.3353	.023	.001
6	60	0.44	.6783	.6619	.6776	.016	.001
		0.36	.5344	.5134	.5337	.021	.001

Table 2. Estimated any-pair power for the Tukey–Kramer (TK) procedure, and the Hayter–Fisher procedure using either the Tukey–Kramer (TK2S) or ANOVA *F* (HF) omnibus test: maximum range means configuration, *k* groups, ν degrees of freedom and effect size *f*

k	***df***	***f***	**Estimated any-pair power**			**Difference**	
			TK2S	**HF**	**TK**	**TK2S–HF**	**TK2S–TK**
4	5	1.60	.8186	.8104	.8186	.008	.000
		1.31	.6588	.6495	.6587	.009	.000
	10	1.31	.8677	.8692	.8676	−.002	.000
		0.73	.4315	.4386	.431	−.007	.000

4	20	0.73	.8059	.8179	.8058	−.012	.000
		0.44	.3977	.4071	.3970	−.009	.001
4	60	0.44	.7898	.7918	.7834	−.002	.006
		0.36	.6373	.6426	.6370	−.005	.000
6	5	1.60	.7747	.7249	.7747	.050	.000
		1.31	.6012	.5526	.6012	.049	.000
6	10	1.31	.8059	.7786	.8059	.027	.000
		0.73	.3404	.3285	.3396	.012	.001
6	20	0.73	.7686	.7419	.7684	.023	.000
		0.44	.3258	.3089	.3246	.017	.001
6	60	0.44	.6979	.6854	.6975	.012	.000
		0.36	.5293	.5151	.5289	.014	.000

Table 3. Estimated any-pair power for the Tukey–Kramer (TK) procedure, and the Hayter–Fisher procedure using either the Tukey–Kramer (TK2S) or ANOVA *F* (HF) omnibus test: minimum range means configuration, *k* groups, ν degrees of freedom and effect size *f*

k	***df***	***f***	**Estimated any-pair power**			**Difference**	
			TK2S	**HF**	**TK**	**TK2S–HF**	**TK2S–TK**
4	5	1.60	.7548	.8199	.7548	−.065	.000
		1.31	.5863	.6529	.5862	−.067	.000
	10	1.31	.8617	.8988	.8617	−.037	.000
		0.73	.4200	.4551	.4195	−.035	.001
4	20	0.73	.8114	.8455	.8113	−.034	.000
		0.44	.3749	.4005	.3739	−.026	.001
4	60	0.44	.7720	.8125	.7717	−.040	.000

		0.36	.5995	.6439	.5988	−.044	.001
6	5	1.60	.6547	.6974	.6547	−.043	.000
		1.31	.4897	.5226	.4897	−.033	.000
6	10	1.31	.7476	.7816	.7476	−.034	.000
		0.73	.2996	.3127	.2990	−.013	.001
6	20	0.73	.6946	.7239	.6943	−.029	.000
		0.44	.0386	.0395	.0301	−.001	.009
6	60	0.44	.6712	.6865	.6707	−.015	.001
		0.36	.4923	.5084	.4914	−.016	.008

Table 4. Estimated average per-pair power for the Tukey–Kramer (TK) procedure, and the Hayter–Fisher procedure using either the Tukey–Kramer (TK2S) or ANOVA *F* (HF) omnibus test: single extreme mean configuration, *k* groups, ν degrees of freedom and effect size *f*

k	**ν**	***f***	**Estimated per-pair power**			**Difference**	
			TK2S	**HF**	**TK**	**TK2S–HF**	**TK2S–TK**
4	5	1.60	.7019	.7058	.6397	−.004	.062
		1.31	.5247	.5323	.4655	−.008	.059
	10	1.31	.7520	.7501	.7026	.002	.049
		0.73	.2989	.3045	.2614	−.006	.038
4	20	0.73	.6130	.6123	.5580	.001	.055
		0.44	.2368	.2421	.2061	−.005	.031
4	60	0.44	.6425	.6447	.5879	−.002	.054
		0.36	.4544	.4592	.4037	−.005	.051
6	5	1.60	.5761	.5654	.5321	.011	.044
		1.31	.4391	.4291	.4011	.010	.038

6	10	1.31	.5930	.5828	.5543	.010	.039
		0.73	.1691	.1644	.1513	.005	.018
6	20	0.73	.5060	.5003	.4665	.006	.039
		0.44	.1457	.1411	.1298	.005	.016
6	60	0.44	.4053	.4015	.3753	.004	.030
		0.36	.2589	.2545	.2353	.004	.024

Table 5. Estimated per-pair power for the Tukey–Kramer (TK) procedure, and the Hayter–Fisher procedure using either the Tukey–Kramer (TK2S) or ANOVA *F* (HF) omnibus test: maximum range means configuration, *k* groups, ν degrees of freedom and effect size *f*

k	***df***	***f***	**Estimated per-pair power**			**Difference**	
			TK2S	**HF**	**TK**	**TK2S–HF**	**TK2S–TK**
4	5	1.60	.4376	.4368	.3720	.001	.066
		1.31	.3150	.3144	.2665	.001	.049
	10	1.31	.4875	.4901	.4213	−.003	.067
		0.73	.1786	.1815	.1507	−.003	.028
4	20	0.73	.3673	.3711	.3146	−.004	.053
		0.44	.1415	.1450	.1199	−.004	.022
4	60	0.44	.2666	.2674	.2386	−.001	.028
		0.36	.2618	.2643	.2221	−.003	.040
6	5	1.60	.3046	.2992	.2729	.005	.032
		1.31	.2053	.1998	.1819	.006	.023
6	10	1.31	.3109	.3080	.2786	.003	.032
		0.73	.0869	.0857	.0766	.001	.010
6	20	0.73	.2426	.2397	.2164	.003	.026

		0.44	.0757	.0740	.0667	.002	.009
6	60	0.44	.1971	.1958	.1770	.001	.020
		0.36	.1265	.1250	.1127	.001	.014

Table 6. Estimated per-pair power for the Tukey–Kramer (TK) procedure, and the Hayter–Fisher procedure using either the Tukey–Kramer (TK2S) or ANOVA *F* (HF) omnibus test: minimum range means configuration, *k* groups, ν degrees of freedom and effect size *f*

k	***df***	***f***	**Estimated per-pair power**			**Difference**	
			TK2S	**HF**	**TK**	**TK2S–HF**	**TK2S–TK**
4	5	1.60	.5872	.6118	.5153	−.025	.072
		1.31	.4134	.4390	.3543	−.026	.059
	10	1.31	.6497	.6627	.5854	−.013	.064
		0.73	.2240	.2373	.1898	−.013	.034
4	20	0.73	.4966	.5093	.4373	−.013	.059
		0.44	.1717	.1816	.1464	−.010	.025
4	60	0.44	.5146	.5288	.4520	−.014	.063
		0.36	.3429	.3585	.2938	−.016	.049
6	5	1.60	.3734	.3806	.3346	−.007	.039
		1.31	.2408	.2463	.2134	−.006	.027
6	10	1.31	.3889	.3941	.3512	−.005	.038
		0.73	.1251	.1277	.1095	−.003	.016
6	20	0.73	.2966	.3015	.2644	−.005	.032
		0.44	.0815	.0826	.0713	−.001	.010
6	60	0.44	.2449	.2473	.2193	−.002	.026
		0.36	.1462	.1489	.1291	−.003	.017

Table 7. Estimated all-pairs power for the Tukey–Kramer (TK) procedure, and the Hayter–Fisher procedure using either the Tukey–Kramer (TK2S) or ANOVA *F* (HF) omnibus test: *k* groups, ν degrees of freedom and effect size *f*

k	***df***	***f***	**Estimated all-pairs power**			**Difference**	
			TK2S	**HF**	**TK**	**TK2S-HF**	**TK2S-TK**
Single extreme mean							
4	5	1.60	.5774	.5789	.4706	−.002	.107
	10	1.31	.5952	.5950	.5058	.000	.089
	20	0.73	.3554	.3540	.2718	.001	.084
	60	0.73	.4483	.4488	.3607	−.001	.088
6	5	1.60	.3425	.3424	.2937	.000	.049
	10	1.31	.3253	.3253	.2834	.000	.042
Maximum Range							
4	5	1.60	.0923	.0923	.0599	.000	.032
	10	1.31	.0912	.0912	.0560	.000	.035
	60	0.44	.1471	.1473	.0984	−.000	.049
Minimum Range							
4	5	1.60	.3783	.3801	.2859	−.002	.092
	10	1.31	.3965	.3973	.3073	−.001	.089
	20	0.73	.1880	.1884	.1268	−.005	.061
	60	0.44	.2288	.2294	.1592	−.001	.070

2.2.1 . Familywise error rate

FWER control for TK and HF has been demonstrated both mathematically and empirically (see, for example, Hayter, 1984; Dunnett, 1980; Ramsey & Ramsey, 2008), and the results of the present simulation were consistent with previous results. Estimated FWERs for TK2S were similar to HF, and no evidence was found of error rates above the nominal level of α= 0.05.

2.2.2 . Any-pair power

Consistent with results of previous studies (see Section 1), both TK and TK2S tended to have higher any-pair power than HF for the single extreme and maximum range mean configurations, especially for *k*= 6 (see Tables 1 and 2). The maximum observed power advantage was .038 (*k*= 6, ν= 10, *f*= 1.31) for the single extreme mean case and .050 (*k*= 6, ν= 5, *f*= 1.60) for the maximum range case. For the minimum range and equally spaced means configurations, HF tended to have slightly higher any-pair power than TK and TK2S: as much as .067 (*k*= 4, ν= 5, *f*= 1.31) for the minimum range case and .020 (*k*= 4, ν= 20, *f*= 0.73) for equally spaced means (Table 3 shows results for the minimum range configuration – results for the equally spaced means configuration were similar).

For individual cases, the maximum observed power advantage for TK2S was .0523 (*k*= 6, *NR*= 11, ν= 10, *f*= 1.31) for the single extreme mean case and .0741 (*k*= 6, *NR*= 2, ν= 5, *f*= 1.31) for the maximum range case; and for HF the advantage was as much as .0782 (*k*= 4, *NR*= 6, ν= 5, *f*= 1.60)for the minimum range case and .0364 (*k*= 4, *NR*= 11, ν= 10, *f*= 1.31) for equally spaced means.

2.2.3 . Per-pair power

Tables 4–6 show estimated per-pair power for the three methods. While TK2S and HF always had substantially higher power than TK, there was little or no difference in power between TK2S and HF. This was true even for cases where either TK2S or HF enjoyed an any-pair power advantage. The maximum per-pair power advantage observed for TK2S over HF was .0024 for the single extreme mean case (*k*= 4, *NR*= 8, ν= 20, *f*= 0.73), and for HF over TK2S was .0019 for the minimum range case (*k*= 4, *NR*= 1.5, ν= 5, *f*= 1.60).

2.2.4 . All-pairs power

Table 7 shows estimated all-pairs power for the three methods for selected cases. Consistent with results of previous studies (e.g., Ramsey & Ramsey, 2008) was that the all-pairs power of TK was always substantially less than that of HF, even in cases where TK had higher any-pair power. However, for all cases, there was little or no difference in all-pairs power between HF and TK2S, with observed differences in all-pairs power less than .005 for all cases considered. As was found for per-pair power, this was true even for cases where either TK2S or HF enjoyed an any-pair power advantage.

3. Discussion

3.1 . Power comparisons

An important result of the simulations is that using the Tukey–Kramer omnibus test instead of the ANOVA *F*-test results in a procedure with both per-pair and all-pairs power essentially equivalent to that of the Hayter–Fisher test. In the single extreme mean and maximum range configurations, situations where pairwise differences are all moderate to large, TK2S was usually more powerful than HF in detecting at least the pair with the largest studentized mean difference. That is, the 'omnibus' TK test returned a significant result more often than the *F*-test. In these cases, since the subsequent pairwise tests for both procedures are identical, this must translate into higher any-pair power for TK2S. However, even for means configurations that tend to favour the *F*-test – configurations where there are small pairwise differences – TK2S had essentially the same per-pair power as HF, and was able to detect all pairwise differences with the same frequency as the HF test. This is due to the fact that although the overall *F*-test returns a significant result more often than does the TK omnibus test, the subsequent TK tests of the HF procedure do not always find at least one pairwise difference significant.

Table 8 gives a particular example for the proportion of rejections for the respective omnibus tests of TK2S and HF for two means configurations: the maximum range configuration, for which TK2S had higher any-pair power, and the minimum range configurations, for which HF had higher any-pair power. For the maximum range means configuration, TK2S showed higher omnibus power, that is, tended to return a significant result more often than HF. Since the subsequent tests for the pairwise differences of the two procedures are identical, this necessarily translates into higher any-pair power for TK2S, since its omnibus test is in fact detecting the largest studentized pairwise difference, which must also be declared significant by the pairwise tests. However, the *F*-test will occasionally reject when TK does not, and pairwise differences may be detected by HF in these cases. This is why the any-pair power advantage enjoyed by TK2S does not translate to a per-pair or all-pairs power advantage. When there tend to be many differences of varying magnitude, non-pairwise contrasts have the potential to be larger than the largest pairwise contrast. Consequently, there might be more samples where the *F*-test rejects, but is not necessarily detecting the largest pairwise difference as the significant contrast. For these cases, HF may not declare any pairwise differences significant and the any-pairs power advantage over TK2S will be less than the omnibus test power advantage. Thus, the fact that the *F*-test rejects more often than TK does not automatically lead to HF having higher average per-pair or all-pairs power than TK2S.

Table 8. Omnibus test, any-pair and average per-pair power for the Tukey–Kramer (TK) procedure, and the Hayter–Fisher procedure using either the Tukey–Kramer (TK2S) or ANOVA *F* (HF) as the omnibus test, *k* groups, ν degrees of freedom and effect size *f*

Means configuration	***k***	**ν**	***f***	**Omnibus power**		**Any-pair power**		**All-pairs power**	
				TK2S	**HF**	**TK2S**	**HF**	**TK2S**	**HF**
Maximum range	6	10	1.31	.8045	.7915	.8043	.7794	.3109	.3080
Minimum range	6	10	1.31	.7478	.8352	.7476	.7816	.3889	.3941

Note also that the any-pair power of TK2S was often slightly higher than that of TK. While it may seem that these two procedures should have exactly the same any-pair power, consider that, in practice, it may happen that the largest observed studentized mean difference is associated with a true null hypothesis, that is, groups whose true means do not differ. When this occurs, both TK2S and TK may also correctly declare other groups different, but TK2S is more likely to do this, since it employs a less conservative critical value for subsequent tests. Thus, the any-pair power of TK2S can be slightly higher than that of TK.

3.2 . Statistical practice

Using TK as the omnibus test results in a consonant procedure (Gabriel, 1969), so that if the omnibus test is significant, then at least one pairwise difference must also be declared significant. This is not the case when using the *F*-test as the omnibus test, however. Recall that the omnibus *F*-test rejects whenever there is at least one significant non-zero contrast of the *k* means, namely that there exist constants, c_1, c_2, …, c_k, such that $c_1\mu_1+c_2\mu_2+\cdots+c_k\mu k\neq 0$, and thus while a significant*F*-test does imply that at least one pair of means differ, it does not guarantee that a pairwise difference will be found significant. (Here by ‘pairwise’ we mean a contrast of the form $\mu i-\mu j$ for $i\neq j$.) For instance, the *F*-test may detect that $(\mu_1+\mu_2)/2 -\mu_3$, or some other non-pairwise contrast of the means, is significantly different from 0, although none of the pairs $\mu i-\mu j$, $i\neq j$, are declared significantly different from 0. This raises a more subtle point regarding the use of the *F*-test as an omnibus test in any multiple pairwise testing procedure.

As has been pointed out many times in the literature, there is no need for a preliminary omnibus test when using the TK procedure. Hsu (1996) states that to consider performing multiple comparisons only if the ANOVA *F*-test rejects is ‘a mistake’. Ramsey (1978, 1981, 2002) and Ramsey and Ramsey (2008) present simulation results illustrating that the power of TK suffers if applied only after a significant omnibus *F*-test. Still, many textbooks (see Ramsey & Ramsey, 2008, p. 116) recommend using TK only after a significant *F*-test, and it is rare to find an example in the applied literature where the TK procedure has been applied without a significant preliminary *F*-test (Ramsey & Ramsey, 2008).

This emphasis on a preliminary *F*-test can only help to perpetuate the myths among many practitioners that a significant *F*-test implies at least one significant pairwise difference using TK; and that a non-significant *F*-test implies it is not possible that TK will find a significant pairwise difference. Certainly rejection of the ‘omnibus’ TK test implies that at least one pairwise difference exists. However, while rejection of the omnibus *F*-test certainly implies that not all the means are equal, it does not guarantee that there is at least one significant *pairwise* contrast of the means. Cohen (2001, Chapter 13) presents numerical examples of such cases, as well as examples with equal sample sizes where using *F* versus TK as the omnibus test with the HF test results in a different conclusion. Textbooks sometimes add to this confusion. For example, Ott and Longnecker (2004, p. 365) state ‘we can safely conclude that all pairs of treatment means are not significantly different, because the AOV *F*-test failed to reject the null hypothesis'. Many statistical consultants and teachers of statistical methods have undoubtedly had to deal with confusion regarding seemingly conflicting results between the *F*-test and a pairwise testing procedure. Although the Hayter–Fisher method is not an example of the incorrect use of the *F*-test as a gateway to a multiple pairwise testing procedure, it may appear to inexperienced practitioners as consistent with incorrect practice. Thus, replacing HF with TK2S would hopefully lead to better practice in general with multiple pairwise comparisons.

4. Conclusion

The Tukey–Kramer procedure continues to be an attractive method for making all pairwise comparisons, especially since confidence intervals are available. However, when greater power to detect pairwise differences is desired and confidence intervals are not required, TK2S is recommended as an alternative to the HF procedure. While HF can have slightly higher any-pair power under certain conditions, TK2S tends to have higher any-pair power for mean configurations where pairwise differences are all moderate to large. In addition, the all-pairs and per-pair power of TK2S is virtually identical to HF for all mean configurations, even those for which HF holds an any-pair power advantage. Finally, TK2S is a consonant procedure for which there cannot be disagreement between the 'omnibus' test and subsequent pairwise tests, so that a significant omnibus test under TK2S guarantees that at least the largest pairwise studentized difference will be declared significant. In contrast, the HF procedure has the undesirable property that it is possible that no pairwise differences are declared significant, even though the omnibus *F*-test is significant.

References

Cohen, B. H. (2001). *Explaining psychological statistics*, 2nd ed. New York : Wiley.

David, H. A., Lachenbruch, P. A., & Brandis, H. S. (1972). The power function of range and Studentized range tests in normal samples.*Biometrika*, **59**, 161–168.

Dunnett, C. W. (1980). Pairwise multiple comparisons in the homogeneous variance, unequal sample size case. *Journal of the American Statistical Association*, **75**, 789–795.

Gabriel, K. R. (1969). Simultaneous test procedures: Some theory of multiple comparisons. *Annals of Mathematical Statistics*, **40**,224–250.

Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *Annals of Statistics*, **12**, 61–75.

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. *Journal of the American Statistical Association*, **81**, 1000–1004.

Hsu, J. C. (1996). *Multiple comparisons: Theory and methods*. London : Chapman & Hall.

Kramer, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics,* **12**, 309–310.

Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis,* 2nd ed. Mahwah , NJ : Lawrence Erlbaum.

Ott, R. L., & Longnecker, M. (2004). *A First Course in Statistical Methods*. Belmont , CA : Brooks/Cole-Thomson Learning.

Peritz, E. (1970). A note on multiple comparisons. Unpublished manuscript, Hebrew University.

Ramsey, P. H. (1978). Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association*, **73**,479–485.

Ramsey, P. H. (1981). Power of univariate pairwise multiple comparison procedures. *Psychological Bulletin,* **90**, 352–366.

Ramsey, P. H. (2002). Comparison of closed procedures for pairwise testing of means. *Psychological Methods,* **7**, 504–523.

Ramsey, P. H., & Ramsey, P. P. (2008). Power of pairwise comparisons in the equal variance and unequal sample size case. *British Journal of Mathematical and Statistical Psychology,* **61**, 115–131.

Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practical procedures. *Psychological Bulletin*, **110**, 577–586.

Shaffer, J. P. (1979). Comparison of means: An F test followed by a modified multiple range test. *Journal of Educational Statistics*, **4**,14–23.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, **81**,826–831.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics,* **5**, 99–114.

Tukey, J. W. (1953). The problem of multiple comparisons. Unpublished manuscript, Department of Statistics, Princeton University.

Welsch, R. E. (1977). Stepwise multiple comparison procedures. *Journal of the American Statistical Association,* **72**, 566–575.

Westfall, P. H. (1997). Multiple testing of general contrasts using logical constraints and correlations. *Journal of the American Statistical Association*, **92**, 299–306.

Appendix

Table A1. Sample sizes used for each case. *k*= number of groups, ν= error degrees of freedom, NR = range of largest to smallest sample size, n_1, … , n_6= sample size associated with the *i*th group

k	**ν**	**NR**	$\boldsymbol{n_1}$	$\boldsymbol{n_2}$	$\boldsymbol{n_3}$	$\boldsymbol{n_4}$	$\boldsymbol{n_5}$	$\boldsymbol{n_6}$
4	5	6.0	6	1	1	1	—	—
		1.5	3	2	2	2	—	—
		3.0	3	3	2	1	—	—
	10	11.0	11	1	1	1	—	—
		5.0	5	5	3	1	—	—
		2.0	4	4	4	2	—	—
4	20	8.0	8	8	7	1	—	—
		5.0	10	10	2	2	—	—
		1.8	7	7	6	4	—	—
4	60	9.8	49	5	5	5	—	—
		1.9	25	13	13	13	—	—
		1.7	20	20	12	12	—	—
6	5	5.0	5	2	1	1	1	1
		3.0	3	3	2	1	1	1
		2.0	2	2	2	2	2	1
6	10	11.0	11	1	1	1	1	1
		9.0	9	3	1	1	1	1
		1.5	3	3	3	3	2	2
6	20	8.0	8	8	3	3	3	1
		2.0	6	5	5	4	3	3
		1.7	5	5	5	4	4	4

6	60	40.0	40	22	1	1	1	1
		2.0	18	12	9	9	9	9
		1.8	16	14	9	9	9	9