

On Detecting a Minimal Important Difference among Standardized Means

Gwown Shieh¹

Published online: 28 December 2016
© Springer Science+Business Media New York 2016

Abstract As a generalization of the standardized mean difference between two independent populations, two different effect size measures have been proposed to represent the degree of disparity among several treatment groups. One index relies on the standard deviation of the standardized means and the second formula is the range of the standardized means. Despite the obvious usage of the two measures, the associated test procedures for detecting a minimal important difference among standardized means have not been well explicated. This article reviews and compares the two approaches to testing the hypothesis that treatments have negligible effects rather than that of no difference. The primary emphasis is to reveal the underlying properties of the two methods with regard to power behavior and sample size requirement across a variety of design configurations. To enhance the practical usefulness, a complete set of computer algorithms for calculating the critical values, *p*-values, power levels, and sample sizes is also developed.

Keywords Effect size · Power · Sample size · Standardized mean difference

The editorial guidelines and methodological recommendations of several prominent educational and psychological journals have extensively recommended effect size reporting

Electronic supplementary material The online version of this article (doi:10.1007/s12144-016-9549-5) contains supplementary material, which is available to authorized users.

✉ Gwown Shieh
gwshieh@mail.nctu.edu.tw

¹ Department of Management Science, National Chiao Tung University, Hsinchu, Taiwan 30010, Republic of China

and interpreting practices for all primary outcomes of empirical studies. Accordingly, numerous practical suggestions for selecting, calculating, and interpreting effect size indices for various types of statistical analyses have been provided in the literature (see Alhija and Levy 2009; Breaugh 2003; Durlak 2009; Ferguson 2009; Fern and Monroe 1996; Grissom and Kim 2012; Huberty 2002; Kirk 1996; Kline 2004; Olejnik and Algina 2000; Richardson 1996; Rosenthal et al. 2000; Rosnow and Rosenthal 2003; Vacha-Haase and Thompson 2004). Overall, group difference and strength of association are two of the major classes of effect sizes in practical applications. It is essential to note that the standardized mean difference is a widely adopted effect size measure when comparing treatment means of two independent groups. On the other hand, in univariate research that compares more than two treatment effects, the correlation ratio, eta squared, measures the proportion of total variance accounted for by any of the treatment effects.

Eta squared is an appropriate generalization of the coefficient of determination and reflects the extent to which knowledge of group membership improves prediction of outcomes, two different extensions of standardized mean difference have also been proposed to represent the degree of difference among several treatment groups (Cohen 1969, 1988, Chapter 8). One index relies on the standard deviation of the standardized means and the second index is the range of the standardized means. To test the null hypothesis of no difference or zero effect, the analysis of variance (ANOVA) *F* test and the studentized range test are the two prominent procedures associated with the effect measures of the standard deviation of the standardized means and the range of the standardized means, respectively. Although there are more sophisticated statistical methods, ANOVA *F* test remains one of the most frequently used approaches to detecting group differences in the behavioral, educational and social sciences

(Skidmore and Thompson 2010; Warne et al. 2012). To facilitate the application of F test for power and sample size calculations in planning research designs, Cohen (1988) provided comprehensive illustrations and extensive tables for the standard deviation of the standardized means.

The studentized range is an intuitively appealing extension of the standardized mean difference under a multiple treatments scenario. However, there are no corresponding demonstrations and explications for the studentized range test in Cohen (1988). It is conceivable that the distributional property of the studentized range statistic is more involved than that of the sample standardized mean difference. The quantiles of the studentized range statistic are commonly available in most statistical textbooks for conducting multiple comparisons between group means. Due to computational demands, the power behavior of a studentized range test is not as widely addressed as the F test for the alternative measure of the standard deviation of the standardized means. Specifically, David et al. (1972) and Hayter and Hurn (1992) conducted various power assessments of the studentized range test and the F test for the equality of several normal means. Hayter and Hurn (1992) found that in most situations the power performance of the F test and the studentized range test is very close to that of an optimal test procedure considered in Hayter and Liu (1992). However, the optimal test procedure is substantially more complicated to use than the F test and the studentized range test. More importantly, the empirical comparisons of David et al. (1972) showed that the studentized range test provides higher power than the F test at some configurations of the treatment means.

It is noteworthy that the prescribed appraisals of the F method and the studentized range procedure were confined to a test of the traditional null hypothesis of no difference in treatment means. However, researchers are often more concerned about whether the treatment effects are large enough to have a practical importance. Essentially, it is more appropriate to test the null hypothesis that the effects of treatments are trivial or negligible as advocated in Cohen (1994), Murphy (1990), and Serlin and Lapsley (1993). The tests of minimal effect null hypotheses require an operational definition of the target minimal reasonable value that corresponds to the threshold for identifying substantial research findings. Moreover, unlike the conventional approaches that test a simple null hypothesis of zero effect, the test statistics of minimum-effect are employed to conclude whether the observed effect sizes would be likely to occur for a range of population values in a composite null hypothesis. Therefore, the notion of least favorable configurations of treatment means is vital to determine the critical values for conducting tests of minimal effect. Such information is generally not available to applied researchers and it is impossible to implement the test procedures without an extensive set of tables or an efficient software package.

More general discussions of the fundamental concept and rationale of the F tests of minimal effect can be found in Murphy and Myors (1999) and Steiger (2004). Alternatively, Bau et al. (1993) and Chen et al. (2011) described the technical development and property for the studentized range test of minimum-effect hypotheses. Although these results justify the execution of the F test and the studentized range test of minimum effect, their formulations are markedly different and demand varying computational efforts. No investigation has systematically compared their distinct characteristics in terms of theoretical principles, power performance, and sample size requirement. Thus it is prudent to examine their unique features and fundamental discrepancies in order to better understand the selection of an appropriate procedure. Equivalence testing is recommended as a better alternative to the traditional difference-based methods for demonstrating the comparability of two or more treatment effects. The related features of both the ANOVA F and the studentized range tests for evaluating the comparability of several standardized effects can be found in Wellek (2010, Chapter 7), Giani and Finner (1991), Cribbie et al. (2010), Chen et al. (2009), and the references therein.

This article seeks to contribute to the literature on detecting a minimal important difference among standardized mean effects in three ways. First, the fundamental distinctions of the standard F test and the studentized range test are reviewed here to provide a clear and concise exposition of their inherent formulations and properties. This report adds to the general understanding of the utilities of effect size measures and enhances the practical importance of the tests of minimal effect hypotheses. Second, comprehensive empirical investigations are conducted to demonstrate the power performance and sample size requirements between the standard F test and the studentized range test under a wide variety of mean structures. The assessments discern not only which method is most suitable under what circumstances but also the actual differences between the contending test procedures. Third, to facilitate the application of the presented procedures, the corresponding computer codes are provided to compute the critical values, observed significance levels, attained power levels, and required sample sizes. Note that the computations of these procedures involve specialized programs not currently available in prevailing statistical packages. This article and the [supplemental files](#) present a unified set of algorithms for design planning and data analysis of the two tests of minimum-effect.

Effect Measures and Test Procedures

Consider the one-way fixed-effects ANOVA model

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad (1)$$

where Y_{ij} is the value of the response variable in the j th trial for the i th factor level, μ_i are treatment means, ε_{ij} are independent $N(0, \sigma^2)$ errors with $i = 1, \dots, G (\geq 2)$ and $j = 1, \dots, N$. To characterize the degree of departure from no treatment effect, two distinctive measures were proposed in Cohen (1969, 1988). The first index is the standard deviation of the standardized means

$$f = \frac{\sigma\mu}{\sigma}, \tag{2}$$

where $\sigma = (\sigma^2)^{1/2}$, $\sigma_\mu = (\sigma_\mu^2)^{1/2}$, $\sigma_\mu^2 = \sum_{i=1}^G (\mu_i - \bar{\mu})^2 / G$ is

the average dispersion between the treatment means, and $\bar{\mu}$

$$= \sum_{i=1}^G \mu_i / G \text{ is the mean of the treatment effects.}$$

It is useful to note that $\sum_{i=1}^G \sum_{l=1}^G (\mu_i - \mu_l)^2 = 2G \sum_{i=1}^G (\mu_i - \bar{\mu})^2$ or $\sigma_\mu^2 = \sum_{i=1}^G \sum_{l=1}^G (\mu_i - \mu_l)^2 / (2G^2)$. Hence, $f = \left\{ \sum_{i=1}^G \sum_{l=1}^G \delta_{il}^2 / (2G^2) \right\}^{1/2}$ is also a function of the sum of all $G(G - 1)$ square standardized mean differences $\delta_{il} = (\mu_i - \mu_l) / \sigma$, i and $l = 1, \dots, G$. The second index is based on the range of the standardized means

$$\delta_R = \frac{\mu_{max} - \mu_{min}}{\sigma}, \tag{3}$$

where μ_{max} and μ_{min} are the maximum and the minimum of the G treatment means, respectively. When $G = 2$, it can be shown that $f = \delta_R / 2 = |\delta| / 2$, where $\delta = (\mu_1 - \mu_2) / \sigma$ is the well-known standardized mean difference. In general, the two effect sizes f and δ_R have no direct functional relationship except for some special cases of the treatment means. Accordingly, the corresponding inferential procedures are also substantially different.

The common F^* is the most widely used test statistic for the null hypothesis that all treatment means are equal:

$$F^* = \frac{SSR / (G - 1)}{SSE / (N_T - G)}, \tag{4}$$

where $SSR = N \sum_{i=1}^G (\bar{Y}_i - \bar{Y})^2$ is the treatment sum of squares,

$SSE = \sum_{i=1}^G \sum_{j=1}^N (Y_{ij} - \bar{Y}_i)^2$ is the error sum of squares,

$N_T = GN$, $\bar{Y}_i = \sum_{j=1}^N Y_{ij} / N$, and $\bar{Y} = \sum_{i=1}^G \sum_{j=1}^N Y_{ij} / N_T$. Here,

the focus is on the minimal effect test of

$$H_0 : f \leq f_0 \text{ versus } H_1 : f > f_0 \tag{5}$$

where $f_0 (> 0)$ is a specified value and denotes a minimal

significant difference among the treatment means. It follows from the model formulation in Eq. 1 that

$$F^* \sim F(G - 1, N_T - G, \Lambda), \tag{6}$$

where $F(G - 1, N_T - G, \Lambda)$ is the noncentral F distribution with $(G - 1)$ and $(N_T - G)$ degrees of freedom, and noncentrality parameter $\Lambda = N_T f^2$ and $f^2 = \sigma_\mu^2 / \sigma^2$ is the signal to noise ratio (Fleishman 1980). Another useful effect size measure in ANOVA is the strength of association η^2 that reflects the proportion of total variance that is attributable to treatment effects. Essentially, the index η^2 is a function of f^2 and can be expressed as $\eta^2 = f^2 / (1 + f^2)$. In addition, the root-mean-square standardized effect Ψ described in Steiger (2004) is a scaled counterpart of f : $\Psi = \left\{ \left[\sum_{i=1}^G (\mu_i - \bar{\mu})^2 / \sigma^2 \right] / (G - 1) \right\}^{1/2}$

$= \{G / (G - 1)\}^{1/2} f$. Under the null hypothesis $H_0: f \leq f_0$, the statistic F^* has the distribution

$$F^* \sim F(G - 1, N_T - G, \Lambda_0), \tag{7}$$

where $\Lambda_0 = N_T f_0^2$. It follows from the monotone property of a noncentral F distribution that $P\{F(G - 1, N_T - G, \Lambda_0) > c\} > P\{F(G - 1, N_T - G, \Lambda) > c\}$ for $\Lambda_0 > \Lambda \geq 0$ and a positive constant c (Ghosh 1973). Thus, the least favorable configuration of $\mu = (\mu_1, \dots, \mu_G)$ is $\mu_0 = (\mu_{10}, \dots, \mu_{G0})$ with the standard deviation of the standardized means $f = f_0$. Hence, H_0 is rejected at the significance level α if $F^* > F_\alpha(G - 1, N_T - G, \Lambda_0)$, where $F_\alpha(G - 1, N_T - G, \Lambda_0)$ is the upper $(100 - \alpha)$ th percentile of the noncentral F distribution $F(G - 1, N_T - G, \Lambda_0)$. Then the corresponding power function of the F^* test is of the form

$$\pi_{F^*}(\Lambda) = P\{F(G - 1, N_T - G, \Lambda) > F_\alpha(G - 1, N_T - G, \Lambda_0)\}. \tag{8}$$

Alternatively, a minimum effect of the standardized means may be assessed with the studentized range test in terms of the range of the standardized means:

$$H_0 : \delta_R \leq \delta_{R0} \text{ versus } H_1 : \delta_R > \delta_{R0} \tag{9}$$

where $\delta_{R0} (> 0)$ is a designated constant and indicates a minimal significant difference among the treatment means. The studentized range is defined as

$$Q^* = \frac{N^{1/2} (\bar{Y}_{max} - \bar{Y}_{min})}{S}, \tag{10}$$

where \bar{Y}_{max} and \bar{Y}_{min} are the maximum and minimum of the G sample means, respectively, $S = (S^2)^{1/2}$, and $S^2 = SSE / (N_T - G)$ is the sample variance. The distribution of the studentized range statistic Q^* depends on all pairwise mean differences $\mu_i - \mu_l$, not just a function of the maximum mean difference $\mu_{max} - \mu_{min}$. Following the model formulation in Eq. 1, the distribution of Q^* does not have a closed-form expression and

a convenient unified notation. For ease of illustration, it is denoted here by

$$Q^* \sim Q(G, N, \boldsymbol{\tau}), \tag{11}$$

where $\boldsymbol{\tau} = (\mu_1, \dots, \mu_G)/\sigma$. However, the cumulative distribution function $\Theta(q)$ of Q^* can be expressed as follows:

$$\Theta(q) = P\{Q^* \leq q\} = E_K \left\{ \sum_{i=1}^G E_{Z_i} \left[\prod_{\substack{l=1 \\ l \neq i}}^G \left(\Phi\{Z_i + N^{1/2}\delta_{il}\} - \Phi\{Z_i + N^{1/2}\delta_{il} - q [K/(N_T - G)]^{1/2}\} \right) \right] \right\}, \tag{12}$$

where $K \sim \chi^2(N_T - G)$ is a chi-square random variable with degrees of freedom $N_T - G$, $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution, $Z_i \sim N(0, 1)$ are independent standard normal random variables, and $E_K\{\cdot\}$ and $E_{Z_i}\{\cdot\}$ are taken with respect to the distribution of K and Z_i , respectively. Under the assumption that all treatment means are equal, the property of Q^* has been discussed in many statistics textbooks, and the cumulative probability and quantile can be readily computed with popular software systems. However, the general distribution of Q^* is relatively more complex and a special purpose algorithm is required to perform the associated calculations.

Unlike the prescribed F^* test, the hypothesis testing of $H_0: \delta_R \leq \delta_{R0}$ with the Q^* statistic is more involved. As noted earlier, the distribution of Q^* is not a simple function of δ_R alone and thus, the critical value cannot be determined for arbitrary treatment means satisfying $\delta_R = \delta_{R0}$. It follows from Bau et al. (1993) that the least favorable configuration of $\boldsymbol{\mu} = (\mu_1, \dots, \mu_G)$ with $\delta_R = \delta_{R0}$ is $\boldsymbol{\mu}_0 = (\mu_{10}, \dots, \mu_{G0})$ where $\mu_{i0} = -\sigma\delta_{R0}/2$, for $i = 1, \dots, [G/2]$, and $\mu_{i0} = \sigma\delta_{R0}/2$, for $i = [G/2] + 1, \dots, G$, and $[G/2]$ is the greatest integer being less than or equal to $G/2$. In other words, $P\{Q(G, N, \boldsymbol{\tau}_0) > c\} > P\{Q(G, N, \boldsymbol{\tau}) > c\}$ for $\boldsymbol{\tau}_0 = \boldsymbol{\mu}_0/\sigma$, any $\boldsymbol{\tau}$ has the range of the standardized means δ_{R0} , and a positive constant c . Hence, the null hypothesis $H_0: \delta_R \leq \delta_{R0}$ is rejected at the significance level α if $Q^* > Q_{\alpha}(G, N, \boldsymbol{\tau}_0)$, where $Q_{\alpha}(G, N, \boldsymbol{\tau}_0)$ is the upper $(100 \cdot \alpha)$ th percentile of the distribution of Q^* when the treatment means have the least favorable configuration $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ and $\boldsymbol{\tau}_0 = \boldsymbol{\mu}_0/\sigma$. Also, the power function of the Q^* test is obtained as

$$\pi_{Q^*}(\boldsymbol{\tau}) = P\{Q(G, N, \boldsymbol{\tau}) > Q_{\alpha}(G, N, \boldsymbol{\tau}_0)\} \tag{13}$$

for all sets of standardized treatment means $\boldsymbol{\tau}$ satisfying $\delta_R > \delta_{R0}$.

To illustrate the application of the suggested procedures, the numerical demonstration of Olejnik and Algina (2000) is reexamined here for the detection of the minimal effect in terms of f and δ_R . Their focus was on the computations and

interpretations of standardized linear contrast of mean differences and the corresponding proportion of variance effect size. The data was obtained from a randomized groups pretest–posttest study reported in Baumann et al. (1992) and they were interested in comparing the relative effectiveness of three types of interventions (Talk-Aloud, Directed Reading-Thinking Activity, and Directed Reading Activity) designed to affect reading comprehension skills of fourth-grade students. Using the 66 outcomes of the posttest error detection task with the number of groups $G = 3$ and group sample size $N = 22$, the ANOVA F and the studentized range statistics are $F^* = 5.3119$ and $Q^* = 4.5443$, respectively. At the significance level $\alpha = 0.05$ and the magnitude $f_0 = 0.10$, the critical value of the minimal effect F test is $F_{0.05} = 4.1011$ and the corresponding p -value is 0.0209. For the studentized range test with a specific effect size $\delta_{R0} = 0.2$ and the least favorable configuration $\boldsymbol{\tau}_0 = \{-0.1, 0.1, 0.1\}$, the computed critical value and p -value are 3.8307 and 0.0154, respectively. As suggested in Cohen (1969, 1988), the particular minimum effect sizes $f_0 = 0.10$ and $\delta_{R0} = 0.2$ represent a small magnitude of the standard deviation of the standardized means f and the difference of the standardized means δ , respectively. Accordingly, the two minimum effect tests of $H_0: f \leq 0.10$ and $H_0: \delta_R \leq 0.2$ are rejected and there exist some essential differences between the three types of intervention programs. The SAS/IML (SAS Institute 2014) and R (R Development Core Team 2014) programs employed to perform the exact critical value and p -value calculations of both tests are presented in supplementary files.

Power Calculations and Sample Size Determinations

The theoretical implications and computational feasibility are important aspects of a test procedure for making statistical inferences. In practice, a research study requires adequate

statistical power and sufficient sample size to detect scientifically credible effects. It is sensible that the corresponding power calculations and sample size determinations must also be considered for a viable procedure to extend the applicability in planning research designs. Consequently, the presented power functions for the F^* and Q^* tests can be useful in conducting power analysis and sample size computation for detecting the differences among standardized mean effects. The following numerical assessment exemplifies a typical research scenario most frequently encountered in the planning stage of a study.

Due to the prospective nature of advance research planning, the general guidelines suggest that typical sources like published findings or expert opinions can offer plausible and reasonable planning values for the model characteristics of mean effects and variance components. To explicate the essential features, the abovementioned posttest error detection task data of the three interventions is employed to provide planning values of the model parameters and effect sizes for a future intervention study. Specifically, the mean effects and variance component are designated as $\mu = \{7.77, 9.77, 6.68\}$ and $\sigma^2 = 10.17$, respectively. With the additional specifications of $G = 3$, $N = 22$, $f_0 = 0.1$, and $\delta_{R0} = 0.2$, the resulting powers for the two tests are $\pi_{F^*} = 0.7109$ and $\pi_{Q^*} = 0.7216$, respectively, when the significance level is $\alpha = 0.05$. The achieved power is slightly less than the fairly common and somehow minimal level of 0.80. Therefore, the power calculation suggests that the designated configurations may not warrant a decent chance of detecting the minimal difference between intervention effects.

With the power formulas and associated algorithms, the sample size N needed to attain the nominal power $(1 - \beta)$ can be found by a simple iterative search for the chosen significance level α and parameter values. Further computations show that a target power of 0.80 necessitates a balanced group size of 28 and 27 for the F^* and Q^* tests, respectively. Due to the underlying metric of integer sample sizes, the corresponding actual powers are 0.8127 and 0.8088, and they are marginally greater than the nominal power level 0.80. Hence, the required sample size is nearly 25% more than is needed for the original design. These configurations are incorporated in the user specifications of the SAS/IML and R programs presented in the [supplemental files](#). Ultimately, users can easily identify the statements containing the key values in the computer code and then modify the program to accommodate their own model specifications.

Although the power performance and sample size requirement between the two procedures are almost identical, the particular results are confined to the specified minimal effect sizes $f_0 = 0.1$ and $\delta_{R0} = 0.2$. It is of theoretical importance and practical interest to evaluate the relative performance between the F^* and Q^* test procedures across a variety of comparable design configurations. However, the distinct formulation of

statistical hypotheses and the distributional complexity of test statistics do not permit a complete analytic examination and technical justification. Instead, a comprehensive empirical appraisal is conducted to assess and compare the power and sample size behavior of the two methods.

Numerical Study

In order to provide a systematic investigation for the power behavior of the F^* and Q^* test procedures, the effect sizes f_0 and f_1 under the null and alternative hypotheses are chosen a priori to facilitate meaningful comparisons. Although δ_R is not a function of f , there exists an intrinsic property for the lower and upper bounds of the range of the standardized means δ_R when the standard deviation of the standardized means f is fixed (Pearson and Hartley 1951). It can be easily shown that, for all sets of standardized treatment means τ with a fixed value f , the maximum range is $\delta_{Rmax} = (2G)^{1/2}f$ for $\tau = \tau_{max}$ where $\tau_{max} = \{-(G/2)^{1/2}f, 0, \dots, 0, (G/2)^{1/2}f\}$. In contrast, when G is even, the minimum range is $\delta_{Rmin} = 2f$ for $\tau = \tau_{Emin}$ where $\tau_{Emin} = \{\tau_1, \dots, \tau_G\}$, $\tau_i = -f$, $i = 1, \dots, G/2$; and $\tau_i = f$, $i = G/2 + 1, \dots, G$. On the other hand, when G is odd, the minimum range is $\delta_{Rmin} = 2Gf/(G^2 - 1)^{1/2}$ for $\tau = \tau_{Omin}$ where $\tau_{Omin} = \{\tau_1, \dots, \tau_G\}$, $\tau_i = -Gf/(G^2 - 1)^{1/2}$, $i = 1, \dots, (G - 1)/2$; and $\tau_i = Gf/(G^2 - 1)^{1/2}$, $i = (G - 1)/2 + 1, \dots, G$. Although similar results were given in David et al. (1972), their formulation of τ_{Omin} differs from the presented expression of the standardized mean structure with a simple location shift. Note that the definition of τ_{Omin} conforms to the least favorable configuration τ_0 for determining the critical value of the Q^* test of minimal effect null hypothesis. Essentially, such standardized mean configurations need to be identified and incorporated in the power evaluation of the Q^* test.

To delineate the underlying features of the two test procedures crossing different model characteristics, the numerical assessments are performed for $(G, N) = (3, 16)$, $(4, 12)$, and $(6, 8)$. Accordingly, the total sample size is fixed as $N_T = 48$ for ease of comparison. Throughout this empirical study, the significance level and the minimum effect size are set as $\alpha = 0.05$ and $f_0 = 0.1$, respectively. Thus, the associated critical values are $F_{0.05} = 3.9308$, 3.2494, and 2.6663 for $G = 3, 4$, and 6, respectively. The corresponding least favorable standardized mean configurations of the Q^* test are $\tau_{min} = \{-0.1061, 0.1061, 0.1061\}$, $\{-0.1, -0.1, 0.1, 0.1\}$, and $\{-0.1, -0.1, -0.1, 0.1, 0.1, 0.1\}$. Moreover, the actual minimum effect sizes and critical values of the Q^* test are $\delta_{R0} = 0.2121$, 0.2, and 0.2, and $Q_{0.05} = 3.7964$, 4.0524, and 4.4125 for $G = 3, 4$, and 6, respectively. Under the designated null hypothesis $H_0: f \leq 0.10$ with the significance level $\alpha = 0.05$, the power π_{F^*} of the F^* test of the alternative hypothesis $H_1: f > 0.10$ are computed for the alternative effect size $f = f_1$ with $f_1 = 0.1$ to 0.9

Table 1 Computed power of the F^* and Q^* tests for $f_0 = 0.1$ and $\alpha = 0.05$

$G = 3$	f_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	π_{F^*}	0.0500	0.1382	0.3180	0.5639	0.7893	0.9267	0.9823	0.9971	0.9997
	δ_{Rmin}	0.2121	0.4243	0.6364	0.8485	1.0607	1.2728	1.4849	1.6971	1.9092
	π_{Q^*min}	0.0500	0.1369	0.3125	0.5525	0.7762	0.9178	0.9786	0.9961	0.9995
	δ_{Rmax}	0.2449	0.4899	0.7348	0.9798	1.2247	1.4697	1.7146	1.9596	2.2045
	π_{Q^*max}	0.0500	0.1377	0.3177	0.5653	0.7921	0.9290	0.9833	0.9973	0.9997
$G = 4$	f_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	π_{F^*}	0.0500	0.1235	0.2786	0.5061	0.7368	0.8963	0.9709	0.9943	0.9992
	δ_{Rmin}	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8
	π_{Q^*min}	0.0500	0.1177	0.2540	0.4542	0.6714	0.8440	0.9436	0.9849	0.9970
	δ_{Rmax}	0.2828	0.5657	0.8485	1.1314	1.4142	1.6971	1.9799	2.2627	2.5456
	π_{Q^*max}	0.0504	0.1239	0.2812	0.5139	0.7477	0.9049	0.9749	0.9955	0.9995
$G = 6$	f_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	π_{F^*}	0.0500	0.1059	0.2271	0.4210	0.6464	0.8330	0.9414	0.9851	0.9973
	δ_{Rmin}	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8
	π_{Q^*min}	0.0500	0.0991	0.1977	0.3509	0.5399	0.7237	0.8635	0.9459	0.9831
	δ_{Rmax}	0.3464	0.6928	1.0392	1.3856	1.7321	2.0785	2.4249	2.7713	3.1177
	π_{Q^*max}	0.0505	0.1068	0.2340	0.4412	0.6773	0.8610	0.9572	0.9908	0.9987

with an increment of 0.1. For each f_1 , the power π_{Q^*} of the Q^* test of $H_0: \delta_R \leq \delta_{R0}$ versus $H_1: \delta_R > \delta_{R0}$ are calculated for the two standardized mean extremes $\{\tau_{max}, \tau_{Qmin}\}$ and $\{\tau_{max}, \tau_{Emin}\}$ when G is odd and even, respectively. For ease of illustration, the attained powers π_{Q^*} of the Q^* test evaluated at the two extremes are denoted by π_{Q^*min} and π_{Q^*max} , respectively. Accordingly, the designated effect sizes ($f_1, \delta_{Rmin}, \delta_{Rmax}$) and resulting powers ($\pi_{F^*}, \pi_{Q^*min}, \pi_{Q^*max}$) are summarized in Table 1 for the three model settings (G, N) = (3, 16), (4, 12), and (6, 8).

It can be seen from the extensive numerical results that the achieved powers have a clear pattern $\pi_{Q^*max} > \pi_{F^*} > \pi_{Q^*min}$ for all $f_1 > 0.1$. The only exceptions occurred with the two cases of $f_1 = 0.2$ and 0.3 when $G = 3$. Specifically, the order between the three power levels is $\pi_{F^*} > \pi_{Q^*max} > \pi_{Q^*min}$. On the other hand, the achieved powers of π_{F^*}, π_{Q^*min} , and π_{Q^*max} are decreasing in the number of groups G when all other factors remain constant including the total sample size. Thus, it requires a larger total sample size to achieve the designated power for the two tests when all other configurations

Table 2 Computed sample size of the F^* and Q^* tests for $f_0 = 0.1$, $\alpha = 0.05$, and $1 - \beta = 0.90$

$G = 3$	f_1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	N_{F^*}	297	78	37	22	15	11	9	7
	δ_{Rmin}	0.4243	0.6364	0.8485	1.0607	1.2728	1.4849	1.6971	1.9092
	N_{Q^*min}	326	83	39	23	16	12	9	7
	δ_{Rmax}	0.4899	0.7348	0.9798	1.2247	1.4697	1.7146	1.9596	2.2045
	N_{Q^*max}	268	76	36	22	15	11	9	7
$G = 4$	f_1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	N_{F^*}	229	62	30	18	13	9	8	6
	δ_{Rmin}	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8
	N_{Q^*min}	299	77	36	21	14	11	9	7
	δ_{Rmax}	0.5657	0.8485	1.1314	1.4142	1.6971	1.9799	2.2627	2.5456
	N_{Q^*max}	180	58	29	18	12	9	7	6
$G = 6$	f_1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	N_{F^*}	160	45	23	14	10	8	6	5
	δ_{Rmin}	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8
	N_{Q^*min}	247	64	30	18	12	9	7	6
	δ_{Rmax}	0.6928	1.0392	1.3856	1.7321	2.0785	2.4249	2.7713	3.1177
	N_{Q^*max}	115	40	21	13	9	7	6	5

are fixed. Instead of fixing the total sample size, the sample sizes required to achieve the nominal power 0.90 are also calculated for the selected group structures and effect settings. The computed sample sizes are presented in Table 2 and they are denoted by N_{F^*} , N_{Q^*min} , and N_{Q^*max} for the power functions π_{F^*} , π_{Q^*min} , and π_{Q^*max} , respectively. As expected, the necessary sample sizes decrease with increasing difference between f_1 and f_0 for all three scenarios. More importantly, the estimated sample sizes demonstrate a general phenomenon that $N_{Q^*min} \geq N_{F^*} \geq N_{Q^*max}$. Also, the discrepancy between the three sample sizes (N_{Q^*min} , N_{F^*} , N_{Q^*max}) substantially reduced as f_1 becomes larger than f_0 . These findings extend those reported in David et al. (1972) for the traditional tests of zero effect. Hence, although the F test is a well-recognized method, the studentized range test may outperform the F test when the standardized mean structures are close to the extreme pattern of τ_{max} . Therefore it is advisable that researchers need to conduct appropriate power analysis and employ suitable techniques based on the best knowledge of the vital model characteristics, especially the plausible and reasonable planning values for the mean effect configurations.

Conclusion

The two effect size measures of the standard deviation of the standardized means and the range of the standardized means among several treatments are direct extensions of the standardized mean difference between two groups. The existing literature on statistical hypothesis testing for comparisons among treatment means often assumes a null hypothesis of no difference. Consequently, the associated procedures and algorithms for data analysis and design planning are confined to the rejection of zero effect instead of the detection of treatment differences that are larger enough to have a practical impact. This study explicates and evaluates the features of the ANOVA F and the studentized range procedures for testing the null hypothesis that the treatment effects have a minimal important difference based on the standard deviation of the standardized means and the range of the standardized means, respectively. The notion of minimal important difference is vital to determine the critical values and power behavior of the ANOVA F and the studentized range procedures. The general guidance of Cohen (1969, 1988) suggests that the small, medium, and large effects in terms of f and δ could be defined as $f = 0.10$, 0.25, and 0.40, and $\delta = 0.2$, 0.5, and 0.8, respectively. It is prudent to emphasize that the particular research question within a concentrated research field determines whether a specific effect size is practically or scientifically important. But in general, it should be based on substantive or empirical grounds, such as results from a prior effect size appraisal. Results in the current report showed that the relative power

performance of the two approaches depends on the variability patterns of the mean effects. The studentized range method can be more or less powerful than the ANOVA F test if the mean structure has a maximum range or minimum range when the standard deviation of the standardized means remains constant. Numerical examples and computer algorithms were presented to illustrate the uses of the suggested methods for data analysis and design planning. The formulation and assessment described in this article gave a unified and rigorous explication of the two major procedures for detecting a minimal important difference among standardized means. The findings update and extend previous studies on the analysis of treatment effects among several groups.

Compliance with Ethical Standards

Funding The author has no support or funding to report.

Conflict of Interest The author declares that he has no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

- Alhija, F. N. A., & Levy, A. (2009). Effect size reporting practices in published articles. *Educational and Psychological Measurement, 69*(2), 245–265.
- Bau, J. J., Chen, H. J., & Xiong, M. (1993). Percentage points of the studentized range test for dispersion of normal means. *Journal of Statistical Computation and Simulation, 44*(3–4), 149–163.
- Baumann, J. F., Seifert-Kessell, N., & Jones, L. A. (1992). Effect of think-aloud instruction on elementary students' comprehension monitoring abilities. *Journal of Reading Behavior, 24*(2), 143–172.
- Breaugh, J. A. (2003). Effect size estimation: factors to consider and mistakes to avoid. *Journal of Management, 29*(1), 79–97.
- Chen, H. J., Wen, M. J., & Wang, M. (2009). On testing the bioequivalence of several treatments using the measure of distance. *Statistics, 43*(5), 513–530.
- Chen, H. J., Wen, M. J., & Chuang, C. J. (2011). On testing the equivalence of treatments using the measure of range. *Computational Statistics and Data Analysis, 55*(1), 603–614.
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Erlbaum.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997–1003.
- Cribbie, R. A., Arpin-Cribbie, C. A., & Gruman, J. A. (2010). Tests of equivalence for one-way independent groups designs. *Journal of Experimental Education, 78*(1), 1–13.
- David, H. A., Lachenbruch, P. A., & Brandis, H. P. (1972). The power function of range and studentized range tests in normal samples. *Biometrika, 59*(1), 161–168.
- Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology, 34*(9), 917–928.

- Ferguson, C. J. (2009). An effect size primer: a guide for clinicians and researchers. *Professional Psychology: Research and Practices*, 40(5), 532–538.
- Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: issues and problems in interpretation. *Journal of Consumer Research*, 23(2), 89–105.
- Fleishman, A. I. (1980). Confidence intervals for correlation ratios. *Educational and Psychological Measurement*, 40(3), 659–670.
- Ghosh, B. K. (1973). Some monotonicity theorems for χ^2 , F and t distributions with applications. *Journal of the Royal Statistical Society, Series B*, 35, 480–492.
- Giani, G., & Finner, H. (1991). Some general results on least favorable parameter configurations with special reference to equivalence testing and the range statistic. *Journal of Statistical Planning and Inference*, 28(1), 33–47.
- Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: univariate and multivariate applications* (2nd ed.). New York: Routledge.
- Hayter, A. J., & Hurn, M. (1992). Power comparisons between the F -test, the studentised range test, and an optimal test of the equality of several normal means. *Journal of Statistical Computation and Simulation*, 42(3–4), 173–185.
- Hayter, A. J., & Liu, W. (1992). Some minimax test procedures for comparing several normal means. In F. M. Hoppe (Ed.), *Multiple comparisons, selection, and applications in biometry, a festschrift in honor of Charles W. Dunnett* (pp. 137–148). New York: Marcel Dekker.
- Huberty, C. (2002). A history of effect size indices. *Educational and Psychological Measurement*, 62(2), 227–240.
- Kirk, R. (1996). Practical significance: a concept whose time has come. *Educational and Psychological Measurement*, 56(5), 746–759.
- Kline, R. B. (2004). *Beyond significance testing: reforming data analysis methods in behavioral research*. Washington: American Psychological Association.
- Murphy, K. R. (1990). If the null hypothesis is impossible, why test it? *American Psychologist*, 45(3), 403–404.
- Murphy, K. R., & Myers, B. (1999). Testing the hypothesis that treatments have negligible effects: minimum-effect tests in the general linear model. *Journal of Applied Psychology*, 84(2), 234–248.
- Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: applications, interpretations, and limitations. *Contemporary Educational Psychology*, 25(3), 241–286.
- Pearson, E. S., & Hartley, H. O. (1951). Charts of the power function for analysis of variance tests, derived from the non-central F -distribution. *Biometrika*, 38(1–2), 112–130.
- R Development Core Team (2014). R: A language and environment for statistical computing [Computer software and manual]. Retrieved from <http://www.r-project.org>.
- Richardson, J. T. E. (1996). Measures of effect size. *Behavior Research Methods, Instruments, & Computers*, 28(1), 12–22.
- Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect size in behavioral research: a correlational approach*. New York: Cambridge University Press.
- Rosnow, R. L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology*, 57(3), 221–237.
- SAS Institute. (2014). *SAS/IML User's Guide, Version 9.3*. Cary: SAS Institute Inc..
- Serlin, R. A., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: methodological issues* (pp. 199–228). Hillsdale: Erlbaum.
- Skidmore, S. T., & Thompson, B. (2010). Statistical techniques used in published articles: a historical review of reviews. *Educational and Psychological Measurement*, 70(5), 777–795.
- Steiger, J. H. (2004). Beyond the F test: effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, 9(2), 164–182.
- Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*, 51(4), 473–481.
- Warne, R. T., Lazo, M., Ramos, T., & Ritter, N. (2012). Statistical methods used in gifted education journals, 2006–2010. *Gifted Child Quarterly*, 56(3), 134–149.
- Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). New York: CRC Press.