British Journal of Mathematical and Statistical Psychology (2014), 67, 72–93
© 2013 The British Psychological Society

The British Psychological Society

www.wileyonlinelibrary.com

Sample size determinations for Welch's test in one-way heteroscedastic ANOVA

Show-Li Jan[1] and Gwowen Shieh[2]*

[1]Chung Yuan Christian University, Taiwan, Republic of China
[2]National Chiao Tung University, Taiwan, Republic of China

For one-way fixed effects ANOVA, it is well known that the conventional *F* test of the equality of means is not robust to unequal variances, and numerous methods have been proposed for dealing with heteroscedasticity. On the basis of extensive empirical evidence of Type I error control and power performance, Welch's procedure is frequently recommended as the major alternative to the ANOVA *F* test under variance heterogeneity. To enhance its practical usefulness, this paper considers an important aspect of Welch's method in determining the sample size necessary to achieve a given power. Simulation studies are conducted to compare two approximate power functions of Welch's test for their accuracy in sample size calculations over a wide variety of model configurations with heteroscedastic structures. The numerical investigations show that Levy's (1978a) approach is clearly more accurate than the formula of Luh and Guo (2011) for the range of model specifications considered here. Accordingly, computer programs are provided to implement the technique recommended by Levy for power calculation and sample size determination within the context of the one-way heteroscedastic ANOVA model.

1. Introduction

The one-way analysis of variance (ANOVA) *F* test is a procedure widely used for testing the equality of means of independent normal distributions with homogeneous variances. The corresponding implications, from the basic diagnostics of underlying assumptions to the required power calculations and sample size determinations, have been extensively addressed in the literature; see, for example, Howell (2010), Kirk (1995), Kutner, Nachtsheim, Neter and Li (2005) and Scheffé (1959). However, the violation of the independence, normality, and homogeneity of variance assumptions either separately or in conjunction with one another has been the target of criticism in applications of ANOVA (Coombs, Algina & Oltman, 1996; Glass, Peckham & Sanders, 1972; Harwell, Rubinstein, Hayes & Olds, 1992; Keselman *et al.*, 1998). Specifically, the *F* test is not robust to all

*Correspondence should be addressed to Gwowen Shieh, Department of Management Science, National Chiao Tung University, Hsinchu, Taiwan 30010, Republic of China (e-mail: gwshieh@mail.nctu.edu.tw).

DOI:10.1111/bmsp.12006

degrees of unequal variances (Brown & Forsythe, 1974; Clinch & Keselman, 1982; De Beuckelaer, 1996; Kohr & Games, 1974; Levy, 1978b; Wilcox, Charlin & Thompson, 1986), and the actual significance level and power can be distorted even when sample sizes are equal (Krutchkoff, 1986; Rogan & Keselman, 1977). Accordingly, various parametric and non-parametric alternatives to the traditional *F* test have been proposed to counter the effects of heteroscedasticity (Lix, Keselman & Keselman, 1996).

Given the extensive Monte Carlo simulation studies conducted in this area, three important aspects of these numerical evidences should be pointed out. First, the non-parametric procedures are also substantially affected by heterogeneous variances and are generally inferior to the parametric approaches (Keselman, Rogan & Feir-Walsh, 1977; Tomarken & Serlin, 1986; Zimmerman, 2000). Second, the parametric tests of Alexander and Govern (1994), Brown and Forsythe (1974), James (1951) and Welch (1951) have been shown to provide accurate control of Type I error rate and competitive power performance (Schneider & Penfield, 1997). However, there appears to be a lack of consensus in the literature on which method is most appropriate. Essentially, there is no one uniformly best alternative to the *F* test under heterogeneity of variance (Dijkstra & Werter, 1981; Grissom, 2000). Third, despite the fact that no approach is ideal, it is still of practical importance to have a reliable and simple test procedure that is sufficiently robust to heteroscedasticity when distributions are normal. On the basis of the comprehensive appraisals by Brown and Forsythe (1974), De Beuckelaer (1996), Grissom (2000), Harwell *et al.* (1992), Levy (1978b), Tomarken and Serlin (1986) and Wilcox *et al.* (1986), the approximation of Welch (1951) is the most widely recommended technique to correct for variance heterogeneity. In short, it has distinct advantages over other competing approaches in its overall performance, computational ease, and general availability in statistical computer packages.

Yet another problem with the common methods for analysing the data from one-way independent groups designs occurs when the distribution of each population is non-normal in form. See Cribbie, Fiksenbaum, Keselman and Wilcox (2012), Lix and Keselman (1998), Wilcox (2003) and Wilcox and Keselman (2003) for modern robust methods and updated strategies when the standard assumptions of normality and homoscedasticity are violated. In particular, the Welch test with robust estimators of trimmed means and Winsorized variances has been shown to provide excellent Type I error control and power performance when data are non-normal and heterogeneous. However, we will restrict our attention to the appropriate procedure for testing the equality of means of independent normal distributions with possibly unequal error variances here.

It is conceivable that a test procedure with robust Type I error control and excellent power performance is not sufficient for the purposes of research design and statistical inference. The corresponding power analysis and sample size computation must also be considered before it can be adopted as a general methodology in practice. Theoretically, the non-null distribution of a test procedure is required in order to evaluate the intrinsic issues of power analysis and sample size assessment. But to our best knowledge, no power function or non-null distribution has been proposed for the prescribed tests of Alexander and Govern (1994), Brown and Forsythe (1974) and James (1951). On the other hand, several approximations have been described for the non-null distribution of Welch's (1951) test in Levy (1978a), Luh and Guo (2011) and Kulinskaya, Staudte and Gao (2003). Although these results permit power and sample size considerations for the well-known Welch (1951) method, no research to date has compared their distinct characteristics in terms of theoretical principles, computational requirements and empirical performance. But in fact their formulations are markedly different and demand varying computational efforts. Thus it is prudent to examine their unique feature and fundamental discrepancy in order to better understand the selection of an appropriate approach to power analysis and sample size determination in one-factor ANOVA studies.

Instead of a non-central F distribution, Kulinskaya *et al.* (2003) presented a chi-square-based power approximation to the non-null distribution of Welch's test. They showed that the shifted and rescaled chi-square approximation is more accurate than the standard chi-square transformation. However, there are two obvious disadvantages of their approximate power function. First, the chi-square-based formulation does not conform to the entrenched F test of homoscedastic ANOVA or Welch's test of heteroscedastic ANOVA. Second, the complexity of their proposed expression is overwhelming. It is worthwhile to consider a more transparent procedure with fewer computational and theoretical hurdles. Thus the approach of Kulinskaya *et al.* (2003) will not be considered further in this paper.

Recently, Luh and Guo (2011) suggested a non-central F distribution to approximate the non-null distribution of Welch's test. The non-centrality of their non-central F distribution is a direct modification of the non-centrality of the usual F test's exact non-null F distribution under balanced design and homogeneity of variance. In particular, the non-centrality derived involves a simple average of the variance and sample size ratios of each group. The adapted formula at first sight provides a convenient approximation and is computationally simple. Notably, Luh and Guo (2011) concluded that their technique is suitable for obtaining the adequate sample sizes in heterogeneous ANOVA. However, a closer inspection of their numerical results reveals that the discrepancy between the nominal power and simulated power (or estimate of true power) is sizeable for several cases considered in their simulation study. Hence, the accuracy of their proposed power function in sample size estimation is questionable. Further examinations are required to demonstrate the underlying drawbacks associated with their approximate procedure.

According to the explication of power and sample size considerations for Welch's procedure presented above, the approximate technique proposed in Levy (1978a) has been given insufficient consideration, though a notable exception is Tomarken and Serlin (1986). Due to the complexity of theoretical justification for Welch's test procedure, no explicit analytic form of the corresponding non-null distribution is available. However, the approximate non-null distribution of Levy (1978a) can be obtained by replacing the sample means and variances in Welch's test statistic with corresponding population parameters. It was shown in the numerical comparisons of the estimated power and simulated power of Levy (1978a) that the suggested non-central F distribution yields an adequate approximation to the non-null distribution of Welch' statistic. Later, Tomarken and Serlin (1986) also strongly recommended the non-central F approximation for conducting power analyses of the Welch procedure. Thus the formula of Levy (1978a) is of great potential use and should be properly recognized. But the explication of Levy's non-central F distribution has been confined to power examination, and no single study has extended the investigation to sample size calculation. In view of the limitations of the existing findings, it is essential to generalize and assess the effectiveness of Levy's (1978a) approximate formula in sample size determination with modern computing facilities and accessible statistical software.

It is important to note that the approximate power functions of Levy's (1978a) and Luh and Guo (2011) both rely on a non-central F distribution, with identical numerator and denominator degrees of freedom. The only difference is in their respective specifications of the non-centrality parameter. Because of the complex nature of the non-central distribution and non-centrality parameter, a complete theoretical treatment and analytical evaluation is not feasible. However, there still remains no simultaneous comparison of the empirical performance of the two approaches. In order to offer well-supported recommendations on desirable sample sizes for heteroscedastic ANOVA models, this paper appraises and compares the two approaches of Levy (1978a) and Luh and Guo (2011) for power calculations and sample size determinations of Welch's test procedure. Since optimal sample size determinations for Welch's (1938) two-group test were presented in Jan and Shieh (2011), this paper focuses on the situations with three or more treatment groups. Comprehensive empirical investigations were conducted to demonstrate the potential advantages and disadvantages between the two methods under a variety of mean structures, variance patterns, as well as equal and unequal sample sizes. Our study reveals unique information that not only demonstrates the fundamental deficiency of existing investigations, but also enhances the usefulness of the Welch test in the context of ANOVA under variance heterogeneity. Moreover, corresponding SAS and R computer codes are presented to facilitate the recommended procedure for computing the achieved power level and required sample size in actual applications.

2. The Welch test

Consider the one-way heteroscedastic ANOVA model in which the observations X_{ij} are assumed to be independent and normally distributed with expected values μ_i and variances σ_i^2:

$$X_{ij} \sim N(\mu_i, \sigma_i^2), \tag{1}$$

where μ_i and σ_i^2 are unknown parameters, $i = 1, \ldots, g(\geq 2)$ and $j = 1, \ldots, N_i$ To test the hypothesis that all treatment means are equal, the classic F test is the most widely used statistical procedure assuming homogeneity of variance ($\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_g^2$). However, it has been shown in extensive studies that the conventional F test is sensitive to the heteroscedasticity formulation defined in (1). Of the numerous alternatives to the ANOVA F test, we focus on the viable approach proposed in Welch (1951) in the form of

$$W = \frac{\sum_{i=1}^{g} W_i(\bar{X}_i - \tilde{X})^2/(g-1)}{1 + 2(g-2)Q/(g^2-1)}, \tag{2}$$

where $W_i = N_i/S_i^2, S_i^2 = \sum_{j=1}^{N_i}(X_{ij} - \bar{X}_i)^2/(N_i - 1), \bar{X}_i = \sum_{j=1}^{N_i} X_{ij}/N_i, \tilde{X} = \sum_{i=1}^{g} W_i\bar{X}_i/U$, $U = \sum_{i=1}^{g} W_i$, and $Q = \sum_{i=1}^{g}(1 - W_i/U)^2/(N_i - 1)$. Under the null hypothesis H_0: $\mu_1 = \mu_2 = \ldots = \mu_g$, Welch (1951) suggests the approximate F distribution for W:

$$W \dot{\sim} F(g-1, \hat{\nu}),$$

where $F(g-1, \hat{\nu})$ is the F distribution with $g - 1$ and $\hat{\nu} = (g^2 - 1)/(3Q)$ degrees of freedom. Hence, H_0 is rejected at the significance level α if $W > F_{(g-1),\hat{\nu},\alpha}$, where $F_{(g-1),\hat{\nu},\alpha}$ is the upper 100 αth percentile of the F distribution $F(g-1, \hat{\nu})$. Although numerical evidence confirms the accurate Type I error control and superior power performance of Welch's test, theoretical justification for the non-null distribution of W has rarely been discussed. Especially, two non-central F approximations are considered in Levy (1978a) and Luh and Guo (2011). Luh and Guo (2011) suggested

$$W \dot{\sim} F(g-1, \nu, \Lambda_{\mathrm{LG}}),$$

where $F(g - 1, \upsilon, \Lambda_{\mathrm{LG}})$ is the non-central F distribution with $(g - 1)$ and $\upsilon = (g^2 - 1)/(3\tau)$ degrees of freedom, where $\tau = \sum_{i=1}^{g}(1 - \omega_i/\upsilon)^2/(N_i - 1), \omega_i = N_i/\sigma_i^2$, and $\upsilon = \sum_{i=1}^{g} \omega_i$, and with non-centrality parameter

$$\Lambda_{\mathrm{LG}} = g\frac{\sum_{i=1}^{g}(\mu_i - \bar{\mu})^2}{\sum_{i=1}^{g}(1/\omega_i)},$$

where $\bar{\mu} = \sum_{i=1}^{g} \mu_i/g$. Then the corresponding power function of Welch's test is of the form

$$\pi(\Lambda_{\mathrm{LG}}) = P\{F(g-1, \nu, \Lambda_{\mathrm{LG}}) > F_{(g-1),\nu,\alpha}\}. \tag{3}$$

On the other hand, Levy (1978a) proposed the approximate non-null distribution for W given by

$$W \dot{\sim} F(g-1, \nu, \Lambda_{\mathrm{L}}),$$

where

$$\Lambda_{\mathrm{L}} = \sum_{i=1}^{g} \omega_i(\mu_i - \tilde{\mu})^2,$$

and $\tilde{\mu} = \sum_{i=1}^{g} \omega_i\mu_i/\upsilon$. In this case, the associated power function $\pi(\Lambda_{\mathrm{L}})$ is expressed as

$$\pi(\Lambda_{\mathrm{L}}) = P\{F(g-1, \nu, \Lambda_{\mathrm{L}}) > F_{(g-1),\nu,\alpha}\}. \tag{4}$$

Note that the non-centrality parameter Λ_{LG} can be expressed as

$$\Lambda_{\mathrm{LG}} = \frac{\sum_{i=1}^{g}(\mu_i - \tilde{\mu})^2}{\sum_{i=1}^{g}(1/\omega_i)/g}.$$

Hence, the heteroscedastic variance property is only accommodated in the quantity $\sum_{i=1}^{g}(1/\omega_i)/g = \sum_{i=1}^{g}(\sigma_i^2/N_i)/g$ as a simple average of variances of group means. Contrast this with the form of the non-centrality parameter of Levy's (1978a) F approximation. The variance heterogeneity directly employed to reflect the weight of each of the group means in Λ_{L} makes a great difference in power performance.

It was demonstrated in Levy (1978a) that the actual power of Welch's test $P\{W > F_{(g-1)\hat{\nu},\alpha}\}$ can be well approximated by $\pi(\Lambda_{\mathrm{L}})$. As noted in Tomarken and Serlin (1986), this procedure may prove useful in conducting power analysis for one-way heteroscedastic ANOVA. Moreover, it is of great interest to extend the approach to sample size determination, just as in the case of Luh and Guo (2011) with the approximate power function $\pi(\Lambda_{\mathrm{LG}})$. In spite of the complexity in the denominator degrees of freedom of the F distribution, the power approximations in equations (3) and (4) closely resemble the power function of the ANOVA F test. But the two non-centrality parameters Λ_{L} and Λ_{LG} differ considerably in their expressions, and thus the resulting behaviours of the two power functions are presumably divergent. We next perform numerical investigations to evaluate and compare the accuracy of the two formulas for computing sample size under various model configurations likely to occur in practice.

3. Simulation studies

In order to enhance the applicability of sample size methodology and the fundamental usefulness of Welch's procedure, two Monte Carlo simulation studies were conducted to investigate the performance of the sample size calculation with respect to the two power functions described in Levy (1978a) and Luh and Guo (2011). With the approximate power formulas given in equations (3) and (4), the sample sizes (N_1, …, N_g) needed to attain the specified power $1 - \beta$ can be found by a simple iterative search for the chosen significance level α and parameter values (μ_i, σ_i^2), i = 1, …, g. Accordingly, the non-centrality parameters Λ_{LG} and Λ_{L} defined in (3) and (4) can be rewritten as

$$\Lambda_{\mathrm{LG}} = N_T \cdot \lambda_{\mathrm{LG}} \text{ and } \Lambda_{\mathrm{L}} = N_{\mathrm{T}} \cdot \lambda_{\mathrm{L}}, \tag{5}$$

respectively, where $N_T = \sum_{i=1}^{g} N_i, \lambda_{\mathrm{LG}} = g \cdot \sum_{g}^{i=1}(\mu_i - \bar{\mu})^2 / \left\{\sum_{i=1}^{g}(\sigma_j^2/q_j)\right\}, \lambda_{\mathrm{L}} = \sum_{i=1}^{g} q_i\{(\mu_i - \tilde{\mu})/\sigma_i\}^2, \tilde{\mu} = \sum_{i=1}^{g}(q_i\mu_i/\sigma_i^2)/\sum_{i=1}^{g}(q_j/\sigma_j^2)$ and $q_i = N_i/N_T$ for i = 1, …, g. Note that λ_{LG} and λ_{L} depend not on the group sizes but rather on the allocation ratio among the groups, and serve as the effect size measures for the approximations in Luh and Guo (2011) and Levy (1978a), respectively. As there may be several possible choices of sample size that satisfy the chosen power level in the process of sample size calculations, it is constructive to consider an appropriate design with a priori designated sample size ratios that leads to a unique and optimal result. For ease of illustration, the sample size ratios (r_1, …, r_g) are specified in advance with $r_i = N_i/N_1 i$ = 1, …, g. Note that $q_i = r_i / \sum_{j=1}^{g} r_j$, where $r_i = N_i/N_1$ for i = 1, …, g. Thus the task is confined to deciding the minimum sample size N_1 (with $N_i = N_1 r_i$, i = 2, …, g) required to achieve the desired power level.

Each of the vital factors of mean pattern, variance characteristic, and sample size structure has been shown to affect the magnitude of non-centrality and power. To provide a systematic demonstration, four patterns of variability in the means were used to assess power and compute sample size: (a) minimum variability (one mean at each extreme of the range, and all other means at the midpoint); (b) intermediate variability (such as means equally spaced through the range); (c) maximum variability (half of the means at each extreme of the range); and (d) extreme variability (one mean at one extreme of the range, and all other means equal and at the other extreme). Similar mean configurations were considered in Alexander and Govern (1994), Cohen (1988), De Beuckelaer (1996) and Tomarken and Serlin (1986). The empirical examination consists of two studies, of which the first re-examines the minimum variability mean patterns in Luh and Guo (2011), and the second evaluates the other cases of intermediate, maximum and extreme variability that were not considered in Luh and Guo (2011).

3.1. Study I

3.1.1. Design

For purposes of comparison, we reconsider the model settings with g = 4 and 6 in Table 1 of Luh and Guo (2011) in which the mean values are of minimum variability with $\mu = \{1, 0, 0, -1\}$ and $\{1, 0, 0, 0, 0, -1\}$, respectively. The corresponding two variance settings, representing homogeneous and heterogeneous structures, are $\sigma^2 = \{1, 1, 1, 1\}$ and $\{1, 4, 9, 16\}$, and $\{1, 1, 1, 1, 1, 1\}$ and $\{1, 1, 4, 4, 9, 9\}$, respectively. Moreover, the sample size ratio is fixed as the variance ratio $r_i = N_i/N_1 = \sigma_i/\sigma_1$ for $i = 1, \ldots, g$. With these specifications, the required sample sizes were computed for the two approaches with the chosen power value and significance level. Throughout this empirical investigation, the significance level is set at $\alpha = .05$. Note that the sample sizes of Luh and Guo's method are calculated with the algorithm presented in Luh and Guo (2011), which involves some further modification when applying the power function $\pi(\Lambda_{LG})$ in equation (3). In contrast, the sample sizes for Levy's procedure are determined with the power function $\pi(\Lambda_L)$ in equation (4). In addition, the actual or approximate powers are calculated with the resulting sample sizes. The SAS/IML (SAS Institute, 2011) and R (R Development Core Team, 2006) programs employed to perform the sample size determination and power calculation for Levy's (1978a) procedure are presented in Appendices A–D. The computed sample sizes and approximate powers are listed in Tables 1–3 for power levels .7, .8 and .9, respectively. Because of the underlying metric of integer sample sizes, the values achieved are marginally larger than the nominal level for both procedures. The only two exceptions occur with the variance homogeneity cases of comparatively small sample sizes in Table 1. Then for both procedures, estimates of the true power associated with given sample size and parameter configuration are computed via Monte Carlo simulation of 10,000 independent data sets. For each replicate, $(N_1, \ldots, N_g)$ normal outcomes are generated with the one-way homoscedastic or heteroscedastic ANOVA model. Next, the test statistic W is computed and the simulated power is the proportion of the 10,000 replicates whose test statistics W exceed the corresponding critical value $F_{g-1,\hat{v},.05}$. For the procedure examined, the adequacy for power and sample size calculation is determined by the difference between the simulated power and approximate power computed earlier. The simulated power and difference are also summarized in Tables 1–3 for the three designated power levels.

3.1.2. Results

An inspection of the reported sample sizes in Tables 1–3 reveals that in general the necessary sample sizes for Luh and Guo's (2011) method are larger than those for Levy's (1978a) approach. There is only one case in Table 3 where the two sets of sample sizes are identical. But even with the same sample sizes, the two power functions $\pi(\Lambda_L)$ and $\pi(\Lambda_{LG})$ still give different approximate power values because of the distinct non-centrality parameter formulations. More importantly, the discrepancies between simulated powers and approximate powers indicate that the performance of Luh and Guo's method is noticeably unstable and in several cases disturbing. Specifically, the resulting errors in Tables 1–3 range from .0131 to .1060. On the other hand, the errors associated with Levy's approach in Tables 1–3 clearly show that the approximate power formula of equation (4) performs extremely well because all absolute errors are less than .01 for the 12 cases examined here.

3.2. Study II

3.2.1. Design

To show a profound implication of the sample size procedures, further numerical assessments were performed with different variability patterns in mean structure. By

Table 1. Computed sample size, approximate power, and simulated power for the approaches of Luh and Guo (2011) and Levy (1978a) when nominal power is .70

	Luh and Guo				Levy[a]			
Mean and variance	Sample sizes structures	Approximate power	Simulated power	Difference	Sample sizes	Approximate power	Simulated power	Difference
$\mu = \{1, 0, 0, -1\}$ $\sigma^2 = \{1, 1, 1, 1\}$	(8, 8, 8, 8)	.7372	.8432	.1060	(7, 7, 7, 7)	.7796	.7760	−.0036
$\mu = \{1, 0, 0, -1\}$ $\sigma^2 = \{1, 4, 9, 16\}$	(12, 24, 36, 48)	.7068	.8080	.1012	(10, 20, 30, 40)	.7129	.7094	−.0035
$\mu = \{1, 0, 0, 0, 0, -1\}$ $\sigma^2 = \{1, 1, 1, 1, 1, 1\}$	(9, 9, 9, 9, 9, 9)	.7509	.8276	.0767	(8, 8, 8, 8, 8, 8)	.7752	.7725	−.0027
$\mu = \{1, 0, 0, 0, 0, -1\}$ $\sigma^2 = \{1, 1, 4, 4, 9, 9\}$	(12, 12, 24, 24, 36, 36)	.7132	.8106	.0974	(10, 10, 20, 20, 30, 30)	.7152	.7082	−.0070

Note. [a]The effect sizes λ_L for the four model configurations are .500, .0980, .3333, and .1010, respectively.

Table 2. Computed sample size, approximate power, and simulated power for the approaches of Luh and Guo (2011) and Levy (1978a) when nominal power is .80

	Luh and Guo				Levy[a]			
Mean and variance	Sample sizes structures	Approximate power	Simulated power	Difference	Sample sizes	Approximate power	Simulated power	Difference
$\mu = \{1, 0, 0, -1\}$ $\sigma^2 = \{1, 1, 1, 1\}$	(9, 9, 9, 9)	.8414	.8997	.0583	(8, 8, 8, 8)	.8529	.8435	−.0094
$\mu = \{1, 0, 0, -1\}$ $\sigma^2 = \{1, 4, 9, 16\}$	(14, 28, 42, 55)	.8069	.8675	.0606	(12, 24, 36, 48)	.8035	.7986	−.0049
$\mu = \{1, 0, 0, 0, 0, -1\}$ $\sigma^2 = \{1, 1, 1, 1, 1, 1\}$	(10, 10, 10, 10, 10, 10)	.8406	.8811	.0405	(9, 9, 9, 9, 9, 9)	.8426	.8360	−.0066
$\mu = \{1, 0, 0, 0, 0, -1\}$ $\sigma^2 = \{1, 1, 4, 4, 9, 9\}$	(15, 15, 29, 29, 43, 43)	.8143	.8957	.0814	(12, 12, 24, 24, 36, 36)	.8127	.8077	−.0050

Note. [a]The effect sizes λ_L for the four model configurations are .500, .0980, .3333, and .1010, respectively.

Table 3. Computed sample size, approximate power, and simulated power for the approaches of Luh and Guo (2011) and Levy (1978a) when nominal power is .90

	Luh and Guo				Levy[a]			
Mean and variance	Sample sizes structures	Approximate power	Simulated power	Difference	Sample sizes	Approximate power	Simulated power	Difference
$\mu = \{1, 0, 0, -1\}$ $\sigma^2 = \{1, 1, 1, 1\}$	(10, 10, 10, 10)	.9146	.9344	.0198	(9, 9, 9, 9)	.9046	.8975	−.0071
$\mu = \{1, 0, 0, -1\}$ $\sigma^2 = \{1, 4, 9, 16\}$	(18, 36, 54, 72)	.9038	.9493	.0455	(16, 32, 48, 64)	.9153	.9143	−.0010
$\mu = \{1, 0, 0, 0, 0, -1\}$ $\sigma^2 = \{1, 1, 1, 1, 1, 1\}$	(11, 11, 11, 11, 11, 11)	.9072	.9203	.0131	(11, 11, 11, 11, 11, 11)	.9282	.9256	−.0026
$\mu = \{1, 0, 0, 0, 0, -1\}$ $\sigma^2 = \{1, 1, 4, 4, 9, 9\}$	(17, 17, 34, 34, 50, 50)	.9053	.9395	.0342	(15, 15, 30, 30, 45, 45)	.9069	.9006	−.0063

Note. [a]The effect sizes λ_L for the four model configurations are .500, .0980, .3333, and .1010, respectively.

way of illustration, we focus on the common situation of $g = 4$ with heterogeneous variance characteristic $\{1, 4, 9, 16\}$. For mean patterns, two treatment structures are examined for each case of the intermediate, maximum, and extreme variability configurations:

- intermediate variability, $\{-3, -1, 1, 3\}/20^{1/2}$ and $\{5, 1, -2, -4\}/46^{1/2}$;
- maximum variability, $\{-1, 1, -1, 1\}/2$ and $\{-1, 1, 1, -1\}/2$;
- extreme variability, $\{3, -1, -1, -1\}/12^{1/2}$ and $\{-1, -1, -1, 3\}/12^{1/2}$.

Note that the average and the sum of the squared deviation for the mean values are $\bar{\mu} = 0$ and $\sum_{i=1}^{g}(\mu_i - \bar{\mu})^2 = 1$ for all six situations. This particular formulation is designed to expose how the non-centrality parameter Λ_{LG} of Luh and Guo (2011) is not sensitive with respect to mean variability pattern. Moreover, the mean patterns are combined with three different sample size ratios, $\{1, 1, 1, 1\}$, $\{1, 2, 3, 4\}$ and $\{4, 3, 2, 1\}$. These three settings not only include both balanced and unbalanced designs, but also create direct and inverse pairing with variance structures. Overall these considerations result in a total of 18 different model configurations. Thus our simulations cover a much broader range of situations than those considered in Luh and Guo (2011). These combinations of different variance structures, mean variability patterns, and sample size allocations were chosen to represent as much as possible the extent of characteristics that are likely to be obtained in actual applications. Moreover, the computed sample sizes associated with these model configurations reveal common and reasonable magnitudes of sample sizes used in typical research study. Similarly to the implementation of the design in Study I, the computed sample sizes, approximate powers, simulated powers, and associated errors of the two competing approaches are presented in Tables 4–6 and Tables 7–9 for power values .8 and .9, respectively.

3.2.2. Results

It is important to note that the sample sizes calculated with the procedure of Luh and Guo (2011) are identical in each of Tables 4–9. In other words, their method does not adequately reflect the actual fluctuation of mean structures in power and sample size computation. As expected, the associated approximate powers also remain the same. In contrast, the corresponding sample sizes of Levy's (1978a) approach vary with different mean variability configurations in combination with variance and sample size structures. With regard to the accuracy of sample size determination, the differences between simulated power and approximate power of Luh and Guo's (2011) formula are substantial and unsatisfactory, especially for cases of extreme variability in means, or circumstances under inverse pairing of sample sizes and variance in Tables 6 and 9. For example, the resulting errors of the two mean patterns $\{3, -1, -1, -1\}/12^{1/2}$ and $\{-1, -1, -1, 3\}/12^{1/2}$ are (.1944, −.2599), (.1615, −.1574), (.1968, −.3700), (.0990, −.2290), (.0884, −.1316), and (.0983, −.3671) in Tables 4–9, respectively. Hence, Luh and Guo's (2011) formula is clearly problematic and their method should not be used. In contrast, Levy's (1978a) method provides excellent performance in that incurred errors are all within the small range of −.0072 to .0098. In short, this numerical evidence demonstrates that Levy's (1978a) approach outperforms the procedure of Luh and Guo (2011) in power and sample size calculations under a wide variety of heteroscedastic model configurations.

Table 4. Computed sample size, approximate power, and simulated power for the approaches of Luh and Guo (2011) and Levy (1978a) with sample size ratio {1, 1, 1, 1} and variance {1, 4, 9, 16} when nominal power is .80

	Luh and Guo				Levy[a]			
Mean structure	Sample sizes	Approximate power	Simulated power	Difference	Sample sizes	Approximate power	Simulated power	Difference
$\{-3, -1, 1, 3\}/20^{1/2}$	(83, 83, 83, 83)	.8048	.9267	.1219	(60, 60, 60, 60)	.8054	.8081	.0027
$\{5, 1, -2, -4\}/46^{1/2}$	(83, 83, 83, 83)	.8048	.9654	.1606	(50, 50, 50, 50)	.8089	.8146	.0057
$\{-1, 1, -1, 1\}/2$	(83, 83, 83, 83)	.8048	.9729	.1681	(47, 47, 47, 47)	.8030	.8052	.0021
$\{-1, 1, 1, -1\}/2$	(83, 83, 83, 83)	.8048	.9844	.1796	(43, 43, 43, 43)	.8060	.8063	.0003
$\{3, -1, -1, -1\}/12^{1/2}$	(83, 83, 83, 83)	.8048	.9992	.1944	(30, 30, 30, 30)	.8084	.8120	.0036
$\{-1, -1, -1, 3\}/12^{1/2}$	(83, 83, 83, 83)	.8048	.5449	−.2599	(139, 139, 139, 139)	.8006	.8007	.0001

Note. [a]The effect sizes λ_L for the six model configurations are .0476, .0579, .0610, .0674, .0992, and .0199, respectively.

Table 5. Computed sample size, approximate power, and simulated power for the approaches of Luh and Guo (2011) and Levy (1978a) with sample size ratio {1, 2, 3, 4} and variance {1, 4, 9, 16} when nominal power is .80

	Luh and Guo				Levy[a]			
Mean structure	Sample sizes	Approximate power	Simulated power	Difference	Sample sizes	Approximate power	Simulated power	Difference
$\{3, 1, 1, 3\}/20^{1/2}$	(28, 55, 83, 110)	.8037	.8543	.0506	(25, 50, 75, 100)	.8131	.8137	.0006
$\{5, 1, -2, -4\}/46^{1/2}$	(28, 55, 83, 110)	.8037	.8954	.0917	(22, 44, 66, 88)	.8027	.8031	.0004
$\{-1, 1, -1, 1\}/2$	(28, 55, 83, 110)	.8037	.8750	.0713	(24, 48, 72, 96)	.8096	.8084	−.0012
$\{-1, 1, 1, -1\}/2$	(28, 55, 83, 110)	.8037	.8863	.0826	(23, 46, 69, 92)	.8082	.8060	−.0022
$\{3, -1, -1, -1\}/12^{1/2}$	(28, 55, 83, 110)	.8037	.9652	.1615	(17, 34, 51, 68)	.8134	.8062	−.0072
$\{-1, -1, -1, 3\}/12^{1/2}$	(28, 55, 83, 110)	.8037	.6463	−.1574	(38, 76, 114, 152)	.8007	.7990	−.0017

Note. [a]The effect sizes λ_L for the six model configurations are .0464, .0517, .0480, .0500, .0693, and .0293, respectively.

Table 6. Computed sample size, approximate power, and simulated power for the approaches of Luh and Guo (2011) and Levy (1978a) with sample size ratio {4, 3, 2, 1} and variance {1, 4, 9, 16} when nominal power is .80

	Luh and Guo				Levy[a]			
Mean structure	Sample sizes	Approximate power	Simulated power	Difference	Sample sizes	Approximate power	Simulated power	Difference
$\{-3, -1, 1, 3\}/20^{1/2}$	(242, 182, 121, 61)	.8032	.9825	.1793	(128, 96, 64, 32)	.8108	.8159	.0051
$\{5, 1, -2, -4\}/46^{1/2}$	(242, 182, 121, 61)	.8032	.9980	.1948	(96, 72, 48, 24)	.8122	.8190	.0068
$\{-1, 1, -1, 1\}/2$	(242, 182, 121, 61)	.8032	1.0000	.1968	(72, 54, 36, 18)	.8177	.8197	.0020
$\{-1, 1, 1, -1\}/2$	(242, 182, 121, 61)	.8032	.9997	.1965	(64, 48, 32, 16)	.8231	.8241	.0010
$\{3, -1, -1, -1\}/12^{1/2}$	(242, 182, 121, 61)	.8032	1.0000	.1968	(48, 36, 24, 12)	.8296	.8394	.0098
$\{-1, -1, -1, 3\}/12^{1/2}$	(242, 182, 121, 61)	.8032	.4332	−.3700	(536, 402, 268, 134)	.8007	.7952	−.0055

Note. [a]The effect sizes λ_L for the six model configurations are .0364, .0495, .0681, .0784, .1096, and .0082, respectively.

Table 7. Computed sample size, approximate power, and simulated power for the approaches of Luh and Guo (2011) and Levy (1978a) with sample size ratio $\{1, 1, 1, 1\}$ and variance $\{1, 4, 9, 16\}$ when nominal power is .90

	Luh and Guo				Levy[a]			
Mean structure	Sample sizes	Approximate power	Simulated power	Difference	Sample sizes	Approximate power	Simulated power	Difference
$\{-3, -1, 1, 3\}/20^{1/2}$	(107, 107, 107, 107)	.9009	.9763	.0754	(77, 77, 77, 77)	.9022	.9000	−.0022
$\{5, 1, -2, -4\}/46^{1/2}$	(107, 107, 107, 107)	.9009	.9929	.0920	(64, 64, 64, 64)	.9044	.9086	.0042
$\{-1, 1, -1, 1\}/2$	(107, 107, 107, 107)	.9009	.9937	.0928	(61, 61, 61, 61)	.9048	.9075	.0027
$\{-1, 1, 1, -1\}/2$	(107, 107, 107, 107)	.9009	.9970	.0961	(55, 55, 55, 55)	.9026	.8961	−.0065
$\{3, -1, -1, -1\}/12^{1/2}$	(107, 107, 107, 107)	.9009	.9999	.0990	(38, 38, 38, 38)	.9026	.9027	.0001
$\{-1, -1, -1, 3\}/12^{1/2}$	(107, 107, 107, 107)	.9009	.6719	−.2290	(180, 180, 180, 180)	.9003	.8967	−.0036

Note. [a]The effect sizes λ_L for the six model configurations are .0476, .0579, .0610, .0674, .0992, and .0199, respectively.

Table 8. Computed sample size, approximate power, and simulated power for the approaches of Luh and Guo (2011) and Levy (1978a) with sample size ratio {1, 2, 3, 4} and variance {1, 4, 9, 16} when nominal power is .90

	Luh and Guo				Levy[a]			
Mean structure	Sample sizes	Approximate power	Simulated power	Difference	Sample sizes	Approximate power	Simulated power	Difference
$\{-3, -1, 1, 3\}/20^{1/2}$	(36, 72, 107, 143)	.9020	.9394	.0374	(32, 64, 96, 128)	.9068	.9045	−.0023
$\{5, 1, -2, -4\}/46^{1/2}$	(36, 72, 107, 143)	.9020	.9581	.0561	(29, 58, 87, 116)	.9089	.9147	.0058
$\{-1, 1, -1, 1\}/2$	(36, 72, 107, 143)	.9020	.9509	.0489	(31, 62, 93, 124)	.9072	.9088	.0016
$\{-1, 1, 1, -1\}/2$	(36, 72, 107, 143)	.9020	.9536	.0516	(30, 60, 90, 120)	.9094	.9152	.0058
$\{3, -1, -1, -1\}/12^{1/2}$	(36, 72, 107, 143)	.9020	.9904	.0884	(22, 44, 66, 88)	.9114	.9130	.0016
$\{-1, -1, -1, 3\}/12^{1/2}$	(36, 72, 107, 143)	.9020	.7704	−.1316	(50, 100, 150, 200)	.9058	.9059	.0001

Note. [a]The effect sizes λ_L for the six model configurations are .0464, .0517, .0480, .0500, .0693, and .0293, respectively.

Table 9. Computed sample size, approximate power, and simulated power for the approaches of Luh and Guo (2011) and Levy (1978a) with sample size ratio {4, 3, 2, 1} and variance {1, 4, 9, 16} when nominal power is .90

	Luh and Guo				Levy[a]			
Mean structure	Sample sizes	Approximate power	Simulated power	Difference	Sample sizes	Approximate power	Simulated power	Difference
$\{-3, -1, 1, 3\}/20^{1/2}$	(315, 236, 158, 79)	.9017	.9973	.0956	(164, 123, 82, 41)	.9062	.9088	.0026
$\{5, 1, -2, -4\}/46^{1/2}$	(315, 236, 158, 79)	.9017	.9999	.0982	(120, 90, 60, 30)	.9002	.9028	.0026
$\{-1, 1, -1, 1\}/2$	(315, 236, 158, 79)	.9017	1.0000	.0983	(92, 69, 46, 23)	.9125	.9154	.0029
$\{-1, 1, 1, -1\}/2$	(315, 236, 158, 79)	.9017	1.0000	.0983	(80, 60, 40, 20)	.9101	.9127	.0026
$\{3, -1, -1, -1\}/12^{1/2}$	(315, 236, 158, 79)	.9017	1.0000	.0983	(60, 45, 30, 15)	.9165	.9206	.0041
$\{-1, -1, -1, 3\}/12^{1/2}$	(315, 236, 158, 79)	.9017	.5346	−.3671	(696, 522, 348, 174)	.9009	.9002	−.0007

Note. [a]The effect sizes λ_L for the six model configurations are .0364, .0495, .0681, .0784, .1096, and .0082, respectively.

4. Conclusions

The problem of heterogeneous error variances in one-way fixed effects ANOVA models has received considerable attention in the literature. Numerous approaches have been suggested to tackle the practical and complicated issue of heteroscedasticity. Notably, the Welch (1951) procedure has proved in several empirical investigations to provide excellent Type I error control and superior power performance. Its ease of computation and inclusion in software packages further enhance the applicability of Welch's (1951) test of the equality of means. But despite the extensive discussions of the selection of viable alternatives to the conventional ANOVA *F* test, the sample size computation has received inadequate attention from researchers. This study thus evaluates the properties of the existing approximate power functions of Welch's test in sample size determination since it is vital that the properties of the rival sample size formulas be clearly understood. Detailed numerical examinations were conducted to compare the procedures of Levy (1978a) and Luh and Guo (2011) under a wide variety of model configurations. The combined frameworks consist of the principle factors of means, variances and sample sizes structures. The present research extends the conditions and findings beyond those previously studied. We conclude that the intuitive approximation in Levy (1978a) provides a feasible and accurate solution to the sample size problem in the heteroscedastic ANOVA model. Considering the importance of power calculation and sample size determination in actual practice and the limited features of available computer packages, corresponding programs are developed to facilitate the use of the suggested approach.

References

Alexander, R. A., & Govern, D. M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. *Journal of Educational Statistics*, *19*, 91–101.

Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, *16*, 129–132.

Clinch, J. J., & Keselman, H. J. (1982). Parametric alternatives to the analysis of variance. *Journal of Educational Statistics*, *7*, 207–214.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.). Hillsdale, NJ: Erlbaum.

Coombs, W. T., Algina, J., & Oltman, D. O. (1996). Univariate and multivariate omnibus hypothesis tests selected to control type I error rates when population variances are not necessarily equal. *Review of Educational Research*, *66*, 137–179.

Cribbie, R. A., Fiksenbaum, L., Keselman, H. J., & Wilcox, R. R. (2012). Effect of non-normality on test statistics for one-way independent groups designs. *British Journal of Mathematical and Statistical Psychology*, *65*, 56–73.

De Beuckelaer, A. (1996). A closer examination on some parametric alternatives to the ANOVA *F*-test. *Statistical Papers*, *37*, 291–305.

Dijkstra, J. B., & Werter, P. S. P. J. (1981). Testing the equality of several means when the population variances are unequal. *Communications in Statistics: Simulation and Computation*, *10*, 557–569.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed-effects analysis of variance and covariance. *Review of Educational Research*, *42*, 237–288.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, *68*, 155–165.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, *17*, 315–339.

Howell, D. C. (2010). *Statistical methods for psychology*. (7th ed.). Belmont, CA: Wadsworth.

James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika*, *38*, 324–329.

Jan, S. L., & Shieh, G. (2011). Optimal sample sizes for Welch's test under various allocation and cost considerations. *Behavior Research Methods*, *43*, 1014–1022.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., … Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research*, *68*, 350–386.

Keselman, H. J., Rogan, J. C., & Feir-Walsh, B. J. (1977). An evaluation of some non-parametric and parametric tests for location equality. *British Journal of Mathematical and Statistical Psychology*, *30*, 213–221.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kohr, R. L., & Games, P. A. (1974). Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variances. *Journal of Experimental Education*, *43*, 1–69.

Krutchkoff, R. G. (1986). One-way fixed effects analysis of variance when the error variances may be unequal. *Journal of Statistical Computation and Simulation*, *30*, 259–271.

Kulinskaya, E., Staudte, R. G., & Gao, H. (2003). Power approximations in testing for unequal means in a one-way ANOVA weighted for unequal variances. *Communications in Statistics: Theory and Methods*, *32*, 2353–2371.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models*. (5th ed.). New York: McGraw-Hill.

Levy, K. J. (1978a). Some empirical power results associated with Welch's robust analysis of variance technique. *Journal of Statistical Computation and Simulation*, *8*, 43–48.

Levy, K. J. (1978b). An empirical comparison of the ANOVA *F*-test with alternatives which are more robust against heterogeneity of variance. *Journal of Statistical Computation and Simulation*, *8*, 49–57.

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality. *Educational and Psychological Measurement*, *58*, 409–429.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance *F* test. *Review of Educational Research*, *66*, 579–620.

Luh, W. M., & Guo, J. H. (2011). Developing the non-centrality parameter for calculating the group sample of the heterogeneous one-way fixed-effect ANOVA. *Journal of Experimental Education*, *79*, 53–63.

R Development Core Team (2006). *R: A language and environment for statistical computing* [computer software and manual]. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

Rogan, J. C., & Keselman, H. J. (1977). Is the ANOVA *F*-test robust to variance heterogeneity when sample sizes are equal? *American Educational Research Journal*, *14*, 493–498.

SAS Institute (2011). *SAS/IML User's Guide, Version 9.2*. Cary, NC: SAS Institute.

Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.

Schneider, R. J., & Penfield, D. A. (1997). Alexander and Govern's approximation: Providing an alternative to ANOVA under variance heterogeneity. *Journal of Experimental Education*, *65*, 271–286.

Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*, 90–99.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350–362.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330–336.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.

Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA *F*, *W*, and *F** statistics. *Communications in Statistics: Simulation and Computation*, *15*, 933–943.

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, *8*, 254–274.

Zimmerman, D. W. (2000). Statistical significance levels of nonparametric tests biased by heterogeneous variances of treatment groups. *Journal of General Psychology*, *127*, 354–364.

Received 26 January 2012; revised version received 28 October 2012

Appendix A: SAS IML program for calculating the required sample sizes of Welch’s test

```
PROC IML;PRINT ”CALCULATE REQUIRED SAMPLE SIZE OF WELCH’S TEST”;
    *USER SPECIFICATIONS;
    *DEGNATED POWER;POWER=0.80;
    *TYPE I ERROR;ALPHA=0.05;
    *GROUP MEANS;MUVEC={-3 -1 1 3}/SQRT(20);
    *GROUP VARIANCES;VARVEC={1 4 9 16};
    *SAMPLE SIZE RATIOS;RVEC={1 1 1 1};
    *END OF SPECIFICATIONS;
    G=NCOL(MUVEC);
    PRINT G,RVEC,MUVEC,VARVEC;
    ORVEC=RVEC/VARVEC;
    MUPR=SUM(ORVEC#MUVEC)/SUM(ORVEC);
    LAMR=SUM(ORVEC#(MUVEC-MUPR)##2);
    DF1=G-1;CCRIT=CINV(1-ALPHA,DF1);
    LAMC=CNONCT(CCRIT,DF1,1-POWER);
    NC=CEIL(LAMC/LAMR);
    N1=NC-1;
    DO UNTIL(NPOWER>POWER);
    N1=N1+1;
    NVEC=N1#RVEC;
    OVEC=NVEC/VARVEC;
    MUP=SUM(OVEC#MUVEC)/SUM(OVEC);
    LAM=SUM(OVEC#(MUVEC-MUP)##2);
    DEL=SUM(((1-OVEC/SUM(OVEC))##2)/(NVEC-1));
    DF2=(G#G-1)/(3#DEL);
    FCRIT=FINV(1-ALPHA,DF1,DF2);
    NPOWER=SDF(‘F’,FCRIT,DF1,DF2,LAM);
    END;*FOR UNTIL;
    PRINT ’SAMPLE SIZES’ NVEC;
    PRINT ’APPROXIMATE POWER’ NPOWER[FORMAT=8.4];
    QUIT;
```

Appendix B: SAS IML program for computing the approximate power of Welch’s test

```
PROC IML;PRINT ”COMPUTE APPROXIMATE POWER OF WELCH’ TEST”;
    *USER SPECIFICATIONS;
    *TYPE I ERROR;ALPHA=0.05;
    *GROUP MEANS;MUVEC={1 0 0 -1};
    *GROUP VARIANCES;VARVEC={1 1 1 1};
    *SAMPLE SIZES;NVEC={9 9 9 9};
    *END OF SPECIFICATIONS;
    G=NCOL(NVEC);
    PRINT G,NVEC,MUVEC,VARVEC;
    OVEC=NVEC/VARVEC;
    MUP=SUM(OVEC#MUVEC)/SUM(OVEC);
    LAM=SUM(OVEC#(MUVEC-MUP)##2);
    DF1=G-1;
    DEL=SUM(((1-OVEC/SUM(OVEC))##2)/(NVEC-1));
    DF2=(G#G-1)/(3#DEL);
    FCRIT=FINV(1-ALPHA,DF1,DF2);
    NPOWER=SDF(‘F’,FCRIT,DF1,DF2,LAM);
    PRINT ’APPROXIMATE POWER’ NPOWER[FORMAT=8.4];
    QUIT;
```

Appendix C: R program for calculating the required sample sizes of Welch’s test

```
function () {
    #REQUIRED USER SPECIFICATIONS PORTION
    power<-0.90 #DESIGNATED POWER
    alpha<-0.05 #TYPE I ERROR
    muvec<-c(-3,-1,1,3)/sqrt(20) #GROUP MEANS
    varvec<-c(1, 4, 9, 16) #GROUP VARIANCES
    rvec<-c(1,1,1,1) #SAMPLE SIZE RATIOS
    #END OF REQUIRED USER SPECIFICATION
    g<-length(muvec)
    orvec<-rvec/varvec
    mupr<-sum(orvec*muvec)/sum(orvec)
    lamr<-sum(orvec*(muvec-mupr)^2)
    df1<-g-1
    n1<-5
    apower<-0
    while (apower<power){
    n1<-n1+1
    nvec<-n1*rvec
    ovec<-nvec/varvec
    mup<-sum(ovec*muvec)/sum(ovec)
    lam<-sum(ovec*(muvec-mup)^2)
    del<-sum(((1-ovec/sum(ovec))^2)/(nvec-1))
    df2<-(g*g-1)/(3*del)
    fcrit<-qf(1-alpha,df1,df2)
    apower<-1-pf(fcrit,df1,df2,lam)
    }
    print(“nvec”)
    print(nvec)
    print(“apower”)
    print(apower,digits=4)
    }
```

Appendix D: R program for computing the approximate power of Welch’s test

```
function () {
    #REQUIRED USER SPECIFICATIONS PORTION
    alpha<-0.05 #TYPE I ERROR
    muvec<-c(1,0,0,-1) #GROUP MEANS
    varvec<-c(1,1,1,1) #GROUP VARIANCES
    nvec<-c(9,9,9,9) #GROUP SAMPLE SIZES
    #END OF REQUIRED USER SPECIFICATION
    g < -length(muvec)
    df1 < -g-1
    ovec<-nvec/varvec
    mup<-sum(ovec*muvec)/sum(ovec)
    lam<-sum(ovec*(muvec-mup)^2)
    del<-sum(((1-ovec/sum(ovec))^2)/(nvec-1))
    df2<-(g*g-1)/(3*del)
    fcrit<-qf(1-alpha,df1,df2)
    apower<-1-pf(fcrit,df1,df2,lam)
    print(“nvec”)
    print(nvec)
    print(“apower”)
    print(apower,digits=4)
    }
```