Psicológica (2015), 36, 367-390.

Power and sample size calculations for testing linear combinations of group means under variance heterogeneity with applications to meta and moderation analyses

Gwowen Shieh[1] & Show-Li Jan[2]

[1]*National Chiao Tung University, Hsinchu, Taiwan*

[2]*Chung Yuan Christian University, Chungli, Taiwan*

The general formulation of a linear combination of population means permits a wide range of research questions to be tested within the context of ANOVA. However, it has been stressed in many research areas that the homogeneous variances assumption is frequently violated. To accommodate the heterogeneity of variance structure, the Welch–Satterthwaite procedure is commonly used as an alternative to the *t* test for detecting the substantive significance of a linear combination of mean effects. This article presents two approaches to power and sample size calculations for the Welch–Satterthwaite test. The usefulness and diversity of the suggested techniques are illustrated with two of the potential applications in meta and moderation analyses. The numerical assessments showed that the proposed approaches outperform the existing methods on the accuracy of power calculations and sample size determinations for meta and moderation studies. Computer algorithms are also developed to implement the recommended procedures in actual research designs.

Within the context of analysis of variance (ANOVA), it is often desirable to perform comparisons among group means to provide specific answers to critical research questions. The general formulation of a linear combination of group means permits a wide range of research hypotheses to be tested. Accordingly, the designated linear comparison represents the substantive hypothesis of interest and reveals essential information that

[1] Corresponding author: Gwowen Shieh. Department of Management Science. National Chiao Tung University. Hsinchu, Taiwan 30010, R.O.C. Email: gwshieh@mail.nctu.edu.tw

cannot be obtained from the omnibus tests. In one-factor designs, the differences between two sets of average group means can be assessed in terms of a linear combination of treatment effects. On the other hand, a linear combination may be employed to evaluate interactions between treatment effects in factorial designs. It follows from the independence, normality, and homogeneity of variance assumptions in ANOVA, that the inference for a linear combination of mean effects can be conducted with a *t* statistic. Comprehensive guidelines are available in Kutner et al. (2005) and Maxwell and Delaney (2004).

Although the homogeneity of variance formulation provides a convenient and useful setup, it is not unusual for the homoscedasticity assumption to be violated in actual applications. Specifically, Grissom (2000) and Ruscio and Roche (2012) emphasized that the existence of heteroscedasticity in clinical and psychological data is more common than most researchers realize. Therefore it is prudent to employ suitable techniques that are superior to the traditional inferential methods under various conditions of unequal variances (Levy, 1978; Tomarken & Serlin, 1986). For testing a hypothesis of a linear combination of group means, the approximation suggested independently by Satterthwaite (1946) and Welch (1947) is the most widely recommended technique to correct for variance heterogeneity. The procedure is sometimes referred to as the Welch–Satterthwaite test and provides a simple and robust *t*-solution with approximate degrees of freedom. Essentially, this problem is a generalization of the well-known Behrens–Fisher problem (Kim and Cohen, 1998) of testing the difference between two population means when population variances are heterogeneous. The technique is also useful for more complex frameworks such as linear mixed models and generalized linear mixed models.

Despite the advantages for tackling the fairly complicated issue of heteroscedasticity, one of the notable issues of the Welch–Satterthwaite procedure is the problem of power and sample size calculations. In view of the considerable practical value of a linear combination in heteroscedastic ANOVA, this article describes two approaches to power and sample size calculations for the Welch–Satterthwaite test. One approach adopts a noncentral *t* approximation to the nonnull distribution of the Welch–Satterthwaite test. Whereas the other approach considers an exact evaluation of the power function of the Welch–Satterthwaite test. The approximate distribution presents a particularly attractive and convenient expression. Alternatively, the exact formulation is noticeably more effective in maintaining the power performance in some situations. Accordingly, the presented two power functions can be utilized to calculate the sample sizes

needed to attain the specified power level for the chosen model configurations. The suggested sample size procedures can be viewed as a heteroscedastic generalization of Wahlsten (1991) for the standard homogeneous variances framework, and a multi-group extension of Jan and Shieh (2011) for the comparison between two population means.

It is noteworthy that all the tests of main and interaction hypotheses relate to linear combinations of group means within the context of an ANOVA in primary research are also applicable in meta analysis. A general treatment of meta analysis can be found in Hedges and Olkin (1985) and Hunter and Schmidt (2004). Moreover, the importance of power calculations in meta analysis have been noted in Aguinis et al. (2011), Hedges and Pigott (2001), Muncer, Taylor, and Craigie (2002), and Valentine, Pigott, and Rothstein (2010). Accordingly, the statistical power analysis should be a standard for articles reporting meta analyses (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008). The most common practice of power computation for contrasts among effect sizes assumes the associated variances are known (Hedges & Pigott, 2001). However, the variance components, just as the mean parameters or effect sizes, are measured with errors from independent studies. Similar notion to accommodate the variability of sample variances in meta analysis has been recommended by Bond, Wiitala, and Richard (2003) and Hartung, Argac, and Makambi (2002). Hence it is of theoretical and practical importance to clarify the adequacy and discrepancy between the proposed approaches and the commonly used technique of Hedges and Pigott (2001).

Another noticeable utility of testing a linear combination of group means is to detect moderating or interactive effects in factorial studies. For example, Aguinis (2004), Cohen et al. (2003), and Frazier, Tix, and Barron (2004) present practical implications for assessing moderation and interaction. However, the tests of hypotheses pertaining to the interaction effects often have very low statistical power in applied psychology and management research (Aguinis et al., 2005 and the references therein). Moreover, Aguinis and Pierce (1998) and Alexander and DeShon (1994) reported that the violation of the homogeneity of error variance assumption has a detrimental impact on the power for the assessment of moderating effects of categorical variables. To remedy the situation, Aguinis, Boik, and Pierce (2001) suggested a generalized solution for approximating the power to detect interaction effects between categorical moderator variables and continuous predictor variables. Alternatively, Guo and Luh (2009) presented a sample size method to identify an interaction effect and main

effects in a heteroscedastic 2 × 2 design. For the ultimate aim of selecting the best approach, it is sensible to explicate the analytical argument and empirical performance of the proposed approaches and the method of Guo and Luh (2009).

In subsequent sections, the suggested exact and approximate formulations for the nonnull distribution of the Welch–Satterthwaite test are presented. Then the considered power functions are employed to compute the power and sample size for detecting a linear combination of population means. Monte Carlo simulation studies were conducted to illustrate the potential advantages and disadvantages between the proposed and available procedures for the meta and moderation analyses. Our study reveals unique information that not only demonstrates the fundamental behavior of existing methodology, but also enhances the usefulness of the Welch–Satterthwaite test in the context of heteroscedastic ANOVA. Moreover, corresponding SAS computer codes are presented as appendixes to facilitate the recommended procedures in planning ANOVA research.

Linear Combinations

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

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

where μ_i and σ_i^2 are unknown parameters, $i = 1, ..., G$ (≥ 2) and $j = 1, \ldots, N_i$. Thus, G denotes the number of groups and N_i is the sample size in the ith group. A linear combination of mean parameters is defined as

$$\psi = \sum_{i=1}^{G} c_i \mu_i, \tag{2}$$

where c_i are the linear coefficients. A contrast is a special case of a linear combination with the mean coefficients $\sum_{i=1}^{G} c_i = 0$. It follows from the model assumption in Equation 1 that a convenient unbiased estimator $\hat{\psi}$ for the combined effect size ψ defined in Equation 2 is of the form

$$\hat{\psi} = \sum_{i=1}^{G} c_i \bar{Y}_i \tag{3}$$

where $\bar{Y}_i = \sum_{j=1}^{N_i} Y_{ij}/N_i$ is the ith group sample mean and is an unbiased estimator of μ_i for $i = 1, ..., G$. Moreover, the linear combination estimator $\hat{\psi}$ given in Equation 3 has the distribution

$$\hat{\psi} \sim N(\psi, \Sigma), \tag{4}$$

where $\Sigma = Var(\hat{\psi}) = \sum_{i=1}^{G} c_i^2 \sigma_i^2/N_i.$ Also, an unbiased estimator $\hat{\Sigma}$ of Σ can be obtained by replacing the variance σ_i^2 in Σ with its unbiased estimator S_i^2 as follows:

$$\hat{\Sigma} = \sum_{i=1}^{G} c_i^2 S_i^2/N_i, \tag{5}$$

where $S_i^2 = \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_i)^2/(N_i - 1)$ is the sample variance for $i = 1, ..., G$. For detecting a linear combination among the mean effects in terms of the hypothesis $H_0: \psi = \psi_0$ versus $H_1: \psi \neq \psi_0$, a useful statistic has the form

$$T^* = \frac{\hat{\psi} - \psi_0}{\hat{\Sigma}^{1/2}} \tag{6}$$

where ψ_0 is a constant. Due to the dependence of $\hat{\Sigma}$ on the sample variances $\{S_1^2, ..., S_G^2\}$, the exact distribution of T^* is fairly complicated. Notably, the sample variances S_i^2 are distributed independently of each other and $(N_i - 1)S_i^2/\sigma_i^2 \sim \chi^2(N_i - 1)$ for $i = 1, ..., G$. It was demonstrated in Satterthwaite (1946) and Welch (1947) by the method of equating moments that $\hat{\Sigma}$ has the approximate distribution

$$\hat{\Sigma} \dot{\sim} \frac{\Sigma}{\nu} \cdot \chi^2(\nu), \tag{7}$$

where $\nu = \{\sum_{i=1}^{G} c_i^2 \sigma_i^2/N_i\}^2/\{\sum_{i=1}^{G} c_i^4 \sigma_i^4/[N_i^2(N_i - 1)]\}.$ Under the null hypothesis $H_0: \psi = \psi_0$, it readily follows from Equations 4 and 7 that the quantity T^* given in Equation 6 has a convenient approximate distribution

$$T^* \dot{\sim} t(\nu),$$

where $t(\nu)$ is a t distribution with degrees of freedom ν. For inferential purposes, the term of degrees of freedom ν is replaced by its counterpart $\hat{\nu}$ with direct substitution of $\{S_1^2, \ldots, S_G^2\}$ for $\{\sigma_1^2, \ldots, \sigma_G^2\}$ in ν, where

$$\hat{\nu} = \{\sum_{i=1}^{G} c_i^2 S_i^2/N_i\}^2/\{\sum_{i=1}^{G} c_i^4 S_i^4/[N_i^2(N_i-1)]\}. \quad (8)$$

Hence, the Welch–Satterthwaite procedure rejects H_0 at the significance level α if $|T^*| > t_{\hat{\nu}, \alpha/2}$ where $t_{\hat{\nu}, \alpha/2}$ is the upper $100(\alpha/2)$ percentile of the t distribution $t(\hat{\nu})$. Accordingly, it can be shown with the same theoretical arguments and analytic derivations that the statistic T^* has the general approximate distribution

$$T^* \sim t(\nu, \delta), \quad (9)$$

where $t(\nu, \delta)$ is a noncentral t distribution with degrees of freedom ν and noncentrality parameter

$$\delta = \frac{\psi - \psi_0}{\Sigma^{1/2}}.$$

It immediately follows from the noncentral t distribution given in Equation 9 that the power function of the Welch–Satterthwaite test can be approximated by

$$\pi_T(\delta) = P\{|t(\nu, \delta)| > t_{\nu, \alpha/2}\}. \quad (10)$$

The suggested formulation of $\pi_T(\delta)$ is referred to as the approximate T approach for ease of exposition.

Alternatively, the exact distribution of test statistic T^* may be expressed in different forms. Note that the linear combination of independent sample variances $\hat{\Sigma}$ given in Equation 5 can be rewritten as

$$\hat{\Sigma} = K \cdot A$$

Where $K = \sum_{i=1}^{G} K_i \sim \chi^2(N_T - G)$, $K_i \sim \chi^2(N_i - 1)$, $N_T = \sum_{i=1}^{g} N_i$, $A = \sum_{i=1}^{G} b_i A_i$, $A_i = K_i/K$, and $b_i = (c_i^2\sigma_i^2)/\{N_i(N_i-1)\}$, $i = 1, \ldots, G$.

The approximate degrees of freedom $\hat{\nu}$ given in Equation 8 can also be expressed as $\hat{\nu} = \{\sum_{i=1}^{G} b_i A_i\}^2/\{\sum_{i=1}^{G} b_i^2 A_i^2/(N_i - 1)\}$. Moreover, it is computationally simple and relatively stable to rewrite the dependence of $\{A_1, \ldots, A_G\}$ on the chi-square random variables in terms of the beta random

variables, see Johnson, Kotz and Balakrishnan (1995, p. 212). Specifically $A_1 = \prod_{i=1}^{G-1} B_i$, $A_2 = (1 - B_1)\prod_{i=2}^{G-1} B_i$, …, $A_{G-1} = (1 - B_{G-2})B_{G-1}$, and $A_G = 1 - B_{G-1}$, where $B_i = \{\sum_{j=1}^{i} K_j\}/\{\sum_{j'=1}^{i+1} K_{j'}\}$ has a beta distribution $B_i \sim \text{Beta}\{\sum_{j=1}^{i} (N_j - 1)/2,\ (N_{i+1} - 1)/2\}$ for $i = 1, \ldots, G - 1$. An important underlying property of the suggested formulations is that the random variables $B_1, \ldots, B_{G-1}$ and K are mutually independent. Hence, both $\hat{\nu}$ and A can be viewed as a function of beta random variables $\{B_1, \ldots, B_{G-1}\}$, and they are independent of K.

With these definitions of transformed variables, the following formulation of T^* is considered:

$$T^* = \frac{T}{V^{1/2}}, \tag{11}$$

where $T = Z/\{K/(N_T - g)\}^{1/2} \sim t(N_T - G, \delta)$, and $V = (N_T - G)A/\Sigma$. Overall, the random variables Z, K and $\{B_1, \ldots, B_{G-1}\}$ are mutually independent. Hence, T and V are independent. With the alternative expression of T^* given in Equation 11, the exact power function of the Welch–Satterthwaite test can be formulated as

$$\pi_E(\delta) = E_B\{P\{|t(N_T - G, \delta)| > t_{\hat{\nu}, \alpha/2} V^{1/2}\}, \tag{12}$$

where the expectation $E_B\{\}$ is taken with respect to the joint distribution of $\{B_1, \ldots, B_{G-1}\}$. Since all related functions are readily embedded in major statistical packages, Monte Carlo integration provides a feasible approach to perform the required assessment of $\pi_E(\delta)$, especially when the number of groups is large.

To determine sample sizes in planning research designs, the power functions can be employed to calculate the sample sizes $\{N_i, i = 1, \ldots, G\}$ needed to attain the specified power $1 - \beta$ for the chosen significance level α, null value ψ_0, mean coefficients $\{c_i, i = 1, \ldots, G\}$ and parameter values $\{\mu_i, \sigma_i^2, i = 1, \ldots, G\}$. It usually involves an iterative process to find the solution. As there may be several possible sets of sample sizes that satisfy the chosen power level, it is constructive to consider an appropriate design with a priori designated sample size ratios $\{r_1, ..., r_G\}$ with $r_i = N_i/N_1$, for $i = 1, \ldots, G$. Thus the process is confined to deciding the minimum sample size N_1 (with $N_i = N_1 r_i$, $i = 2, ..., G$) required to achieve the selected power level with the power functions in Equations 10 and 12, respectively. In order to

explicate the applicability of power and sample size methodology for the Welch–Satterthwaite procedure, in subsequent sections this study considers design configurations under the contexts of meta and moderation analyses.

Power calculations in meta analysis

The most appealing reason for conducting a meta analysis is that a collection of related studies have higher statistical power than any single one of those studies (Hunter & Schmidt, 2004). However, as with prospective power analysis in a primary study, the statistical power in a meta-analysis depends on the joint impact of responsible factors including the population effect size, the associated variance component, Type I error rate, and sample size. Without a detailed appraisal, the actual power of the collection of studies may still not be high enough to detect the effect size of importance and to support the research question of interest. It is essential to note that the comparison of mean effect sizes between two sets of studies in meta analysis corresponds to testing a linear combination of mean effects in a one-way heteroscedastic ANOVA. To demonstrate the contrasting behavior of the alternative power functions of the Welch-Satterthwaite test in the context of meta analysis, a numerical investigation was conducted in two stages. The first stage presented power calculations for Hedges and Pigott's (2001, Equation 31) method and the proposed exact and approximate approaches described in Equations 10 and 12, respectively, under several model configurations. Then, a Monte Carlo simulation was performed to explicate the accuracy of the competing procedures under the design characteristics specified in the first step.

To reveal the potential extent of characteristics that an applied work may reflect in meta studies, the examined frameworks consist of the principle factors of sample sizes, variance components and linear coefficients. Following the model formulation of Hedge and Pigott (2001, Equation 1), the numbers of studies are set as $G = 4$ and 12 with an average study size $N_T/G = 10$. The sample size patterns are deliberately varied with variance components to have three different characteristics: balanced, direct-pairing and inverse-pairing structures. For the case of $G = 4$, the heterogeneous variances are chosen as $\{1, 4, 9, 16\}$. Accordingly, the three study size designs $\{N_i, i = 1, ..., 4\}$ for a total $N_T = 40$ are

Balanced design: {10, 10, 10, 10};

Direct-pairing design: {4, 8, 12, 16};

Inverse-pairing design: {16, 12, 8, 4}.

Moreover, the three sample size schemes are combined with three different sets of linear coefficients: $\{c_1, c_2, c_3, c_4\} = \{1, -1/3, -1/3, -1/3\}$, $\{1/3, 1/3, 1/3, -1\}$ and $\{1/2, 1/2, -1/2, -1/2\}$.

On the other hand, the prescribed configurations for $G = 4$ are extended to $G = 12$ by replicating each element three times. Specifically, the variance components are {1, 1, 1, 4, 4, 4, 9, 9, 9, 16, 16, 16} and the corresponding sample size designs $\{N_i, i = 1, ..., 12\}$ are

Balanced design: {10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10};

Direct-pairing design: {4, 4, 4, 8, 8, 8, 12, 12, 12, 16, 16, 16};

Inverse-pairing design: {16, 16, 16, 12, 12, 12, 8, 8, 8, 4, 4, 4}.

Although these sample sizes may be smaller than would be likely in many ANOVA studies, it is plausible that if problems or deficiencies were to be seen with power calculations, they would be most apparent with small study sizes. In this case, the three settings of linear coefficients $\{c_i, i = 1, ..., 12\}$ are denoted by

LC1: {1/3, 1/3, 1/3, –1/9, –1/9, –1/9, –1/9, –1/9, –1/9, –1/9, –1/9, –1/9};

LC2: {1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, –1/3, –1/3, –1/3};

LC3: {1/6, 1/6, 1/6, 1/6, 1/6, 1/6, –1/6, –1/6, –1/6, –1/6, –1/6, –1/6}.

Without loss of generality, the mean effects are set as $\mu_1 = \mu$ and $\mu_i = 0$ for $i = 2$ to G, where μ is properly selected such that the resulting power level of the approximate T procedure is near 0.90. Specifically, the selected values of μ for the nine settings in Table 1 are 2.18, 14.21, 5.87, 2.53, 11.05, 5.27, 3.15, 29.42, and 9.38. The corresponding mean values in Table 2 are 3.69, 23.02, 9.87, 4.10, 18.50, 8.96, 4.84, 38.34, and 14.03. Throughout this empirical study, the significance level is set as $\alpha = 0.05$ and the null value is $\psi_0 = 0$. Overall these considerations result in a total of 18 different model configurations.

With these specifications, the attained power level for the designated power function can be readily computed. In addition, a simple expression has been described in Hedges and Pigott (2001) for the nonnull distribution of the Welch–Satterthwaite statistic T^*. Specifically, under the assumption that the variances are known, Hedges and Pigott (2001, Equation 31) suggested an approximate Z formulation

$$\pi_Z(\delta) = P\{|N(\delta, 1)| > z_{\alpha/2}\}, \tag{13}$$

where $N(\delta, 1)$ is a normal distribution with mean δ and variance 1, and $z_{\alpha/2}$ is the upper $100(\alpha/2)$ percentile of the standard normal distribution. The performance of power functions $\pi_E(\delta)$, $\pi_T(\delta)$ and $\pi_Z(\delta)$ are examined for each of the 18 combined settings of 2 numbers of studies, 3 study size

structures, and 3 linear coefficient sets. The computed powers for the selected model configurations are listed in Tables 1 and 2 for $G = 4$ and 12, respectively. An inspection of the power calculations reported in Tables 1 and 2 shows that a consistent order among the achieved power levels: $\pi_E(\delta) < \pi_T(\delta) < \pi_Z(\delta)$ for all the cases considered here. The power outcome given by the approximate Z formula may be the largest; however, this does not imply that it is the best method. The accuracy of the alternative formulas is further evaluated through the following Monte Carlo simulation.

In the second step, estimates of the true power associated with the given model configurations for all three procedures were computed via Monte Carlo simulations of 10,000 independent data sets. For each replicate, $\{N_1, ..., N_G\}$ normal outcomes are generated with the designated configurations for the one-way heteroscedastic ANOVA model. Next, the test statistic T^* is computed and the simulated power is the proportion of the 10,000 replicates whose test statistics T^* exceed the corresponding critical value $|T^*| > t_{\hat{\nu}, 0.025}$. Adequacy of the examined procedure for power calculation is determined by the error between the simulated power and computed power presented above. The simulated power and errors are also summarized in Tables 1 and 2.

According to the extensive numerical results, the approximate Z method of Hedges and Pigott (2001) is not consistently accurate because only 7 out of 18 cases have absolute error less than or equal to 0.02. The differences of the remaining 11 cases are substantial and unsatisfactory, especially for the circumstances under inverse pairing of variance heterogeneity and sample sizes. Specifically, the results associated with the inverse-pairing condition incur the sizeable errors of –0.0919, –0.1008 and –0.1003 for the three linear combinations in Table 1. Also, the corresponding errors are –0.0402, –0.0765 and –0.0482 for the three comparisons presented in Table 2. Thus, the absence to incorporate uncertainty associated with variance estimation is a disadvantage of the approximate Z power function proposed in Hedges and Pigott (2001).

On the other hand, the computed powers of the noncentral t function $\pi_T(\delta)$ appear to maintain a reasonable range near the simulated outcomes. For the balanced and direct-pairing designs, the approximate T method generally gives reliable results with absolute errors mostly less than 0.01. The only exception is the error –0.0144 associated with the first comparison of the direct-pairing scheme in Table 1. However, the performance is less satisfactory in the inverse-pairing situations. When the number of studies is $G = 4$, the induced errors for the first and third linear combinations of the inverse-pairing settings are –0.0222 and –0.0219, respectively.

Table 1. Simulated power and computed power for the test of linear combination H_0: $\psi = 0$ versus H_1: $\psi \neq 0$ with $\alpha = 0.05$ and $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 4, 9, 16)$

			Exact approach		Approximate *T*		Hedges and Pigott (2001)	
Sample sizes	Linear coefficients	Simulated power	Computed power	Error	Computed power	Error	Computed power	Error
Balanced design: {10, 10, 10, 10}								
	{1, –1/3, –1/3, –1/3}	0.8970	0.8987	–0.0017	0.9007	–0.0037	0.9185	–0.0215
	{1/3, 1/3, 1/3, –1}	0.9035	0.8992	0.0043	0.9002	0.0033	0.9468	–0.0433
	{1/2, 1/2, –1/2, –1/2}	0.8977	0.8979	–0.0002	0.9004	–0.0027	0.9235	–0.0258
Direct-pairing design:{4, 8, 12, 16}								
	{1, –1/3, –1/3, –1/3}	0.8874	0.8850	0.0024	0.9018	–0.0144	0.9472	–0.0598
	{1/3, 1/3, 1/3, –1}	0.8952	0.8992	–0.0040	0.9002	–0.0050	0.9265	–0.0313
	{1/2, 1/2, –1/2, –1/2}	0.8962	0.8980	–0.0018	0.9002	–0.0040	0.9151	–0.0189
Inverse-pairing design: {16, 12, 8, 4}								
	{1, –1/3, –1/3, –1/3}	0.8788	0.8792	–0.0004	0.9010	–0.0222	0.9707	–0.0919
	{1/3, 1/3, 1/3, –1}	0.8970	0.8950	0.0020	0.9001	–0.0031	0.9978	–0.1008
	{1/2, 1/2, –1/2, –1/2}	0.8786	0.8784	0.0002	0.9005	–0.0219	0.9789	–0.1003

Table 2. Simulated power and computed power for the test of linear combination H_0: $\psi = 0$ versus H_1: $\psi \neq 0$ with $\alpha = 0.05$ and variance components: {1, 1, 1, 4, 4, 4, 9, 9, 9, 16, 16, 16}

		Exact approach			Approximate T		Hedges and Pigott (2001)	
Sample sizes	Linear coefficients	Simulated power	Computed power	Error	Computed power	Error	Computed power	Error
Balanced design: {10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10}								
	LC1[a]	0.8952	0.8993	–0.0041	0.9003	–0.0051	0.9064	–0.0112
	LC2[b]	0.8968	0.8979	–0.0011	0.9002	–0.0034	0.9168	–0.0200
	LC3[c]	0.8984	0.8990	–0.0006	0.9004	–0.0020	0.9082	–0.0098
Direct-pairing design:{4, 4, 4, 8, 8, 8, 12, 12, 12, 16, 16, 16}								
	LC1	0.8926	0.8945	–0.0019	0.9013	–0.0087	0.9174	–0.0248
	LC2	0.8984	0.8994	–0.0010	0.9003	–0.0019	0.9093	–0.0109
	LC3	0.8965	0.8993	–0.0028	0.9002	–0.0037	0.9052	–0.0087
Inverse-pairing design: {16, 16, 16, 12, 12, 12, 8, 8, 8, 4, 4, 4}								
	LC1	0.8872	0.8887	–0.0015	0.9006	–0.0134	0.9274	–0.0402
	LC2	0.8744	0.8769	–0.0025	0.9000	–0.0256	0.9509	–0.0765
	LC3	0.8834	0.8858	–0.0024	0.9000	–0.0166	0.9316	–0.0482

Note: [a]LC1: {1/3, 1/3, 1/3, –1/9, –1/9, –1/9, –1/9, –1/9, –1/9, –1/9, –1/9, –1/9}, [b]LC2: {1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, –1/3, –1/3, –1/3}, [c]LC3: {1/6, 1/6, 1/6, 1/6, 1/6, 1/6, 1/6, –1/6, –1/6, –1/6, –1/6, –1/6}

Under the extended inverse-pairing consideration of $G = 12$, the second comparison yields the largest error –0.0256 of all 18 cases. Based on the numerical evidence, the approximate T approach is slightly vulnerable under the circumstance that the sample sizes are inversely paired with heterogeneous variances. Conversely, the exact approach performs extremely well because all absolute errors are less than or equal to 0.0043 for the 18 cases examined here. In addition to the reported assessments with the nominal power 0.90, the accuracy was further justified for the same model configurations with a smaller target power 0.80. To conserve space, the details are not given here. Thus the methodology of exact power calculation is of great potential use. Although it is more computationally intensive than the approximate T approach, it is of little consequence if a computer is employed.

Sample size calculations in moderation analysis

It is an important problem in moderation research to clarify the impact of a moderator on the direction and/or strength of the relationship between a predictor and a criterion variable (Baron & Kenny, 1986). Accordingly, the simplest situation of the moderation analysis is to measure a dichotomous independent variable's effect on the dependent variable varies as a function of another dichotomy. The particular moderation phenomenon is conceptually equivalent to the interaction effect in a 2 × 2 factorial design. Two of the vital factors known to affect power are the sample size and error variance heterogeneity. Hence there is a need to understand the inherent relationship that exists between the desired power performance and the necessary sample size conditional on the heteroscedastic model structure.

For ease of explication, the statistical model of a 2 × 2 heteroscedastic ANOVA design is written as:

$$X_{stl} \sim N(\mu_{st}, \sigma^2_{st}),$$

where X_{stl} represents the independent and normally distributed response variable with expected values μ_{st} and variances σ^2_{st}, μ_{st} is the population mean, and σ^2_{st} is the error variance at level s of A and level t of B for s and t = 1 and 2, and $l = 1, \ldots, M_{st}$. Accordingly, the interaction or moderation effect size between the two factors A and B can be expressed as

$$\psi_I = \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}. \tag{14}$$

The linear contrast

$$\hat{\psi}_I = \bar{X}_{11} - \bar{X}_{12} - \bar{X}_{21} + \bar{X}_{22} \tag{15}$$

is an unbiased estimator of ψ_I where $\bar{X}_{st} = \sum_{l=1}^{M_{st}} X_{stl}/M_{st} \sim N(\mu_{st}, \sigma_{st}^2/M_{st})$ for s and $t = 1$ and 2. It is easily seen that there exists a close resemblance between the linear formulations and statistical properties of $\hat{\psi}_I$ and $\hat{\psi}$ defined in Equations 15 and 3, respectively. Hence, the techniques for obtaining power and sample size for the test of ψ can immediately be applied to compute power and sample size for the test of ψ_I. A detailed account of the related methodology is presented next to document their distinct characteristics in terms of theoretical principles and computational requirements.

The hypothesis testing of H_0: $\psi_I = \psi_{I0}$ versus H_1: $\psi_I \neq \psi_{I0}$ can be conducted with the following statistic

$$T_I^* = \frac{\hat{\psi}_I - \psi_{I0}}{\hat{\Sigma}_I^{1/2}}, \tag{16}$$

where ψ_{I0} is a specified constant, $\hat{\Sigma}_I = \sum_{s=1}^{2}\sum_{t=1}^{2} S_{st}^2/M_{st}$ is the typical estimator of $\Sigma_I = Var(\hat{\psi}_I) = \sum_{s=1}^{2}\sum_{t=1}^{2} \sigma_{st}^2/M_{st}$ and $S_{st}^2 = \sum_{l=1}^{M_{st}} (X_{stl} - \bar{X}_{st})^2/(M_{st} - 1)$ is the sample variance estimator of σ_{st}^2 for s and $t = 1$ and 2. The test procedure rejects H_0 at the significance level α if $|T_I^*| > t_{\hat{\nu}_I, \alpha/2}$ where

$$\hat{\nu}_I = \{\sum_{s=1}^{2}\sum_{t=1}^{2} S_{st}^2/M_{st}\}^2/\{\sum_{s=1}^{2}\sum_{t=1}^{2} S_{st}^4/[M_{st}^2(M_{st}-1)]\}. \tag{17}$$

It follows from Equation 10 that the corresponding power can be approximated by

$$\pi_{TI}(\delta_I) = P\{|t(\nu_I, \delta_I)| > t_{\nu_I, \alpha/2}\}, \tag{18}$$

where

$$\delta_I = (\psi_I - \psi_{I0})/\Sigma_I^{1/2} \text{ and } \nu_I = \{\sum_{s=1}^{2}\sum_{t=1}^{2} \sigma_{st}^2/M_{st}\}^2/\{\sum_{s=1}^{2}\sum_{t=1}^{2} \sigma_{st}^4/[M_{st}^2(M_{st}-1)]\}.$$

Also, the exact power function in Equation 12 is modified as

$$\pi_{EI}(\delta_I) = E_B\{P\{|t(M_T - 4, \delta_I)| > t_{\hat{\nu}_I, \alpha/2} V_I^{1/2}\}, \tag{19}$$

where $M_T = \sum_{s=1}^{2}\sum_{t=1}^{2} M_{st}$, and V_I is the counterpart of V defined in Equation 11. In this case, the expectation $E_B\{\}$ is taken with respect to the joint distribution of $\{B_1, B_2, B_3\}$.

In contrast to the proposed formulations, Guo and Luh (2009, pp. 420-421) exploited Welch's (1938) two-sample statistic for the Behrens–Fisher problem to obtain a distinct method for the detection of an interaction effect H_0: $\psi_I = \psi_{I0}$ versus H_1: $\psi_I \neq \psi_{I0}$. With the definitions of $S_a^2 = (S_{11} + S_{22})^2$ and $S_b^2 = (S_{12} + S_{21})^2$, Guo and Luh's (2009) test procedure rejects H_0: $\psi_I = \psi_{I0}$ at the significance level α if $|T_I^*| > t_{\hat{\nu}_{GL}, \alpha/2}$ where

$$\hat{\nu}_{GL} = (S_a^2/M_a + S_b^2/M_b)^2/[(S_a^2/M_a)^2/(M_a - 1) + (S_b^2/M_b)^2/(M_b - 1)].$$

For notational simplicity, let $\sigma_a^2 = (\sigma_{11} + \sigma_{22})^2$, $\sigma_b^2 = (\sigma_{12} + \sigma_{21})^2$, $M_a = M_{11} + M_{22}$, and $M_b = M_{12} + M_{21}$. In general, they suggested the approximate noncentral t distribution for T_I^*:

$$T_I^* \sim t(\nu_{GL}, \delta_{GL}), \tag{20}$$

where $\nu_{GL} = (\sigma_a^2/M_a + \sigma_b^2/M_b)^2/[(\sigma_a^2/M_a)^2/(M_a - 1) + (\sigma_b^2/M_b)^2/(M_b - 1)]$, $\delta_{GL} = (\psi_I - \psi_{I0})/\Sigma_{GL}^{1/2}$ and $\Sigma_{GL} = \sigma_a^2/M_a + \sigma_b^2/M_b$. Hence, the corresponding power function is of the form

$$\pi_{GL}(\delta_{GL}) = P\{|t(\nu_{GL}, \delta_{GL})| > t_{\nu_{GL}, \alpha/2}\}. \tag{21}$$

To determine sample sizes for testing an interaction effect within the context of heteroscedastic ANOVA, the power functions $\pi_{TI}(\delta_I)$, $\pi_{EI}(\delta_I)$, and $\pi_{GL}(\delta_{GL})$ defined in Equations 18, 19 and 21, respectively, can be employed to calculate the sample sizes (M_{11}, M_{12}, M_{21}, M_{22}) needed to attain the specified power $1 - \beta$ for the chosen significance level α, null effect ψ_{I0}, mean effects $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\}$, error variances $\{\sigma_{11}^2, \sigma_{12}^2, \sigma_{21}^2, \sigma_{22}^2\}$ and designated sample size ratios $\{r_{11}, r_{12}, r_{21}, r_{22}\}$ where $r_{st} = M_{st}/M_{11}$ for s and t = 1 and 2.

To reveal the underlying robustness and deficiency of the contending techniques, numerical assessments were carried out for the model settings in Guo and Luh (2009) in which the mean values and error variances are $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\} = \{71.3, 93.9, 77.1, 93.3\}$ and $\{\sigma_{11}^2, \sigma_{12}^2, \sigma_{21}^2, \sigma_{22}^2\} = \{146.41, 129.96, 207.36, 153.76\}$, respectively. Moreover, seven patterns of sample size ratios were used to assess power and sample size calculations: $\{r_{11}, r_{12}, r_{21}, r_{22}\} = \{1, 1, 1, 1\}, \{1, 1, 2, 2\}, \{1, 2, 1, 2\}, \{2, 1, 2, 1\}, \{2, 2, 1, 1,\}, \{2, 1, 4, 3\}$, and $\{3, 4, 1, 2\}$. Essentially, these designated allocation schemes produce a wide variety of balanced, mixed-pairing, direct-pairing, and inverse-pairing settings with the heterogeneous variances and thus cover a

broader range of situations than the single case considered by Guo and Luh (2009).

With these specifications, the required sample sizes were computed for the abovementioned three approaches with the chosen power value $(1 - \beta) = 0.80$, significance level $\alpha = 0.05$ and null value $\psi_{I0} = 0$. Accordingly, the results associated with the exact and approximate T approaches are essentially identical. Therefore, only the empirical outcomes of the exact approach and Guo and Luh's method are presented in Table 3. The actual powers or attained powers associated with the required sample sizes are computed with the power functions $\pi_{EI}(\delta_I)$ and $\pi_{GL}(\delta_{GL})$. Similar to the empirical study for the meta analysis, simulated powers were obtained by a Monte Carlo simulation study and the results are also presented in Table 3. A numerical analysis was also conducted for the same model configurations with the modified heteroscedastic magnitudes $\{\sigma_{11}^2, \sigma_{12}^2, \sigma_{21}^2, \sigma_{22}^2\}$ = {146.41, 129.96, 207.36, 153.76}/9 ={16.27, 14.44, 23.04, 17.08}. The corresponding results are presented in Table 4.

It follows from the comprehensive results in Tables 3-4 that the necessary sample sizes for Guo and Luh's (2009) method are equal to or slightly smaller than those of the exact approach. However, the powers given by the power function $\pi_{GL}(\delta_{GL})$ seems to be markedly larger than the nominal value 0.80 with two exceptions in the balanced sample size designs. In addition, the errors of their procedure are sizable, especially for the two cases of –0.1199 and –0.1296 associated with the inverse pairing between sample size and error variance in Tables 3 and 4, respectively. Also, it can be shown that $\Sigma_{GL} = \Sigma_I$ and $\delta_{GL} = \delta_I$ when $r_{st} = M_{st}/M_{11} = \sigma_{st}/\sigma_{11}$ for s and $t = 1$ and 2. Due to the dominant role of noncentrality in the power function, the power function $\pi_{GL}(\delta_{GL})$ will give the proper value only when the specific condition is satisfied. On the other hand, the behavior of the exact approach appears to be excellent for the range of model specifications considered here. In particular, the incurred errors of the 14 cases are all within the small range of –0.0096 to 0.0091. Hence the proposed procedure possesses the advantage of general applicability and good accuracy without any imposed restriction to the model configurations. In short, these analytic clarification and numerical evidence show that the suggested approach outperforms Guo and Luh's (2009) method in power calculations and sample size determinations for the Welch-Satterthwaite test of interaction effect within the 2 × 2 heteroscedastic ANOVA framework.

Table 3. Computed sample size, computed power, and simulated power for the test of interaction effect H_0: $\psi_I = 0$ versus H_1: $\psi_I \neq 0$ with $\alpha = 0.05$, nominal power $(1 - \beta) = 0.80$, mean effects $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\} = \{71.3, 93.9, 77.1, 93.3\}$, and error variances $\{\sigma^2_{11}, \sigma^2_{12}, \sigma^2_{21}, \sigma^2_{22}\} = \{146.41, 129.96, 207.36, 153.76\}$

	Exact approach				Guo and Luh (2009)			
Sample size ratios	Sample sizes	Simulated power	Computed power	Error	Sample sizes	Simulated power	Computed power	Error
{1, 1, 1, 1}	(123, 123, 123, 123)	0.7978	0.8010	–0.0032	(123, 123, 123, 123)	0.8030	0.8039	–0.0009
{1, 1, 2, 2}	(88, 88, 176, 176)	0.7993	0.8000	–0.0007	(88, 88, 176, 176)	0.7979	0.8308	–0.0329
{1, 2, 1, 2}	(96, 192, 96, 192)	0.8025	0.8020	–0.0005	(96, 191, 96, 191)	0.7950	0.8604	–0.0654
{2, 1, 2, 1}	(178, 89, 178, 89)	0.7991	0.8013	–0.0022	(178, 89, 178, 89)	0.7972	0.8350	–0.0378
{2, 2, 1, 1}	(194, 194, 97, 97)	0.7999	0.8032	–0.0033	(193, 193, 97, 97)	0.8050	0.8639	–0.0589
{2, 1, 4, 3}	(120, 60, 240, 180)	0.8023	0.8059	–0.0036	(118, 59, 236, 177)	0.7974	0.8696	–0.0722
{3, 4, 1, 2}	(213, 284, 71, 142)	0.8053	0.8012	0.0041	(211, 281, 71, 141)	0.8009	0.9208	–0.1199

Table 4. Computed sample size, computed power, and simulated power for the test of interaction effect H_0: $\psi_I = 0$ versus H_1: $\psi_I \neq 0$ with $\alpha = 0.05$, nominal power $(1 - \beta) = 0.80$, mean effects $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\} = \{71.3, 93.9, 77.1, 93.3\}$, and error variances $\{\sigma^2_{11}, \sigma^2_{12}, \sigma^2_{21}, \sigma^2_{22}\} = \{16.27, 14.44, 23.04, 17.08\}$

	Exact approach				Guo and Luh (2009)			
Sample size ratios	Sample sizes	Simulated power	Computed power	Error	Sample sizes	Simulated power	Computed power	Error
{1, 1, 1, 1}	(15, 15, 15, 15)	0.8244	0.8233	0.0011	(15, 15, 15, 15)	0.8254	0.8280	–0.0026
{1, 1, 2, 2}	(11, 11, 22, 22)	0.8275	0.8282	–0.0007	(11, 11, 21, 21)	0.8267	0.8520	–0.0253
{1, 2, 1, 2}	(12, 24, 12, 24)	0.8208	0.8270	–0.0062	(11, 22, 11, 22)	0.8028	0.8628	–0.0600
{2, 1, 2, 1}	(22, 11, 22, 11)	0.8344	0.8253	0.0091	(21, 11, 21, 11)	0.8244	0.8520	–0.0276
{2, 2, 1, 1}	(24, 24, 12, 12)	0.8154	0.8250	–0.0096	(22, 22, 11, 11)	0.7942	0.8628	–0.0686
{2, 1, 4, 3}	(16, 8, 32, 24)	0.8438	0.8498	–0.0060	(14, 7, 27, 21)	0.8099	0.8774	–0.0675
{3, 4, 1, 2}	(27, 36, 9, 18)	0.8162	0.8193	–0.0031	(24, 32, 8, 16)	0.7912	0.9208	–0.1296

DISCUSSION AND CONCLUSIONS

In view of its practical value in the context of heteroscedastic ANOVA designs, this article presents two approaches to power and sample calculations for the Welch-Satterthwaite test of linear combinations of group means. The approximate method provides a transparent formulation and relies on a noncentral *t* distribution, whereas the exact procedure is of theoretical importance and involves a Beta mixture of noncentral *t* distributions. It can be justified that the approaches are asymptotically equivalent as sample size goes to infinity. However, their finite-sample properties can be substantially different and the respective power functions may yield markedly different results for relative small samples and certain model settings. It is shown here that while computation is slightly involved when using the exact procedure, the extra complexity is outweighed by its superiority in accuracy.

It is vital to ensure that the underlying properties of the power and sample size procedure are well understood so that a well-supported and useful recommendation can be offered for empirical studies. The extensive usefulness and great diversity of the suggested power and sample size procedures are illustrated with two applications in meta and moderation analyses. Detailed analytic explication and numerical assessment are presented to demonstrate the prominent advantage of the proposed procedures and the potential deficiency of existing methods. In particular, the failure to accommodate the stochastic nature of error variances and the absence to embed the diverse structure of sample sizes are restrictions of the current methods of Hedges and Pigott (2001) and Guo and Luh (2009) for meta analysis and moderation analysis, respectively. Consequently, the suggested power and sample size procedures update and expand upon current work in the literature and the developed computer programs can facilitate the application of the suggested algorithms.

This study focuses on the appropriate procedure for testing the linear combination of group means of independent normal distributions with possibly unequal error variances. Moreover, according to the findings of Vallejo, Ato, and Fernandez (2010), the Welch-James procedure is robust to departures from normality assumption when the distribution type is symmetric with moderate degree of kurtosis. Therefore, the related Welch–Satterthwaite test procedure is still of practical interest and usefulness. On the other hand, the established class of generalized linear models offers an excellent alternative for analyzing data when the normality and homogeneity assumptions are not tenable. Related details and follow-up

procedures can be found in McCullagh and Nelder (1989) and McCulloch, Searle, and Neuhaus (2008).

REFERENCES

Aguinis, H. (2004). *Regression analysis for categorical moderators*. New York, NY: Guilford Press.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-Year Review. *Journal of Applied Psychology*, 90, 94-107.

Aguinis, H., Boik, R. J., & Pierce, C. A. (2001). A generalized solution for approximating the power to detect effects of categorical moderator variables using multiple regression. *Organizational Research Methods*, 4, 291-323.

Aguinis, H., & Pierce, C. A. (1998). Heterogeneity of error variance and the assessment of moderating effects of categorical variables: A conceptual review. *Organizational Research Methods*, 1, 296-314.

Aguinis, H., Pierce, C. A., Bosco, F. A., Dalton, D. R., & Dalton, C. M. (2011). Debunking myths and urban legends about meta-analysis. *Organizational Research Methods*, 14, 306-331.

Alexander, R. A., & DeShon, R. P. (1994). Effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin,* 115*,* 308-314.

American Psychological Association Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: why do we need them? What might they be? *American Psychologist*, 63, 839-851.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173-1182.

Bond, C. F., Jr., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods*, 8, 406-418.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Mahwah, NJ: Erlbaum.

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology*, 51, 115-134.

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

Guo, J. H., & Luh, W. M. (2009). On sample size calculation for 2 × 2 fixed-effect ANOVA when variances are unknown and possibly unequal. *British Journal of Mathematical and Statistical Psychology*, 62, 417-425.

Hartung, J., Argac, D., & Makambi, K. H. (2002). Small sample properties of tests on homogeneity in one-way ANOVA and meta-analysis. *Statistical Papers*, 43, 197-235.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, 6, 203-217.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd.). New York, NY: Academic Press.

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.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd ed., Vol. 2). New York, NY: Wiley.

Kim, S. H., & Cohen, A. S. (1998). On the Behrens-Fisher problem: A review. *Journal of Educational and Behavioral Statistics*, 23, 356-377.

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

Levy, K. J. (1978). A priori contrasts under conditions of variance heterogeneity. *Journal of Experimental Education*, 47, 42-45.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall/CRC.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. W. (2008). *Generalized, linear, and mixed models* (2nd ed.). New York: Wiley.

Muncer, S., Taylor, S., & Craigie, M. (2002). Power dressing and meta-analysis: Incorporating power analysis into meta-analysis. *Journal of Advanced Nursing*, 38, 274-280.

Ruscio, J., & Roche, B. (2012). Variance heterogeneity in published psychological research: A review and a new index. *Methodology*, 8, 1-11.

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

Satterthwaite, F. E. (1946). An approximate distribution of estimate of variance components. *Biometrics Bulletin*, 2, 110-114.

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

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, 35, 215-247.

Vallejo, G., Ato, M., & Fernandez, M. P. (2010). A robust approach for analyzing unbalanced factorial designs with fixed levels. *Behavior Research Methods*, 42, 607-617.

Wahlsten, D. (1991). Sample size to detect a planned contrast and a one degree-of-freedom interaction effect. *Psychological Bulletin*, 110, 587-595.

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. (1947). The generalization of Students' problem when several different population variances are involved. *Biometrika*, 34, 28-35.

APPENDIX A

SAS IML program for computing the attained power for Welch-Satterthwaite's test

```
PROC IML;PRINT "POWER CALCULATIONS";
*USER SPECIFICATIONS;
*TYPE I ERROR;           ALPHA=0.05;
*GROUP MEANS;                 MUVEC={2.18 0 0 0};
*GROUP VARIANCES;        VARVEC={1 4 9 16};
*GROUP SAMPLE SIZES;          NVEC={10 10 10 10};
*COEFFICIENTS;           CVEC={1 -1 -1 -1}/{1 3 3
3};
*END OF SPECIFICATIONS;

G=NCOL(VARVEC);NT=SUM(NVEC);PSI=CVEC*MUVEC`;
VARPSI=(CVEC##2)*(VARVEC/NVEC)`;
DELTA=PSI/SQRT(VARPSI);DF=NT-G;DFVEC=NVEC-1;
*APPRO METHOD;
KV=(CVEC##2)#VARVEC/NVEC;V1=SUM(KV)##2;
V2=SUM((KV##2)/(NVEC-1));DFAP=V1/V2;CRIT=TINV(1-
ALPHA/2,DFAP);
APP=CDF('T',-
CRIT,DFAP,DELTA)+SDF('T',CRIT,DFAP,DELTA);
PRINT 'APPROXIMATE T: POWER' APP[FORMAT=8.4];
*EXACT METHOD;
SEED=1001;CALL STREAMINIT(SEED);REPN=10000;
DF1=CUSUM(DFVEC[1,1:G-1]);DF2=DFVEC[1,2:G];EP=0;
DO I=1 TO REPN;BVEC=RAND('BETA',DF1`/2,DF2`/2);
AVEC=J(G,1,0);AVEC[1,1]=EXP(SUM(LOG(BVEC)));
DO IG=2 TO G-1;
AVEC[IG,1]=(1-BVEC[IG-1,1])#EXP(SUM(LOG(BVEC[IG:G-
1,1])));END;
AVEC[G,1]=1-BVEC[G-1,1];
LBVEC=(CVEC##2)#VARVEC/(NVEC#DFVEC);
DFV=(LBVEC*AVEC)##2/(((LBVEC##2)/DFVEC)*(AVEC##2))
;
CRIT=TINV(ALPHAU,DFV);H=DF#(LBVEC*AVEC)/VARPSI;
EP=EP+SDF('T',CRIT#SQRT(H),DF,DELTA)+CDF('T',-
CRIT#SQRT(H),DF,
DELTA);
```

```
END;EXP=EP/REPN;PRINT 'EXACT METHOD: POWER'
EXP[FORMAT=8.4];
QUIT;
```

APPENDIX B

SAS IML program for computing the required sample size for Welch-Satterthwaite's test

```
PROC IML;PRINT "SAMPLE SIZE CALCULATIONS";
*USER SPECIFICATIONS;
*TYPE I ERROR;           ALPHA=0.05;
*NOMINAL POWER;               POWER=0.80;
*GROUP MEANS;                 MUVEC={71.3 93.9 77.1
93.3};
*GROUP VARIANCES;        VARVEC={146.41 129.96
207.36 153.76};
*SAMPLE SIZE RATIOS;          RVEC={1 1 1 1};
*COEFFICIENTS;           CVEC={1 -1 -1 1};
*END OF SPECIFICATIONS;

PSI=CVEC*MUVEC`;G=NCOL(VARVEC);
*APPRO METHOD;
N=3;DO UNTIL
(EPAP>POWER);N=N+1;NVEC=N#RVEC;DFVEC=NVEC-1;
VARPSI=(CVEC##2)*(VARVEC/NVEC)`;KV=(CVEC##2)#VARVE
C/NVEC;
V1=SUM(KV)##2;V2=SUM((KV##2)/DFVEC);DFAP=V1/V2;
CRIT=TINV(1-ALPHA/2,DFAP);DELTA=PSI/SQRT(VARPSI);
EPAP=CDF('T',-
CRIT,DFAP,DELTA)+SDF('T',CRIT,DFAP,DELTA);END;
PRINT 'APPROXIMATE T: POWER & N' EPAP[FORMAT=8.4]
NVEC[FORMAT=4.0];
*EXACT METHOD;
SEED=1001;CALL STREAMINIT(SEED);REPN=10000;
N=MAX(NVEC[1,RVEC[1,>:<]]-5,3);DO UNTIL
(EPEX>POWER);
N=N+1;NVEC=N#RVEC;NT=SUM(NVEC);DFVEC=NVEC-1;DF=NT-
G;
VARPSI=(CVEC##2)*(VARVEC/NVEC)`;DELTA=PSI/SQRT(VAR
PSI);
```

```
DF1=CUSUM(DFVEC[1,1:G-1]);DF2=DFVEC[1,2:G];EP=0;
DO I=1 TO
REPN;BVEC=RAND('BETA',DF1`/2,DF2`/2);AVEC=J(G,1,0)
;
AVEC[1,1]=EXP(SUM(LOG(BVEC)));DO IG=2 TO G-1;
AVEC[IG,1]=(1-BVEC[IG-1,1])#EXP(SUM(LOG(BVEC[IG:G-
1,1])));END;
AVEC[G,1]=1-BVEC[G-
1,1];LBVEC=(CVEC##2)#VARVEC/(NVEC#DFVEC);
DFV=(LBVEC*AVEC)##2/(((LBVEC##2)/DFVEC)*(AVEC##2))
;
CRIT=TINV(1-ALPHA/2,DFV);H=DF#(LBVEC*AVEC)/VARPSI;
EP=EP+CDF('T',-
CRIT#SQRT(H),DF,DELTA)+SDF('T',CRIT#SQRT(H),DF,
DELTA);END;EPEX=EP/REPN;END;
PRINT 'EXACT METHOD: POWER & N' EPAP[FORMAT=8.4]
NVEC[FORMAT=4.0];
QUIT;
```

(Manuscript received: 1 September 2014; accepted: 3 December 2014)