Behav Res (2013) 45:25–37
DOI 10.3758/s13428-012-0228-7

Confidence intervals and sample size calculations for the weighted eta-squared effect sizes in one-way heteroscedastic ANOVA

Gwowen Shieh

Published online: 18 July 2012

Abstract Effect size reporting and interpreting practices have been extensively recommended in academic journals when primary outcomes of all empirical studies have been analyzed. This article presents an alternative approach to constructing confidence intervals of the weighted eta-squared effect size within the context of one-way heteroscedastic ANOVA models. It is shown that the proposed interval procedure has advantages over an existing method in its theoretical justification, computational simplicity, and numerical performance. For design planning, the corresponding sample size procedures for precise interval estimation of the weighted eta-squared association measure are also delineated. Specifically, the developed formulas compute the necessary sample sizes with respect to the considerations of expected confidence interval width and tolerance probability of interval width within a designated value. Supplementary computer programs are provided to aid the implementation of the suggested techniques in practical applications of ANOVA designs when the assumption of homogeneous variances is not tenable.

Keywords Heteroscedasticity · Precision · Welch's statistic

The analysis of variance (ANOVA) compares the impact of categorical design factors on a continuous response variable in order to determine whether differences exist among the treatment groups. To indicate how much the knowledge of a treatment group improves prediction of the response variable, several strength-of-association measures have been suggested in the literature, such as the estimators of $\widehat{\eta}^2$, $\widehat{\varepsilon}^2$, and $\widehat{\omega}^2$ (Grissom & Kim, 2005; Hays, 1994; Keppel, 1991; Kline, 2004; Maxwell & Delaney, 2004). They can be interpreted as a proportion that reflects how much variability in the response variable is associated with the variation in the treatment levels. The underlying rationale and discrepancy of these three association measures have been discussed in Fern and Monroe (1996), Glass and Hakstian (1969), Maxwell, Camp, and Arvey (1981), and Richardson (1996). Accordingly, the sample eta-squared $\widehat{\eta}^2$, is one of the most commonly reported association indices in practical applications of ANOVA. Detailed discussion and related issues can be found in Cohen (1973), Haase (1983), Levine and Hullett (2002), Olejnik and Algina (2003), Pierce, Block, and Aguinis (2004), and Richardson (2011).

One important assumption underlying the ANOVA designs is that of equal population variances. Violation of the homogeneity-of-variance assumption has been the target of criticism in applications of ANOVA. For example, Grissom (2000) emphasized that there are theoretical reasons to expect, and empirical results to document, the existence of heteroscedasticity in clinical studies. Moreover, Grissom and Kim (2005, pp. 10–14) provided additional explanations for the intrinsic causes of variance heterogeneity in real data. The practical importance and methodological complexity of the problem has incurred numerous attempts to develop various parametric and nonparametric alternative procedures to counter the effects of heteroscedasticity (Keselman et al., 1998; Lix, Keselman, & Keselman, 1996). It follows from the comprehensive reviews of Grissom (2000), Harwell, Rubinstein, Hayes, and Olds (1992), and Tomarken and Serlin (1986) that the Welch (1951) procedure is a widely accepted technique for correcting for variance heterogeneity. Specifically, it has advantages over other, contending methods in its overall

Electronic supplementary material The online version of this article (doi:10.3758/s13428-012-0228-7) contains supplementary material, which is available to authorized users.

G. Shieh (✉)
Department of Management Science,
National Chiao Tung University,
Hsinchu, Taiwan 30010, Republic of China
e-mail: gwshieh@mail.nctu.edu.tw

performance, computational ease, and general availability in statistical computer packages.

According to the general discussions of Breaugh (2003), Ferguson (2009), Fern and Monroe (1996), Kirk (1996), Richardson (1996), and Vacha-Haase and Thompson (2004), group difference and strength of association (or correlation ratio) are two of the major classes of effect sizes in practical applications. It should be noted that these conventional measures and conversion formulas of effect size make the standard assumption of homoscedasticity. However, Grissom and Kim (2001) were concerned by the frequent occurrence of variance heterogeneity in many areas of research. Therefore, they advised caution regarding the robustness and appropriateness to heteroscedasticity of current effect size measures that assume homogeneous variances. Ultimately, the prevailing heteroscedastic phenomenon has prompted different conceptions and definitions of effect size. This issue has important implications for interpreting the meaning of effect sizes, but it has received relatively limited attention in the methodological literature. In the case of comparing the means for two different groups, Keselman, Algina, Lix, Wilcox, and Deering (2008) emphasized that the well-known Cohen's (1988) standardized mean difference is not appropriate when the homogeneity-of-variance assumption is violated and discussed several alternative definitions of a standardized mean difference effect size to circumvent the untenable assumption of equal variances. Accordingly, the diversity of suggested measures in Grissom and Kim (2001) and Keselman et al. (2008) implies that there is no firm consensus as to the definition of a standardized mean difference effect size in the presence of heteroscedasticity. The various indices of standardized mean difference simply represent different quantities, each with their unique features, and may prove to be useful in a given application. Also, see Bonett (2008) for a discussion of standardized linear contrasts of means with different standardizers in heteroscedastic ANOVA.

Although numerous approaches have been suggested to tackle the practical and complex issue of heteroscedasticity, Keselman et al. (2008) presented a unified formulation of approximate degrees of freedom (ADF) procedures within the context of general linear models. Essentially, the prescribed Welch (1951) method for comparing mean equality can be obtained from the general ADF perspective. Thorough treatment and related applications of the Welch statistic and other ADF methods are also described in Lix and Keselman (1995). To further circumvent the sensitivity of traditional methods for comparing mean equality with nonnormality, in additional to heteroscedasticity, Keselman et al. (2008) emphasized the applications of ADF procedures with robust estimators of both central tendency and variability. As was noted earlier, Keselman et al. (2008) pointed out the apparently problematic outcome of Cohen's standardized mean difference when variance heterogeneity is present. More important, they also explicated the vital differences and merits of various definitions and estimators of standardized mean difference effect size. It is important to note that Bonett (2008) has suggested several useful standardized linear contrasts of means within the heteroscedastic ANOVA setting. However, no explicit formulations were provided for a population strength of association effect size measure such as the counterpart of eta-squared η^2 in traditional ANOVA designs.

In contrast, a weighted formulation of effect size θ was proposed in Kulinskaya and Staudte (2006, Equation 2) to accommodate the underlying characteristics of possibly unequal error variances and unbalanced group sizes for a one-way heteroscedastic ANOVA. It was also noted in Kulinskaya and Staudte that when variances are equal, the weighted effect size θ reduces to the widely recognized effect size index f^2 in traditional one-way ANOVA (Cohen, 1988, p. 274). Moreover, the weighted effect size can be readily transformed to a weighted coefficient of determination $\rho^2 = \theta/(1 + \theta)$, just as the prevalent strength of association measure of eta-squared $\eta^2 = f^2/(1 + f^2)$ is a one-to-one function of effect size f^2. Hence, the coefficient of determination ρ^2 resembles the eta-squared index η^2 for representing the proportion of explained variance within the context of a one-way heteroscedastic ANOVA. In view of the appealing features and versatile usefulness for the definitions of weighted effect size θ and weighted coefficient of determination ρ^2, Kulinskaya and Staudte presented an approximate interval estimation procedure for θ on the basis of a shifted and rescaled chi-square transformation of Kulinskaya, Staudte, and Gao (2003). Clearly, it follows from the monotone transformation $\rho^2 = \theta/(1 + \theta)$ that a desired confidence interval of ρ^2 can be immediately constructed from the obtained interval estimate of θ. Moreover, several simulation studies were conducted in Kulinskaya and Staudte to examine the performance of the suggested technique. According to the numerical results, they concluded that the interval procedure is surprisingly accurate in terms of the nominal coverage probability, except for very small sample sizes. Also, the coverage probability tends to exceed the nominal level when the magnitude of the weighted effect size is small.

Despite the aforementioned arguments and findings in Kulinskaya and Staudte (2006), the following four caveats to their interval estimation method should be noted. First, their confidence interval of θ is constructed from a shifted and rescaled chi-square approximate distribution for an estimator of the explained sum of squares (Kulinskaya & Staudte, 2006, Equation 11) as an alternative method for computing the distribution function of Welch's statistic (Kulinskaya et al., 2003, Equation 6). Since they have not successfully obtained a pivotal quantity with the shifted and rescaled chi-square distribution, further approximations are

made to the shifted and rescaled parameters in order to compute the involved confidence limits. It is notable that the statistical presentations and algebraic expressions for their interval estimators of θ are fairly complicated and the calculation of confidence intervals requires a special-purpose computer program for performing the necessary computation. Therefore, the complexity may result in limited acceptance in application. Second, the exact interval procedure for the association strength effect size η^2 in homoscedastic ANOVA was repeatedly described in Fleishman (1980), Kelley (2007), Kline (2004), Odgaard and Fowler (2010), Smithson (2001), and Steiger (2004). Specifically, the exact approach employs a noncentrality inversion technique of F distributions and is called the "cumulative distribution function" pivotal method in Casella and Berger (2002, Section 9.2.3) and Mood, Graybill, and Boes (1974, Section 4.2). Corresponding routines and scripts for the computations of noncentral F distributions and exact confidence intervals are available in popular software packages such as R, SAS, SPSS, and STATISTICA. Kulinskaya and Staudte's approximate interval estimation method deals with the more general target effect size of the weighted coefficient of determination ρ^2, which subsumes the association strength eta-squared η^2 as a special case. However, the shifted and rescaled chi-square transformation of Kulinskaya and Staudte does not conform to the established noncentrality inversion procedure. Thus, the failure to embed the confidence intervals of ρ^2 and η^2 in a unified principle is an obvious limitation of the existing method of Kulinskaya and Staudte.

Third, the empirical investigation in Kulinskaya and Staudte (2006) seems to give practically acceptable results for a wide range of two-sample settings in Tables 1–4. But a closer inspection of their numerical performance for three-group situations in Tables 5–7 suggests that the coverage probability tends to increase with decreasing weighted effect size θ. In other words, the resulting two-sided confidence interval may be too wide when the population weighted effect size is small, whereas the reported interval estimate is probably not wide enough to attain the desired confidence level if the magnitude of underlying weighted effect size is large. Consequently, the unknown magnitude of the underlying population weighted effect size could distort the coverage performance of the interval estimates. Potential users should be aware of the robustness problem associated with the approximate formula of Kulinskaya and Staudte. Fourth, they particularly remarked that the actual distribution of the principal statistic proposed in Kulinskaya et al. (2003) is highly skewed and does not converge rapidly enough to a noncentral chi-square distribution. This implies that their interval procedure gives rise to asymmetric confidence intervals for θ or that the resulting two-sided interval estimates are not equidistant around the principal statistic. However, the accuracy of the one-sided confidence intervals and the sensitivity to heteroscedasticity and unbalanced structures of Kulinskaya and Staudte are essentially unknown. The existing results for two-sided confidence intervals in Kulinskaya and Staudte are not detailed enough to elucidate these fundamental issues. It seems prudent, therefore, to confirm that the properties of their technique are well clarified before it can be adopted as a general procedure.

According to the editorial guidelines and methodological recommendations of several prominent educational and psychological journals, it is necessary to include some measures of effect size and confidence intervals for all primary outcomes (Alhija & Levy, 2009; Odgaard & Fowler, 2010; Sun, Pan, & Wang, 2010). Furthermore, Maxwell, Kelley, and Rausch (2008) advocated the desirability of achieving required precision in parameter estimation and emphasized the importance of sample size planning in constructing precise

Table 1 Simulated coverage probability, error, and average width of the approximate confidence intervals for weighted signal-to-noise ratio λ^* when $\left(\sigma_1^2, \sigma_2^2, \sigma_3^2\right) = (1, 1/2, 4)$, $(N_1, N_2, N_3) = (10, 10, 10)$, $(\mu_1, \mu_2, \mu_3) = (0, 1, 1)$, $(0, 1, 2)$, $(0, 1, 3)$, and $(0, 1, 4)$

	The proposed approach							Kulinskaya and Staudte (2006)						
λ^*	Upper 95 % CI	Error	Lower 95 % CI	Error	Two-sided 90 % CI	Error	Average width	Upper 95 % CI	Error	Lower 95 % CI	Error	Two-sided 90 % CI	Error	Average width
0.23	.9597	.0097	.9485	−.0015	.9082	.0082	0.7202	.9625	.0125	.9870	.0370	.9495	.0495	0.8637
0.36	.9542	.0042	.9522	.0022	.9064	.0064	0.9460	.9541	.0041	.9817	.0317	.9358	.0358	1.1230
0.64	.9461	−.0039	.9538	.0038	.8999	−.0001	1.3907	.9383	−.0117	.9816	.0316	.9199	.0199	1.6316
1.08	.9387	−.0113	.9517	.0017	.8904	−.0096	2.0274	.9199	−.0301	.9790	.0290	.8989	−.0011	2.3555
λ^*	Upper 97.5 % CI	Error	Lower 97.5 % CI	Error	Two-sided 95 % CI	Error	Average width	Upper 97.5 % CI	Error	Lower 97.5 % CI	Error	Two-sided 95 % CI	Error	Average width
0.23	.9799	.0049	.9744	−.0006	.9543	.0043	0.8609	.9760	.0010	.9999	.0249	.9759	.0259	1.0393
0.36	.9774	.0024	.9764	.0014	.9538	.0038	1.1287	.9718	−.0032	.9939	.0189	.9657	.0157	1.3461
0.64	.9724	.0026	.9772	.0022	.9496	−.0004	1.6606	.9575	−.0175	.9934	.0184	.9509	.0009	1.9604
1.08	.9672	.0078	.9753	.0003	.9425	−.0075	2.4228	.9383	−.0367	.9919	.0169	.9302	−.0198	2.8437

Table 2 Simulated coverage probability, error, and average width of the approximate confidence intervals for weighted eta-squared η^{2*} = 1/6 when $\left(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2\right)$ = (1, 4, 9, 16), (N_1, N_2, N_3, N_4) = (15, 15, 15, 15), (6, 12, 18, 24), (24, 18, 12, 6), and mean structures (μ_1, μ_2, μ_3, μ_4) = {−1, 0, 0, 1}/c, {−3, −1, 1, 3}/c, and {−1, −1, 1, 1}/c denoted by μ = 1, 2, and 3, respectively

	The proposed approach							Kulinskaya and Staudte (2006)						
μ	Upper 95 % CI	Error	Lower 95 % CI	Error	Two-sided 90 % CI	Error	Average width	Upper 95 % CI	Error	Lower 95 % CI	Error	Two-sided 90 % CI	Error	Average width
(N_1, N_2, N_3, N_4) = (15, 15, 15, 15)														
1	.9550	.0050	.9526	.0026	.9076	.0076	0.2771	.9442	−.0058	.9892	.0392	.9334	.0334	0.3168
2	.9555	.0055	.9517	.0017	.9072	.0072	0.2779	.9454	−.0046	.9892	.0392	.9346	.0346	0.3192
3	.9509	.0009	.9529	.0029	.9038	.0038	0.2769	.9468	−.0032	.9903	.0403	.9371	.0371	0.3229
(N_1, N_2, N_3, N_4) = (6, 12, 18, 24)														
1	.9453	−.0047	.9493	−.0007	.8946	−.0054	0.2778	.9381	−.0119	.9892	.0392	.9273	.0273	0.3276
2	.9461	−.0039	.9512	.0012	.8973	−.0027	0.2778	.9364	−.0136	.9894	.0394	.9258	.0258	0.3238
3	.9558	.0058	.9501	.0001	.9059	.0059	0.2783	.9439	−.0061	.9880	.0380	.9319	.0319	0.3191
(N_1, N_2, N_3, N_4) = (24, 18, 12, 6)														
1	.9579	.0079	.9532	.0032	.9111	.0111	0.2977	.9384	−.0116	.9918	.0418	.9302	.0302	0.3409
2	.9497	−.0003	.9529	.0029	.9026	.0026	0.2984	.9357	−.0143	.9926	.0426	.9283	.0283	0.3506
2	.9389	−.0111	.9494	−.0006	.8883	−.0117	0.2987	.9282	−.0218	.9930	.0430	.9212	.0212	0.3601

confidence intervals. It is worthwhile to note that the notion of coefficient of determination ρ^2 in multiple linear regression is more commonly referred to as the eta-squared index η^2 to represent the strength of association in the context of ANOVA settings. For clarity, the weighted coefficient of determination in Kulinskaya and Staudte (2006) is therefore referred to as the weighted eta-squared in the remainder of this article. In an effort to improve the quality of research analysis and design, this article presents interval estimation and sample size procedures for the weighted eta-squared effect sizes in one-way heteroscedastic ANOVAs. On the basis of the approximate noncentral F distribution for Welch's statistic in Levy (1978), we apply the cumulative distribution function pivotal method to construct well-supported confidence intervals for the weighted eta-squared effect sizes. The proposed general methodology not only enables a transparent and concise exposition of the inherent statistical arguments and properties, but also combines the interval procedures for both homoscedastic and heteroscedastic ANOVA designs into one unified framework. The accuracy of the suggested approach is evaluated by the computed confidence interval corresponding to the nominal coverage probability and the actual probability of coverage it achieves. Extensive numerical examinations

Table 3 Simulated coverage probability, error, and average width of the approximate confidence intervals for weighted eta squared η^{2*} = 1/6 when $\left(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2\right)$ = (1, 4, 9, 16), (N_1, N_2, N_3, N_4) = (15, 15, 15, 15), (6, 12, 18, 24), (24, 18, 12, 6), and mean structures (μ_1, μ_2, μ_3, μ_4) = {−1, 0, 0, 1}/c, {−3, −1, 1, 3}/c, and {−1, −1, 1, 1}/c denoted by μ = 1, 2, and 3, respectively

	The proposed approach							Kulinskaya and Staudte (2006)						
μ	Upper 97.5 % CI	Error	Lower 97.5 % CI	Error	Two-sided 95 % CI	Error	Average width	Upper 97.5 % CI	Error	Lower 97.5 % CI	Error	Two-sided 95 % CI	Error	Average width
(N_1, N_2, N_3, N_4) = (15, 15, 15, 15)														
1	.9787	.0037	.9774	.0024	.9561	.0061	0.3212	.9697	−.0053	.9976	.0226	.9673	.0173	0.3647
2	.9795	.0045	.9740	−.0010	.9535	.0035	0.3222	.9709	−.0041	.9973	.0223	.9682	.0182	0.3675
3	.9728	−.0022	.9775	.0025	.9503	.0003	0.3210	.9674	−.0076	.9974	.0224	.9648	.0148	0.3711
(N_1, N_2, N_3, N_4) = (6, 12, 18, 24)														
1	.9716	−.0034	.9749	−.0001	.9465	−.0035	0.3221	.9642	−.0108	.9971	.0221	.9613	.0113	0.3759
2	.9719	−.0031	.9740	−.0010	.9459	−.0041	0.3221	.9624	−.0126	.9970	.0220	.9594	.0094	0.3718
3	.9804	.0054	.9736	−.0014	.9540	.0040	0.3226	.9709	−.0041	.9963	.0213	.9672	.0172	0.3673
(N_1, N_2, N_3, N_4) = (24, 18, 12, 6)														
1	.9789	.0039	.9754	.0004	.9543	.0043	0.3437	.9605	−.0145	.9987	.0237	.9592	.0092	0.3895
2	.9750	.0000	.9769	.0019	.9519	.0019	0.3447	.9577	−.0173	.9987	.0237	.9564	.0064	0.3992
3	.9664	−.0086	.9744	−.0006	.9408	−.0092	0.3452	.9495	−.0255	.9989	.0239	.9484	−.0016	0.4090

Table 4 Simulated coverage probability, error, and average width of the approximate confidence intervals for weighted eta-squared $\eta^{2*} = 1/6$ when $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 1, 1, 1)$, $(N_1, N_2, N_3, N_4) = (15, 15, 15, 15)$, $(6, 12, 18, 24)$, $(24, 18, 12, 6)$, and mean structures $(\mu_1, \mu_2, \mu_3, \mu_4) = \{-1, 0, 0, 1\}/c$, $\{-3, -1, 1, 3\}/c$, and $\{-1, -1, 1, 1\}/c$ denoted by $\mu = 1, 2$, and 3, respectively

	The proposed approach							Kulinskaya and Staudte (2006)						
μ	Upper 95 % CI	Error	Lower 95 % CI	Error	Two-sided 90 % CI	Error	Average width	Upper 95 % CI	Error	Lower 95 % CI	Error	Two-sided 90 % CI	Error	Average width
$(N_1, N_2, N_3, N_4) = (15, 15, 15, 15)$														
1	.9484	−.0016	.9482	−.0018	.8966	−.0034	0.2737	.9446	−.0054	.9868	.0368	.9314	.0314	0.3174
2	.9509	.0009	.9531	.0031	.9040	.0040	0.2747	.9429	−.0071	.9894	.0394	.9323	.0323	0.3170
3	.9528	.0028	.9552	.0052	.9080	.0080	0.2748	.9444	−.0056	.9907	.0407	.9351	.0351	0.3151
$(N_1, N_2, N_3, N_4) = (6, 12, 18, 24)$														
1	.9339	−.0161	.9521	.0021	.8860	−.0140	0.2887	.9252	−.0248	.9918	.0418	.9170	.0170	0.3428
2	.9364	−.0136	.9524	.0024	.8888	−.0112	0.2895	.9256	−.0244	.9929	.0429	.9185	.0185	0.3443
3	.9468	−.0032	.9510	.0010	.8978	−.0022	0.2896	.9301	−.0199	.9893	.0393	.9194	.0194	0.3392
$(N_1, N_2, N_3, N_4) = (24, 18, 12, 6)$														
1	.9397	−.0103	.9512	.0012	.8909	−.0091	0.2888	.9308	−.0192	.9914	.0414	.9222	.0222	0.3423
2	.9367	−.0133	.9509	.0009	.8876	−.0124	0.2891	.9285	−.0215	.9919	.0419	.9204	.0204	0.3445
3	.9469	−.0031	.9512	.0012	.8981	−.0019	0.2896	.9315	−.0185	.9921	.0421	.9236	.0236	0.3388

were conducted to reveal the advantages in coverage probability and interval width of the proposed approach over the approximate transformation method of Kulinskaya and Staudte under a variety of group mean configurations, variance patterns, and sample size structures. Moreover, sample size calculations for precise interval estimation of weighted eta-squared effect sizes are also demonstrated in two different perspectives. One approach gives the minimum sample size, such that the expected confidence interval width is within the designated bound. The other approach provides the sample size needed to guarantee, with a given tolerance probability, that the width of a confidence interval will not exceed the planned range. To facilitate the recommended procedures in empirical applications, SAS computer programs are developed for computing the confidence intervals of the weighted eta-squared association strength and the necessary sample sizes for designated interval precision criteria in planning research designs.

Interval estimation of weighted eta-squared

Consider the one-way heteroscedastic ANOVA model in which the observations X_{ij} are assumed to be independent

Table 5 Simulated coverage probability, error, and average width of the approximate confidence intervals for weighted eta-squared $\eta^{2*} = 1/6$ when $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 1, 1, 1)$, $(N_1, N_2, N_3, N_4) = (15, 15, 15, 15)$, $(6, 12, 18, 24)$, $(24, 18, 12, 6)$, and mean structures $(\mu_1, \mu_2, \mu_3, \mu_4) = \{-1, 0, 0, 1\}/c$, $\{-3, -1, 1, 3\}/c$, and $\{-1, -1, 1, 1\}/c$ denoted by $\mu = 1, 2$, and 3, respectively

	The proposed approach							Kulinskaya and Staudte (2006)						
μ	Upper 97.5 % CI	Error	Lower 97.5 % CI	Error	Two-sided 95 % CI	Error	Average width	Upper 97.5 % CI	Error	Lower 97.5 % CI	Error	Two-sided 95 % CI	Error	Average width
$(N_1, N_2, N_3, N_4) = (15, 15, 15, 15)$														
1	.9716	−.0034	.9740	−.0010	.9456	−.0044	0.3175	.9650	−.0100	.9968	.0218	.9618	.0118	0.3653
2	.9773	.0023	.9772	.0022	.9545	.0045	0.3187	.9699	−.0051	.9972	.0222	.9671	.0171	0.3651
3	0.9775	.0025	.9780	.0030	.9555	.0055	0.3187	.9695	−.0055	.9979	.0229	.9674	.0174	0.3630
$(N_1, N_2, N_3, N_4) = (6, 12, 18, 24)$														
1	.9621	−.0129	.9757	.0007	.9378	−.0122	0.3341	.9476	−.0274	.9979	.0229	.9455	−.0045	0.3910
2	.9621	−.0129	.9757	.0007	.9378	−.0122	0.3351	.9482	−.0268	.9991	.0241	.9473	−.0027	0.3928
3	.9723	−.0027	.9738	−.0012	.9461	−.0039	0.3351	.9569	−.0181	.9978	.0228	.9547	.0047	0.3881
$(N_1, N_2, N_3, N_4) = (24, 18, 12, 6)$														
1	.9662	−.0088	.9757	.0007	.9419	−.0081	0.3342	.9521	−.0229	.9981	.0231	.9502	.0002	0.3906
2	.9628	−.0122	.9752	.0002	.9380	−.0120	0.3346	.9496	−.0254	.9980	.0230	.9476	−.0024	0.3928
3	.9733	−.0017	.9753	.0003	.9486	−.0014	0.3351	.9555	−.0195	.9982	.0232	.9537	.0037	0.3875

Table 6 Computed sample size, expected width and tolerance probability for 95 % two-sided confidence interval of weighted eta-squared $\eta^{2*} = .15$ with interval bound $b = \omega = .3$ and tolerance probability $1-\gamma = .90$, when $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 4, 9, 16)$, $(q_1, q_2, q_3, q_4) = (1/4, 1/4, 1/4, 1/4)$, $(1/10, 2/10, 3/10, 4/10)$, $(4/10, 3/10, 2/10, 1/10)$, and mean structures $(\mu_1, \mu_2, \mu_3, \mu_4) = \{-1, 0, 0, 1\}/c$, $\{-3, -1, 1, 3\}/c$, and $\{-1, -1, 1, 1\}/c$ denoted by $\mu = 1$, 2, and 3, respectively

	Expected width				Tolerance probability			
μ	Sample sizes	Simulated $E[H]$	Approximate $E[H]$	Error	Sample sizes	Simulated $P\{H < \omega\}$	Approximate $P\{H < \omega\}$	Error
$(q_1, q_2, q_3, q_4) = (1/4, 1/4, 1/4, 1/4)$								
1	(17, 17, 17, 17)	0.2942	0.2941	0.0001	(24, 24, 24, 24)	.9244	.9170	0.0074
2	(17, 17, 17, 17)	0.2939	0.2941	−0.0002	(24, 24, 24, 24)	.9166	.9170	−0.0004
3	(17, 17, 17, 17)	0.2955	0.2941	0.0013	(24, 24, 24, 24)	.9187	.9170	0.0017
$(q_1, q_2, q_3, q_4) = (1/10, 2/10, 3/10, 4/10)$								
1	(7, 14, 21, 28)	0.2914	0.2905	0.0009	(10, 20, 30, 40)	.9782	.9733	0.0049
2	(7, 14, 21, 28)	0.2916	0.2905	0.0011	(10, 20, 30, 40)	.9779	.9733	0.0046
3	(7, 14, 21, 28)	0.2910	0.2905	0.0005	(10, 20, 30, 40)	.9739	.9733	0.0006
$(q_1, q_2, q_3, q_4) = (4/10, 3/10, 2/10, 1/10)$								
1	(32, 24, 16, 8)	0.2920	0.2923	−0.0003	(48, 36, 24, 12)	.9743	.9593	0.0150
2	(32, 24, 16, 8)	0.2931	0.2923	0.0008	(48, 36, 24, 12)	.9708	.9593	0.0115
3	(32, 24, 16, 8)	0.2925	0.2923	0.0002	(48, 36, 24, 12)	.9629	.9593	0.0036

and normally distributed with expected values μ_i and variances σ_i^2:

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

where μ_i and σ_i^2 are unknown parameters, i=1, . . . , g (≥ 2) and j=1, . . . , N_i. For testing the hypothesis that all treatment means are equal, the classic F^* statistic is the most widely used statistical procedure assuming homogeneity of variance $\left(\sigma_1^2 = \sigma_2^2 = ... = \sigma_g^2 = \sigma^2\right)$:

$$F^* = \frac{SSR/(g-1)}{SSE/(N_T - g)}, \tag{2}$$

where SSR is the treatment sum of squares, SSE is the error sum of squares, and $N_T = \sum_{i=1}^{g} N_i$. It follows that

$$F^* \sim F(g-1, N_T - g, \Lambda), \tag{3}$$

where $F(g-1, N_T-g, \Lambda)$ is the noncentral F distribution with $(g-1)$ and (N_T-g) degrees of freedom, and noncentrality parameter $\Lambda = N_T\lambda$,

$$\lambda = \sum_{i=1}^{g} q_i \left\{ \left(\mu_i - \tilde{\mu}\right)/\sigma \right\}^2, \tag{4}$$

$q_i = N_i/N_T$, and $\tilde{\mu} = \sum_{i=1}^{g} q_i \mu_i$. Furthermore, λ can be alternatively expressed as $\lambda = f^2 = \sigma_\mu^2/\sigma^2$ with $\sigma_\mu^2 = \sum_{i=1}^{g} q_i \left(\mu_i - \tilde{\mu}\right)^2$, and is called the signal-to-noise ratio (Fleishman, 1980). The measure of strength of association or correlation ratio η^2 is a one-to-one function of λ:

$$\eta^2 = \frac{\lambda}{1+\lambda}. \tag{5}$$

Accordingly, the widely used index of the association effect size η^2 is the sample eta-squared:

$$\widehat{\eta}^2 = \frac{SSR}{SSR + SSE} = \frac{F^*}{F^* + (N_T - g)/(g-1)} \tag{6}$$

where F^* is defined in Equation 2. Moreover, exact confidence intervals of η^2 can be constructed with the noncentrality inversion technique of the noncentral F distribution of F^* given in Equation 3 (e.g., Odgaard & Fowler, 2010).

However, it has been shown in numerous studies that the conventional F^* test statistic is sensitive to the heteroscedasticity formulation defined in Equation 1. Of the numerous alternatives to the ANOVA F test, we focus on the approach proposed in Welch (1951) in the form of

$$W = \frac{\sum_{i=1}^{g} W_i \left(\overline{X}_i - \tilde{X}\right)^2/(g-1)}{1 + 2(g-2)Q/(g^2-1)}, \tag{7}$$

where $W_i = N_i/S_i^2, S_i^2 = \sum_{j=1}^{N_i} (X_{ij} - \overline{X}_i)^2/(N_i - 1), \overline{X}_i = \sum_{j=1}^{N_1} X_{ij}/N_i, \tilde{X} = \sum_{i=1}^{g} W_i \overline{X}_i/U, U = \sum_{i=1}^{g} W_i$, and $Q = \sum_{i=1}^{g} (1 - W_i/U)^2/(N_i - 1)$. In contrast to the well-documented results of F^* under homoscedasticity, the statistical properties of Welch's statistic are more complex, and no explicit analytic form of the

corresponding distribution is available. It was presented in Levy (1978) that an approximate noncentral F distribution can be obtained by replacing the sample means and variances in Welch's statistic with corresponding population parameters. The numerical comparisons of the estimated power and simulated power of Levy (1978) suggest that the noncentral F distribution yields an adequate approximation for the underlying distribution of Welch's statistic. Specifically, the approximate distribution for W in Levy is

$$W \dot{\sim} F(g-1, v, \Lambda^*), \tag{8}$$

where the denominator degrees of freedom $v = (g^2-1)/(3\tau)$, $\tau = \sum_{i=1}^{g} (1-\omega_i/\upsilon)^2/(N_i-1)$, $\omega_i = N_i/\sigma_i^2$, $\upsilon = \sum_{i=1}^{g} \omega_i$, noncentrality parameter $\Lambda^* = N_T\lambda^*$,

$$\lambda^* = \sum_{i=1}^{g} q_i \left\{ \left(\mu_i - \tilde{\mu}^*\right)/\sigma_i \right\}^2, \tag{9}$$

and $\tilde{\mu}^* = \sum_{i=1}^{g} \omega_i\mu_i/\upsilon$. It is essential to note that the formulation of λ^* is the direct extension of the signal-to-noise ratio λ given in Equation 4 under the heterogeneity-of-variance assumption. For ease of reference, λ^* is termed as the weighted signal-to-noise ratio for its recognizable relationship with λ. An analogue application of the monotone transformation between λ and η^2 in Equation 5 can arguably be recommended to arrive at a weighted eta-squared effect size with λ^* in Equation 9 under the heterogeneity of variance setting,

$$\eta^{2*} = \frac{\lambda^*}{1+\lambda^*}. \tag{10}$$

Accordingly, the weighted eta-squared effect size η^{2*} was presented in Kulinskaya and Staudte (2006) as a weighted coefficient of determination with the notation ρ^2. In addition to the supporting arguments in Kulinskaya and Staudte, the weighted eta-squared effect size η^{2*} provides a natural generalization of the simple index η^2, and it reflects the proportion of total variance accounted for by the effect of treatment means, heterogeneous variance components, and sample size allocation ratios. Notably, the alternative expressions of $\omega_i/\upsilon = \left(q_i/\sigma_i^2\right) \Big/ \left(\sum_{j=1}^{g} \left(q_j/\sigma_j^2\right)\right)$ and $\tilde{\mu}^* = \sum_{i=1}^{g} \left(q_i\mu_i/\sigma_i^2\right) \Big/ \left(\sum_{j=1}^{g} \left(q_j/\sigma_j^2\right)\right)$ imply that both λ^* and η^{2*} do not depend on the group sizes but, rather, on the allocation ratio among the groups.

To indicate the actual level of the strength of association in a study, a sample estimate of the weighted eta-squared η^{2*} may be obtained as

$$\hat{\eta}^{2*} = \frac{W}{W + (N_T - g)/(g-1)}, \tag{11}$$

where W is the Welch statistic given in Equation 7. Clearly, $\hat{\eta}^{2*}$ is a heteroscedastic extension of the common effect size measure $\hat{\eta}^2$ given in Equation 6. Unlike the degrees of freedom for the distribution of F^*, the denominator degrees of freedom ν in the noncentral F distribution for W given in Equation 8 depends on the unknown variances. For inferential purposes, a further modification of the noncentral F distribution can be obtained by substituting the respective sample estimates for the variances in ν, and the resulting approximation is

$$W \dot{\sim} F(g-1, \hat{v}, \Lambda^*), \tag{12}$$

where the denominator degrees of freedom $\hat{v} = (g^2-1)/(3Q)$ and Q is defined in Equation 7. Ultimately, we propose to compute the confidence intervals of η^{2*} with the noncentrality inversion principle through the approximate noncentral F distribution presented in Equation 12. This is useful because $\Lambda^* = N_T\lambda^*$ can be viewed as a one-to-one function of η^{2*} in terms of $\Lambda^* = N_T\eta^{2*}/(1-\eta^{2*})$ with the equality between λ^* and η^{2*} in Equation 10. Explicitly, the upper $100(1-\alpha_1)\%$ confidence interval of η^{2*} is of the form $\left(\hat{\eta}_L^{2*}, 1\right)$, in which $\hat{\eta}_L^{2*}$ satisfies

$$P\left\{F\left(g-1, \hat{v}, N_T\hat{\eta}_L^{2*}/\left(1-\hat{\eta}_L^{2*}\right)\right) < W_{OL}\right\} = 1-\alpha_1, \tag{13}$$

where $W_{\mathrm{OL}} = \max\left(W_O, F_{(g-1),\hat{v},1-\alpha_1}\right)$ and W_{O} is the observed value of the W statistic defined in Equation 7. Likewise, the lower $100(1-\alpha_2)\%$ confidence interval of η^2 is of the form $\left(0, \hat{\eta}_U^{2*}\right)$, in which $\hat{\eta}_U^{2*}$ satisfies

$$P\left\{F\left(g-1, \hat{v}, N_T\hat{\eta}_U^{2}/\left(1-\hat{\eta}_U^{2}\right)\right) > W_{OU}\right\} = 1-\alpha_2, \tag{14}$$

where $W_{\mathrm{OU}} = \max\left(W_O, F_{(g-1),\hat{v},1-\alpha_2}\right)$. Typically, a $100(1-\alpha)\%$ two-sided confidence interval $\left(\hat{\eta}_L^{2*}, \hat{\eta}_U^{2*}\right)$ of weighted eta-squared association effect size η^{2*} can be obtained by jointly applying Equations 13 and 14 with $\alpha_1 = \alpha_2 = \alpha/2$. Since the noncentrality parameter of a noncentral F distribution is always nonnegative, it is necessary to use W_{OL} and W_{OU}, instead of W_{O}, to give proper results for the confidence limits. The particular adjustments not only have theoretical implications, but also yield appropriate arguments to prevent computational error. Although the noncentrality inversion procedure was also presented in confidence interval calculations of η^2, such as Odgaard and Fowler (2010), their algorithm did not entail subtle modifications of the observed F^* statistic. Note that the calculation of confidence intervals $\left(\hat{\eta}_L^{2*}, \hat{\eta}_U^{2*}\right)$ needs to be performed merely for the value of the statistic W_{O} actually observed. In addition, even though Equations 13 and 14 cannot

be solved analytically, it is really only necessary to compute them numerically, since a 100(1−α)% confidence level does not require a closed-form solution. In short, with the desired confidence level, observed value W_O, and estimated degrees of freedom $\widehat{\nu}$, the numerical computation of confidence limits $\widehat{\eta}_L^{2*}$ and $\widehat{\eta}_U^{2*}$ involves the evaluation of the noncentrality distribution function of a noncentral F variable, such as the SAS noncentrality function FNONCT. Accordingly, a SAS/IML (SAS Institute, 2011) program has been developed to perform the confidence interval calculations and is available as supplementary material. In contrast, Kulinskaya and Staudte (2006) also presented an approximate confidence interval procedure for η^2* based on a shifted and rescaled chi-square transformation. It should be emphasized, however, that their method differs markedly from the noncentrality inversion technique. More important, their analytical arguments and derived formulas are noticeably more involved than the prescribed justification and methodology. Thus, it is of both practical value and theoretical interest to explicate the underlying properties of the two distinct interval procedures. But due to the complex nature of the interval estimation formulas under study, a complete analytical treatment is not possible. Hence, a detailed simulation study is performed next to evaluate and compare their accuracy under a variety of treatment effect configurations, heterogeneous variance patterns, and sample size allocation structures.

Numerical comparison of interval estimation procedures

To demonstrate the performance of the two alternative procedures under ANOVA settings, the following empirical examination consists of two studies, of which the first one reexamines the interval estimation of weighted signal-to-noise ratio λ* for the three-group case in Kulinskaya and Staudte (2006), and the second study evaluates the confidence intervals of weighted eta-squared η^2* for the case of four groups that were not considered in Kulinskaya and Staudte.

First, we consider the model settings in Table 6 of Kulinskaya and Staudte (2006) with g=3. Specifically, the sample sizes and error variances are chosen as (N_1, N_2, N_3)=(10, 10, 10) and ($\sigma_1^2, \sigma_2^2, \sigma_3^2$) = (1, 1/2, 4), respectively. Moreover, four mean effect settings are considered: (μ_1, μ_2, μ_3)=(0, 1, 1), (0, 1, 2), (0, 1, 3), and (0, 1, 4), and the resulting weighted signal-to-noise ratio λ* values are 0.23, 0.36, 0.64 and 1.08, respectively. With the given sample sizes and parameter configurations, estimates of the true coverage probability are computed through Monte Carlo simulation of 10,000 independent data sets. For each replicate, the confidence limits associated with one-sided upper and lower 100(1−α/2)% confidence intervals are computed for both (1−α/2)=.95 and .975. These confidence limits are also employed to construct the two-sided 90 % and 95 % confidence intervals. Accordingly, a total of six different sets of confidence intervals are obtained. Thus, our simulations cover a much broader range of situations than those considered in Kulinskaya and Staudte, which examined only the performance of two-sided 95 % confidence intervals. In each case, the simulated coverage probability is the proportion of the 10,000 replicates whose intervals contain the population-weighted effect size λ*. The accuracy of the examined procedure is determined by the difference between the simulated coverage probability and the designated coverage probability as error=simulated coverage probability−nominal coverage probability. In addition, the average interval width of λ* is also computed for the 10,000 replicated widths of both 90 % and 95 % two-sided confidence intervals. The simulated results of coverage probabilities, errors, and average widths for Kulinskaya and Staudte's method and the suggested approach are presented in Table 1. For a concise visualization of these results, the simulated coverage probabilities of one-sided upper and lower 95 % confidence intervals and two-sided 90 % confidence intervals are plotted for the proposed approach and Kulinskaya and Staudte's method in Figs. 1 and 2, respectively. It appears that the discrepancy between simulated and nominal coverage probabilities of Kulinskaya and Staudte's two-sided confidence intervals tend to decrease for larger λ*. Although this general pattern agrees with the findings of Kulinskaya and Staudte, the notable errors of the associated one-sided confidence intervals reveal that the results of their two-sided interval estimates remain problematic even for large values of λ*. Specifically, the simulated coverage

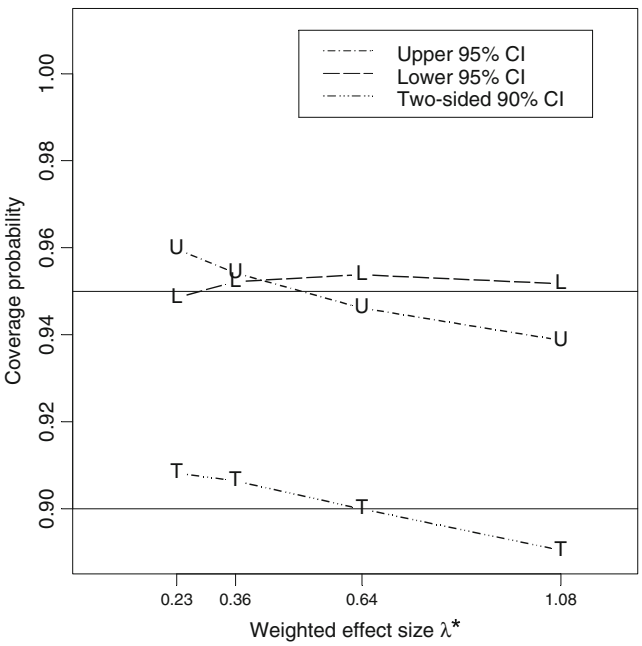

Fig. 1 Simulated coverage probabilities of the proposed confidence intervals

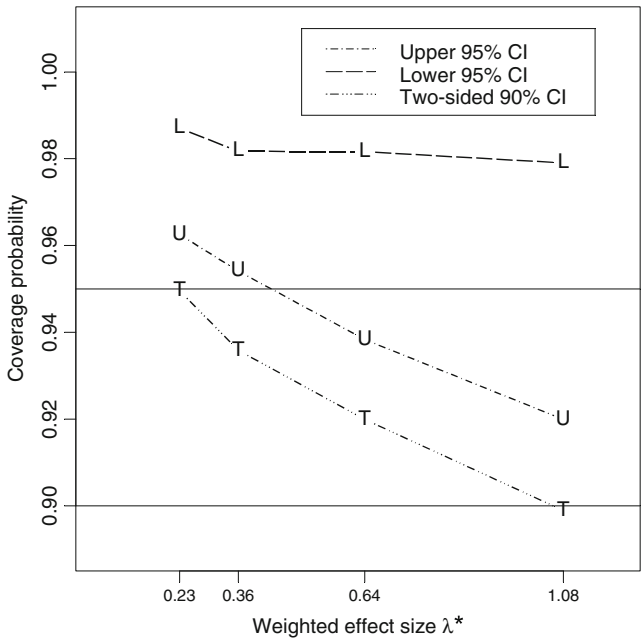

Fig. 2 Simulated coverage probabilities of Kulinskaya and Staudte's (2006) confidence intervals

probability of their 90 % two-sided confidence interval is .8989 with error−.0011 for $\lambda^*=1.08$. But the resulting coverage probabilities of the upper and lower 95 % one-sided confidence intervals are .9199 and .9790 with substantial errors−.0301 and .0290, respectively. In addition, the best performance of the 95 % two-sided confidence intervals is associated with $\lambda^*=0.64$ and has a simulated coverage probability of .9509, with error of .0009. In this case, the corresponding upper and lower 97.5 % one-sided confidence intervals incur the simulated coverage probabilities of .9575 and .9934, with sizable errors of−.0175 and .0184, respectively. Note that the confidence limits of the $(1-\alpha)$% two-sided confidence interval are constructed with the respective lower and upper limit of the one-sided upper and lower $(1-\alpha/2)$% confidence intervals. Thus, it is misleading to report that a two-sided interval procedure is accurate on the basis of a combination of some noticeable under- and overestimated one-sided coverage probabilities. Consequently, a mere coverage probability assessment of two-sided confidence intervals may obscure systematic overestimation in confidence limits that might have existed in the shifted and rescaled chi-square transformation of Kulinskaya and Staudte. In contrast, the simulated coverage probabilities of the suggested one- and two-sided confidence intervals closely agree with the nominal confidence levels for all 24 combined cases in Table 1. Although the case of the upper 95 % confidence interval for $\lambda^*=1.08$ yields a coverage probability .9387 and induces the largest error − .0113, this result still outperforms that of Kulinskaya and Staudte, which yields a coverage probability .9199 and error − .0301. Moreover, in terms of the average widths of the simulated two-sided confidence intervals for the weighted signal-to-noise ratio λ^*, it is apparent that the average width of the proposed approach is consistently smaller than that computed by the method of Kulinskaya and Staudte for each of the eight combinations of two confidence levels $(1-\alpha)$ and four values of weighted effect size λ^*.

To demonstrate that the previous contrasting behaviors between the two interval procedures continue to exist in other heteroscedastic ANOVA situations, further numerical investigations were conducted with a wide range of different model configurations. In the second study, we focus on the interval estimation of weighted eta-squared η^2* with $g=4$ under both settings of heterogeneous variances $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 4, 9, 16)$ and homogeneous variances $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 1, 1, 1)$. For sample size structures, three allocation schemes are examined to represent diverse patterns: $(N_1, N_2, N_3, N_4)=(15, 15, 15, 15)$, $(6, 12, 18, 24)$, and $(24, 18, 12, 6)$. These three settings not only include both balanced and unbalanced designs, but also create direct- and inverse-pairing with heteroscedastic structure. Moreover, the three sample size allocation schemes are cross-combined with three different mean variability settings: $(\mu_1, \mu_2, \mu_3, \mu_4)=\{-1, 0, 0, 1\}$, $\{-3, -1, 1, 3\}$, and $\{-1, -1, 1, 1\}$. For ease of comparison, the actual mean structure is further modified as $(\mu_1, \mu_2, \mu_3, \mu_4)/c$ with a constant c for adjustment so that the resulting weighted eta-squared η^2* remains the same as $\eta^2*=1/6$ $(\lambda^*=1/5)$ for each case of a total of 18 different model configurations. Similar variance structures, mean variability patterns, and sample size allocations were considered in Cohen (1988) and Tomarken and Serlin (1986). These combinations of model configurations are selected to reveal the extent of characteristics that are likely to be obtained in actual applications. General guidelines of design and implementation of Monte Carlo experiments can be found in Paxton, Curran, Bollen, Kirby, and Chen (2001). Similar to the implementation of the preceding examination, the simulated results of coverage probabilities, errors, and average widths for $(1-\alpha/2)$% one-sided and $(1-\alpha)$% two-sided confidence intervals for heterogeneous variances $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 4, 9, 16)$ are presented in Tables 2 and 3 for $\alpha=.10$ and .05, respectively. In addition, Tables 4 and 5 contain the corresponding numerical results under homogeneous structure $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 1, 1, 1)$ for $\alpha=.10$ and .05, respectively.

According to the extensive numerical results in Tables 2–5, the coverage probabilities of the proposed interval procedure maintain a close range near the nominal levels. Although some of the absolute errors are slightly larger than .01, the performance seems generally acceptable. It is noteworthy that the suggested approach is developed under the possibly unequal variances assumption. In view of the stable and adequate performance of the resulting confidence intervals under both homogeneous and heterogeneous variance settings, the

proposed interval procedure has great potential usefulness in practical situations where the extent of underlying variance heterogeneity is rarely known and may be nearly trivial. Unfortunately, the interval method of Kulinskaya and Staudte (2006) does not provide satisfactory results, even though the coverage performance of some two-sided confidence intervals in Tables 3 and 5 are reasonably good. The sign and magnitude of the errors associated with upper and lower one-sided confidence intervals show that the simulated coverage probabilities are consistently lower or higher than the nominal levels throughout Tables 2–5. The poor performance implies that their approach fails to produce accurate confidence limits under most of the conditions examined here. Furthermore, all of the average widths of two-sided confidence intervals of the suggested interval procedure are less than those computed by the method of Kulinskaya and Staudte. Consequently, the noncentrality inversion approach is recommended over the existing shifted and rescaled transformation of Kulinskaya and Staudte for its overall performance in the accuracy of coverage probability and the tightness of interval width corresponding to the nominal confidence level.

Sample size determination for precise interval estimation

With the emphasis on greater use of summary measure and confidence intervals in the sixth edition of the *Publication Manual of the American Psychological Association* (American Psychological Association, 2009), it is prudent to facilitate this research practice by determining the necessary sample sizes to satisfy the desired precision of interval estimation in the planning stage of research design. Hence, the sample size determination for precise confidence intervals of the weighted eta-squared effect size is considered.

According to the formulation of the approximate noncentral F distribution of Welch's statistic and the application of the noncentrality inversion technique, an approximate 100 $(1-\alpha)$% two-sided interval estimate $\left(\widehat{\eta}_L^{2*}, \widehat{\eta}_U^{2*}\right)$ of η^2* can be computed from Equations 13 and 14 with equal tail confidence probability, $\alpha_1=\alpha_2=\alpha/2$. To ensure that the confidence interval is narrow enough to produce meaningful findings, researchers must recognize the stochastic nature of confidence intervals due to the inherent randomness in Welch's statistic W and the degrees of freedom estimator $\widehat{\nu}$. However, the property of the approximate degrees of freedom $\widehat{\nu}$ involves the joint consideration of g heterogeneous sample variances, and consequently, the complexity of the exact distribution of $\widehat{\nu}$ can be overwhelming. To provide a feasible solution, the random feature of $\widehat{\nu}$ is ignored in the proposed sample size calculations. This simplification is a small price to pay for developing a sample size framework that is informative and useful for precise interval estimation. The empirical examinations presented later reveal that sampling fluctuations in $\widehat{\nu}$ are minimal and the associated effects may be negligible. Hence, the width of a confidence interval $\left(\widehat{\eta}_L^{2*}, \widehat{\eta}_U^{2*}\right)$, denoted by $H = \widehat{\eta}_U^{2*} - \widehat{\eta}_L^{2*}$, can be viewed as a function of the Welch statistic, degrees of freedom ν, weighted eta-squared η^2*, sample sizes (N_1, . . . , N_g), and confidence coefficient $(1-\alpha)$. Specifically, the approximate noncentral F distribution suggested in Levy (1978) is utilized to determine the sample sizes required to achieve the specified precision properties of a confidence interval. Two useful principles concerning the control of the expected width and the tolerance probability of the width within a preassigned value are presented here. First, it is necessary to determine the required sample size such that the expected width $E[H]$ of a 100$(1-\alpha)$% confidence interval $\left(\widehat{\eta}_L^{2*}, \widehat{\eta}_U^{2*}\right)$ is within the given bound

$$E[H] \leq b, \tag{15}$$

where b (>0) is a constant. Second, one may compute the sample size needed to guarantee, with a given tolerance probability $(1-\gamma)$, that the width H of a 100$(1-\alpha)$% interval estimate $\left(\widehat{\eta}_L^{2*}, \widehat{\eta}_U^{2*}\right)$ will not exceed the planned value

$$P\{H \leq \omega\} \geq 1-\gamma, \tag{16}$$

where ω (> 0) is a constant. Both the expectation $E[H]$ and probability $P\{H \leq \omega\}$ are evaluated with respect to the approximate distribution of W presented in Equation 8.

For ease of numerical computation, the sample size allocation ratios (q_1, . . . , q_g) are rewritten as $q_i = r_i \Big/ \sum_{j=1}^{g} r_j$ where $r_i=N_i/N_1$ for $i=1, \ldots, g$. Equivalently, $r_i=q_i/q_1$, $i=1, \ldots, g$. Thus, with the initially specified sample size allocation ratios ($q_1, \ldots, q_g$) or sample size ratios ($r_1, \ldots, r_g$), the task is reduced to deciding the minimum sample size N_1 (with $N_i=N_1 r_i$, $i=2, \ldots, g$) required to attain the desired precision level. With the computational formulas of expected width and tolerance probability in Equations 15 and 16, the sample sizes ($N_{\mathrm{EW}1}, \ldots, N_{\mathrm{EW}g}$) needed for the expected width of a 100$(1-\alpha)$% two-sided confidence interval $\left(\widehat{\eta}_L^{2*}, \widehat{\eta}_U^{2*}\right)$ to fall within the designated bound b are the minimum integers (N_1, . . . , N_g)=$N_1(r_1, \ldots, r_g)$ such that $E[H] \leq b$. On the other hand, the sample size ($N_{\mathrm{TP}1}, \ldots, N_{\mathrm{TP}g}$) required to guarantee with a given tolerance probability $(1-\gamma)$ that the width of a 100$(1-\alpha)$% two-sided confidence interval $\left(\widehat{\eta}_L^{2*}, \widehat{\eta}_U^{2*}\right)$ will not exceed the planned range ω are the smallest integers ($N_1, \ldots, N_g$)=$N_1(r_1, \ldots, r_g)$ such that $P\{H \leq \omega\} \geq 1-\gamma$. The computation of expected width and tolerance probability requires the numerical integration and noncentrality inversion with respect to a noncentral F probability distribution function. To enhance the applicability of these sample size methodologies, supplementary SAS/IML (SAS Institute, 2011)

computer programs have been written to aid researchers with the suggested techniques, and empirical illustrations are presented next to demonstrate their usefulness in sample size calculations.

Numerical investigation of sample size procedures

Due to the approximate nature of the suggested sample size procedures for precise interval estimation of the weighted eta-squared effect sizes, their features and performances need to be delineated and examined through numerical investigations. To demonstrate the sample size methodology, an empirical study was conducted in two steps. The first step involved extensive sample size calculations for the two precision measures of expected width and tolerance probability across a wide range of model configurations. In the second step, a Monte Carlo simulation study was performed to provide insights into the precision behavior for the recommended sample size formulas under the design characteristics specified in the first step.

Note that the determination of sample sizes needed for the chosen precision of the confidence interval procedures requires detailed specifications of the confidence level, sample size allocation ratio, and the magnitudes of mean effects and variance components. To demonstrate the potential extent of characteristics that an applied work may cover in heteroscedastic ANOVA research, a systematic numerical investigation of four-group design is conducted by fixing the confidence level $(1-\alpha)$=.95 and heterogeneous error variances $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2)$ = (1, 4, 9, 16) and varying the other two factors of sample size allocation ratio and mean variability pattern for the selected magnitude of weighted eta-squared η^2*=.15. Accordingly, to represent balanced and unbalanced patterns, three sample size allocation settings are considered: (q_1, q_2, q_3, q_4)=(1/4, 1/4, 1/4, 1/4), (1/10, 2/10, 3/10, 4/10), and (4/10, 3/10, 2/10, 1/10). As in the empirical illustration presented above, the sample size allocation schemes are cross-combined with three different mean spread settings: (μ_1, μ_2, μ_3, μ_4)= {−1, 0, 0, 1}/c, {−3, −1, 1, 3}/c, and {−1, −1, 1, 1}/c. Note that different values of constant c are used for adjustment so that the weighted eta-squared is kept constant as η^2*=.15 throughout this numerical study. Moreover, the interval bound $b=\omega$=.3, and tolerance probability $(1-\gamma)$=.90 are selected for the two precision criteria of expected width and tolerance probability. These levels were selected to reflect common sample sizes used in typical research settings. Accordingly, the necessary sample sizes (N_{EW1}, N_{EW2}, N_{EW3}, N_{EW4}) and (N_{TP1}, N_{TP2}, N_{TP3}, N_{TP4}) are computed with respect to the selected precision requirements of expected width and of tolerance probability, respectively. The resulting sample sizes are presented in Table 6 for all nine joint model configurations of varying sample size allocation and mean dispersion structure.

An inspection of the sample sizes reported in Table 6 shows that the computed sample sizes are identical for all three mean patterns when the sample size allocation ratio is fixed due to the restriction of a constant weighted eta-squared η^2*=.15. Accordingly, the actual sample sizes under the expected width consideration are (N_{EW1}, N_{EW2}, N_{EW3}, N_{EW4})=(17, 17, 17, 17), (7, 14, 21, 28), and (32, 24, 16, 8) for the three sample size allocation settings (q_1, q_2, q_3, q_4)=(1/4, 1/4, 1/4, 1/4), (1/10, 2/10, 3/10, 4/10), and (4/10, 3/10, 2/10, 1/10), respectively. On the other hand, the corresponding sample sizes associated with the assurance of tolerance probability principle are (N_{TP1}, N_{TP2}, N_{TP3}, N_{TP4})=(24, 24, 24, 24), (10, 20, 30, 40), and (48, 36, 24, 12) for the three sample size allocation schemes, respectively. Also, it is important to note that the total sample sizes, N_T, of the balanced structure are less than those of the unbalanced structure for both types of interval precisions. The case with inverse pairing of heterogeneous variance and unbalanced allocation incurs the largest number of total sample sizes. Since the two precision criteria impose unique and distinct precision characteristics on the resulting confidence intervals, the required sample sizes are different. Although the results are not completely comparable, it typically requires a larger sample size to meet the necessary precision of tolerance probability than the control of a designated expected width, as was noted in Kupper and Hafner (1989). More important, the sample size procedures and empirical results presented here enable researchers to better understand the underlying relationship that exists between the designated interval precision and the required sample size given the fundamental information of model configurations.

In the process of sample size determination, the attained precision levels associated with the listed sample sizes (N_{EW1}, N_{EW2}, N_{EW3}, N_{EW4}) and (N_{TP1}, N_{TP2}, N_{TP3}, N_{TP4}) should be less than or greater than the nominal level for width bound b=.3 and tolerance probability $(1-\gamma)$=.90, respectively. The achieved expected width $E[H]$ and tolerance probability $P\{H\leq\omega\}$ computed with the approximate noncentral F distribution in Equation 8 are also summarized in Table 6. It appears that the resulting approximate expected widths .2941, .2905, and .2923 for the three sample size allocation schemes are marginally smaller than the selected width, b=.3. However, the approximate tolerance probabilities .9170 of equal allocation ratio or balanced design are slightly greater than the nominal level .90. For the two unbalanced designs, the approximate tolerance probabilities are .9733 and .9593 for direct- and inverse-pairing of allocation ratio with heteroscedastic structure, respectively. It is conceivable that the substantial differences between the

actual tolerance probabilities and the target level $(1-\gamma)=.90$ are due to the underlying metric of integer sample sizes and the constraint of a designated sample size allocation ratio. Since it is not possible to compute exact expected width or tolerance probability with the specified sample sizes, we then evaluate the accuracy of the sample size calculations through the following Monte Carlo simulation study. Under the computed sample sizes, parameter configurations and precision settings described in Table 6, estimates of the true expected width or tolerance probability are computed through Monte Carlo simulation of 10,000 independent data sets. For each replicate, the confidence limits and corresponding interval width of the two-sided 95 % confidence intervals of η^2* are calculated. Then the simulated expected width is the mean of the 10,000 replicates of interval widths, whereas the simulated tolerance probability is the proportion of the 10,000 replicates whose values of interval width are less than or equal to the specified bound $\omega=.3$.

The adequacy of the sample size procedure for precise interval estimation is determined by one of the following formulas: error=simulated expected width−approximate expected width or error=simulated tolerance probability−approximate tolerance probability. Both the simulated and corresponding errors of expected width and tolerance probability are also summarized in Table 6. It can be seen from the results that the performance of the proposed approaches appears to be good for the range of model specifications considered here. In particular, the absolute errors of the expected width are less than .002 for the nine cases examined here. Also, the absolute discrepancies in tolerance probability are smaller than .01, with the two exceptions of .0150 and .0115, associated with inverse pairing of heterogeneous variance and unbalanced allocation. Overall, this empirical evidence demonstrates that the proposed sample size procedures provide feasible and accurate solutions to precise interval estimation of the weighted eta-squared under a wide variety of heteroscedastic model configurations.

Conclusions

To extend and fortify the use of effect sizes and associated confidence intervals in empirical studies, this article has focused on the interval estimation and sample size determination for the weighted eta-squared effect sizes in one-way heteroscedastic ANOVA. Although existing studies have shown several interesting and fundamental results, this research contributes to the effect sizes literature by considering three methodological issues with analytical and numerical expositions. First, in connection with the well-known signal-to-noise ratio and eta-squared effect sizes in the homoscedastic ANOVA framework, we have provided enhanced interpretations and supportive usages for the notions of weighted effect size and weighted coefficient of determination in Kulinskaya and Staudte (2006), as the weighted signal-to-noise ratio and weighted eta-squared effect size within the extended context of heteroscedastic ANOVA. Second, for the interval estimation of weighted eta-squared, we have addressed the potential deficiencies of the shifted and rescaled chi-square transformation approach of Kulinskaya and Staudte and have proposed an improved procedure that has the advantages of theoretical justification, computational simplicity, and numerical performance over the existing method of Kulinskaya and Staudte. Third, the corresponding sample size procedures for precise interval estimation of weighted eta-squared have been developed for both the expected width and tolerance probability considerations. The performance of the suggested sample size calculations appears to be sufficiently accurate for practical purposes within the range of model specifications considered in the present article. Overall, the recommended methodology facilitates the advocated practice of confidence intervals for effect sizes, and it reinforces the potential usefulness of ANOVA models under heterogeneity of variance.

Author Note The authors thank the editor, Gregory Francis, and the two anonymous reviewers for their helpful comments that substantially improved the presentation. Correspondence concerning this article should be addressed to G. Shieh, Department of Management Science, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan 30050 (e-mail: gwshieh@mail.nctu.edu.tw).

References

Alhija, F. N. A., & Levy, A. (2009). Effect size reporting practices in published articles. *Educational and Psychological Measurement, 69,* 245–265.

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Bonett, D. G. (2008). Confidence intervals for standardized linear contrasts of means. *Psychological Methods, 13,* 99–109.

Breaugh, J. A. (2003). Effect size estimation: Factors to consider and mistakes to avoid. *Journal of Management, 29,* 79–97.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Cohen, J. (1973). Eta-squared and partial eta squared in fixed factor ANOVA designs. *Educational and Psychological Measurement, 33,* 107–112.

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

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practices, 40,* 532–538.

Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research, 23,* 89–105.

Fleishman, A. I. (1980). Confidence intervals for correlation ratios. *Educational and Psychological Measurement, 40,* 659–670.

Glass, G. V., & Hakstian, A. R. (1969). Measures of association in comparative experiments: Their development and interpretation. *American Educational Research Journal, 6,* 403–414.

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

Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods, 6,* 135–146.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Erlbaum.

Haase, R. F. (1983). Classical and partial eta square in multifactor ANOVA designs. *Educational and Psychological Measurement, 43,* 35–39.

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.

Hays, W. L. (1994). *Statistics* (5th ed.). Belmont, CA: Wadsworth.

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods, 39,* 979–984.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods, 13,* 110–129.

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.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kulinskaya, E., & Staudte, R. G. (2006). Interval estimates of weighted effect sizes in the one-way heteroscedastic ANOVA. *British Journal of Mathematical and Statistical Psychology, 59,* 97–111.

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.

Kupper, L. L., & Hafner, K. B. (1989). How appropriate are popular sample size formulas? *The American Statistician, 43,* 101–105.

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research, 28,* 612–625.

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

Lix, L. M., & Keselman, H. J. (1995). Approximate degrees of freedom tests: A unified perspective on testing for mean equality. *Psychological Bulletin, 117,* 547–560.

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.

Maxwell, S. E., Camp, C. J., & Arvey, R. D. (1981). Measures of strength of association: A comparative examination. *Journal of Applied Psychology, 66,* 525–534.

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

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59,* 537–563.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill.

Odgaard, E. C., & Fowler, R. L. (2010). Confidence intervals for effect sizes: Compliance and clinical significance in the *Journal of Consulting and Clinical Psychology*. *Journal of Consulting and Clinical Psychology, 78,* 287–297.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect sizes for some common research designs. *Psychological Methods, 8,* 434–447.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8,* 287–312.

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement, 64,* 916–924.

Richardson, J. T. E. (1996). Measures of effect size. *Behavior Research Methods, Instruments, & Computers, 28,* 12–22.

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6,* 135–147.

SAS Institute. (2011). *SAS/IML user's guide, version 9.3*. Cary, NC: Author.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurements, 61,* 605–632.

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9,* 164–182.

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology, 102,* 989–1004.

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

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51,* 473–481.

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