Problems with Probabilistic Hindsight: A Comparison of Methods for Retrospective Statistical Power Analysis

Jeffrey Kromrey, University of South Florida
Kristine Y. Hogarty, University of South Florida

In contrast to prospective uses of power analysis, retrospective power analysis provides an estimate of the statistical power of a hypothesis test after an investigation has been conducted. The purpose of this research was to empirically investigate the bias and sampling errors of three point estimators of retrospective power and the confidence band coverage of an interval estimate approach. Monte Carlo methods were used to investigate a broad range of research designs and population effect sizes that may be encountered in field research. The results suggest that none of the retrospective power estimation techniques were effective across all of the conditions examined. For point estimates, the "unbiased" and "median unbiased" estimators showed improved performance relative to the plug-in estimator, but these procedures were not completely free from bias except under large sample sizes and large effect sizes (as the statistical power approaches unity). Further the RMSE of these estimates suggests large amounts of sampling error for all three of the point estimators. The interval estimates showed good confidence band coverage under most conditions examined, but the width of the bands suggests that they are relatively uninformative except for large sample and large effect size conditions.

Statistical power analysis is useful from both prospective and retrospective viewpoints. Prospectively, power analysis is used in the planning of inquiry, typically to provide an estimate of the sample size required to obtain a desired level of statistical power under an assumed population effect size, experimental design and nominal alpha level. In contrast, retrospective uses of power analysis involve a consideration of statistical power after inquiry has been completed. This important application of power analysis is somewhat more complicated than the prospective uses.

Two Views on Retrospective Power

Recent literature suggests that retrospective power analysis is conceptualized in two very different forms. Characteristic of one approach, Zumbo and Hubley (1998) and Ottenbacher and Maas (1999) present Bayesian power estimation techniques directed at determining the probability of the null hypothesis being false, given that the null has been rejected, that is *Pr* (H_o=*false*|*rejected* H_o). While this probability is of importance in applied research, it's practical applications appear to be limited because of the unknown proportions of true and false null hypotheses in any field of inquiry (Zumbo & Hubley, 1998). This approach also introduces a different formal definition of "power" than is typically considered in inferential statistics (i.e., power usually represents *Pr* (H_o *will be rejected*|H_o=*false*) which is equal to 1 – β). These two probabilities are often very different. Because this conceptualization of retrospective power is not practical, it will not be further addressed here.

The second approach to retrospective power analysis (Gerard, Smith & Weerakkody, 1998; Steiger & Fouladi, 1997; Brewer & Sindelar, 1987) aims to estimate the statistical power of a hypothesis test after the test has been conducted. That is, information obtained from a particular study may be used to estimate the population effect size, which in turn may be used (in concert with the study's sample size and nominal alpha level) to estimate the power under which the research was conducted. This approach to retrospective power analysis appears to satisfy a practical need in applied research and retains the familiar formal definition of power (i.e., 1 – β). As applied researchers, we have been urged to consider the effect sizes associated with our data (e.g., Kirk, 1996; Harlow, Muliak & Steiger, 1997), in conjunction with the reject/fail-to-reject decisions of our hypothesis tests. The second approach to retrospective power analysis simply extends our use of sample effect sizes to provide estimates of power. However, the estimation of statistical power based on a sample effect size is characterized by considerable controversy.

Estimation Procedures for Retrospective Power

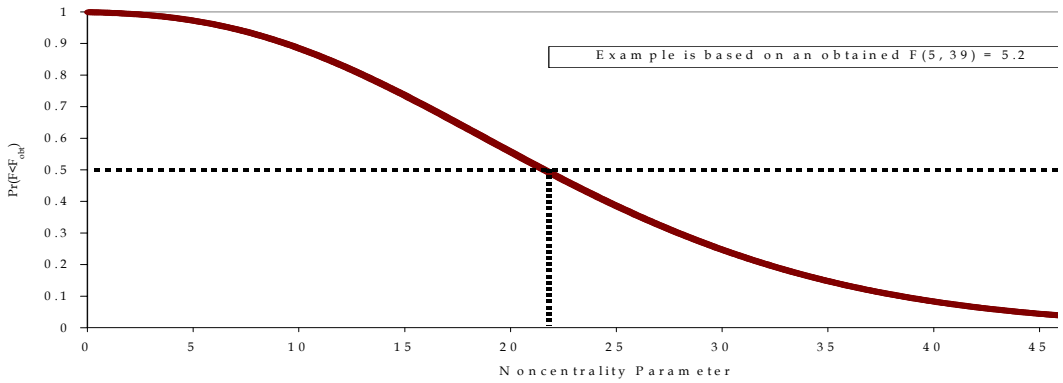
Several techniques for the second approach to retrospective power analysis have been suggested in the literature. Gerard, Smith and Weerakkody (1998) describe three statistics (estimates of noncentrality) that lead to point estimates of retrospective power: a "plug-in estimator" (λ_p), an "unbiased estimator" (λ_{ub}), and a "median unbiased" or "percentile estimator" (λ_{50}) of the noncentrality parameter.

The plug-in estimator simply represents the use of the sample noncentrality parameter (λ_p) as if it were the same as the population parameter. For the F distribution, the sample noncentrality parameter is given by

$$\lambda_p = \nu_1 F$$

where ν_1 = numerator degrees of freedom for the sample *F*, and *F* = obtained sample *F*statistic.

Figure 1
Probability of F < F_{obt} as a Function of Population Noncentrality



The obtained sample noncentrality parameter is then used to estimate the statistical power of the test

$$Power = \Pr(F_{\nu_1, \nu_2, \lambda_p} \geq F_{\nu_1, \nu_2, 1-\alpha})$$

where $F_{\nu_1,\nu_2,\lambda_p}$ = the noncentral F distribution with ν_1 and ν_2 degrees-of-freedom and a noncentrality parameter λ_p,

and $F_{\nu_1,\nu_2,1-\alpha}$ = the $(1-\alpha)$ percentile of central *F*-distribution (i.e., the critical value of *F* with ν_1 and ν_2 degrees of freedom).

The use of λ_p is known to produce biased estimates of power with a distinct positive bias in conditions of low power (Johnson et al., 1995). Johnson et al. suggested an alternative estimator (λ_{ub}) intended to reduced the bias inherent in λ_p. This "unbiased" estimator of noncentrality is given by

$$\lambda_{ub} = \frac{\nu_1(\nu_2 - 2)F}{\nu_2} - \nu_1$$

Although λ_{ub} may provide an unbiased estimate of the population noncentrality, estimates of power derived from unbiased noncentrality estimates are not necessarily unbiased themselves, because power is a nonlinear function of noncentrality (Gerard et al., 1988).

A third point estimate of noncentrality was suggested by Taylor and Muller (1996). This approach (λ_{50}) is reported to underestimate noncentrality 50% of the time and overestimate it 50% of the time (hence, Gerard et al., 1998, refer to the method as "median unbiased"). This method makes use of the cumulative distribution function of *F* and seeks the value of noncentrality for which the obtained value of F in a particular study (i.e., with a given ν_1 and ν_2) is expected 50% of the time (see Figure 1). Because analytical formulae for solving this problem are not available, the value of noncentrality must be obtained by numerical methods (see, for example, Press, Teukolsky, Vetterling & Flannery, 1992).

In contrast to the point estimates suggested by Gerard et al. (1998), Steiger and Fouladi (1997) presented an interval estimation approach based on the earlier work of Hedges and Olkin (1985). This approach provides confidence bands on the noncentrality parameter (noncentrality interval estimates) which subsequently may be used to obtain confidence bands on statistical power. Using logic analogous to that used to obtain the λ_{50} point estimate, the approach involves the inversion of percentiles from noncentral sampling distributions to obtain confidence bands around the noncentrality parameter. That is, instead of seeking the value of noncentrality expected 50% of the time, a 95% confidence band is obtained by seeking the value of noncentrality (λ) for which $\Pr(F_{\nu_1, \nu_2, \lambda} < F_{obt}) = .025$ and the value for which $\Pr(F_{\nu_1, \nu_2, \lambda} < F_{obt} = .975$. This provides a confidence band for noncentrality, the endpoints of which are transformed into the endpoints of a 95% confidence band for statistical power.

Purpose of the Study

Neither the point nor the interval estimation methods for retrospective power analysis have been thoroughly investigated in terms of their operating characteristics. The purpose of this research was to empirically investigate the bias and standard errors of the three point estimators of retrospective power and the confidence band coverage of the noncentrality interval estimate approach. The investigation covered a broad range of research designs and population effect sizes that may be encountered in field research.

Method

A Monte Carlo study was conducted to investigate the bias and standard errors of the three point estimators of retrospective power, and the confidence band coverage of the interval estimation technique. Data were simulated from linear models and sample effect size estimates were used to obtain power estimates. The Monte Carlo study included three factors in the design. These factors were (a) the experimental design simulated, including one factor designs with 2, 4, and 8 levels of the independent variable and three factorial designs (2X2, 2X4 and 3X3), (b) the sample size of the study, with sample sizes ranging from 5 to 100 per cell, including equal and unequal cell sizes, and (c) population effect sizes, with f^2 values (Cohen, 1988) of .01, .02. .15, .35 and .50, as well as a null condition ($f^2 = 0$). The combination of population effect sizes and sample sizes provides conditions with power values ranging from α to nearly 1.00. For each sample generated, the power of the hypothesis test was estimated using the three point estimators and the interval estimate.

The Monte Carlo study was conducted using SAS/IML version 6.12, running on Windows 95 and 98 platforms. The RANNOR random number generator was used to generate normally distributed variables for the observations in each study, and a different seed value for the random number generator was used in each execution of the program. The program code was verified using benchmark datasets.

Fifty thousand replications were conducted for each condition. The use of 50,000 samples provides adequate precision for estimating the relative success of the procedures investigated. For example, the maximum width of a 95% confidence interval around a sample proportion based on 50,000 samples is $\pm$ .0044 (Robey & Barcikowski, 1992).

Results

The results are presented in terms of statistical bias and root mean squared error (RMSE) for the point estimates of power. Statistical bias of the power estimates was estimated as

$$Bias = \frac{\sum_{k=1}^{K}\left(\hat{\theta}_k - \theta\right)}{K}$$

where $\hat{\theta}_k$ = power estimate for the kth sample,
θ = population power, and
K = number of samples simulated.

This statistic represents the difference between the mean sample estimate of power and the true population power for the condition examined.

RMSE of the power estimates was estimated as

$$RMSE = \sqrt{\frac{\sum_{k=1}^{K}\left(\hat{\theta}_k - \theta\right)^2}{K}}$$

This statistic represents the standard deviation of the sample estimates in which deviation is computed from the population parameter rather than from the mean of the sample estimates.

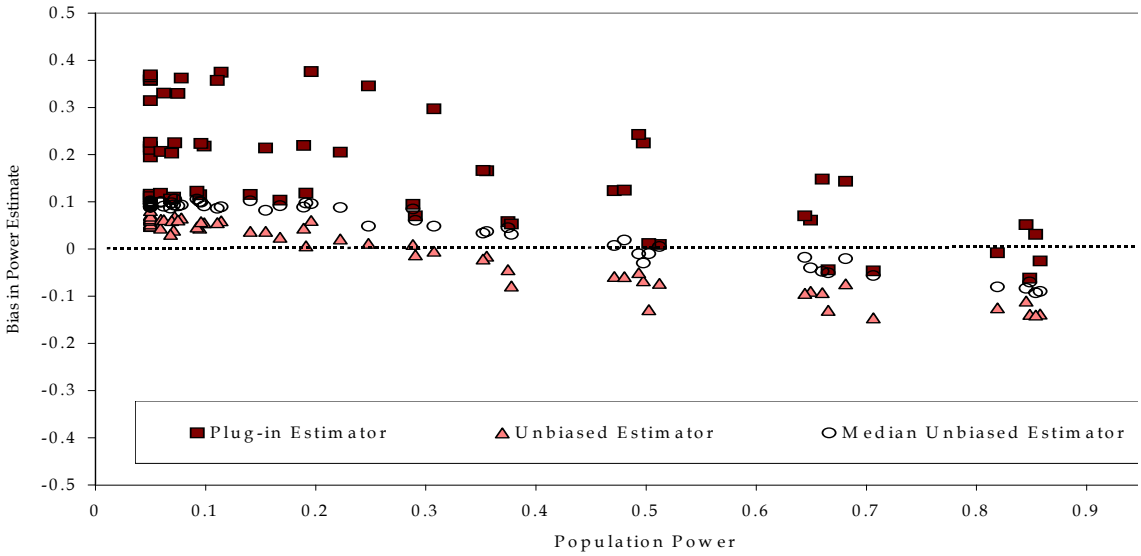
For the interval estimates of power, the proportion of sample confidence bands that contained the parameter were calculated to provide an estimate of the accuracy of the bands. Further, the average width of the confidence bands for each condition was calculated.

To conserve space, results are presented for a subset of the conditions examined (conditions that correspond to Cohen's, 1988, small, medium and large effect sizes in addition to the null condition). Complete results are available from the authors.

Single Factor Designs. Estimates of statistical bias in the point estimates of power for single factor designs are presented in Table 1. Graphs of these bias estimates are provided in Figures 2 and 3. To construct the figures, the population effect size, sample size and number of groups were translated into a population power value which is plotted on the abscissa of each figure. For the null condition ($f^2 = 0$), in balanced designs, all of the estimates evidenced positive bias, with the plug-in estimator presenting the greatest amount of bias (reaching as high as 0.37 for the 8-group design with large samples). Bias evidenced by the plug-in estimator, for a small effect size, was greatest for designs with larger numbers of groups, but the other two estimators did not show such a pattern. The bias in all three of the estimators was reduced as the population effect size increased and many conditions evidenced an underestimate of the power (negative bias). For example, with a medium effect size ($f^2 = .15$), the unbiased estimator evidenced negative bias large as –0.12, with n = 20 in 2-group and 4-group designs. With large samples and a large effect size, all of the estimators converged to the true power (i.e., showing zero bias).

For unbalanced designs, the same pattern was maintained, but the bias estimates were, in general, slightly larger in magnitude. For the null conditions and conditions with a small effect size, a positive bias was evident in most cases, while all of the estimators provided unbiased power estimates for large samples and a large effect size.

Figure 2
Statistical Bias in Point Estimates of Retrospective Power
Balanced Designs



The root mean squared errors (RMSEs) of these point estimates are provided in Table 2. Graphs of these error estimates are provided in Figures 4 and 5. These statistics reflect sampling variability in terms of squared deviations from the population parameter. If a statistic is unbiased, the RMSE is the same as the standard error. Because these statistics reflect sampling error, in many conditions the RMSEs become smaller with larger sample sizes (e.g., for conditions with a large effect size). When estimators are biased, however, the RMSE may not decrease with larger sample sizes. In general the magnitudes of the RMSE associated with these point estimates of retrospective power are quite large for conditions with a small or medium population effect size and small sample size. However, with large samples and large effect sizes, the sampling error is substantially reduced. Further, the magnitude of the RMSE does not appear to be systematically larger with unbalanced designs.

For the interval power estimates, the proportion of confidence bands that contained the true value of power and the confidence interval width are presented in Table 3 and illustrated in Figure 6. For balanced designs, the intervals showed 95% coverage across all non-null conditions, but performance decreased with the unbalanced designs. For the unbalanced designs, confidence band coverage decreased with increasing effect sizes and increasing sample sizes.

As with the RMSE for the point estimates, for both balanced and unbalanced designs, the average width of the confidence bands (Table 3) suggests that the bands are relatively uninformative for small samples and even for large samples if the effect size is small. Only for those conditions with large samples and medium and large effect sizes did the width of the bands become small enough to be considered informative in a practical sense.

Factorial Designs. Estimates of statistical bias in the point estimates of power for factorial designs are presented in Table 4 and illustrated in Figure 7. Consideration of bias for factorial designs must include an examination of row, column and interaction effects. For the null condition ($f^2 = 0$), all of the estimates evidenced positive bias for all three effects, with the plug-in estimator presenting a greater amount of bias for both the column and interaction effects for the 2 X 4 factorial design (approximately .22 across all sample sizes). The greatest amount of statistical bias was seen for the interaction effect for 3 X 3 factorial designs (reaching .26 for all but the smallest sample size). A similar pattern was evidenced for the smallest effect size ($f^2 = .02$) for all but the largest sample sizes. That is, bias in the plug-in estimator, for small effect sizes, was greater for column effects with the 2 X 4 designs and for the interaction effect for both the 2 X 4 and 3 X 3 factorial designs, but the other two estimators did not show such a pattern. Similar to the single factor designs, the bias in all three of the estimators was reduced as the population effect size increased and many conditions evidenced an under-estimate of power (negative bias). For example, with

Table 1. Statistical Bias of Three Point Estimates of Retrospective Power in One Factor Designs.

Balanced Designs

		$f^2 = 0.00$			$f^2 = 0.02$			$f^2 = 0.15$			$f^2 = 0.35$		
Groups	N	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}
	5	0.11	0.04	0.09	0.11	0.04	0.09	0.11	0.00	0.09	0.07	-0.07	0.05
	10	0.12	0.05	0.10	0.11	0.04	0.10	0.05	-0.05	0.04	-0.04	-0.14	-0.05
2	20	0.12	0.05	0.10	0.11	0.03	0.09	-0.04	-0.12	-0.04	-0.07	-0.11	-0.07
	50	0.12	0.06	0.10	0.07	-0.02	0.06	-0.06	-0.09	-0.06	0.00	0.00	0.00
	100	0.12	0.06	0.10	0.00	-0.08	-0.01	-0.01	-0.01	-0.01	0.00	0.00	0.00
	5	0.19	0.06	0.09	0.20	0.05	0.10	0.20	0.01	0.08	0.12	-0.07	0.01
	10	0.21	0.06	0.09	0.22	0.06	0.10	0.12	-0.05	0.01	-0.02	-0.13	-0.08
4	20	0.22	0.07	0.10	0.22	0.04	0.09	-0.01	-0.12	-0.08	-0.02	-0.04	-0.03
	50	0.22	0.07	0.10	0.17	-0.01	0.04	-0.01	-0.02	-0.02	0.00	0.00	0.00
	100	0.22	0.07	0.10	0.05	-0.10	-0.05	0.00	0.00	0.00	0.00	0.00	0.00
	5	0.32	0.07	0.09	0.33	0.06	0.09	0.30	0.01	0.06	0.13	-0.10	-0.05
	10	0.35	0.07	0.09	0.36	0.06	0.09	0.15	-0.09	-0.04	-0.01	-0.08	-0.06
8	20	0.36	0.07	0.09	0.37	0.05	0.08	0.00	-0.09	-0.07	0.00	0.00	0.00
	50	0.37	0.07	0.10	0.25	-0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.37	0.07	0.09	0.06	-0.11	-0.08	0.00	0.00	0.00	0.00	0.00	0.00

Unbalanced Designs

		$f^2 = 0.00$			$f^2 = 0.02$			$f^2 = 0.15$			$f^2 = 0.35$		
Groups	N	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}
	5	0.11	0.04	0.09	0.11	0.03	0.09	0.06	-0.03	0.04	-0.02	-0.14	-0.04
	10	0.11	0.05	0.10	0.10	0.03	0.08	-0.03	-0.13	-0.05	-0.18	-0.28	-0.19
2	20	0.12	0.05	0.10	0.08	0.01	0.06	-0.17	-0.26	-0.18	-0.20	-0.27	-0.20
	50	0.12	0.05	0.10	0.00	-0.08	-0.01	-0.18	-0.24	-0.19	-0.03	-0.04	-0.03
	100	0.12	0.06	0.10	-0.11	-0.19	-0.12	-0.04	-0.06	-0.04	0.00	0.00	0.00
	5	0.20	0.06	0.09	0.21	0.06	0.10	0.24	0.05	0.13	0.19	0.01	0.09
	10	0.21	0.06	0.10	0.23	0.07	0.11	0.19	0.03	0.09	0.04	-0.04	0.00
4	20	0.22	0.07	0.10	0.25	0.07	0.12	0.06	-0.03	0.01	0.00	-0.01	-0.01
	50	0.22	0.07	0.10	0.23	0.05	0.10	0.00	-0.01	0.00	0.00	0.00	0.00
	100	0.22	0.07	0.10	0.13	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00
	5	0.32	0.07	0.09	0.35	0.08	0.11	0.38	0.10	0.16	0.21	0.06	0.10
	10	0.35	0.07	0.09	0.40	0.09	0.12	0.23	0.07	0.11	0.02	0.00	0.00
8	20	0.36	0.07	0.09	0.42	0.11	0.15	0.03	0.00	0.01	0.00	0.00	0.00
	50	0.37	0.07	0.09	0.34	0.10	0.14	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.37	0.07	0.10	0.12	0.04	0.05	0.00	0.00	0.00	0.00	0.00	0.00

Note. Estimates are based on 50,000 samples.

Table 2. RMSE of Three Point Estimates of Retrospective Power in One Factor Designs.

Balanced Designs

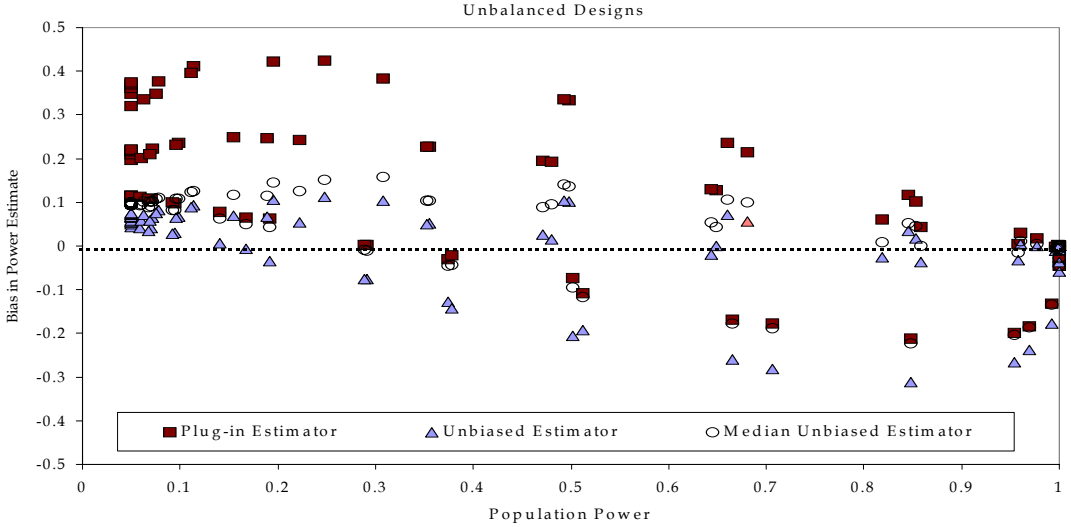
		$f^2 = 0.00$			$f^2 = 0.02$			$f^2 = 0.15$			$f^2 = 0.35$		
Groups	N	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}
	5	0.20	0.13	0.19	0.22	0.14	0.20	0.28	0.22	0.27	0.30	0.28	0.29
	10	0.20	0.13	0.19	0.22	0.17	0.22	0.29	0.28	0.29	0.27	0.33	0.28
2	20	0.19	0.14	0.19	0.24	0.20	0.24	0.28	0.33	0.29	0.18	0.23	0.18
	50	0.19	0.14	0.19	0.27	0.26	0.28	0.16	0.20	0.16	0.02	0.02	0.02
	100	0.19	0.14	0.19	0.29	0.31	0.29	0.03	0.04	0.03	0.00	0.00	0.00
	5	0.26	0.14	0.18	0.28	0.16	0.20	0.32	0.23	0.27	0.28	0.30	0.29
	10	0.27	0.15	0.18	0.30	0.18	0.22	0.28	0.30	0.29	0.18	0.29	0.24
4	20	0.28	0.15	0.19	0.32	0.21	0.25	0.19	0.29	0.26	0.05	0.10	0.08
	50	0.28	0.15	0.18	0.30	0.27	0.28	0.04	0.07	0.06	0.00	0.00	0.00
	100	0.28	0.15	0.18	0.24	0.31	0.29	0.00	0.00	0.00	0.00	0.00	0.00
	5	0.38	0.15	0.18	0.39	0.17	0.20	0.38	0.26	0.28	0.22	0.31	0.28
	10	0.40	0.16	0.18	0.42	0.20	0.23	0.23	0.30	0.28	0.06	0.17	0.15
8	20	0.41	0.16	0.18	0.43	0.24	0.26	0.08	0.19	0.17	0.00	0.01	0.01
	50	0.42	0.16	0.18	0.32	0.29	0.29	0.00	0.01	0.01	0.00	0.00	0.00
	100	0.42	0.16	0.18	0.14	0.27	0.25	0.00	0.00	0.00	0.00	0.00	0.00

Unbalanced Designs

		$f^2 = 0.00$			$f^2 = 0.02$			$f^2 = 0.15$			$f^2 = 0.35$		
Groups	N	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}
	5	0.20	0.13	0.19	0.21	0.13	0.20	0.24	0.19	0.24	0.28	0.28	0.28
	10	0.19	0.13	0.19	0.21	0.15	0.20	0.27	0.28	0.27	0.34	0.41	0.35
2	20	0.19	0.14	0.19	0.22	0.18	0.22	0.33	0.40	0.34	0.31	0.38	0.32
	50	0.19	0.14	0.19	0.24	0.24	0.25	0.29	0.35	0.29	0.07	0.09	0.07
	100	0.19	0.14	0.19	0.30	0.34	0.31	0.10	0.13	0.10	0.00	0.01	0.00
	5	0.27	0.14	0.18	0.29	0.16	0.21	0.35	0.26	0.30	0.31	0.30	0.30
	10	0.28	0.15	0.19	0.32	0.19	0.23	0.31	0.30	0.30	0.14	0.21	0.18
4	20	0.28	0.15	0.18	0.34	0.24	0.27	0.16	0.23	0.20	0.02	0.05	0.04
	50	0.28	0.15	0.18	0.34	0.29	0.31	0.02	0.03	0.02	0.00	0.00	0.00
	100	0.28	0.15	0.18	0.25	0.28	0.27	0.00	0.00	0.00	0.00	0.00	0.00
	5	0.38	0.15	0.18	0.41	0.19	0.22	0.44	0.30	0.33	0.25	0.25	0.25
	10	0.40	0.16	0.18	0.45	0.22	0.25	0.27	0.26	0.26	0.03	0.06	0.05
8	20	0.41	0.16	0.18	0.48	0.28	0.30	0.04	0.08	0.07	0.00	0.00	0.00
	50	0.42	0.16	0.18	0.38	0.31	0.31	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.42	0.16	0.18	0.14	0.17	0.16	0.00	0.00	0.00	0.00	0.00	0.00

Note. Estimates are based on 50,000 samples.

Figure 3
Statistical Bias in Point Estimates of Retrospective Power
Unbalanced Designs



a medium effect size (f^2 = .15), the unbiased estimator evidenced negative bias of –0.10, for the column, row, and interaction effects with n = 10 for all factorial designs. Once again, with large samples and large effect sizes, all of the estimators converged to the true power (i.e., showing zero bias). For the 2 X 2 factorial designs (in which each effect is tested with a single degree of freedom), trends in bias were similar for all of the power estimates across all effect sizes. However, for the 2 X 4 factorial designs, more striking similarities were witnessed for the column and interaction effects (each tested with three degrees of freedom). While maintaining a similar pattern, in general, the bias estimates were slightly smaller for the row effects than for the column and interaction effects.

The root mean squared errors (RMSEs) of the point estimates are provided in Table 5 and illustrated in Figure 8. An examination of these statistics revealed a considerable amount of error associated with small effect sizes and small samples for all effects examined (i.e. row, column and interaction effects). Substantially less error was evidenced when medium and large effect sizes were paired with larger sample sizes. Additionally, the magnitude of the RMSE did not appear to differ systematically across the row, column, or interaction effects.

For the interval power estimates, the proportion of confidence bands that contained the true value of power are presented in Table 6. For all effects, the intervals showed 95% coverage across all conditions. In general, the average width of confidence bands (Table 6) suggests that these bands are relatively uninformative, that is they provide very little information on true power for small samples and small effect sizes. Only when medium or large effect sizes were paired with large samples sizes, did the width of the bands become small enough to be considered useful.

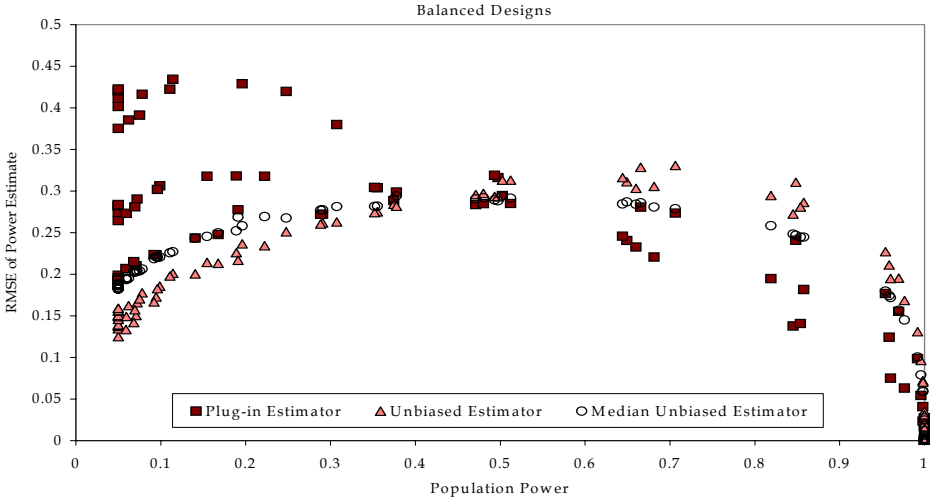
Discussion

The results suggest that none of the retrospective power estimation techniques evaluated were effective across the conditions examined. For point estimates, the "unbiased" and "median unbiased" estimators showed improved performance relative to the plug-in estimator, but these procedures were not completely free from bias except under large sample sizes and large effect sizes (as the statistical power approaches unity). Further, the sampling error in these estimates, reflected in the RMSE, suggests large sampling deviations for all three of the point estimators. These sampling deviations are greatly reduced with large sample estimates of retrospective power.

The confidence band approach suggested by Steiger and Fouladi (1997) provided excellent coverage of the parameter across most of the conditions examined. The coverage problems observed under extreme conditions (i.e., $f^2 = 0$ for both balanced and unbalanced designs, and $f^2 = .35$ with large sample, unbalanced designs) represent research contexts in which power is either zero or very close to one. The calculation of a one-sided confidence interval (e.g., "I am 95% sure that the power is greater than .986") rather than a two-sided band should improve the performance of the confidence bands and may be more useful than a two-sided interval at these extremes.

The coverage results obtained from the confidence band approach suggest that the method appears to be a wise choice (because it is unbiased). However, the width of the resulting confidence bands that provide such excellent coverage were typically so broad that they provided little information about

Figure 4
RMSE of Point Estimates of Retrospective Power
Balanced Designs



the true power of the study. Only with relatively large samples (e.g., n = 100 per cell for one-factor designs) and large effect sizes did the band width become small enough that it appears to be useful for research applications. As with the RMSE associated with the point estimates, the width of these confidence bands reflects the large amount of sampling error that appears to be inherent in retrospective power analysis. For researchers who have the luxury of working with very large samples, these bands appear to be the best approach to power analyses.

Although prospective power analysis is of critical importance in the planning of empirical investigations, retrospective power analysis is important for both the interpretation of research results and the planning of subsequent studies, hence it is a logical extension of the substantive interpretation of sample effect sizes. However, retrospective power analysis has received little attention in the research methods literature. Our results suggest that the currently available methods for retrospective power analysis evidence severe limitations (except for studies with large sample sizes) in terms of statistical bias and large sampling errors. Such results highlight the magnitude of the caveats that should be employed when researchers use retrospective power estimates. Additionally, these results suggest that improved methods of estimation appear to be necessary to supply researchers with an important tool that can be trusted to provide unbiased and precise estimates of retrospective power across conditions typically encountered in applied research.

References

Brewer, J. K. & Singular, P. T. (1987). Adequate sample size: A priori and post hoc considerations. *Journal of Special Education, 21*, 74 - 84.

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

Gerard, P. D., Smith, D. R. & Weerakkody, G. (1998). Limits of retrospective power analysis. *Journal of Wildlife Management, 62*, 801 - 807.

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

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

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

Ottenbacher, K. J. & Maas, F. (1998). How to detect effects: Statistical power and evidence-based practice in occupational therapy research. *American Journal of Occupational Therapy, 53*, 181 - 188.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical recipes in FORTRAN: The art of scientific computing* (2nd Ed.). New York: Cambridge.

Steiger, J. H. & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Muliak & J. H. Steiger (Eds.), *What if there were no significance tests?* Mahweh, NJ: Erlbaum.

Taylor, D. J. & Muller, K. E. (1996). Bias in linear model power and sample size cacluation due to estimating noncentrality. *Communications in Statistics: Theory and Methods, 25*, 1595-1610.

Zumbo, B. D. & Hubley, A. M. (1998). A note on misconceptions concerning prospective and retrospective power. *The Statistician, 47*, 385 - 388.