

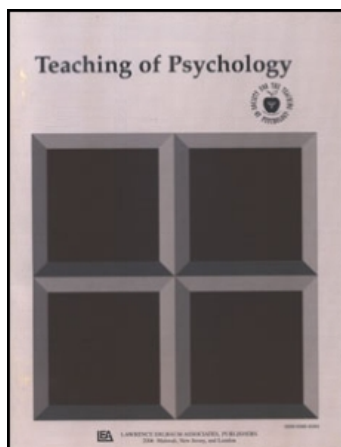
This article was downloaded by: [Florida International University]

On: 25 September 2009

Access details: Access Details: [subscription number 907746653]

Publisher Routledge

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



## Teaching of Psychology

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title-content=t775653707>

### The Power of Teaching Activities: Statistical and Methodological Recommendations

Thomas J. Tomcho <sup>a</sup>; Rob Foels <sup>b</sup>

<sup>a</sup> Salisbury University, <sup>b</sup> Amherst College,

Online Publication Date: 01 April 2009

**To cite this Article** Tomcho, Thomas J. and Foels, Rob(2009)'The Power of Teaching Activities: Statistical and Methodological Recommendations',Teaching of Psychology,36:2,96 — 101

**To link to this Article:** DOI: 10.1080/00986280902739743

**URL:** <http://dx.doi.org/10.1080/00986280902739743>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

# The Power of Teaching Activities: Statistical and Methodological Recommendations

Thomas J. Tomcho  
Salisbury University

Rob Foels  
Amherst College

*Researchers rarely mention statistical power in Teaching of Psychology teaching activity studies. Insufficiently powered tests promote uncertainty in the decision to accept or reject the tested null hypothesis and influence the interpretation of results. We analyzed the a priori power of statistical tests from 197 teaching activity effectiveness studies published from 1974 through 2006. We found that two thirds of the studies were powerful enough to detect only large effects. We compared observed sample sizes with expert recommendations and found that studies typically used sample sizes that were too small. We discuss limitations of underpowered statistical tests in evaluating teaching activity effectiveness and make design-related recommendations for improving power.*

Statistical experts recommend reporting a priori power analyses for published studies (e.g., Bailar & Mosteller, 1988; Cohen, 1962, 1988, 1992a, 1992b; Wilkinson & the Task Force on Statistical Inference, 1999). The power of a test is  $1 - \beta$ , or its probability of correctly rejecting a false null hypothesis (Cohen, 1992a), and researchers who a priori set parameters to achieve sufficient power can have greater confidence in their results (Cohen, 1988). By convention (Cohen, 1969), power of .8 is minimally sufficient and equates to an 80% likelihood of a hypothesis test being statistically significant. Of practical importance, when researchers conduct research with sufficient power (i.e.,  $> .8$ ), failure to reject the null hypothesis warrants high researcher confidence that they have made the correct decision; when conducted with insufficient power (i.e.,  $< .8$ ), low confidence in the correctness of the decision is warranted (Cohen, 1988; Rossi, 1990). Therefore, researchers using underpowered studies might hamper

efforts to evaluate the effectiveness of psychology teaching activities and methods.

## Power Parameters

Statistical power is primarily a function of effect size, significance level, and sample size. Of these parameters, researchers minimally control a teaching activity's effect size. Likewise, researchers typically have little impact on significance level, as convention dictates that researchers set an alpha level equal to .05. However, researchers can readily estimate sample size. Depending on the type of statistical test, Cohen (1988, 1992a) has recommended guidelines to detect large, medium, and small effects. For example, for correlation coefficients, correlations of .5, .3, and .1 represent large, medium, and small effects, respectively. Researchers can use commercial software (e.g., Power and Precision; Borenstein, Rothstein, & Cohen, 2001) and freeware (e.g., G\*Power 3; Faul, Erdfelder, Lang, & Buchner, 2007) to easily estimate sample sizes for combinations of significance level and power for various statistical tests.

## Importance of Sample Size

Adequate sample size enables accurate inferences about the population. However, statistical inference involves attendant probability that either Type I or Type II error will occur. A priori estimation of sample

sizes maximizes power and minimizes risks of Type I error (Asraf & Brewer, 2004; Cohen, 1988). Typically, *Teaching of Psychology (ToP)* authors rely on samples made up of students enrolled in their courses.

Although psychological researchers are increasing their attention to power, we conducted an electronic search of the terms *statistical power* or *power* in *ToP* and found only three studies in which researchers addressed power (i.e., Harlow, Burkholder, & Morrow, 2006; Lakin & Wichman, 2005; LoSchiavo & Roberts, 2005). Not addressing power when evaluating teaching activities and methods limits the detection of true effects due to the likelihood of increased Type I error.

Based on the importance of power estimation, we examined studies on teaching activities and methods published in *ToP* to answer three questions. Did researchers' statistical tests have sufficient a priori power to detect large, medium, or small effects? Did researchers use adequate sample sizes to have sufficiently powered statistical tests of hypotheses? Is *ToP* researchers' power to detect effects comparable to findings in other studies of psychology and teaching?

## Method

We identified *ToP* teaching activity and methods studies from 1974 through 2006 in which researchers (a) reported statistical significance test results, (b) reported sample size data, and (c) employed a commonly used significance test from which we could calculate power (see Cohen, 1988). Our sample included 193 articles with a total of 197 independent studies (four articles each contained two studies using independent samples; i.e., Madigan & Brosamer, 1990; Sagarin & Lawler-Sagarin, 2005; Thomas & McDaniel, 2004; Waschull, 2001) and a total of 497 statistical significance test results, which were primarily *t* tests for independent and dependent samples, correlation coefficients, and chi-square tests.

We used Borenstein et al.'s (2001) Power and Precision software to calculate power estimates. We assumed a two-tailed hypothesis test for all studies. To adhere to methodological approaches used in power analysis (e.g., Rossi, 1990; Sedlmeier & Gigerenzer, 1989) we calculated an average power across hypothesis tests within study. We used an alpha level equal to .05 for each study and corrected alpha for articles employing multiple hypothesis tests to obtain a study-level experiment-wise error rate equal to .05. In the four articles that each reported two studies involving

independent samples, we treated the study as the unit of analysis, correcting the study hypothesis tests to obtain a study-level experiment-wise error rate equal to .05.

We extracted the total *N* for each independent study ( $n = 197$ ). Where relevant, and where researchers reported data, we identified sample sizes per condition (e.g., teaching activity intervention group vs. comparison group). In those studies in which researchers conducted multiple hypothesis tests with varying sample sizes, we calculated an average sample size per study. In 17 studies, researchers employed more than one type of design analysis, most notably a between groups independent *t* test design along with a pretest–posttest dependent *t* test design. Therefore, for studies in which researchers employed multiple types of designs, we derived an averaged estimate of sample size for each type of design analysis (i.e., independent *t* test, dependent *t* test, chi-square test, or correlation). Thus, we analyzed 214 sample sizes.

We next conducted power analyses using expert effect size recommendations (Borenstein et al., 2001; Cohen, 1988, 1992a) to assess studies' a priori power to detect large, medium, and small effects. To detect large, medium, and small effects, we used, respectively, Cohen's *d* equal to .8, .5, and .2 to evaluate the power of *t* tests for independent and dependent samples and *r* equal to .5, .3, and .1 to evaluate the power of correlational tests. We used differences equal to .38, .25, and .1 to evaluate the power of chi-square tests.

## Results

### *Statistical Power*

**Descriptive statistics.** Across the 197 studies and 497 statistical tests, researchers reported an average of 2.5 ( $SD = 2.3$ ) statistical tests per study. The modal number of statistical tests per study was 1 ( $n = 81$ ). The range of statistical tests per study was 1 to 12.

**A priori power analyses.** Table 1 contains the power levels for all quartiles calculated by combining hypothesis tests within a study. We conducted a priori power analyses based on the reported sample size, experiment-wise error rate-corrected alpha level equal to .05 per study as the unit of analysis, and expected power of .8 to detect each of Cohen's (1988) large, medium, and small effects. The average power to detect a large effect was .81 ( $SD = .23$ ), a medium effect

**Table 1. A Priori Statistical Power ( $N = 197$  Studies)**

	Effect Size		
	Small $\alpha$	Medium $\alpha$	Large $\alpha$
<i>M</i>	.14	.53	.81
<i>SD</i>	.13	.28	.23
25th percentile	.06	.31	.72
50th percentile	.10	.52	.92
75th percentile	.17	.77	.99
Studies with power $\geq .8$	1	44	130

*Note.* Results for two-tailed hypothesis tests using experiment-wise error rate equal to .05.

was .53 ( $SD = .28$ ), and a small effect was .14 ( $SD = .13$ ). We also examined the percentage of studies in which researchers had power equal to or greater than .8. Researchers were able to detect large effects in 66% of the studies. However, both Osborne, Christensen, and Gunter (2001) in educational psychology, and also Tomcho and Foels (2008) in teaching activity and methods research published in *ToP*, identified medium effects as the norm. *ToP* researchers detected medium effects in 22% of the studies and small effects in only 0.5% of studies.

### Sample Size

**Descriptive statistics.** Sample sizes ranged from 8 to 631 with an average of 79.1 ( $SD = 75.2$ ) and quartiles of 34 (25th quartile), 63 (median), and 111 (75th quartile) participants. Given the large sample size standard deviations, we included quartiles for sample size analyses in Table 2.

**Sample size analyses.** Table 2 contains our examination of observed versus recommended sample sizes from 180 studies for *t* test for independent samples, *t* test for dependent samples, chi-square test, and correlations. We deleted 34 studies because researchers failed to provide sufficient information to determine conclusively the number of participants per condition. For *t* tests for independent samples ( $k = 90$ ), approximately 25% of *ToP* researchers had sample sizes sufficient to detect medium effects, whereas approximately 67% had sample sizes sufficient to detect large effects. For *t* tests for dependent samples ( $k = 68$ ), approximately 50% of *ToP* researchers had sample sizes sufficient to detect medium effects, whereas approximately 90% had sample sizes sufficient to detect large effects.

For chi-square test samples with 1 degree of freedom in the numerator ( $k = 14$ ), less than 5% of *ToP* researchers had sample sizes sufficient to detect medium effects, whereas approximately 50% had sample sizes

**Table 2. Sample Size Analyses by Type of Statistical Test ( $N = 180$  Studies)**

	Statistical Test			
	<i>t</i> Test		Chi-Square	Correlation
	Independent	Dependent		
Recommended sample				
Medium effect	64 per condition	34	124	82
Large effect	26 per condition	15	52	26
Current study sample				
<i>M</i>	49.6 experimental 51.7 comparison	55.0	62.9	46.7
<i>SD</i>	38.2 experimental 71.0 comparison	56.2	36.6	39.6
25th percentile	22.0 experimental 22.0 comparison	20.0	28.6	20.0
50th percentile	34.0 experimental 34.0 comparison	32.4	56.5	28.0
75th percentile	69.5 experimental 59.8 comparison	70.2	80.8	54.5
Percentile needed to obtain recommended medium effect	72nd experimental 76th comparison	53rd	97th	87th
Percentile needed to obtain recommended large effect	33rd experimental 34th comparison	10th	48th	47th

sufficient to detect large effects. For correlational analysis samples ( $k = 8$ ), less than 15% of *ToP* researchers had sample sizes sufficient to detect medium effects, whereas approximately 50% had sample sizes sufficient to detect large effects.

### *Comparison With Other Power Studies*

Comparable to findings in the field of psychology (e.g., Cohen, 1962; Rossi, 1990; Sedlmeier & Gigerenzer, 1989), *ToP* researchers more likely detected large than medium or small effects. In a secondary analysis, *ToP* researchers' power to detect effects was generally similar to teaching-related journals in other disciplines (e.g., *Journal of Research in Science Teaching* [JRST]) for medium and large effect sizes, but less for small effect sizes (e.g., Daly & Hexamer, 1983; Penick & Brewer, 1972; Woolley, 1983; Woolley & Dawson, 1983). For example, Woolley and Dawson (1983) used a two-tailed alpha of .05 for each hypothesis test in examining 192 articles published in JRST and found statistical power estimates of .23, .63, and .85, for small, medium, and large effects, respectively. We recalculated statistical power for each hypothesis test in our sample using a two-tailed alpha equal to .05 for each test (rather than controlling for experiment-wise error rates) to compare our results with Woolley and Dawson. In comparison, *ToP* researchers ( $M = .19$ ) were significantly less likely than JRST researchers ( $M = .23$ ) to be able to detect small effects,  $t(387) = 2.27$ ,  $p = .03$ . However, *ToP* researchers ( $M = .61$ ;  $M = .87$ ) were as likely as JRST researchers ( $M = .63$ ;  $M = .85$ ) to detect medium effects,  $t(387) = 0.76$ , *ns*, and large effects,  $t(387) = 1.09$ , *ns*, respectively.

## Discussion

We examined the power of research published in *ToP* with three goals in mind. We sought to (a) determine whether the level of a priori power associated with statistical tests from teaching activity and methods studies was sufficient to detect large, medium, or small effects; (b) assess whether researchers were using adequate sample sizes to have sufficiently powered statistical tests of hypotheses reported in *ToP*; and (c) compare *ToP* researchers' statistical power with the power evidenced by other researchers in psychology and science education.

In terms of a priori power, *ToP* researchers conducted research that, averaged across all types of tests,

had a 22% probability that the studies' hypothesis tests would be statistically significant if the underlying effect was medium. In terms of sample size, *ToP* researchers used sample sizes that were sufficient in nearly two thirds of the studies if the underlying effect was large. However, to detect medium effects, researchers used sample sizes that were sufficient in only one fifth of the studies. The exception for medium effects was *t* tests for dependent samples, in which approximately half of the samples were sufficient to detect medium effects.

Relative to findings in the field of science teaching, *ToP* researchers were as likely to detect large or medium effects as were other researchers (e.g., Woolley & Dawson, 1983). Given Osborne et al.'s (2001) and Tomcho and Foels's (2008) findings that psychology teaching research results in a medium effect, *ToP* researchers appear just as likely to be able to detect effects as researchers teaching in other areas of science. The detection of small effects was the only area where *ToP* researchers were statistically less likely than other science teachers to detect effects, but this finding might be an artifact of the large number of hypothesis tests ( $n = 3,556$ ) examined by Woolley and Dawson.

### *Recommendations to Increase Power*

1. *ToP* researchers must attend to power in planning their teaching activity effectiveness research. Researcher failure to attend to power seriously limits the instructional development efforts in the field of teaching of psychology.
2. Researchers are generally employing suboptimal sample sizes to detect anything short of a large effect. Based on Osborne et al. (2001) and Tomcho and Foels (2008), researchers have a reasonable starting estimate of a likely medium effect, which they should adjust by either theoretical rationale or previous research on the psychological phenomenon being taught (Hallahan & Rosenthal, 1996; Kraemer, Mintz, Noda, Tinklenberg, & Yesavage, 2006).
3. Researchers are conducting multiple hypothesis tests of the effectiveness of their teaching activities, often without mentioning adjustments to control for experiment-wise error rates. Although researchers conventionally choose a significance level equal to .05, researchers can reset the level of Type I error (e.g., Cohen, 1990; Rosnow & Rosenthal, 1989). For exploratory research, they might consider increasing their alpha (Cohen, 1988; Hallahan & Rosenthal, 1996) or employ focused tests of the limited hypotheses of interest, rather than

conducting omnibus tests or multiple hypothesis tests that will result in decreased power (Hallahan & Rosenthal, 1996).

4. Given that *ToP* researchers rarely randomly assign participants when comparing independent samples, the use of more homogeneous participant populations or blocking variables can increase power (Hallahan & Rosenthal, 1996). For example, if a researcher examines the effectiveness of a teaching activity in a statistics course, another statistics course is a more homogeneous sample than an introductory psychology course as a comparison group. Researchers who account for additional within-group variability by using more homogeneous samples increase their power. For blocking variables, researchers expecting their teaching activity to increase examination performance might consider blocking on students' existing grade point average as a way to reduce error variance and increase power. Repeated measures designs that block on the individual can provide more power than is possible from independent groups comparison approaches (Hallahan & Rosenthal, 1996).

#### *Power and the Evaluation of Teaching Activities*

We have presented a case for researchers to address power in demonstrating the quality of teaching activities and methods. However, three potential caveats to our findings warrant consideration: representativeness of our sample, class size, and the role of traditional significance testing in evaluating teaching.

We examined all teaching activity effectiveness research published in *ToP* between 1974 and 2006. However, only 193 articles (i.e., 28% of all articles in Tomcho et al., 2008) reported statistical tests of hypotheses. Moreover, in an ancillary analysis we found that 62% of the 193 published teaching effectiveness articles including statistical tests of effectiveness have been published since 2000. Thus, readers should view our results as a preliminary snapshot of the issue of power in *ToP* teaching activity and methods effectiveness research.

Concerning class size, teachers might counter that in contrast to laboratory-based examinations of psychological phenomena, classroom-based examinations of teaching effectiveness have limited access to adequate samples. In response to this concern we suggest that classroom-based researchers increase power by testing their hypotheses across multiple sections of a course to ensure adequate sample size. This approach is not conceptually different from laboratory-based researchers

who conduct multiple sessions of their experiment to ensure adequate sample sizes.

A more serious concern that some teachers might raise is that traditional significance testing might not provide effectiveness evidence for teaching activities that otherwise have a phenomenal in-class response or other unquantifiable experiential qualities. Without joining the controversial fray surrounding the need for rigorous statistical evaluation of *all* in-class teaching activities, we encourage researchers to estimate power and sample size when they design evaluations of their teaching activities. If researchers fail to use adequate power, the field of psychology might miss some good teaching activities because researchers will be unable to publish them due to nonsignificant findings, whereas researchers using more powerful tests will enhance the likelihood of significant findings. At the very least, researchers should estimate and report statistical power so that readers can better evaluate the results.

#### References

- Asraf, R. M., & Brewer, J. K. (2004). Conducting tests of hypotheses: The need for an adequate sample size. *The Australian Educational Researcher*, 31, 79–94.
- Bailar, J. C., III, & Mosteller, F. (1988). Guidelines for statistical reporting in medical journals: Amplifications and explanations. *Annals of Internal Medicine*, 108, 266–273.
- Borenstein, M., Rothstein, H., & Cohen, J. (2001). Power and precision [Computer software]. Englewood, NJ: Biostat, Inc.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145–153.
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304–1312.
- Cohen, J. (1992a). A power primer. *Psychological Bulletin*, 112, 155–159.
- Cohen, J. (1992b). Statistical power analysis. *Current Directions in Psychological Science*, 1, 98–101.
- Daly, J. A., & Hexamer, A. (1983). Statistical power in research in English education. *Research in the Teaching of English*, 17, 157–164.
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191.

- Hallahan, M., & Rosenthal, R. (1996). Statistical power: Concepts, procedures, and applications. *Behaviour Research and Therapy*, 34, 489–499.
- Harlow, L. L., Burkholder, G. J., & Morrow, J. A. (2006). Engaging students in learning: An application with quantitative psychology. *Teaching of Psychology*, 33, 231–235.
- Kraemer, H. C., Mintz, J., Noda, A., Tinklenberg, J., & Yesavage, J. A. (2006). Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives of General Psychiatry*, 63, 484–489.
- Lakin, J. L., & Wichman, A. L. (2005). Applying social psychological concepts outside the classroom. *Teaching of Psychology*, 32, 110–113.
- LoSchiavo, F. M., & Roberts, K. L. (2005). Testing pseudo-scientific claims in research methods courses. *Teaching of Psychology*, 32, 177–180.
- Madigan, R., & Brosamer, J. (1990). Improving the writing skills of students in introductory psychology. *Teaching of Psychology*, 17, 27–30.
- Osborne, J. W., Christensen, W. R., II, & Gunter, J. S. (2001, April). *Educational psychology from a statistician's perspective: A review of the quantitative quality of our field*. Paper presented at the 2001 meeting of the American Educational Research Association. Retrieved May 15, 2006, from <http://www4.ncsu.edu/~jwosbor2/otherfiles/MyResearch/POWER.pdf>
- Penick, J. E., & Brewer, J. K. (1972). The power of statistical tests in science teaching research. *Journal of Research in Science Teaching*, 9, 377–379.
- Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276–1284.
- Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, 58, 646–656.
- Sagarin, B. J., & Lawler-Sagarin, K. A. (2005). Critically evaluating competing theories: An exercise based on the Kitty Genovese murder. *Teaching of Psychology*, 32, 167–169.
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309–316.
- Thomas, J. H., & McDaniel, C. R. (2004). Effectiveness of a required course in career planning for psychology majors. *Teaching of Psychology*, 31, 22–27.
- Tomcho, T. J., & Foels, R. (2008). Assessing effective teaching of psychology: A meta-analytic integration of learning outcomes. *Teaching of Psychology*, 35, 286–296.
- Tomcho, T. J., Foels, R., Rice, D., Johnson, J., Moses, T. P., Warner, D. J., Wetherbee, R., & Amalfi, T. (2008). A review of ToP teaching strategies: Links to students' scientific inquiry skills development. *Teaching of Psychology*, 35, 147–159.
- Waschull, S. B. (2001). The online delivery of psychology courses: Attrition, performance, and evaluation. *Teaching of Psychology*, 28, 143–147.
- Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604.
- Woolley, T. W. (1983). A comprehensive power-analytic investigation of research in medical education. *Journal of Medical Education*, 58, 710–715.
- Woolley, T. W., & Dawson, G. O. (1983). A follow-up power analysis of the statistical tests used in the *Journal of Research in Science Teaching*. *Journal of Research in Science Teaching*, 20, 673–681.

## Notes

1. Please contact the authors for the list of articles used in the current analyses.
2. Thomas J. Tomcho collected some of these data while he was a faculty member at Delaware State University.
3. Send correspondence to Thomas J. Tomcho, Department of Psychology, Salisbury University, Salisbury, MD 21802; e-mail: [tjtomcho@salisbury.edu](mailto:tjtomcho@salisbury.edu).