Running head: POWER ANALYSIS SOFTWARE (Revised 10-31-2011)

Power Analysis Software for Educational Researchers

Chao-Ying Joanne Peng

Haiying Long

Serdar Abaci

Indiana University

Author Note

An earlier version of this paper was presented at the 2010 annual meeting of American Educational Research Association in Denver, CO.

Correspondence concerning this manuscript should be addressed to C.-Y. Joanne Peng, Department of Counseling and Educational Psychology and Department of Statistics, Indiana University, Bloomington, IN 47405. E-mail: peng@indiana.edu.

Abstract

Given the importance of statistical power analysis in quantitative research and the repeated emphasis on it by AERA/APA journals, we examined the reporting practice of power analysis by the quantitative studies published in 12 education/psychology journals between 2005 and 20~~09~~10. It was surprising to uncover that less than 2% of the studies conducted prospective power analysis. Another 3.5~~4~~% computed observed power, a practice not endorsed by the literature on power analysis. In this paper, we clarify these two types of power analysis and discuss functionalities of eight programs/packages (G*Power 3.1.3, PASS 11, SAS/STAT 9.3, Stata 12, SPSS 19, SPSS/Sample Power 3.0.1, Optimal Design Software 2.01, and MLPowSim 1.0 BETA) to encourage proper and planned power analysis. Based on our review, we recommend two programs (SPSS/Sample Power and G*Power) for general-purpose univariate/multivariate analyses, and one (Optimal Design Software) for hierarchical/multilevel modeling and meta-analysis. Recommendations are also made for reporting power analysis results and exploring additional software. The paper concludes with an examination of the role of statistical power in research and viable alternatives to hypothesis testing.

Keywords: statistical power analysis, prospective power, observed power, G*Power, PASS, Optimal Design Software, SAS, Stata, SPSS, Sample Power, MLPowSim, HLM

Power Analysis Software for Educational Researchers

Throughout the methodological literature, there has been no shortage of papers and books on power analysis, nor debates of various definitions of statistical power, especially since the publication of Cohen's seminal paper in 1962 (Cohen, 1962; Finch, Cumming, & Thomason, 2001; Hallahan & Rosenthal, 1996; Hoenig & Heisey, 2001; Jennions & Moller, 2003; Levine & Ensom, 2001; Murphy & Myors, 1998; O'Keefe, 2007; Ortiz, 2002; Sedlmeier & Gigerenzer, 1989; Thomas & Juanes, 1996; Yuan & Maxwell, 2005; Zumbo & Hubley, 1998). Many papers offered helpful guide on estimating desirable sample size and assessing power without relying on statistical software (e.g., Feldt & Mahmoud, 1958; Fox, 1956; Gillett, 1994a; Keselman, 1976; Koele, 1982; Levin, 1997; Miller & Knapp, 1972; Severo & Zelen, 1960). Likewise, popular statistical software, such as SAS, SPSS, responded to users' needs by providing procedures/modules for power and sample size calculations. Indeed, starting with the report written by the American Psychological Association (APA) Task Force on Statistical Inference (Wilkinson and the Task Force on Statistical Inference, 1999), the need to ensure sufficient statistical power with a suitable sample size for quantitative studies has been repeatedly emphasized. Specifically, the APA Task Force Report, under **Power and sample size**, recommends that researchers

> [P]rovide information on sample size and the process that led to sample size decisions. Document the effect sizes, sampling and measurement assumptions, as well as analytic procedures used in power calculations. [...]
>
> Largely because of the work of Cohen (1969, 1988), psychologists have become aware of the need to consider power in the design of their studies, before they collect data. The intellectual exercise required to do this stimulates authors to

> take seriously prior research and theory in their field, and it gives an opportunity, with incumbent risk, for a few to offer the challenge that there is no applicable research behind a given study. If exploration were not disguised in hypothetico-deductive language, then it might have the opportunity to influence subsequent research constructively.
>
> Computer programs that calculate power for various designs and distributions are now available. One can use them to conduct power analyses for a range of reasonable alpha values and effect sizes. Doing so reveals how power changes across this range and overcomes a tendency to regard a single power estimate as being absolutely definite.
>
> Many of us encounter power issues when applying for grants. Even when not asking for money, think about power. Statistical power does not corrupt. (Wilkinson and the Task Force on Statistical Inference, 1999, pp. 596-597)

These cogent words were reinforced by similar phrases in the 5th APA Publication Manual, and again in the most recent 6th edition under **Sample size, power, and precision** (APA Publication Manual, 2010, pp. 30-31). AERA's Standards for Reporting on Empirical Social Science Research in AERA Publications (AERA, 2006) is less specific about recommending power analysis and sample size estimation. It nonetheless emphasizes the need to report "any considerations that are identified during the data analysis (e.g., violations of assumptions of statistical procedures, failure of iterative statistical procedures to converge, changes in data analysis models necessitated by unexpected data patterns) that might compromise the validity of the statistical analyses or inference should be reported." (AERA, 2006, p. 37). Insufficient statistical power and sample size are two such considerations that could

compromise the validity of any statistical analysis, just as violations of assumptions, missing data, and other considerations cited in the Standards could.

The importance of statistical power and precise estimation of sample size is viewed not only as "a necessary condition of achieving success in scientific research," but also as "a procedural facet which is largely under the individual scientist's control" (Bausell & Li, 2002, p.2). Yet, these important, recommended practices have not been universally adopted by educational researchers who published in mainstream journals we reviewed. Furthermore, using statistical software to implement these practices is not as straightforward as the APA Task Force on Statistical Inference claimed (Wilkinson and the Task Force on Statistical Inference, 1999, p. 597 top). Thus, this paper seeks to encourage proper and planned power analysis by (1) clarifying confusion surrounding two types of power analysis (prospective versus observed) and by (2) reviewing functionalities of eight accessible programs/packages for power analysis. Included in our review are three specialized power analysis programs (G*Power, PASS, SPSS/Sample Power), three general-purpose statistical packages (SAS, Stata, SPSS/Statistics), and two suitable for hierarchical linear modeling (Optimal Design Software, MLPowSim). Recommendations are provided for selecting power analysis software, reporting power analysis results, and for exploring additional software. We conclude the paper with a discussion of the role of statistical power in inference-making in educational research and several viable alternatives to hypothesis testing.

The remainder of this paper is divided into five sections: (1) The State of Power Analysis in Published Studies, (2) Prospective versus Observed Power, (3) Software for Power Analysis, (4) Recommendations, and (5) Discussion.

The State of Power Analysis in Published Studies

To gauge the impact of APA Task Force Report on Statistical Inference (1999) and of AERA Report on Standards (2006) on the research practice of power analysis, we conducted a review of quantitative studies published in 12 journals between 2005 and 20~~09~~10. The 12 journals reviewed were *American Educational Research Journal*, *Educational Researcher*, *Journal of Counseling Psychology*, *Journal of Educational Psychology*, *Journal for Research in Mathematics Education*, *Journal of Research in Science Teaching*, *Journal of Research on Technology in Education*, *Journal of Special Education*, *Journal of School Psychology*, *The Modern Language Journal*, *Research in Higher Education*, and *Theory and Research in Social Education*. These journals were selected because of their emphasis on research, broad coverage of research topics, relevance to subfields in education, and reputable editorial policies. We assumed that the research reported in these 12 journals reflected the mainstream topics and methodologies practiced by educational researchers.

Our review included studies that employed inferential statistical tests, including those that used a mixed-methods approach, and for which statistical power was a relevant methodological consideration. Studies without inferential statistical analyses of empirical data (e.g., historical, qualitative, descriptive, philosophical, or review in nature) were excluded. Methodological papers without empirical data, or without an objective to answer real-world questions with empirical data, were likewise excluded. The type of statistical analyses employed ranged from one- and two-samples *z*- or *t*-test, to *F*-test for univariate and multivariate ANOVA or general linear models, and χ^2 test of proportions/frequencies, of goodness of fit for maximum-likelihood factor analysis, hierarchical linear modeling (HLM), structural equation modeling (SEM), or growth mixture modeling (GMM). Each article was read by one of the authors and the coding of its power analysis (or lack of) was cross-validated by another author. Differences in coding were

resolved through discussion and re-reading of the article. The final agreement between the two readers reached 100% for each journal. The unit of our review was article, not study or statistical analysis.

Results in Table 1 show that 1.7~~6~~7% (or 2~~0~~4) of 1~~13~~4357 articles published in 12 journals between 2005 and 20~~09~~10 estimated a desirable sample size during the planning stage—called the prospective (or "a priori," "planned") power analysis recommended by the APA Task Force. Another 3.467% (= 2.~~82~~58%+0.~~62~~59%+0.~~26~~29%; or 4~~7~~2) conducted power analysis after collecting and analyzing data, and 1~~2~~4.~~43~~15% (or 1~~41~~92) merely mentioned power and sample size without computation or estimation. The computation of power based on data (called the observed, or achieved, computed, estimated, post-hoc, posterior, or retrospective, power analysis) is not endorsed by the APA Task Force; its reporting should be discouraged. Thus, power analysis reported in 12 refereed journals between 2005 and 20~~09~~10 demonstrated a serious lack of adoption of APA's recommendations. Consequences of failing to adopt this recommendation are multifold, including the possibility of missing effects of interests due to inadequate sample sizes, poor use of resources, and inability to expand prior research or to constructively influence subsequent research, just to name a few.

In the next two sections, we seek to clarify the conceptual and computational differences between prospective and observed power analyses, and to provide a review of eight programs/packages that can assist educational researchers with prospective power analysis for a variety of research designs.

Prospective versus Observed Power

Statistical power (hereafter abbreviated as power) is a concept derived from the Neyman-Pearson null hypothesis testing paradigm. Within this paradigm, power is defined as the

conditional probability of rejecting the false null hypothesis (H_0), given a specific Type I error rate (α), the sample size (n), the directionality of the statistical test, and the population effect size (ES), namely, the degree of falsehood of H_0 (Kirk, 1995, 2008; Murphy & Myors, 1998; Sedlmeier & Gigerenzer, 1989). When all else is held as a constant, power increases as α, or n, or ES, increases. If the correct directionality is specified in the alternative hypothesis, a one-tailed test is more powerful than a two-tailed test. Apart from the directionality, the other four variables (power, α, n, and ES) are inter-related; if three of these four are specified, the fourth is automatically determined (Cohen, 1988, p. 14). Other factors, such as the reliability of the measurement or data, also impact the magnitude of power (Cohen, 1988, pp. 535-537). For the purpose of this paper, we assume that the reliability of measurements is perfect, hence, not an issue.

The literature identifies several different types of power analysis (Peng, Long, & Abaci, 2010). Of all the types, only ***prospective*** and ***observed*** power analyses were reported in 12 journals we reviewed. The goal of the prospective power analysis is to estimate a desirable sample size for a given power, α, and population ES, whereas the observed power analysis is conducted to estimate the power, given the sample size, α, and sample ES (see Table 2 for further differentiation and examples published in journals we reviewed). As the APA Task Force Report on Statistical Inference (1999, pp. 596-597) and the APA Publication Manual (2010, pp. 30-31) emphasized, sufficient power and sample size should be considered during the planning stage of a study, not afterwards. Therefore, the prospective power analysis is the power analysis that educational researchers should employ when making inferences within the Neyman-Pearson framework.

In contrast, the observed power analysis is conducted after data have been collected and analyzed using the observed or estimated ES from a sample to substitute for the population ES (Thomas, 1997). According to the study by Yuan and Maxwell (2005), the observed power—called estimated or post-hoc power in their paper—is positively biased, especially when the true power is small. The true power is defined in Yuan and Maxwell (2005) as the probability of rejecting H_0, given a specified α, the population ES, and the directionality of the test. The magnitude of bias cannot be offset even by a large n. When the true power is 0.5, the observed power is distributed as a uniform distribution with a maximal variance. When the true power is close to 0 or 1, the distribution of the observed power is highly skewed. Yuan and Maxwell consistently obtained these results from analytical, numerical, or Monte Carlo methods for one-sample *t*- and *z*-tests and two-sample *t*- and *z*-tests with equal but unknown variance. They therefore concluded that the calculation of the observed power was not useful when the population ES is small, regardless of the sample size. When the population ES is greater than 0.78, Yuan and Maxwell (2005, p. 163) suggested that "the observed power may provide some useful information."

Gerard, Smith, and Weerakkody (1998) also discussed the bias and precision of three estimators of the true power. They defined the precision to be the average width of the 95% confidence intervals, averaged over 500 replications, for $\alpha = .05$ and the true power of .05, .11, .34, .66, .90, and .98 respectively. Using simulations, they demonstrated that all three power estimators, i.e., the observed powers, were extremely variable and severely bounded. They tend to overestimate, especially when the true power is low. Furthermore, the observed power is monotonically inversely related to the *p*-value; the smaller the *p*-value is, the greater is the observed power, and vice versa.

In addition, observed power analysis tells us nothing about the ability of a statistical test detecting an ES of interest, or of importance. What's worse, it incurs at least two misinterpretations from researchers (Hoenig & Heisey, 2001). First, the magnitude of observed power is often misinterpreted as the support for the retention of H_0. Thus, a higher observed power (say, .60) is mistakenly interpreted as evidence of a stronger support for the retention of H_0 than a lower observed power (say, .30) (Hoenig & Heisey, 2001). Second, observed power is misused in computing "detectable ES" that is taken to be the upper bound for the true ES when H_0 is not rejected. Several numerical examples were provided in Hoenig and Heisey (2001) to expose the logical flaws in these two misinterpretations of observed power, which they termed the "power approach paradox." (Hoenig & Heisey, 2001, p. 21).

Finally, the concept of observed power as a conditional probability is deemed illogical, once a H_0 is rejected. Because, with this decision, the conditional probability of a Type II error (β) is 0; hence, the statistical power ($1-\beta$) is 1, conditioned on H_0 being false. If H_0 is not rejected, statistical power is defined to be α, the level of significance, because in this case, the H_0 distribution is taken to be the distribution of H_1, on which statistical power is defined. For these reasons, observed power should not be computed or reported. In our review, 32 articles made one of these two logical errors (Table 1).

Software for Power Analysis

We reviewed eight programs/packages for power analysis. Three are stand-alone/specialized programs for power analysis only (G*Power, PASS, SPSS/Sample Power), three are general-purpose statistical packages (SAS/STAT, Stata, SPSS/Statistics), and two are suitable for HLM and/or meta-analysis (Optimal Design Software, MLPowSim). They were chosen for the following reasons: (i) educational researchers reported using one of these for

power analysis between 2005 and 20~~09~~10 in 12 journals we reviewed; (ii) they are available in Windows or Mac operating system; (iii) they are accessible to and/or popular among educational researchers; and (iv) they are reputable in statistical computing. Of the eight software, G*Power, Optimal Design Software, and MLPowSim are free; PASS and SPSS/Sample Power can be downloaded for a trial use of limited period of time. All others are leased for a fee. Table 3 presents information, current as of October, 2011, on each product's version, price, and relevant websites.

These eight programs/packages differ in terms of their capability to conduct either prospective or observed power analysis, or both. They also differ in terms of functionalities and the richness of the output [e.g., one estimate or a range of estimates, with or without written interpretation beyond the estimate(s)]. These features and output formats are summarized in Table 4. The computation of observed power is not always made explicit by the program/package. Thus, Table 4 distinguishes these two kinds of power in describing each program/package's features. All conduct power analysis based on either ES, means and *SD*s, or others (such as proportions, odds ratio, relative risk). An ES is defined as the standardized mean or mean difference (e.g., Cohen's *d*, Hedges' *g*), the variance explained (e.g., ω^2, Cohen's *f* or f^2), regression slope, or odds ratio (Huberty, 2002). Any departure from these ES definitions is noted in the software descriptions below and also in Table 4. Readers are advised to pay special attention to the definition of ES when planning a study and estimating a desirable sample size to ensure sufficient statistical power. An ES of .8 for a between-subject factor in a split-block factorial design requires more participants to detect than the ES of the same magnitude for a within-subject design.

The free Optimal Design Software and MLPowSim are for sample size determination for

hierarchical/multi-level modeling and/or meta-analysis. Because of their unique purposes and functionality, these two programs are described in this section, but not contrasted with the other six in Table 4. Appendix (pp. 2 to 25) demonstrates the syntax or interface of all program/package, except for SPSS/Statistics as it computes observed power, but not the prospective power. The appendix is available from https://oncourse.iu.edu/access/content/user/peng/Appendix-power-2011.pdf.

Obviously, there are other programs that can perform power analysis, such as Lenth's web site with multiple calculators (Anonymous, 2003; Lenth, 2001), Dupont and Plummer's free PS program for multiple statistical procedures (http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize), Wheeler's programs for hand-held personal organizers (Wheeler, 2000), or those well suited for a particular field, such as medicine, ecology, biology (cited in Thomas & Krebs, 1997). And many have been evaluated elsewhere (Dattalo, 2009; Lewis, 2006; Thomas & Krebs, 1997). Yet these programs are either highly specialized, limited in functionalities/computing environment, did not distinguish different power definitions, or have not been updated regularly or recently. For these reasons, we did not include them in this paper. We also decided to not review software for precision, uncertainty, or sensitivity analysis, though we noted these capacities in the software we reviewed. Below is a brief description of each program/package; readers are advised to go directly to the web link provided in Table 3 for updated information on each program/package. Readers are also advised to explore emerging programs for power analysis; some are discussed in the section titled "**Recommendations**."

G*Power

G*Power is a stand-alone power analysis software for a variety of statistical tests (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007). It does not perform descriptive or inferential statistical analysis of any kind. Its current version, G*Power 3.1.3, is capable of performing both prospective and observed power analyses. G*Power 3.1.3 accepts either ES, or means and *SD*s, as input to initialize the power analysis. G*Power 3.1.3 provides a built-in ES calculator to convert hypothesized/conjectured means and *SD*s, or variance explained, or partial η^2, to one of the ES indices published in the literature and accepted by G*Power 3.1.3.

PASS

PASS stands for Power Analysis and Sample Size, licensed and sold by NCSS as a standalone software. It performs both prospective and observed power analyses. The Guide section on the data specification window provides users with information and recommendations for each input specification. The output is informative containing graphic and verbal explanations to enhance the numeric results.

Users initialize the power analysis by selecting the type of analysis, followed by specifying power, α level, means and *SD*s. Means and *SD*s are specified in profiles, such as (1 2 3) or (10 20 30) for means, and (1) or (1000) for *SD*s. From the means and *SD*s specified by users, PASS 11 computes ES for each statistical test. Unlike G*Power 3.1.3 or SPSS/Sample Power 3.0.1, PASS 11 does not accept ES directly from users.

SAS

Two statistical procedures from SAS/STAT 9.3 perform the prospective power analysis: **POWER** and **GLMPOWER**. These two procedures require researchers to specify conjectured/hypothesized group means, or cell means (for factorial, block, and Latin-square

ANOVA), and the *SD*s within groups or cells. From the conjectured means and *SD*s, SAS proceeds to estimate a desirable sample size based on user-specified power, α, and the directionality of the test, if appropriate. SAS 9.3 and PASS 11 are similar in this requirement, although SAS is more flexible than PASS 11 with user's specification of means and *SD*s.

A separate desktop module developed and sold by SAS for power analysis is called **PSS** (**Power and Sample Size Application**)**.** Its functionalities are similar to those of PROC POWER. It does not perform power analysis for complex ANOVA designs. As a menu-driven program, PSS offers the advantages of easy-to-follow commands, displaying results in narratives, and producing an array of output formats (e.g., HTML, RTF, PDF, PostScript or PCL), that are supported by SAS's Output Delivery System.

Stata

Stata is an integrated statistical package sold by Stata Corp. The current version, Stata 12, has two built-in functions for prospective and observed power analyses: **SAMPSI** and **STPOWER.** SAMPSI is for one- or two-sample *t*-test, proportions, and correlations whereas STPOWER is for survival models. Three sub-commands in the STPOWER function are for three different survival analyses: **LOGRANK** for the log-rank test, **EXPONENTIAL** for the exponential test, and **COX** for the Cox proportional hazards model.

In addition, several functions developed by users can be used for power analysis including, **FPOWER** for *F*-test of fixed-effects in one-way ANOVA, **POWERREG** for linear regression, **POWERLOG** for logistic regression, **CHI2POWER** for Chi-square test. A researcher can use the FINDIT command to add these functions to the command line (e.g. FINDIT FPOWER). The resulting page will also provide a link to the documentation for the function.

FPOWER and POWERLOG do not ask users to specify power in order to estimate a sample size. Both functions output a range of sample sizes corresponding to a range of power. In contrast, POWERREG estimates a sample size corresponding to a unique power value specified by users. Finally, CHI2POWER estimates a sample size or a range of sample sizes corresponding to user's specified power or a range of powers, and vice versa.

Furthermore, FPOWER defines ES as the standardized difference between the largest hypothesized mean and the smallest hypothesized mean—a definition comparable with those established in the literature (e.g., Cohen's *d* and *f*).

SPSS

SPSS/Statistics 19 performs observed power analysis for *F*-tests of fixed-effects in ANOVA, ANCOVA, and MANOVA models. A separate standalone product licensed and sold by IBM Inc., **Sample Power 3.0.1**, performs prospective power analysis for six general procedures: means, proportions, correlations, ANOVA and ANCOVA designs up to three factors, multiple regression, and general cases, in which the user can specify non-centrality parameters directly. Users initialize the prospective power analysis by selecting one of the two interfaces (classic versus step-by-step guide), followed by a selection of the procedure (design) and then the specification of power, α, ES or means and *SD*s, and the directionality of the test, if appropriate. The output includes numerical estimation for sample size, given a desired power, as well as graph and matrix of 'power as a function of sample size'. One unique output format is a report of estimated sample size, along with the study design, assumptions, and user's specifications. This feature is available also in PASS 11.

Software for HLM

Optimal Design Software. This free program is developed by Spybrook, Raudenbush, Congdon, and MartÍnez (2009) for prospective power analysis of hierarchical/multilevel modeling and meta-analytic research. This program can be used for individual and group randomized trials conducted in a variety of settings with or without blocking variables/covariates: single-level, multi-site trial, repeated measures, two-level cluster randomized trial, three-level cluster randomized trial, three- and four-level multi-site cluster randomized trial, and cluster randomized trial with repeated measures. The outcome measure can be continuous or binary. Thus, it is very useful for power analysis of single- or multi-site studies.

Optimal Design Software 2.01 initializes the power analysis with user's specifications of power, α, and ES in order to estimate a desirable sample size. Alternatively, it calculates the smallest ES that can be detected using power, α, and a sample size specified by users. The definition of ES follows those established in the literature (e.g., Cohen's *d* and *f*).

MLPowSim. The free MLPowSim can be used for sample size determination in complex random-effects models (Browne, Golalizadeh Lahi, & Parker, 2009). The software is still under development in its beta 1.0 version. Therefore, authors do not provide any warranty and they do not take any responsibility for the results. The program was developed to generate MLwiN macro or R commands to run the simulations to calculate power for user-defined scenarios. As a result, it works in conjunction with one of these programs.

MLPowSim runs in command mode in Windows platforms and has a text-based interface, which is not entirely user-friendly, admitted by the three developers. Despite all drawbacks, MLPowSim has something novel to offer for power analysis: "it can create scripts to perform sample size calculations for models which have more than two levels of nesting, for models with crossed random effects, for unbalanced data, and for non-normal responses." (Browne,

Golalizadeh, & Parker, 2009, p.1). In particular, it can generate simulation scripts in R or MLwiN for sample size calculation for the following models: (a) single-level models, (b) two- or three-level balanced and unbalanced nested models, and (c) three-level cross-classified balanced and unbalanced models. In addition, it can handle models with continuous (normal), binary, and Poisson outcomes. Script files are generated in the folder where MLPowSim program is located. Additional demonstrations and explanations for different multi-level models can be found in the program manual (see web link in Table 3).

Recommendations

In this section, we present our recommendations for selecting power analysis software, reporting power analysis results, and exploring additional software. First, we present results of a comparative evaluation of accuracy in sample size estimation.

Comparative Evaluation

To evaluate the accuracy of the power analysis software, we estimated desirable sample sizes for 12 research scenarios using G*Power 3.1.3, PASS 11, SPSS/Sample Power 3.0.1, SAS 9.3, and Stata 12. SPSS 19 was excluded because it cannot perform sample size estimation. The 12 scenarios and their corresponding null hypotheses were suggested by 12 examples in Kirk (1995, 2008). Raw data, along with the scenarios and null hypotheses, are shown in the Appendix (pp. 326 to 448). For each of 12 examples, a desirable sample size was estimated to test an appropriate null hypothesis at $\alpha = .05$ with a power of .80 or .90. Assuming fixed effects and balanced designs, sample sizes were estimated for four levels of ES. The first three levels of ES were Cohen's small, medium, and large ESs (Cohen, 1988, pp. 20-27; pp. 284-288); the fourth level was a reference level for comparison purposes.

Results in terms of the total sample (*N*), each group (*n*) of a factor, or each cell (per cell), are presented in Table A for power = .80 and Table B for power = .90 of the Appendix (pp. 4~~5~~59 to 5~~1~~15), both available from the web link given above. These results agreed almost entirely across the software, with *n*, per group or per cell, differing by no more than 2. Although differences in estimates could be due to algorithmic errors, we believe that a more plausible reason for these differences is the different rounding rules used in the algorithms. We arrived at this conclusion after an independent calculation based on the noncentrality parameter of noncentral *t*- and *F*-distributions. In other words, results shown in Tables A and B of the Appendix did not cast doubt on the accuracy of these programs/packages. To ensure accurate sample size estimates and, therefore, sufficient power in an empirical study, we recommend that~~, to the extent possible,~~ readers (i) pay attention to how ES is defined by the program/package of choice (see our general and specific comments on the ES issue on pp 26-31 of the Appendix) ~~obtain sample size estimates from at least two reputable programs/packages~~, and (ii) include at least one extra participant per group or condition in the actual study, if it is feasible.

Selecting Power Analysis Software

Several factors were considered in formulating our recommendations, including cost, user-friendliness, versatility of functionalities, requirements for hardware, graphical capabilities, interface with other software applications, output results, and availability of support from the software company or developers. These considerations are similar to those used by Dattalo (2009) in his software evaluations. Speed was not an issue with any of the programs/packages we evaluated in the Windows 7 platform. And accuracy did not appear to be an issue either.

For *t*-tests, *z*-tests, *F*-tests, χ^2-tests of means, proportions, variances, correlations, general linear models, survival analysis, and nonparametric techniques, we recommend SPSS/Sample

Power 3.0.1 (pp. 12 to 20 of Appendix) and the free G*Power 3.1.3 (pp. 2 to 5 of Appendix). Both can accept either ES or means and *SD*s specified by users. With its helpful output explanations and scaffolding during the analysis, SPSS/Sample Power 3.0.1 is a comprehensive and user-friendly power analysis program. It runs on a graphical user interface; a trial version can be used for 14 days (Table 3). G*Power 3.1.3 is a free alternative to SPSS/Sample Power 3.0.1 as both are comparable in terms of comprehensiveness, flexibility, user-friendliness, and graphical capabilities. Unlike SPSS/Sample Power 3.0.1, G*Power 3.1.3 can handle repeated-measures ANOVA designs. However, G*Power 3.1.3 lacks documentation and guidance for power analysis of complex designs.

For hierarchical/multi-level modeling techniques and meta-analysis, we recommend Optimal Design Software 2.01 (pp. 23 to 24 of Appendix) due to its versatility and ease of use. After selecting the multi-level design, user can change the design parameters (e.g., cluster size, effect size, variability) independent of each other and each change in parameters is directly reflected in the power curve. With regard to each program/package, we made several observations that may facilitate educational researchers' usage of these computing tools (see Appendix, pp. 2 to ~~25~~35).

Reporting Power Analysis Results

In accordance with recommendations and reporting standards set forth in the APA Task Force Report on Statistical Inference (1999, pp. 596-597), the APA Publication Manual [2010, pp. 30-31, and Table 1 of Journal Article Reporting Standards (JARS) on p. 248], we recommend the following two principles in disseminating power analysis results in quantitative studies:

1) To report the program or package that performs the power analysis, along with its version.
2) To refrain from computing/reporting the observed power, or justifying sample size based on the observed power, for reasons stated earlier under **Prospective versus Observed Power Analysis**.

Exploring Additional Software

Our investigation thus far reveals a critical need for expanding power analysis software, particularly for multiple comparison procedures of means, trend analysis in within-subject designs, and for comparisons of models. *The Journal of Experimental Education* has recently published several papers dealing with sample size estimation in specialized contexts, such as, estimating the noncentrality parameter for testing contrasts in one-way fixed-effects ANOVA (Liu, 2009) or in heterogeneous ANOVA (Luh & Guo, 2010), for testing trimmed means (Luh & Guo, 2009; Luh, Olejnik & Guo, 2008), or for one or two-level unbalanced designs (Konstantopoulos, 2010). Whereas some of these methods and algorithms are published in SAS codes (i.e., Luh, Olejnik & Guo, 2008; Luh & Guo, 2009, 2010), others are not published (Konstantopoulos, 2010; Liu, 2009).

The free software PinT (stands for Power in Two-level designs) considers budget constraints while estimating sample sizes for two-level HLM models (Bosker, Snijders, & Guldemond, 1996; Snijders & Bosker, 1993). Its current version, PinT 2.12 (since September, 2007) can be downloaded from http://stat.gamma.rug.nl/multilevel.htm#progPINT. Various modules in R are appropriate for prospective power analysis: **asypow**, **powerpkg**, **pwr**, **MBESS**. These modules were written in open-source codes for others to modify. To download a copy of R and its modules, go to http://cran.r-project.org/. For users of SAS, SPSS, and Stata, a helpful

website (http://www.statmethods.net/stats/power.html) provides detailed information on the module pwr.

An area for improvement for all software we reviewed is to allow ES to be specified in terms of both (a) hypothesized or conjectured means, and (b) standardized means or mean differences (e.g., Cohen's *d*), or variance explained (e.g., Cohen's *f*). At the present, SPSS/Sample Power 3.0.1 and G*Power 3.1.3 are the only two programs that are capable of accepting either specification. It would be more researcher-friendly, if the program/package could accept both specifications of ES for a variety of study designs.

Discussion

With an increasing emphasis on providing evidence for statistical inference making with information on power and sample size, researchers need to know how to specify power and how to estimate sample sizes before carrying out a study. Yet many researchers deal with these two issues as a post-hoc exercise. As Aguinis and Harden (2009) pointed out, in order to reach a satisfactory conclusion in studies, researchers either increase a priori α to compensate for a small sample size, or use Cohen's (1962, 1988) rules of thumb to justify their conclusions. We too observed these actions by authors who published in the 12 journals we reviewed.

The serious neglect of prospective power analysis may be explained by two reasons (Sedlmeier & Gigerenzer, 1989): (1) the α-adjusted procedures for multiple comparisons of means, and (2) the confusion over the hybrid of Fisherian and Neyman-Pearsonsian conceptualization of null hypothesis significance testing (NHST). This confusion has persisted in almost all of the applied statistical textbooks since World War II (e.g., Henson, Hull, & Williams, 2010; Huberty, 1993; Rodgers, 2010; Sedlmeier & Gigerenzer, 1989).

Historically, Fisher and Neyman and Pearson developed two different approaches to testing statistical hypotheses (Carlson, 1976; Chow, 1996; Cowles, 1989; Harlow, Mulaik, & Steiger, 1997; Huberty, 1987; Oakes, 1986; Spielman, 1974). Fisher employed an inductive inference method that focused merely on objective phenomena (i.e., data) and assumed the effects of strict randomization (Huberty, 1993; Lehmann, 1993; Mulaik, Raju, & Harshman, 1997). The significance test by the Fisherian approach included only a natural (or null) hypothesis that indicated no effect, or no difference, existed between/among the expected means of treatments. Within this framework, H_0 can be disproved but "never proved or established." (Fisher, 1935, p. 19, cited in Mulaik et al., 1997). According to the Fisherian conceptualization, power is a nonexistent concept because there is no alternative hypothesis. "Fisher viewed the hypothesis testing process as incremental, driven by replication, improving with each NHST decision, and potentially self-correcting" (Rodgers, 2010, p. 2). Yet the Neyman and Pearsonian conceptualization of NHST included the null and alternative hypotheses and the corresponding Type I (α) and II (β) error rates, hence, power ($1 - \beta$) (refer back to the section titled "**Prospective versus Observed Power**"). Their goal was to reach a dichotomous decision (reject or do not reject H_0) at the conclusion of each statistical test, as in clinical diagnosis or quality control context. Thus, the question, "Does statistical power matter?" should be answered with, "It depends." It depends on the methodological framework within which a study is conducted. It may be correct to state that the epistemological goal of Fisherian approach to NHST was to evaluate a substantive theory and to answer research questions probabilistically, whereas Neyman and Pearsonian approach aimed to reach a dichotomous conclusion regarding H_0, in light of data. These two conceptualizations are not compatible, so they should not have been combined into one methodological framework (Gigerenzer, 1993).

Given this historical background for NHST and the current practice of merging the two opposing methodological frameworks, it is therefore not surprising that many conceptual and interpretative errors have prevailed in the literature, such as interpreting a nonsignificant result as confirming the null hypothesis (Sedlmeier & Gigerenzer, 1989). Virtually all studies with statistically insignificant findings attributed the findings to small sample sizes, hence insufficient power. For these studies, power appeared to be an after-thought, rather than a planning tool. Even studies with hundreds or thousands of participants could still suffer from insufficient power, when data were analyzed by techniques such as HLM with a small number of units at Level 2 or higher (e.g., two schools). Yet the issue of low power for tests beyond Level 1 in HLM or GMM analyses was not addressed in any of the studies we reviewed. Furthermore, for studies with hundreds of participants, a more pertinent consideration than power is the stability of parameter estimates across time, regions, cultures, or countries, as hundreds of participants surely lead to a powerful statistical test of even a negligible true ES.

Can there be too much power? The answer is yes, especially with goodness of fit type of statistical tests (e.g., χ^2) for which the null hypothesis is in the form of a good fit of the model to data. Therefore, too much power will lead to the rejection of the null hypothesis, hence, the model under consideration. In these cases, a careful balance needs to be achieved between sufficient power and a good model fit (e.g., magnitude of standard error, Akaike's information criterion). Likewise, in comparing/assessing competing models for the same data set, a thoughtful balance between power for statistical tests of a null hypothesis (usually in the form of no difference between competing models) and the magnitude of various model comparison indices should be considered.

To advance knowledge through research, educational researchers need to practice sound methodologies. In this paper, we documented a lack of adoption of APA's and AERA's recommended practices for performing prospective power analysis by authors of 12 journals between 2005 and 20~~09~~10. Our review revealed that merely 1.7~~6~~7% of all quantitative articles conducted prospective power analysis; another ~~4~~3.5% conducted observed power analysis; the majority of the articles made no mention of the power issue. To tackle this continued neglect of power issue, we demonstrated the conceptual and computational differences between prospective and observed power analyses and reviewed eight accessible and popular programs/packages for power analysis. We recommend SPSS/Sample Power 3.0.1 and G*Power 3.1.3 for general-purpose univariate analyses, multivariate analyses of variance, survival analyses, and regression. Optimal Design Software 2.01 was found to be versatile for hierarchical/multilevel modeling and meta-analysis, hence, recommended for these purposes. We also recommend that researchers always report the version of the computing program or package that performed the power analysis and refrain from computing/reporting the magnitude of observed power.

Of course, power is a relevant issue only to the methodological framework of NHST. Still there are viable alternatives to NHST for making inferences about population characteristics. Many (e.g., precision analysis, interval estimation, equivalence testing) are already mentioned in the 6th edition of the APA publication manual (APA, 2010, pp. 30-35). It is our belief that, regardless of the methodological framework a researcher subscribes to, there is no substitute for a careful and proper planning for research.

References

Aguinis, H., & Harden, E. E. (2009). Sample size rules of thumb: Evaluating three common practices. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends* (pp. 267-286). New York: Routledge.

American Educational Research Association (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33-40. doi: 10.3102/0013189X035006033

American Psychological Association (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Anonymous (2003). Statistics calculators. Retrieved from http://calculators.stat.ucla.edu.

Bausell, R. B., & Li, Y-F. (2002). *Power analysis for experimental research: A practical guide for the biological, medical and social sciences*. Cambridge: Cambridge University Press.

Bosker, R. J., Snijders, T.A.B., & Guldemond, H. (1996). PINT (Power IN Two-level designs) User Manual. Retrieved from http://stat.gamma.rug.nl/Pint21_UsersManual.pdf

Browne, W. J.,Golalizadeh, M. & Parker, R.M.A (2009). *A guide to sample size calculations for random effect models via simulation and the MLPowSim software package.* University of Bristol. Retrieved from http://www.bristol.ac.uk/cmm/software/mlpowsim/mlpowsim-manual.pdf

Carlson, R. (1976). The logic of tests of significance. *Philosophy of Science, 43*(1), 116-128. Retrieved from http://www.jstor.org/stable/187338

Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility.* Thousand Oaks, California: Sage Publications.

Chronister, K. M., & McWhirter, E. H. (2006). An experimental examination of two career interventions for battered women. *Journal of Counseling Psychology, 53*, 151-164. doi:10.1037/0022-0167.53.2.151

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145-153. doi: 10.1037/h0045186

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. San Diego, CA: Academic Press.

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

Cowles, M. P. (1989). *Statistics in psychology: A historical perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dattalo, P. (2009). A review of software for sample size determination. *Evaluation & the Health Professions, 32*, 229-248. doi:10.1177/0163278709338556

Elbaum, B. (2007). Effects of an oral accommodation on the mathematics performance of secondary students with and without learning disabilities. *Journal of Special Education, 40*(4), 218-229. doi: 10.1177/002246690704000403

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160. doi:10.3758/BRM.41.4.1149

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. Retrieved from http://brm.psychonomic-journals.org/

Feldt, L. S., & Mahmoud, M. W. (1958). Power function charts for specifying numbers of observations in analyses of variance of fixed effects. *The Annals of Mathematical Statistics, 29,* 871-877. Retrieved from http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aoms

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement, 61,* 181-210. doi:10.1177/0013164401612001

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Fox, M. (1956). Charts of the power of the F-Test. *The Annals of Mathematical Statistics, 27,* 484-497. Retrieved from http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aoms

Gerard, P. D., Smith, D. R., & Weerakkody, G. (1998). Limits of retrospective power analysis. *The Journal of Wildlife Management, 62,* 801-807. Retrieved from http://joomla.wildlife.org/index.php?option=com_content&task=view&id=43

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences*: *Vol. 1. Methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum.

Gillett, R. (1994a). The average power criterion for sample size estimation. *The Statistician, 43,* 389-394.

Gillett, R. (1994b). Post hoc power analysis. *Journal of Applied Psychology, 79,* 783-785. doi: 10.1037/0021-9010.79.5.783

Hallahan, M., & Rosenthal, R. (1996). Statistical power: Concepts, procedures, and applications. *Behavioral Research Therapy, 34,* 489-499. doi:10.1016/0005-7967(95)00082-8

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Ed.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Henson, R. K., Hull, D. M., & Williams, C. S. (2010). Methodology in our education research culture: Toward a stronger collective quantitative proficiency. *Educational Researcher, 39,* 229-240. doi:10.3102/0013189X10365102

Hoenig, J. M. & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55*(1), 19-24. Retrieved from http://proquest.umi.com/pqdweb?did=68156036&sid=1&Fmt=4&clientId=12010&RQT=309&VName=PQD

Huberty, C. J. (1987). On statistical testing. *Educational Researcher, 16*(8), 4-9. Retrieved from http://www.jstor.org/stable/1175368

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *The Journal of Experimental Education, 61(4),* 317-333. Retrieved from http://www.jstor.org/stable/20152384

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62,* 227-240. doi:10.1177/0013164402062002002

Jennions, M. D., & Moller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology, 14,* 438-445. Retrieved from http://beheco.oxfordjournals.org/

Keselman, H. J. (1976). A power investigation of the Tukey multiple comparison statistic. *Educational and Psychological Measurement, 36,* 97-104. doi:10.1177/001316447603600108

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

Kirk, R. E. (2008). *Statistics: An introduction* (5th ed.). Belmont, CA: Thomson Wadsworth.

Koele, P. (1982). Calculating power in analysis of variance. *Psychological Bulletin, 92,* 513-516. doi:10.1037/0033-2909.92.2.513

Konstantopoulos, S. (2010). Power analysis in two-level unbalanced designs. *The Journal of Experimental Education*, *78*(3), 291-317. doi: 10.1080/00220970903292876

Lee, R. M. (2005). Resilience against discrimination: Ethnic identity and other-group orientation as protective factors for Korean Americans. *Journal of Counseling Psychology, 52,* 36-44. doi:10.1037/0022-0167-52.1.36

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, *88*(424), 1242-1249. Retrieved from http://www.jstor.org/stable/pdfplus/2291263.pdf

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician, 55,* 187-193. Retrieved from http://proquest.umi.com/pqdweb?did=78329191&sid=2&Fmt=3&clientId=12010&RQT=309&VName=PQD

Levin, J. R. (1997). Overcoming feelings of powerlessness in "aging" researchers: A primer on statistical power in analysis of variance designs. *Psychology and Aging, 12,* 84-106. doi:10.1037/0882-7974.12.1.84

Levine, M., & Ensom, M. H. H. (2001). Post hoc power analysis: An idea whose time has passed? *Pharmacotherapy*, *21*, 405-409. Retrieved from http://www.pharmacotherapy.org/

Lewis, K. P. (2006). Statistical power, sample sizes, and the software to calculate them easily. *BioScience, 56,* 607-612. Retrieved from http://caliber.ucpress.net/loi/bio

Liow, S. J. R., & Lau, L. H. (2006). The development of bilingual children's early spelling in English. *Journal of Educational Psychology*, *98*, 868-878. doi:10.1037/0022-0663.98.4.868

Liu, X. S. (2009). A note on noncentrality parameters for contrast tests in a one-way analysis of variance. *The Journal of Experimental Education*, *78*(1), 53-59. doi: 10.1080/00220970903224669

Luh, W.- M. & Guo, J.-H. (2009). The sample size needed for the trimmed *t* test when one group size is fixed. *The Journal of Experimental Education*, *78*(1), 14-25. doi: 10.1080/00220970903224578

Luh, W.- M. & Guo, J.-H. (2010). Developing the noncentrality parameter for calculating group sample sizes in heterogeneous analysis of variance. *The Journal of Experimental Education*, *79*(1), 53-63. doi: 10.1080/00220970903292942

Luh, W.- M., Olejnik, S., & Guo, J.-H. (2008). Sample size determination for one- and two-sample trimmed mean tests. *The Journal of Experimental Education*, *77*(2), 167-184. doi:10.3200/JEXE.77.2.167-184

Miller, J. K., & Knapp, T. R. (1972). *The importance of statistical power in educational research* (Occasional paper No. 13). Bloomington, IN: Phi Delta Kappa Research Service Center.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik & J.H. Steiger (Eds.), *What if there were no significance tests*? (pp. 65-115). Mahwah, NJ: Erlbaum.

Murphy, K. R. & Myors, B. (1998). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Mahwah, NJ: Lawrence Erlbaum Associates.

Oakes, M. (1986). *Statistical inference*. New York: Wiley.

O'Keefe, D. J. (2007). Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of statistical power analyses. *Communication Methods and Measures*, *1*, 291-299. Retrieved from http://www.tandf.co.uk/journals/HCMS

Onwuegbuzie, A. J., & Leech, N. L. (2004). Post hoc power: A concept whose time has come. *Understanding Statistics*, *3*, 201-230. doi:10.1207/s15328031us0304_1

Ortiz, M. (2002). Optimum sample size to detect perturbation effects: The importance of statistical power analysis--A critique. *Marine Ecology*, *23*, 1-9. Retrieved from http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291439-0485

Peng, C.-Y. J., Long, H., Abaci, S. (2010, May). *Statistical power and sample size estimation in quantitative studies*. Paper presented at the annual meeting of American Educational Research Association, Denver, CO.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*, 1-12. doi:10.1037/a0018326

SAS Institute Inc. (2010). *SAS/STAT(R) 9.22 user's guide*. Cary, NC: SAS Institute Inc.

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309-312. doi:10.1037/0033-2909.105.2.309

Severo, N. C., & Zelen, M. (1960). Normal approximation to the chi-square and non-central F probability functions. *Biometrika, 47,* 411-416. Retrieved from http://biomet.oxfordjournals.org/content/by/year

Snijders, T.A.B. & Bosker, R. J. (1993). Standard errors and sample sizes in two-level research. *Journal of Educational Statistics, 18,* 3, 237-260. Retrieved from http://www.jstor.org/stable/1165134

Spielman, S. (1974). The logic of tests of significance. *Philosophy of Science, 41,* 211-226. Retrieved from http://www.jstor.org/stable/pdfplus/187132.pdf

Spybrook, J., Raudenbush, S.W., Congdon, R., & Martinez, A. (2009). Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software V.2.0. Available at www.wtgrantfoundation.org.

Thomas, L. (1997). Retrospective power analysis. *Conservation Biology, 11*(1), 276-280. Retrieved from http://www.cienciasmarinas.com/index.php/cmarinas

Thomas, L., & Juanes, F. (1996). The importance of statistical power analysis: An example from *Animal Behavior*. *Animal Behavior, 52,* 856-859. Retrieved from http://www.elsevier.com/wps/find/journaldescription.cws_home/622782/description#description

Thomas, L., & Krebs, C. J. (1997). A review of statistical power analysis software. *Bulletin of the Ecological Society of America, 78,* 126-139. Retrieved from http://www.esajournals.org/loi/ebul

Vansteenkiste, M., Timmermans, T., Lens, W., Soenens, B., & Van den Broeck, A. (2008). Does extrinsic goal framing enhance extrinsic goal-oriented individuals' learning and performances? An experimental test of the match perspective versus self-determination theory. *Journal of Educational Psychology, 100,* 387-397. doi:10.1037/0022-0663.100.2.387

Wheeler, R. E. (2000). ECHIP Sample Size Estimator (Version 1.0) [Software]. Available from http://www.bobwheeler.com/statistics/SSize/ssize.html.

Wilkinson, L., and the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604. doi:10.1037/0003-066X.54.8.594

Xin, P. Y., Jitendra, A. K., & Deatline-Buchman, A. (2005). Effects of mathematical word problem-solving instruction on middle school students with learning problem. *Journal of Special Education, 39,* 181-192. Retrieved from http://sed.sagepub.com/

Yuan, K.-H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics, 30,* 141-167. Retrieved from http://www.jstor.org/action/showPublication?journalCode=jeducbehastat

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

Table 1

Power Calculation and Sample Size (n) Estimation Reported in 12 Education Journals from 2005 to 2010

Journal[a]	Articles[b,c]	Estimated *n* during the planning phase of a study (%)		**Conducted observed power analysis**						Mentioned power and/or *n* without actual computation or estimation (%)	
				To compute observed power (%)		To justify *n* **after** data collection (%)		To calculate *n* for future studies (%)			
AERJ	87	2	(2.30)	1	(1.15)	2	(2.30)	0	(0)	9	(10.34)
ER	19	0	(0)	0	(0)	0	(0)	0	(0)	3	(15.79)
JCP	230	13	(5.65)	20	(8.70)	3	(1.30)	3	(1.30)	74	(32.17)
JEP	360	3	(.83)	5	(1.39)	1	(.28)	1	(.28)	51	(14.17)
JRME	26	0	(0)	0	(0)	0	(0)	0	(0)	2	(7.69)
JRST	135	0	(0)	3	(2.22)	0	(0)	0	(0)	4	(2.96)
JRTE	63	0	(0)	2	(3.17)	0	(0)	0	(0)	1	(1.59)
JSE	44	3	(6.82)	0	(0)	0	(0)	0	(0)	6	(13.64)
JSP	128	2	(1.56)	3	(2.34)	1	(.78)	0	(0)	35	(27.34)
MLJ	66	0	(0)	1	(1.52)	0	(0)	0	(0)	3	(4.55)
RHE	181	1	(.55)	0	(0)	1	(.55)	0	(0)	4	(2.21)
TRSE	18	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
Total	**1357**	**24**	**(1.77)**	**35**	**(2.58)**	**8**	**(.59)**	**4**	**(.29)**	**192**	**(14.15)**

Note: Percentages listed in parentheses are row percents.

[a] Journal abbreviations

AERJ: *American Educational Research Journal*
ER: *Educational Researcher*
JCP: *Journal of Counseling Psychology*
JEP: *Journal of Educational Psychology*
JRME*:* *Journal for Research in Mathematics Education*
JRST: *Journal of Research in Science Teaching*
JRTE: *Journal of Research on Technology in Education*
JSE: *Journal of Special Education*
JSP: *Journal of School Psychology*
MLJ: *The Modern Language Journal*
RHE: *Research in Higher Education*
TRSE: *Theory and Research in Social Education*

[b] Articles were the unit of analysis, even though several articles reported multiple studies.

[c] Detailed review of each article may be obtained from the first author.

Table 2

Prospective versus Observed Power Analysis

Power Analysis (selected reference)	When to conduct	How to conduct	Alternative names (selected references)	Comments
Prospective (SAS, 2010)	During the planning phase of a study; before data collection and analysis	(1) A researcher specifies a population ES, an α, power, and the directionality of the test, if appropriate. (2) He/She estimates a sample size based on the sampling distribution derived from (1) above. Examples: Elbaum (2007); Xin, Jitendra, and Deatline-Buchman (2005)	*A priori* power analysis (Faul, Erdfelder, Lang & Buchner, 2007), *Planned* power analysis (Yuan & Maxwell, 2005)	This is the only power analysis recommended by APA Task Force (1999), AERA reporting standards (2006), and the sixth edition of APA *Publication Manual* (2010).
Observed (Thomas & Krebs, 1997)	After data collection and analysis	(1) A researcher computes an observed ES based on data. (2) He/She computes the observed power based on the sampling distribution derived from the observed ES obtained in (1), the α specified for the statistical test, and the sample size actually used in the study. Examples: Chronister and McWhirter (2006); Liow and Lau (2006); Vansteenkiste, Timmermans,	*Achieved*, *Computed*, *Estimated*, *Post Hoc*, *Posterior*, or *Retrospective* power analysis (Gillett, 1994b; Hoenig & Heisey, 2001; Levine & Ensom, 2001; O'Keefe, 2007; Onwuegbuzie & Leech, 2004; Yuan & Maxwell, 2005)	

	Lens, Soenens, and Van den Broeck, (2008); Lee (2005).

Table 3

Software Version and Pricing

Software	Current Version	Pricing	Website
G*Power	3.1.3	Free	Product & download at http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/download-and-register
PASS	11	$575.00/year	Product at http://www.ncss.com/download_freetrial.html Academic Price at http://www.ncssorders.com/ncss_pricelist_annual.asp?Pricing=Academic
SAS/ STAT	9.3	Available upon request	Product at www.sas.com Academic Price at https://www.sas.com/order/product.jsp?code=PERSANLBNDL
Stata	12	Gradplan prices vary by university and STATA module	Product at http://www.stata.com/ Academic Price at http://www.stata.com/order/educational.html
SPSS/ Statistics	19	Faculty pack: $254.99/year Grad pack: $99.99/year	Product at http://www-01.ibm.com/software/analytics/spss/products/statistics/base/ Academic Price at http://www.onthehub.com/spss (Faculty Pack and Premium Gradpack includes **Sample Power**)
SPSS/ Sample Power	3.0.1	See SPSS above	Product at http://www-01.ibm.com/software/analytics/spss/products/statistics/samplepower/ A trail SPSS Sample Power 3.0.1 download at http://www14.software.ibm.com/download/data/web/en_US/trialprograms/U741655I36057W80.html?S_CMP=rnav
Optimal Design Software	2.01	Free	Product & download at http://sitemaker.umich.edu/group-based/optimal_design_software
MLPowSim	1.0 BETA	Free	Product & download at http://seis.bris.ac.uk/~frwjb/esrc.html

Note. Information present in this table is current as of October, 2011.

Table 4

*Functionalities in G*Power, PASS, SAS, Stata, and SPSS for Prospective (P) or Observed (O) Power analysis, or Power Curve (C)/ for a Single Sample Size (S) or a Range of Sample Sizes (RS) Estimation in Prospective Power Analysis/ by Specifying Effect Size (ES)[a] or Means, SDs, or Others (MSO)[b]*

			SAS 9.3 [c,d]			SPSS	
Statistical Test	**G*Power 3.1.3[c,d]**	**PASS 11[c,d]**	**POWER**	**GLMPOWER**	**Stata 12 (function or command)**	**Statistics 19**	**Sample Power 3.0.1[c,d]**
A. Means							
1-sample and 2-samples *z*- or *t*-test (1 and 2-tailed)	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO	P, C*/ S/ MSO		P, O/ S/ MSO (**SAMPSI**)		P, C/ S, RS/ MSO
F-test of fixed-effects in 1-way ANOVA	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO	P, C*/ S/ MSO	P, C/ S/ MSO	P, O, C/ RS/ ES (**FPOWER**)	O/ / MSO	P, C/ S, RS/ ES, MSO

F-test of fixed main and interaction effects in factorial and randomized block factorial ANOVA	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO	P, C/ S/ MSO	O/ / MSO	P, C/ S, RS/ ES, MSO
F-test of fixed main and block effects in randomized block and generalized randomized block ANOVA	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO	P, C/ S MSO	O/ / MSO	P, C/ S, RS/ ES, MSO
F-test of fixed main effect in Latin-square and Latin-square fractional factorial ANOVA	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO	P, C/ S/ MSO	O/ / MSO	P, C/ S, RS/ ES, MSO
F-test of fixed between-subject, fixed within-subject, and fixed interaction effects in split plot factorial (repeated measures) ANOVA	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO	P, C/ S/ MSO	O/ / MSO /	
fixed contrast effects		P, O, C/ S/ MSO	P, C/ S/ MSO		
ANCOVA with continuous or categorical covariate	P, O, C/ S/ ES, MSO		P, C/ S/ MSO		P, C/ S, RS/ ES, MSO
Test of trimmed means		P, O, C/			

		S/ MSO			
Means in cross-over designs		P, O, C/ S/ MSO	P, C*/ S/ MSO		
MANOVA	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO			
B. Mixed models		P, O, C/ S/ MSO			
C. Regression					
Simple & Multiple	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO	P, C*/ S/ MSO	P/ S/ / **(POWERREG)**	P, C/ S, RS/ MSO
Logistic regression	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO	P, C*/ S/ MSO	P/ RS/ / **(POWERLOG)**	P, C/ S, RS/ MSO

Poisson regression	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO			
D. Correlation					
1 Pearson correlation coefficient = (constant)	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO	P, C*/ S/ MSO	P,O/ S/ / **(SAMPSI)**	P, C/ S, RS/ MSO
Equality of 2 Pearson correlation coefficients	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO		P,O/ S/ / **(SAMPSI)**	P, C/ S, RS/ MSO
1 Partial correlation coefficient=(constant)		P, O, C/ S/ MSO	P, C*/ S/ MSO	P,O/ S/ / **(SAMPSI)**	
Equality of 2 partial correlation coefficient		P, O, C/ S/ MSO		P,O/ S/ / **(SAMPSI)**	

Cronbach's alpha		P, O, C/ S/ MSO			
Intraclass correlation coefficient		P, O, C/ S/ MSO			
Kappa test for agreement		P, O, C/ S/ MSO			
Point biserial	P, O, C/ S/ ES, MSO				
Tetrachoric correlation	P, O, C/ S/ ES, MSO				
E. **Proportions**					
1 proportion= (constant)	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO	P, C*/ S/ MSO	P,O/ S/ / **(SAMPSI)**	P, C/ S, RS/ MSO

Equality of 2 independent or correlated proportions	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO	P, C*/ S/ MSO	P,O/ S/ / **(SAMPSI)**	P, C/ S, RS/ MSO
(Risk) Ratio of 2 proportions		P, O, C/ S/ MSO	P, C*/ S/ MSO		
Odds ratio		P, O, C/ S/ MSO	P, C*/ S/ MSO		
Multinomial test	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO			
Chi-square test of goodness of fit or t in contingency table	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO	P, C*/ S/ MSO	P,O/ S, RS/ / **(CHI2POWER)**	P, C/ S, RS/ ES, MSO
Longitudinal time-average difference in 2 proportions		P, O, C/ S/ MSO			

F.	**Survival analysis**				
	2-sample survival rank test		P, O, C/ S/ MSO	P, C*/ S/ MSO	P,O/ S/ / **(STPOWER LOGRANK)**	P, C/ S, RS/ MSO
	Exponential means (1 or 2)		P, O, C/ S/ MSO		P,O/ S/ / **(STPOWER EXPONENTIAL)**	P, C/ S, RS/ MSO
	Group sequential tests		P, O, C/ S/ MSO			P, C/ S, RS/ MSO
	Cox regression		P, O, C/ S/ MSO		P,O/ S/ +/ **(STPOWER COX)**	
G.	**Nonparametric**					
	1 or 2 group location(s)	P, O, C/ S/	P, O, C/ S/	P, C*/ S/		P, C/ S, RS/

	ES, MSO	MSO	MSO	MSO
Multiple group locations	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO		
Exact test of frequencies (proportions)	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO		P, C/ S, RS/ MSO
H. Variance				
1 variance=(constant)	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO		
Equality of 2 variances	P, O, C/ S/ ES, MSO	P, O, C/ S/ MSO		
J. Normality test		P, O, C/ S/ MSO		
K. Designs of experiments		P, O, C/		
(Block, factorial, fractional factorial,		S/		
Latin-square, optimal, two-level,		MSO		

response surface designs, design generator)

Note: All effects tested under H_0 are assumed to be fixed.

[a] ES for *t*-test was defined as Cohen's *d*; ES for *F*-test based on ANOVA was defined as Cohen's *f* in G*Power 3.1.3 and Sample Power 3.0.1 whereas it is defined as the difference between the largest and the smallest means divided by the pooled *SD* for FPOWER in Stata 12.

[b] MS for tests of proportions is simply the proportion in a group or in a row, because the SD for proportions is a function of the proportion itself.

[c] The software can also perform precision analysis in terms of confidence intervals.

[d] The software has built-in probability calculators based on binomial, normal, central and noncentral *t*-, *F*-, χ^2 distributions and/or odds ratio.

***** Power curve is obtained using X=effect (effect-size-like specification) in PROC POWER.

[+] As an option, STPOWER COX can use *SD* as one of the input parameters for standard deviation of covariate of interest. Default value is 0.5.