Behavior Research Methods, Instruments, & Computers
1996, 28 (1), 1-11

GPOWER: A general power analysis program

EDGAR ERDFELDER
University of Bonn, Bonn, Germany

FRANZ FAUL
University of Kiel, Kiel, Germany

and

AXEL BUCHNER
University of Trier, Trier, Germany

GPOWER is a completely interactive, menu-driven program for IBM-compatible and Apple Macintosh personal computers. It performs high-precision statistical power analyses for the most common statistical tests in behavioral research, that is, t tests, F tests, and χ^2 tests. GPOWER computes (1) power values for given sample sizes, effect sizes and α levels (post hoc power analyses); (2) sample sizes for given effect sizes, α levels, and power values (a priori power analyses); and (3) α and β values for given sample sizes, effect sizes, and β/α ratios (compromise power analyses). The program may be used to display graphically the relation between any two of the relevant variables, and it offers the opportunity to compute the effect size measures from basic parameters defining the alternative hypothesis. This article delineates reasons for the development of GPOWER and describes the program's capabilities and handling.

Following Jacob Cohen's (1962) pioneering work on the power of statistical tests in behavioral research, many authors have stressed the necessity of statistical power analyses. Textbooks and articles have appeared that provide more or less extensive tables of power and sample sizes (e.g., Cohen, 1969, 1977, 1988, 1992; Cohen & Cohen, 1983; Hager & Möller, 1986; Kraemer & Thiemann, 1987; Lipsey, 1990). In addition, several computer programs for performing a variety of power analyses have become available during the past few years (for a review, see Goldstein, 1989). Given this state of affairs, does it make sense to publish yet another power analysis program?

In the first part of this article, we present reasons as to why the answer to this question is "yes." We begin with an analysis of the probable causes for the unchanged low level of statistical power in behavioral research. We argue that this might, to some extent, be a consequence of the weaknesses of existing power analysis tools. In the second part of this article, GPOWER, a new power analysis program, is presented as an alternative. We report GPOWER's algorithms and their precision. The final part of the paper describes the scope, handling, and availability of the program.

WEAKNESSES OF EXISTING POWER ANALYSIS TOOLS

Sedlmeier and Gigerenzer (1989) investigated the impact of power analysis studies and textbooks on the power of recent psychological studies. Surprisingly, these authors found no significant increase in power values since 1962 when Cohen published his power study of the 1960 volume of the *Journal of Abnormal & Social Psychology* (JASP). In fact, the average power of studies published in the 1984 volume of the *Journal of Abnormal Psychology* (a successor to the JASP) had dropped slightly compared with Cohen's (1962) results. Rossi (1990) conducted a similar study based on the 1982 volume of the *Journal of Abnormal Psychology* and other journals. He found power values slightly larger than Cohen's (1962) results. He commented, however, that "these increases are no cause for joy" (Rossi, 1990, p. 650).

In an attempt to explain this discouraging state of affairs, Cohen (1988, 1992) referred to the generally slow methodological advances in psychology. Sedlmeier and Gigerenzer (1989), in contrast, focused on persistent shortcomings in the statistical education of psychologists as reflected in ambiguities and errors in textbooks on statistical methods in behavioral research. As Bredenkamp (1972), Gigerenzer and Murray (1987), Oakes

The authors would like to thank S. Dilger for performing the evaluation of the GPOWER accuracy mode calculations and J. Bredenkamp as well as several anonymous reviewers for helpful comments on earlier drafts of this paper. We also are grateful to V. Fischer, P. Frensch, J. Funke, P. Onghena, R. Pohl, E. Stirner, and I. Wegener, who served as beta testers of GPOWER. Correspondence should be addressed to E. Erdfelder, Psychologisches Institut der Universität Bonn, Römerstraße 164, D-53117 Bonn, Germany (e-mail: erdfelder@uni-bonn.de). Correspondence concerning the MS-DOS version of GPOWER should be addressed to E. Erdfelder (address above) or to F. Faul, Institut für Psychologie an der Universität Kiel, Olshausenstraße 40, D-24098 Kiel, Germany (e-mail: ffaul@psychologie.uni-kiel.de). Correspondence concerning the Macintosh version should be addressed to A. Buchner, FB I-Psychologie, Universität Trier, Universitätsring 15, D-54286 Trier, Germany (e-mail: buchner@cogpsy.uni-trier.de).

(1986), Pollard and Richardson (1987), Tversky and Kahneman (1971), and others have shown, there is obviously some confusion about the notion of statistical significance and the role of sample size among both students of psychology and professional psychologists. According to Sedlmeier and Gigerenzer (1989), a major reason for this confusion is the "hybridization" of the Fisherian and the Neyman-Pearson theories of statistical inference in the psychological literature (see also Gigerenzer, 1993).

We agree with Sedlmeier and Gigerenzer's (1989) diagnosis. Nevertheless, errors and ambiguities in textbooks are probably not the only and perhaps not even the most important reasons for the persistence of low statistical power in behavioral research. Basically, there are only two ways to raise the power if the null hypothesis (H_0), the alternative hypothesis (H_1), and the test statistic have already been specified: One must increase either the sample size N or the Type 1 error probability α.[1] However, as will be discussed in more detail below, both ways are associated with serious practical problems. These problems could be the major reasons for the negative results obtained by Sedlmeier and Gigerenzer (1989) and by Rossi (1990).

Let us first consider a priori power analyses, which are considered the ideal type of power analysis by most authors. In an a priori power analysis, researchers specify the size of the effect to be detected (i.e., a measure of the "distance" between H_0 and H_1), the α level, and the desired power level $(1 - \beta)$ of the test. Given these specifications it is possible to compute the necessary sample size N. In standard applications, the selection of the effect size and of the error probabilities is based on conventions. There is a long tradition of using either $\alpha = .05$ or $\alpha = .01$ as Type 1 error probability (Cowles & Davis, 1982), and it is common to select effect sizes that are "small," "medium," or "large" as defined by Cohen (1962, 1969, 1977, 1988, 1992). No unique conventions have been established with respect to the Type 2 error probability β. Cohen (1977, 1988) suggested using $\beta = .20$ as a standard level, whereas other researchers prefer α and β levels to be equal (e.g., Bredenkamp, 1980).

A priori power analyses are ideal in that low error probabilities α and β can be achieved for any specification of the effect size. Unfortunately, however, the calculated sample sizes are usually much larger than what is considered manageable in behavioral research. Time constraints, financial constraints, and methodological reasons (e.g., sample heterogeneity in case of data aggregation across studies) prohibit the use of "ideal" sample sizes.

Let us assume, then, that a behavioral scientist has arrived at some maximum N that can be achieved given the institutional constraints of the research. This N will most likely be smaller than the ideal N as determined by an a priori power analysis. Thus, the only way to arrive at a reasonable power level is to increase the chances of committing an α error (Cohen, 1965). Unfortunately, however, power tables are typically based on conventional α levels (i.e., $\alpha \leq .1$) exclusively and therefore do not provide the information necessary to arrive at a reasonable power value.[2] Power analysis programs, in contrast, allow for nonstandard α levels in principle but do not encourage researchers to make use of them. All programs we know of are restricted to a priori and post hoc power analyses. A priori power analyses are useless when N is fixed. In post hoc power analyses, researchers specify α, the effect size, and the sample size N to compute the power of a test.[3] However, the mere possibility of specifying any α value is of little use, because there is no clue as to which α level is *reasonable* given the limited sample size and the size of the effect to be detected. In this confusing situation, researchers might be tempted to rely on some standard α value and to ignore the power problem entirely.

From this perspective, it is not at all surprising that the power of psychological studies seems immune to criticisms of low-power research. If researchers stick to standard α levels and, at the same time, face difficulties in increasing the effect size and the sample size, a stable low level of statistical power is the unavoidable consequence. The hope for future developments (Cohen, 1988), the publication of simplified sample size tables (Cohen, 1992), or improvements in the methodological literature (Sedlmeier & Gigerenzer, 1989) cannot be expected to remedy the problem. What behavioral researchers need is the means of planning rationally the level of α, taking into account the available resources.

Compromise power analyses (Erdfelder, 1984) have been designed especially for this purpose. In compromise power analyses, researchers specify the size of the effect to be detected, the maximum possible sample size, and the ratio $q := \beta/\alpha$, which defines the relative seriousness of both types of errors (Cohen, 1965, 1988). Given these specifications, an optimum critical value for the test statistic and the associated α and β values is computed. This optimum critical value is a rational compromise between the demands for a low α risk and a large power level, given a fixed sample size, a fixed effect size, and an error ratio of q.

It goes without saying that compromise power analyses may produce nonstandard levels of α and β. Given a relatively small sample size, a compromise analysis might, for instance, suggest the use of $\alpha = \beta = .168$. Although unusual, these error probabilities may certainly be reasonable. To illustrate, consider the case of a substantive hypothesis that implies as H_0 the hypothesis of no interaction. Does it make more sense to choose $\alpha = \beta = .168$ rather than to insist on the standard level $\alpha = .05$ associated with $\beta = .623$? Obviously, the standard α level makes no sense in this situation because it implies a very high risk to falsely accept the hypothesis of interest.

The reverse problem arises in those rare cases in which researchers can make use of extremely large samples. In such cases, a compromise analysis might suggest using $\alpha = \beta = .003$. It is again much better to follow this advice rather than choosing $\alpha = .05$, which is associated with a power of $(1 - \beta) > .999$, even for negligible deviations from H_0. Usually, one is not interested in a test indicating tiny effects. In most applications, effect sizes must be at least "small" (Cohen, 1977, 1988) to be of practical importance.

In principle, compromise power analyses can be approximated by repeatedly performing post hoc power analyses until the desired ratio of α and β is found with a sufficient degree of precision. However, with existing power analysis tools, this is troublesome and time-consuming. That was one major reason why we developed the GPOWER program.

At a more general level, GPOWER was designed to serve as an efficient, broadly applicable, and easy-to-use research tool. Therefore, options that are useful primarily in an educational context (e.g., Monte Carlo simulations or illustrations of the relation between mean differences, error variances, and effect sizes) were omitted. Good programs for these purposes have already become available (e.g., Borenstein & Cohen, 1988; Borenstein, Cohen, Rothstein, Pollack, & Kane, 1990, 1992; Rothstein, Borenstein, Cohen, & Pollack, 1990). In developing GPOWER, we gave priority to providing for a variety of power analyses for most of the common statistical tests in behavioral research. It appears that *t* tests, *F* tests, and χ^2 tests characterize this class sufficiently.[4] Moreover, we aimed at high-precision power calculations that are offered by only a few of the available power programs (see Goldstein, 1989). A high level of precison is especially important for power analyses based on small α and β values (as they occur, for instance, when α or β are adjusted in order to control for the cumulation of error probabilities; see Westermann & Hager, 1986).

THE GPOWER PROGRAM

GPOWER is available in two computationally equivalent versions for IBM-compatible PCs (written in Turbo-Pascal 6.0; Faul & Erdfelder, 1992) and Apple Macintosh PCs (written in Think-Pascal; Buchner, Faul, & Erdfelder, 1992), both of which have similar user interfaces. Therefore, we will describe the MS-DOS version and the Macintosh version simultaneously.

GPOWER users can select either an accuracy mode or a speed mode for computing a priori, post hoc, and compromise power analyses. The accuracy mode is based on the actual noncentral distributions of the relevant test statistics, while the speed mode calculations approximate the noncentral distributions by other distribution types. We first describe the numerical algorithms of GPOWER. Next, we compare GPOWER results with results obtained by other power analysis tools. Finally, the program handling and the hardware and software requirements are described briefly.

Algorithms

GPOWER's a priori, post hoc, and compromise power analyses are all based on the same subroutines. These subroutines compute (or approximate) power values for a certain noncentral distribution type (depending on the degrees of freedom, the noncentrality parameter, and on the α level), which is what is needed for post hoc power analyses. In a priori power analyses, however, *N* must be adjusted to fit a prespecified power level. GPOWER does this by first searching for an arbitrary *u*pper *b*ound N_{ub} to the solution. If N_{lb} denotes the smallest possible sample size, then the solution must be an integer element of the real interval $[N_{lb}, N_{ub}]$. This interval is iteratively dissected, using a slight modification of the Van Wijngaarden-Dekker-Brent method (see Press, Flannery, Teukolsky, & Vetterling, 1988, chap. 9.3): The smallest integer value $N \in [N_{lb}, N_{ub}]$ yielding a power value larger than or equal to the prespecified power level is regarded as the solution.[5]

Almost the same procedure is used in compromise power analyses. Here, GPOWER searches for a value of $\alpha \in [10^{-6}, (1 - 10^{-6})]$, which fits the prespecified ratio $q := \beta/\alpha$. Again, this interval is dissected by means of the Van Wijngaarden-Dekker-Brent method using an interval width of 10^{-6} as the criterion of convergence.

Six subroutines are used for power calculations, these being both approximate and precise routines for the noncentral *t*, *F*, and χ^2 distributions. All speed mode calculations are based on the approximate routines. The noncentral *t* distribution is approximated using Formula 12.2.1 in Cohen (1988, p. 544), which is based on Dixon and Massey (1957, p. 253). Laubscher's (1960) cube root normal approximation is used for the noncentral *F* distribution (see Cohen, 1988, p. 550, Formula 12.8.4), and a Pascal adaptation of Milligan's (1979) program is used for an approximation of the noncentral χ^2 distribution.

The precise routines are used in all accuracy mode calculations of GPOWER. They are slightly modified PASCAL adaptations of the subroutines NCTX (noncentral *t* integrals), NCFX (noncentral *F* integrals), and NCHI (noncentral χ^2 integrals) published by Bargmann and Ghosh (1964) in FORTRAN-II code. Our modifications of these subroutines do not change the basic algorithms. Rather, they make the program faster and render the program source code more readable.

Routines to compute the incomplete beta function and the incomplete gamma function play a key role in calculating exact probabilities for the central *t*, *F*, and χ^2 distributions. These routines were not adapted from Bargmann and Ghosh (1964). Instead, PASCAL adaptations of the more efficient C routines published by Press et al. (1988, chap. 6) were used.

Evaluation of the GPOWER Algorithms

According to Bargmann and Ghosh (1964), the FORTRAN-II subroutines on which the accuracy mode calculations of GPOWER are based should be correct to at least five significant digits for all input values, provided the parameters of the noncentral distributions remain within the range $[10^{-8}, 10^{+8}]$. We decided to test this for our implementation by comparing the accuracy mode post hoc power analyses of GPOWER with the "exact" χ^2 and *F* power values published by Patnaik (1949), and with a sample of results from the SAS routines TPROB, FPROB, and CPROB, which are known to be highly accurate (see Hardison, Quade, & Langston, 1983). We obtained perfect 4-digit agreement with Patnaik's (1949, Tables 1–5) "exact" χ^2 power values in 61 of 65 cases and no disagreements on the first 2 digits. A similar picture emerged for Patnaik's (1949, Table 6) "exact" F power values. We observed perfect 3-digit agreement in 22 of 24 cases and a difference of .001 in the remaining two cases.

An even closer agreement was observed with respect to the SAS routines for noncentral t, F, and χ^2 integrals. The 6-digit power values for the noncentral t distribution agreed perfectly in 599 of 600 cases. The disagreement in the remaining case was .000001. More disagreements were obtained for F power values. Again, however, none of the 32 differences from a total of 1,440 comparisons concerned the first 5 digits. Absolutely no differences in 140 comparisons were observed for 6-digit χ^2 power values.

We conclude that the power values obtained by GPOWER's accuracy mode calculations are indeed correct up to 5 significant digits, provided the input parameters are not too extreme. Since Patnaik's (1949) "exact" values are based on highly complex and laborious calculations by hand, the rare differences between his values and the GPOWER results are probably due to occasional rounding errors in his tables.

Although the accuracy mode and the speed mode calculations of GPOWER produce quite similar results for most of the standard analyses, significant differences may sometimes occur. For example, the speed mode of GPOWER calculates a power of .8340 for one-tailed correlation t tests based on $N = 8$ pairs of values (thus, $df = 6$), $\alpha = .01$, and a very large population correlation ($\rho = 0.9$). The accuracy mode computes a power of .9805 for the same set of parameters. These differences are due to the fact that speed mode results may be very misleading for extreme values of the parameters. Therefore, we recommend the speed mode only for taking a first glance at the problem. Publications of power values and final decisions concerning sample sizes or critical values should always be based on accuracy mode calculations.

We also investigated the agreement between GPOWER results and the tables published by Cohen (1988), because power analyses have often been conducted based on Cohen's books. In general, Cohen (1988) and GPOWER agree quite well. Of course, perfect 2-digit agreement with GPOWER's accuracy mode results cannot be expected because most of the power values and sample size tables in Cohen (1988) are based on approximations. Nevertheless, we found perfect agreement quite often, and power differences larger than .03 were rare. If such large differences appeared, it was usually for extreme values of the parameters.

Noteworthy exceptions to this are power analyses for special F tests in complex analysis of variance (ANOVA) designs, for example, F tests for main effects or interactions (i.e., Cases 2 and 3 of ANOVA F tests in Cohen, 1977 and 1988, pp. 364–379). As already noted by Koele and Hoogstraten (1980, see also Koele, 1982, p. 514, note 1), Cohen (1977, 1988) systematically underestimated the power and overestimated the sample sizes if the total sample size N and the term $v + u + 1$ differ, where v and u denote the numerator and the denominator degrees of freedom of the F test, respectively. In order to reduce the number of tables necessary to perform power analyses, Cohen provided readers with tables for global F tests only (i.e., his Cases 0 and 1 of ANOVA F tests, see Cohen, 1977 and 1988, pp. 356–364). These tables are based on the premises that

$$v = (n - 1)(u + 1) \tag{1}$$

and

$$\lambda = f^2 n(u + 1), \tag{2}$$

where n denotes the average sample size per cell of the ANOVA design, f denotes Cohen's (1977, 1988, chap. 8.2) effect size index, and λ is the noncentrality parameter of the noncentral F distribution (see Johnson & Kotz, 1970, chap. 30). These formulas are correct for global F tests, because here the number k of cells is equal to $u + 1$ (see Equation 7 below). However, as noted by Cohen (1977, 1988, p. 365), Formula 1 is incorrect for special F tests in factorial designs in which the relation between k and u breaks down. To cope with this problem, Cohen suggested adjusting n so that

$$n' := v/(u + 1) + 1 \tag{3}$$

is used in his tables instead of n. Substituting n' for n in Equation 1 shows that this adjustment indeed leads to the correct denominator degrees of freedom (v) in all possible cases. Unfortunately, the adjustment has an undesirable side effect in Formula 2, in which λ is replaced by $\lambda' = f^2(v + u + 1)$. In general, $\lambda' \leq \lambda$, with $\lambda' = \lambda$ if and only if $v + u + 1 = N$. Actually, this problem can be solved by simultaneously adjusting f so that $f' := f\,(N/(u + v + 1))^{1/2}$ is used instead of f (see Koele & Hoogstraten, 1980, p. 9). If f is not adjusted, the power is underestimated. The underestimation is negligible for small effect sizes f, but it becomes substantial for large effect sizes and large differences between N and $v + u + 1$. To illustrate, Cohen (1977 and 1988, p. 375, Table 8.3.34) reported a power of .66 for the B × C two-way interaction test in a 3 × 4 × 5 ANOVA design (thus, $u = 12$), given a large effect size ($f = .40$), $\alpha = .01$, and $n = 3$ per cell (thus, $v = 120$). The GPOWER accuracy mode calculates a power of .8531 for the same situation.

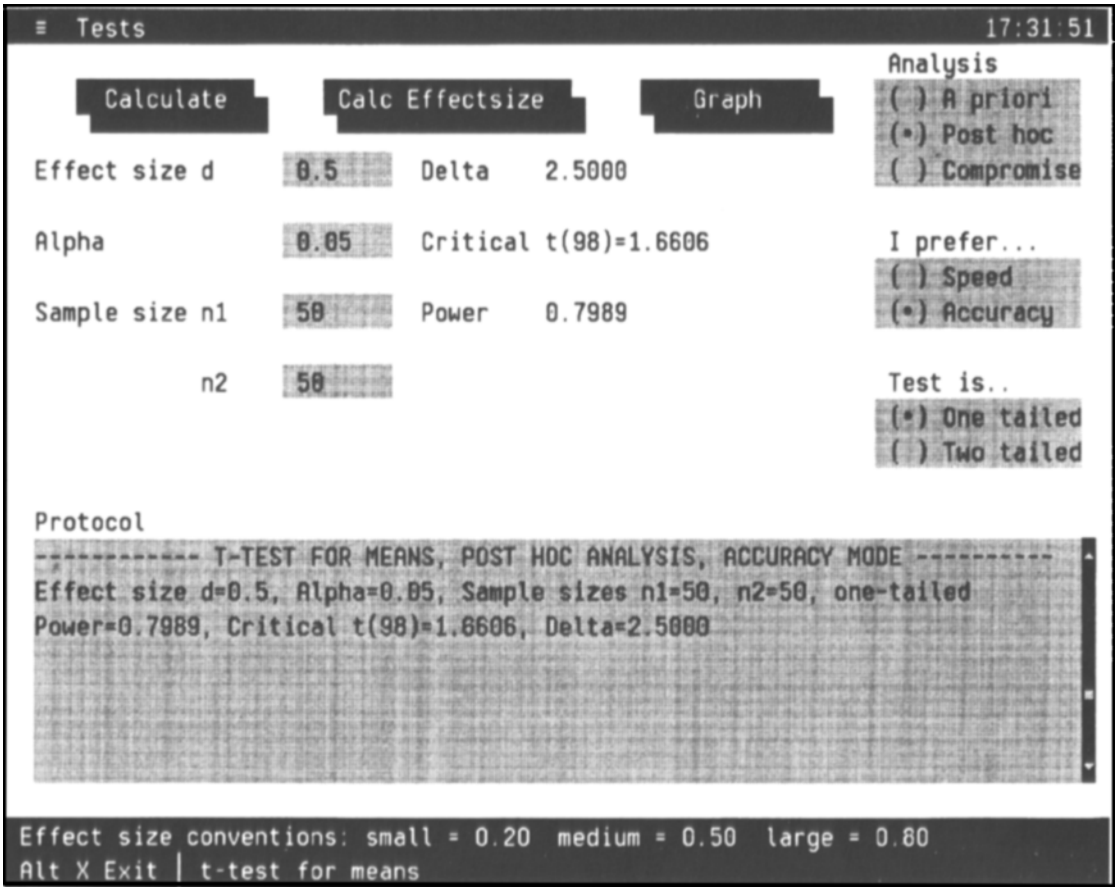
Program Handling

The present versions of GPOWER assume that users are familiar with the basic concepts of statistical power analyses. Moreover, it is useful if users know about Cohen's effect size measures and the definitions of "small," "medium," and "large" effect sizes. The relevant background information may be found in Cohen (1988). However, GPOWER also allows calculation of the effect size measures from basic parameters such as means, variances, and probabilities.

The first three steps in GPOWER applications are (1) the selection of the statistical test to be considered, (2) the specification of the desired type of power analysis, and (3) the selection of the accuracy level of the computations. This is done by choosing the appropriate items in the "Test" menu (Macintosh version: "Type of Test"), the "Analysis" menu (Macintosh version: "Type of Power Analysis"), and the "I prefer..." menu, respectively. GPOWER offers both accurate and approximate a priori, post hoc, and compromise power analyses for seven types of tests: (1) *t* tests for means based on two independent samples (Cohen, 1977, 1988, chap. 2); (2) *t* tests for correlations (Cohen, 1977, 1988, chap. 3); (3) other *t* tests; (4) *F* tests in fixed-effects ANOVAs (Cohen, 1977, 1988, chap. 8); (5) *F* tests in multiple regression/correlation (MRC) analyses (Cohen, 1977, 1988, chap. 9); (6) other *F* tests; and (7) χ^2 tests (Cohen, 1977, 1988, chap. 7). The χ^2 test option is general in the sense that power analyses can be conducted for all χ^2 tests on discrete data. The "Other *t* Tests" and "Other *F* Tests" items were added to allow for power analyses of nonstandard *t* tests and *F* tests. One-sample and matched-pairs *t* tests are examples of the former, while approximate *F* tests for fixed factors in mixed models (Koele, 1982) and approximate multivariate analysis of variance (MANOVA) *F* tests (Bredenkamp & Erdfelder, 1985; Cohen, 1988, chap. 10; O'Brien & Muller, 1993) are examples of the latter. Last but not least, power analyses for *z* tests based on the standardized normal distribution can also be conducted with GPOWER, because the noncentral *t* distribution with noncentrality parameter δ and the normal distribution with mean δ and standard deviation 1 converge for $df \rightarrow \infty$ (Johnson & Kotz, 1970, p. 207). Thus, in order to compute the power of the *z* test, one selects the "Other *t* Tests" item and specifies a very large *df* value (e.g., $df = 32000$).

GPOWER always adjusts the display to the selected type of test and the type of power analysis. For example, if a post hoc power analysis for *t* tests for means is selected and performed for the default input values (by clicking on the "Calculate" button or pressing the return key), the MS-DOS version and the Macintosh version of GPOWER present the displays shown in Figures 1 and 2, respectively. The visible parameters are either input or output parameters. Input parameters can be manipulated by users, while output parameters correspond to power analytic implications of the input parameters.

Input parameters. One part of the obligatory input is determined by the selected type of test. For example,

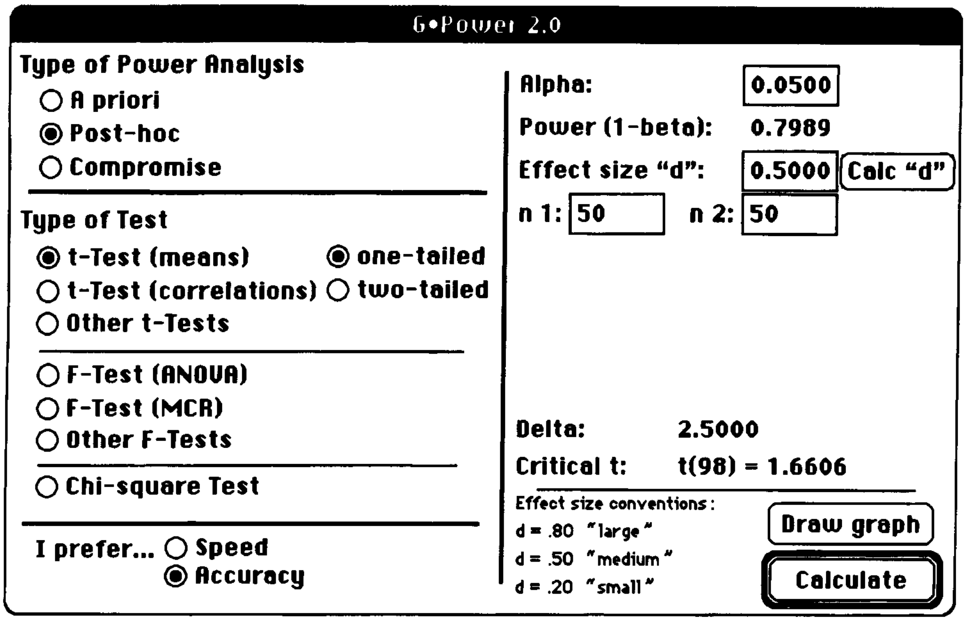


```
Effect size d    0.5     Delta     2.5000
Alpha            0.05    Critical  t(98)=1.6606
Sample size n1   50      Power     0.7989
            n2   50
```

```
----------- T-TEST FOR MEANS, POST HOC ANALYSIS, ACCURACY MODE ---------
Effect size d=0.5, Alpha=0.05, Sample sizes n1=50, n2=50, one-tailed
Power=0.7989, Critical t(98)=1.6606, Delta=2.5000
```

```
Effect size conventions: small = 0.20  medium = 0.50  large = 0.80
Alt X Exit | t-test for means
```

Figure 1. GPOWER display for post hoc power analyses in *t* tests for means (MS-DOS version).



G•Power 2.0

Type of Power Analysis

○ A priori
● Post-hoc
○ Compromise

Type of Test

● t-Test (means) ● one-tailed
○ t-Test (correlations) ○ two-tailed
○ Other t-Tests

○ F-Test (ANOVA)
○ F-Test (MCR)
○ Other F-Tests

○ Chi-square Test

I prefer... ○ Speed
● Accuracy

Alpha: 0.0500
Power (1-beta): 0.7989
Effect size "d": 0.5000
n 1: 50 n 2: 50

Delta: 2.5000
Critical t: t(98) = 1.6606

Effect size conventions:
d = .80 "large"
d = .50 "medium"
d = .20 "small"

Figure 2. GPOWER display for post hoc power analyses in *t* tests for means (Macintosh version).

for all three types of *t* tests, users must specify whether they consider a one-tailed or a two-tailed test. In addition, the degrees of freedom are needed in the "Other *t* Tests" procedure. If, alternatively, ANOVA *F* tests are selected, then it is necessary to specify whether the analysis refers to global or special *F* tests. A *global F test* is a test of the hypothesis that all means are equal (i.e., the H_0 in one-way ANOVAs or the hypothesis of neither main effects nor interactions in multi-way ANOVAs). *Special F tests* refer to subsets of linear contrasts (e.g., trend tests or planned comparisons in one-factorial designs and tests for interactions in multifactorial designs). For both types of *F* tests, it is necessary to specify the total number of groups or treatment combinations. For special *F* tests, it is also mandatory to determine the numerator degrees of freedom.

The distinction between global and special tests is also necessary for *F* tests in MRC analyses. For MRC analyses, global tests refer to the hypothesis that the multiple correlation is zero, whereas special tests refer to the hypothesis that the regression weights are zero for some proper subset of the predictors. For both types of tests, the total number of predictors in the regression model must be specified. In addition, again, the numerator degrees of freedom are needed in order to perform power analyses for special *F* tests.

GPOWER offers no separate option for *F* tests in analyses of covariance (ANCOVAs) because these are easily expressed as MRC *F* tests (see, e.g., Cohen & Cohen, 1983). In order to perform power analyses for ANCOVAs, one simply selects the MRC special *F* test option, enters the appropriate number of numerator degrees of freedom, and specifies the number of covariates plus the total number of groups minus 1 as the total number of predictor variables.

Although the ANOVA and MRC *F* test options cover a considerable number of *F* test applications, not all *F* tests fit into these two frames. Therefore, the "Other *F* Tests" item was added, which allows for power analyses of *any F* test (including those handled more conveniently by the ANOVA and MRC options). Both the numerator and the denominator degrees of freedom are requested as input parameters with this option.

For all types of tests, the remaining obligatory input is determined by the type of power analysis selected. In a priori power analyses, the desired α and β levels as well as a test-specific effect size measure must be specified. In post hoc analyses, the α level, the sample size, and the effect size need to be determined. Finally, compromise analyses require the specification of the sample size, the effect size, and the error ratio $q := \beta/\alpha$.

For most of the tests covered by GPOWER, all three types of power analyses are available. The exceptions are that no a priori analyses can be selected in the "Other *t* Tests" and "Other *F* Tests" procedures. This restriction is necessary because the parameter *N* is not linked to *df*, the (denominator) degrees of freedom of the test, in the "Other *t* Tests" or "Other *F* Tests" procedures. The problem is that the parameters *N* and *df must* be specified independently for the "Other *t* Tests" and "Other *F* Tests" items to be applicable to all possible *t* tests and *F* tests, respectively. Taking a small detour, it is nevertheless possible to compute a priori power analyses. One simply performs post hoc analyses repeatedly, adjusting *N* and the corresponding *df* value until the desired power level is found.

Cohen's (1969, 1977, 1988, 1992) effect size measures are well known and his conventions of "small," "medium," and "large" effects proved to be useful. For these reasons, we decided to render GPOWER completely compatible with Cohen's measures and to display the effect size conventions appropriate for the type of test selected. Effect size values can either be entered directly or they can be calculated from basic parameters characterizing H_1 (e.g., means, variances, and probabilities). To use the latter option, users must click on the "Calc Effectsize" button (Macintosh version: "Calc '*x*'," with *x* representing the effect size parameter).

In order to prepare the appropriate GPOWER input, it may sometimes be necessary to know the relation between sample sizes and effect size measures on the one hand and the noncentrality parameters of the noncentral distributions on the other hand. In *t* tests for means, the noncentrality parameter δ is

$$\delta = d \cdot \sqrt{\frac{n_1 \cdot n_2^2 + n_2 \cdot n_1^2}{(n_1 + n_2)^2}}, \qquad (4)$$

where $d := |\mu_1 - \mu_2|/\sigma$ is Cohen's (1977 and 1988, p. 40) effect size parameter for *t* tests for means, and n_1 and n_2 are the sample sizes in Groups 1 and 2, respectively. In *t* tests for correlations,

$$\delta = \sqrt{\frac{\rho^2}{1 - \rho^2}} \cdot \sqrt{N}\,, \qquad (5)$$

where N is the total sample size (i.e., the number of pairs of values) and ρ is the population correlation coefficient according to H_1 (i.e., Cohen's *r*; see Cohen, 1977, 1988, pp. 77–81).

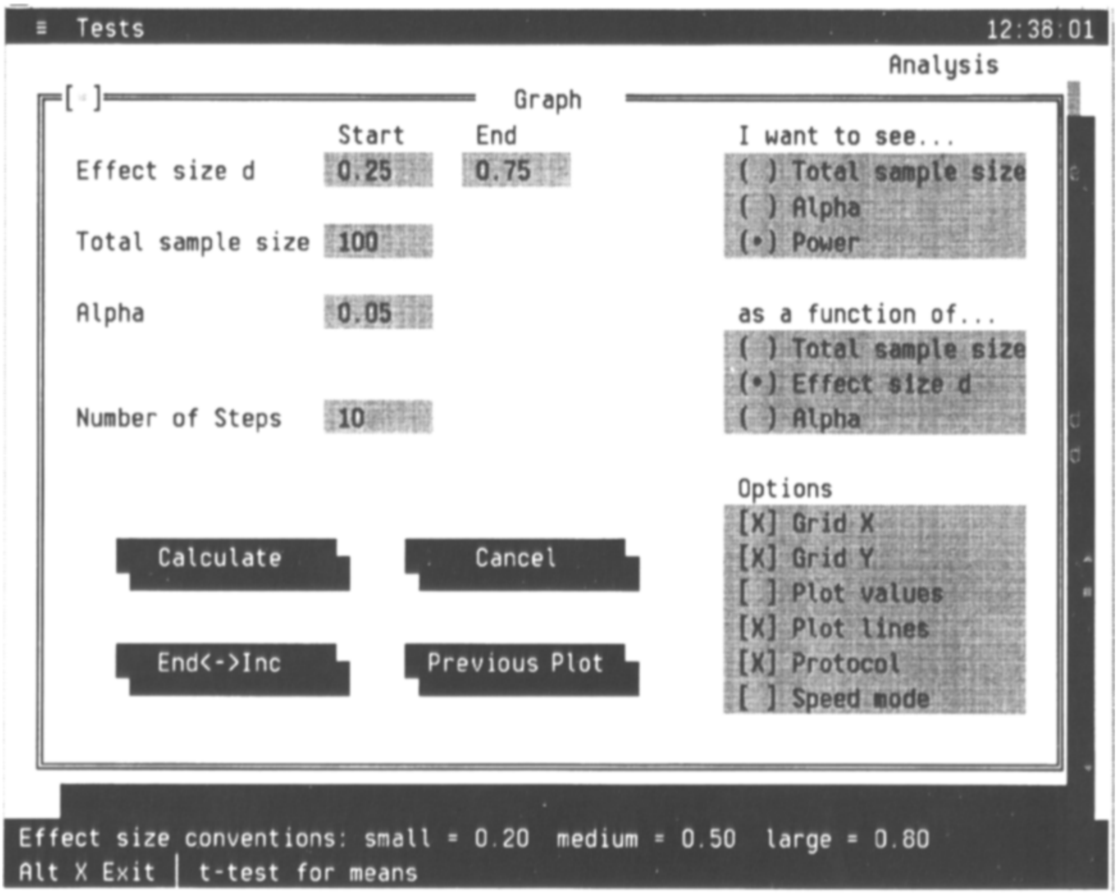
In the "Other *t* Tests" option, we used f as an effect size measure (Cohen, 1977 and 1988, chap. 8.2). The relation between δ and f is simply

$$\delta = f \cdot \sqrt{N}\,. \qquad (6)$$

The standardized effect size measures f or f^2 are also used in power analyses for *F* tests. Their relation to the noncentrality parameter λ of the noncentral *F* distribution is given by

$$\lambda = f^2 \cdot N, \qquad (7)$$

where $f^2 := \rho^2/(1-\rho^2)$, and ρ^2 denotes the coefficient of determination in the population according to H_1 (see, e.g., Koele, 1982, p. 514).[6]



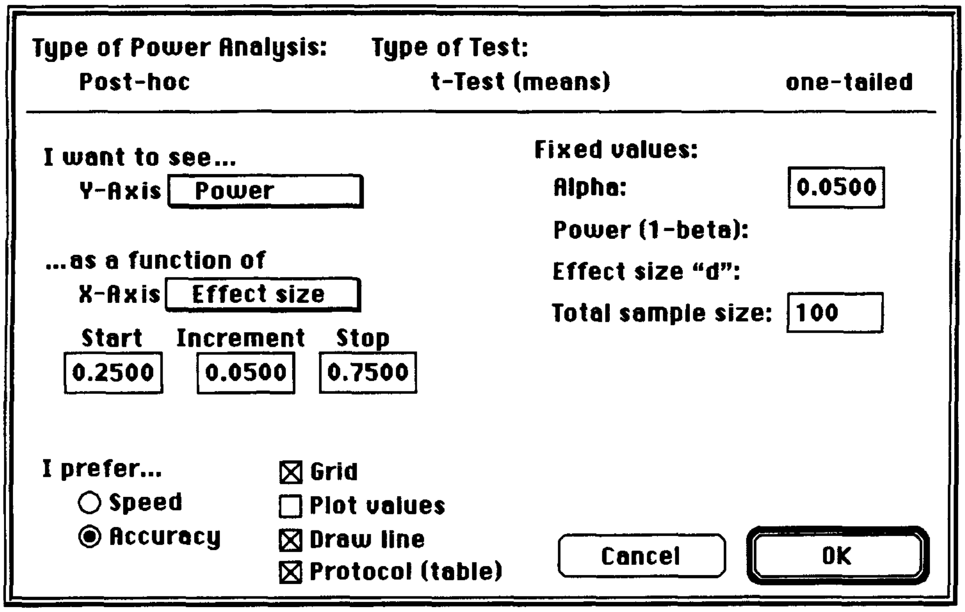
```
≡ Tests                                                    12:38:01
                                                           Analysis
[ ]                         Graph
                    Start     End              I want to see...
Effect size d       0.25      0.75             ( ) Total sample size
                                               ( ) Alpha
Total sample size   100                        (•) Power

Alpha               0.05                       as a function of...
                                               ( ) Total sample size
                                               (•) Effect size d
Number of Steps     10                         ( ) Alpha

                                               Options
                                               [X] Grid X
                                               [X] Grid Y
                                               [ ] Plot values
                                               [X] Plot lines
                                               [X] Protocol
                                               [ ] Speed mode

Effect size conventions: small = 0.20  medium = 0.50  large = 0.80
Alt X Exit | t-test for means
```

Figure 3. GPOWER graph display for *t* tests for means (MS-DOS version).



Type of Power Analysis: Post-hoc

Type of Test: t-Test (means) one-tailed

I want to see...
Y-Axis Power

...as a function of
X-Axis Effect size

Start 0.2500 Increment 0.0500 Stop 0.7500

Fixed values:
Alpha: 0.0500
Power (1-beta):
Effect size "d":
Total sample size: 100

I prefer...
○ Speed
● Accuracy

☒ Grid
□ Plot values
☒ Draw line
☒ Protocol (table)

Figure 4. GPOWER graph display for *t* tests for means (Macintosh version).

For χ^2 tests based on m-cell contingency tables ($m \in \mathbb{N}$), Cohen (1977, 1988, chap. 7) used

$$w := \sqrt{\sum_{i=1}^{m} \frac{(p_{1i} - p_{0i})^2}{p_{0i}}} \tag{8}$$

as an effect size measure, where p_{0i} and p_{1i} denote the cell probabilities for the ith cell according to H_0 and H_1, respectively. Then

$$\lambda = w^2 \cdot N \tag{9}$$

is the noncentrality parameter of the noncentral χ^2 distribution (Cohen, 1988, p. 549).

Output parameters. Pressing the return key or clicking on the "Calculate" button initiates the GPOWER calculations. The output consists of (1) sample sizes in a priori analyses, power values in post hoc analyses, and α as well as β values in compromise analyses; (2) the noncentrality parameter of the reference distribution as implied by N and the effect size specification; (3) the critical value of the test statistic defining the boundary of the rejection region of H_0; and (4) the degrees of freedom of the test. We recommend comparing the degrees of freedom output with the degrees of freedom as reported by the computer program used for statistical data analysis. If the reported degrees of freedom mismatch, either the GPOWER input or the input to the data analysis package has been misspecified. In either case, the GPOWER results do not apply to the test reported by the data analysis program. If the reported degrees of freedom match and if, in addition, users make sure that they base their statistical decision on the critical value as reported by GPOWER, then there is only one possible source of error left, namely, the noncentrality parameter of the reference distribution. By carefully specifying the effect sizes and the sample sizes, errors can be ruled out completely for most of the tests offered in the "Test" menu. However, special care must be taken when using the "Other *t* Tests" and "Other *F* Tests" options. Only users who are familiar with the definition of the noncentrality parameter for their special type of test should make use of these options. Equations 6 and 7 will help to specify the input parameters N and f correctly.

Each power calculation conducted by GPOWER is automatically copied to a protocol window. The contents of this window can be saved to a file.

Graph options. GPOWER results can be displayed graphically by clicking on the "Graph" button (Macintosh version: "Draw graph"). Starting from the main windows shown in Figures 1 or 2, the graph parameters are specified in windows as displayed in Figure 3 (MS-DOS version) or Figure 4 (Macintosh version). For most of the tests covered by GPOWER, each of the variables α, $1 - \beta$, effect size, and sample size can be plotted as a function of any other of these variables. However, for the reasons already discussed, sample sizes may not be selected as variables when using "Other *t* Tests" or "Other *F* tests." Plots can be generated with several display options turned on or off, and a table containing the plotted values can be copied to the protocol window. Users can specify the lowest ("Start") and the largest value ("End") on the abscissa and the number of data points to be calculated. Both speed mode and accuracy mode calculations are available in the graph window.

The plots shown in Figures 5 and 6 were generated with the MS-DOS and the Macintosh versions, respectively. They may be obtained by preparing the inputs shown in Figures 3 or 4 and pressing the return key or clicking on the "Calculate" button (Macintosh version: "OK" button). In the Macintosh version, the graph can

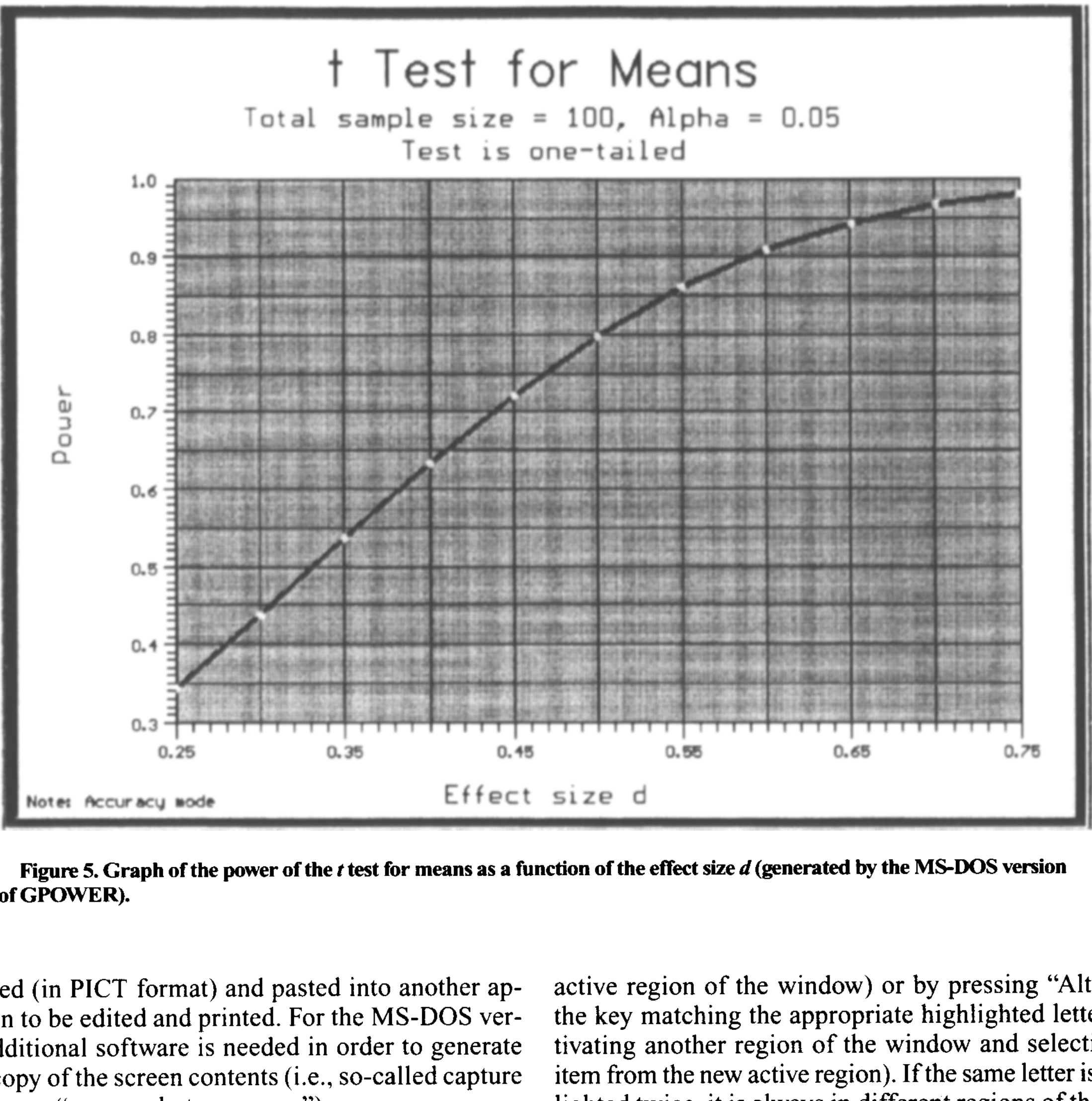


Figure 5. Graph of the power of the *t* test for means as a function of the effect size *d* (generated by the MS-DOS version of GPOWER).

be copied (in PICT format) and pasted into another application to be edited and printed. For the MS-DOS version, additional software is needed in order to generate a hard copy of the screen contents (i.e., so-called capture programs or "screen-shot programs").

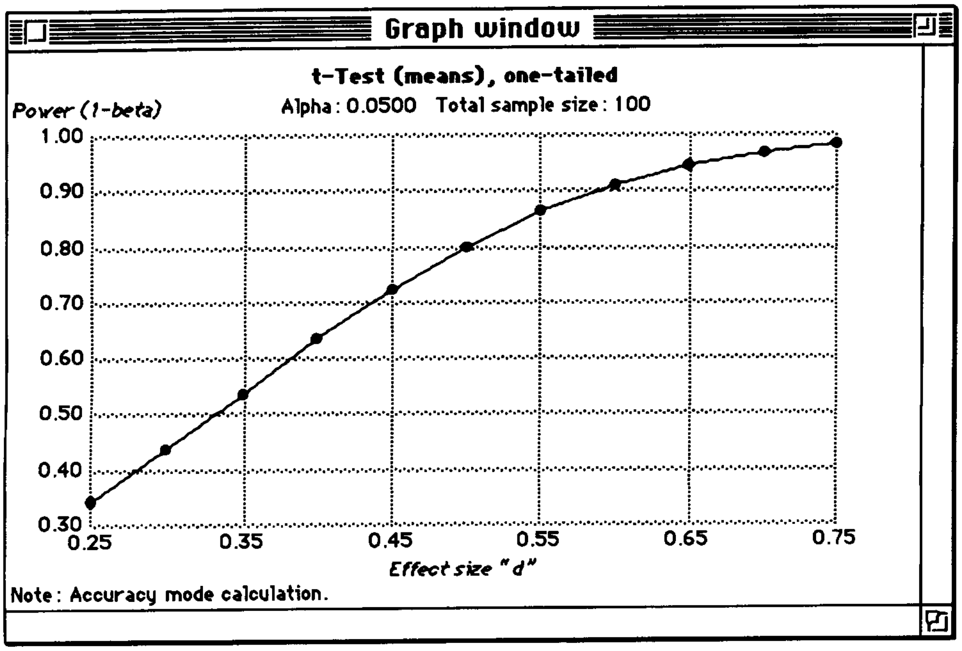
Hardware and Software Requirements

The Macintosh version of GPOWER should run on any 68K Macintosh using system 6.0.7 or higher. It has also been tested successfully on some PowerMacintosh models where it runs in emulation mode. Two different implementations are available, one that requires and takes advantage of an arithmetic coprocessor (GPOWER/FPU), and one that does not (GPOWER).

The MS-DOS GPOWER version requires an IBM-compatible PC with MS-DOS 3.31 or higher and a graphic card. GPOWER may also be used in the DOS windows of Windows 3.1 or OS/2 2.0. We recommend installing the program on a 386 (or better) PC with an arithmetic coprocessor. To take full advantage of all GPOWER options, a VGA graphic card and a color monitor are necessary. A mouse is not necessary but it is very helpful. When using the MS-DOS version without a mouse, one selects options by pressing the key corresponding to the appropriate highlighted letter (selecting items within the active region of the window) or by pressing "Alt" plus the key matching the appropriate highlighted letter (activating another region of the window and selecting an item from the new active region). If the same letter is highlighted twice, it is always in different regions of the window. Within parts of the regions, items can also be selected with the arrow keys.

AVAILABILITY OF THE PROGRAM

GPOWER 2.0 can be obtained free of charge. The most convenient way to get a copy of GPOWER is to download the program from the public FTP server at the University of Trier, Germany (ftp.uni-trier.de; user ID: anonymous; password: your e-mail address). The self-extracting archive "gpower2i.exe" contains all necessary files for the MS-DOS version of GPOWER. It is located in the directory "/pub/pc/msdos." Both Macintosh implementations may be obtained by downloading the StuffIt archive "gpower202.sit" from the directory "/pub/mac/local." Alternatively, Macintosh users may download "gpower202.sit" from the "MacPsych archive for psychology concerning the Macintosh computer" (see Huff & Sobiloff, 1993, for details). Transfer of the programs via regular mail is also possible. Interested readers should



Graph window

t-Test (means), one-tailed

Alpha: 0.0500 Total sample size: 100

Note: Accuracy mode calculation.

Figure 6. Graph of the power of the *t* test for means as a function of the effect size *d* (generated by the Macintosh version of GPOWER).

write to the first or second author if they want to receive the MS-DOS version, and to the third author if they want the Macintosh version. A completely new, unformatted floppy disc must be enclosed.

In publications involving the use of GPOWER, users are expected to cite the program version used (i.e., Faul & Erdfelder, 1992, for the MS-DOS version and Buchner et al., 1992, for the Macintosh version).

Although considerable effort has been directed toward making GPOWER error free, there is no warranty. Users are kindly asked to communicate any problems encountered with the program to the authors.

REFERENCES

Bargmann, R. E., & Ghosh, S. P. (1964). *Noncentral statistical distribution programs for a computer language* (IBM Research Report RC-1231). Yorktown Heights, NY: IBM Watson Research Center.

Borenstein, M., & Cohen, J. (1988). *Statistical power analysis: A computer program*. Hillsdale, NJ: Erlbaum.

Borenstein, M., Cohen, J., Rothstein, H. R., Pollack, S., & Kane, J. M. (1990). Statistical power analysis for one-way analysis of variance: A computer program. *Behavior Research Methods, Instruments, & Computers*, **22**, 271-282.

Borenstein, M., Cohen, J., Rothstein, H. R., Pollack, S., & Kane, J. M. (1992). A visual approach to statistical power analysis on the microcomputer. *Behavior Research Methods, Instruments, & Computers*, **24**, 565-572.

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung* [The test of significance in behavioral research]. Frankfurt, Germany: Akademische Verlagsgesellschaft.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente* [Theory and design of psychological experiments]. Darmstadt, Germany: Steinkopff.

Bredenkamp, J., & Erdfelder, E. (1985). *Multivariate Varianzanalyse nach dem V-Kriterium* [Multivariate analysis of variance using the V-criterion]. *Psychologische Beiträge*, **27**, 127-154.

Buchner, A., Faul, F., & Erdfelder, E. (1992). *GPOWER. A priori-, post hoc-, and compromise power analyses for the Macintosh* [Computer program]. Bonn, Germany: Bonn University.

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

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York. McGraw-Hill.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, **112**, 155-159.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, **37**, 553-558.

Dixon, W. F., & Massey, F. J., Jr. (1957). *Introduction to statistical analysis* (2nd ed.). New York: McGraw-Hill.

Erdfelder, E. (1984). Zur Bedeutung und Kontrolle des β-Fehlers bei der inferenzstatistischen Prüfung log-linearer Modelle [On significance and control of the β error in statistical tests of log-linear models]. *Zeitschrift für Sozialpsychologie*, **15**, 18-32.

Erdfelder, E., & Bredenkamp, J. (1994). Hypothesenprüfung [Evaluation of hypotheses]. In T. Herrmann & W. H. Tack (Eds.), *Methodologische Grundlagen der Psychologie* (pp. 604-648). Göttingen, Germany: Hogrefe.

Faul, F., & Erdfelder, E. (1992). *GPOWER. A priori-, post hoc-, and compromise power analyses for MS-DOS* [Computer program]. Bonn, Germany. Bonn University.

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. Methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.

Goldstein, R. (1989). Power and sample size via MS/PC-DOS computers. *American Statistician*, **43**, 253-260.

Hager, W., & Möller, H. (1986). Tables and procedures for the determination of power and sample sizes in univariate and multivariate analyses of variance and regression. *Biometrical Journal*, **28**, 647-663.

Hardison, C. D., Quade, D., & Langston, R. D. (1983). Nine functions for probability distributions. In SAS Institute, Inc. (Ed.), *SUGI supplemental library user's guide, 1983 edition* (pp. 229-236). Cary, NC: SAS Institute, Inc.

Huff, C., & Sobiloff, B. (1993). MacPsych: An electronic discussion list and archive for psychology concerning the Macintosh computer. *Behavior Research Methods, Instruments, & Computers*, **25**, 60-64.

Johnson, N. L., & Kotz, S. (1970). *Distributions in statistics Continuous univariate distributions–2*. New York: Wiley.

Koele, P. (1982). Calculating power in analysis of variance *Psychological Bulletin*, **92**, 513-516.

Koele, P., & Hoogstraten, J. (1980). *Power and sample size calculations in analysis of variance* (Révész Berichten No. 12). Amsterdam. University of Amsterdam.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage.

Laubscher, N. F. (1960). Normalizing the noncentral *t* and *F* distributions. *Annals of Mathematical Statistics*, **31**, 1105-1112.

Lehmann, E. L. (1975). *Nonparametrics. Statistical methods based on ranks*. San Francisco: Holden-Day.

Lipsey, M. W. (1990). *Design sensitivity Statistical power for experimental research*. Newbury Park, CA: Sage.

Milligan, G. W. (1979). A computer program for calculating power of the chi-square test. *Educational & Psychological Measurement*, **39**, 681-684.

Oakes, M. (1986). *Statistical inference A commentary for the social and behavioral sciences*. New York: Wiley.

O'Brien, R. G., & Muller, K. E. (1993). Unified power analysis for *t*-tests through multivariate hypotheses. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 297-344). New York. Marcel Dekker.

Onghena, P. (1994). *The power of randomization tests for single-case designs*. Unpublished doctoral dissertation. Leuven, Belgium: Katholieke Universiteit Leuven.

Onghena, P., & Van Damme, G. (1994). SCRT 1.1: Single-case randomization tests. *Behavior Research Methods, Instruments, & Computers*, **26**, 369

Patnaik, P. B. (1949). The non-central χ^2- and *F*-distributions and their applications. *Biometrika*, **36**, 202-232.

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making type I errors. *Psychological Bulletin*, **102**, 159-163.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1988). *Numerical recipes in C The art of scientific computing*. Cambridge: Cambridge University Press.

Rossi, J. S. (1990). Statistical power of psychological research. What have we gained in 20 years? *Journal of Consulting & Clinical Psychology*, **58**, 646-656.

Rothstein, H., Borenstein, M., Cohen, J., & Pollack, S. (1990). Statistical power analysis for multiple regression/correlation: A computer program. *Educational & Psychological Measurement*, **50**, 819-830.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, **105**, 309-316.

Singer, B., Lovie, A. D., & Lovie, P. (1986). Sample size and power In A. D. Lovie (Ed.), *New developments in statistics for psychology and the social sciences* (pp. 129-142). London: British Psychological Society and Methuen.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, **76**, 105-110.

Westermann, R., & Hager, W. (1986). Error probabilities in educational and psychological research. *Journal of Educational Statistics*, **11**, 117-146.

NOTES

1. If H_1 is not determined uniquely, increasing the effect size may be a third way to raise statistical power (Rossi, 1990). Techniques to increase effect sizes aim at controlling the various sources of error variance, for example, by using highly reliable measures as dependent variables (Erdfelder & Bredenkamp, 1994). Where such possibilities exist, they should of course be used. However, it appears that substantial gains in statistical power cannot be achieved along these lines.

2. The tables by Hager and Möller (1986) are a pleasing exception, covering selected α values in the range $.002 \leq \alpha \leq .40$. However, these tables are based on the noncentral χ^2 distributions exclusively, which allow only rough approximations of the power of the *F* test.

3. One anonymous reviewer pointed out that the term *post hoc power analysis* might possibly be misunderstood. Therefore, we want to emphasize that this term does *not* mean that the power is computed for an effect size *as estimated from a sample*. We use the term in the same way as Cohen (1969, 1977, 1988) did. Thus the effect size is a population parameter to be specified a priori in *all* types of power analyses. This specification should not depend on a sample of data but on theoretical considerations. In fact, it is erroneous to assume that a post hoc power analysis applied to effect sizes equated with their sample estimates yields something like the "true power level" in all applications of a statistical test. This assumption would be valid only if one could make sure that the population effect size were equal to the effect size estimate, irrespective of sample size. Obviously, it is impossible to show this, and if it would be possible, then there would be no need to conduct statistical tests.

4. Some variants of power analyses for nonparametric tests can be conducted by adjusting the result obtained for the corresponding parametric test (Bredenkamp, 1980; Singer, Lovie, & Lovie, 1986). For example, an a priori power analysis for the Wilcoxon-Mann-Whitney *U* test can be conducted by first performing an a priori power analysis for the *t* test for means. If the *t* test model is valid, and N_t designates the sample size necessary for the *t* test to achieve some given power $(1 - \beta)$, then the sample size $N_U = N_t$/A.R.E. yields approximately the same power for the *U* test. A.R.E. denotes the *a*symptotic *r*elative *e*fficiency (or Pitman efficiency) of the *U* test relative to the *t* test, which is $3/\pi = .955$ (see Lehmann, 1975). The same procedure may often be used to approximate the power of randomization tests (Onghena, 1994). In this case, the A.R.E. of the randomization test relative to the corresponding parametric test is 1. For power analyses in randomization tests that do not have a corresponding parametric test, special computer software is in preparation (Onghena, 1994; Onghena & Van Damme, 1994).

5 To be precise: In the procedures "Other *t* Tests" and "Other *F* Tests," no rounding to integer values is performed. In *t* tests for means and analysis of variance (ANOVA) *F* tests, in contrast, the solution is always the smallest multiple of the number *k* of groups that yields a power value as large or larger than the prespecified value.

6. For global ANOVA *F* tests, ρ^2 is just η^2. For special *F* tests of main effects or interactions in complex ANOVA designs, ρ^2 equals the partial η^2. Analogously, ρ^2 coincides with the (partial) squared multiple correlation in multiple regression/correlation *F* tests (Cohen, 1988, chap. 9.2.1).

(Manuscript received October 12, 1993;
revision accepted for publication October 27, 1994.)