Perspect Behav Sci (2019) 42:133–152
https://doi.org/10.1007/s40614-018-0167-4

Tutorial: Small-N Power Analysis

Elizabeth G. E. Kyonka[1]

Published online: 22 May 2018

Abstract Power analysis is an overlooked and underreported aspect of study design. A priori power analysis involves estimating the sample size required for a study based on predetermined maximum tolerable Type I and II error rates and the minimum effect size that would be clinically, practically, or theoretically meaningful. Power is more often discussed within the context of large-N group designs, but power analyses can be used in small-N research and within-subjects designs to maximize the probative value of the research. In this tutorial, case studies illustrate how power analysis can be used by behavior analysts to compare two independent groups, behavior in baseline and intervention conditions, and response characteristics across multiple within-subject treatments. After reading this tutorial, the reader will be able to estimate just noticeable differences using means and standard deviations, convert them to standardized effect sizes, and use G*Power to determine the sample size needed to detect an effect with desired power.

Keywords Experimental design · A priori power analysis · Effect size · Sample size · Tests of statistical significance · Hypothesis testing · G*Power

Behavior analysts have a longstanding history of skepticism when it comes to the necessity and utility of statistical inference. Sidman (1960, p. 44) described statistical tests as a "curious negation of the professed aims of science," inferior to techniques that establish experimental control because statistical tests rely on comparisons against an unknown parent distribution. In a similar vein, Michael (1974, p. 650) described statistical inference as a "weak solution to an artificial problem" in single-organism research, arguing that even in applied settings the problem of uncontrolled variance in observations can be eliminated with appropriate experimental controls. Behavior analysts continue to argue that adequate experimental control obviates the need for statistical tests (e.g., Cohen, Feinstein, Masuda, & Vowles, 2014; Fisher & Lerman, 2014; Perone, 1999). In spite of many eloquent arguments against their use, a survey of

✉ Elizabeth G. E. Kyonka
liz.kyonka@une.edu.au

[1] Psychology, University of New England, Armidale, NSW 2351, Australia

almost any poster session at a large behavior analysis conference will show that this skepticism does not prevent behavior analysts from relying on statistical tests.

Null hypothesis significance testing (NHST), the orthodox statistical procedure in psychology for most of the 20th century, has been criticized by behavior analysts and others for as long as it has been practiced. It involves rejecting or failing to reject a particular "null" hypothesis based on whether the probability of the obtained result is less than some value (usually 5%) if the hypothesis were true. Branch (1999, 2014) outlined the logical fallacy and malignant consequences of NHST for behavior analysts in at least two publications. To wit, *p*-values reported as outcomes of NHST are generally misinterpreted, and reliance on NHST suppresses genuine scientific advancement. Many social and behavioral scientists who are not behavior analysts also call for an end to the statistical ritual of NHST (e.g., Gigerenzer, 2004). Some researchers blame the misuse of NHST for a replicability crisis in psychology (Pashler & Harris, 2012). The *American Statistical Association* recently released a statement on statistical significance and *p*-values (Wasserstein & Lazar, 2016). It clarifies that a *p*-value is a measure of how incompatible data are with a specified statistical model, not of the probability that a particular hypothesis is true, the size of an effect or the importance of a result. It also provides some recommendations for good statistical practice that emphasize the importance of using an analytic approach that is compatible with study design and caution against relying solely on *p*-values when drawing scientific conclusions. The *American Psychological Association* also published guidelines on the use of statistical methods in psychology journals (Wilkinson & the Task Force on Statistical Inference, 1999). Among other recommendations, it prohibited some of the language associated with NHST and mandated the inclusion of effect sizes and estimates of reliability (i.e., confidence intervals) when reporting statistical results of any hypothesis test. At least one journal reacted by banning *p*-values altogether (Trafimow & Marks, 2015).

At a time when other fields seem to be abandoning certain types of statistical inference in favor of other ways of evaluating evidence, behavior analysts' use of inferential statistics is increasing. The number of randomized controlled trials appearing in applied journals has increased over time. A search of the Wiley Online Library on January 22, 2018 for "randomized controlled trial" in the *Journal of Applied Behavior Analysis* produced 40 results. Two articles were published prior to 2000, 8 between 2000 and 2009, and the remaining 30 in were published in published 2010 or later. Zimmermann, Watkins, and Poling (2015, p. 209) reported that the proportion of articles published in the *Journal of the Experimental Analysis of Behavior* that include "an inferential statistic" increases by approximately 8% every five years. A nontrivial portion of the research that behavior analysts do involves group-average comparisons and statistical inferences. Considering statistical power when determining sample sizes can help behavior analysts use these techniques correctly.

Behaviorist Constructions of Statistical Induction

Critiques of NHST are valid, but NHST is not the only form of statistical inference and there is a place for tests of statistical significance in behavioral research (Haig, 2017). NHST is a hybrid of Fisherian significance testing and Neyman and Pearson's hypothesis testing, approaches that are mathematically similar to each other and to NHST, but involve different objectives, procedures, and philosophies of statistics. Although the amalgamation of Fisher's insights about *p*-values with Neyman and Pearson's ideas about error rates embeds a logical fallacy (post hoc ergo propter hoc) in NHST, other perspectives (e.g., Bayesian, Neo-Fisherian, and error-statistical) are logically consistent. Within those approaches, tests of statistical significance can be used in combination with other analyses to answer certain research questions (Haig, 2017; Wasserstein & Lazar, 2016).

The error-statistical approach (Mayo & Spanos, 2006) is related to Neyman-Pearsonian hypothesis testing (Neyman & Pearson, 1928) and incorporates statistical methods of testing hypotheses that are based on Neyman and Pearson's inductive behaviorist philosophy of science. For both techniques, empirical research can contrast null and alternative hypotheses about data generating mechanisms. Hypotheses can be directional (e.g., response rates will be higher in this condition than in that condition), nondirectional (mean scores for this group will be different than for that group), or nil (this independent variable will have no effect on that behavior). The null and alternative hypotheses must exhaust the parameter space, such that one or the other must be correct (e.g., if the null hypothesis is that the response rates in two conditions are equal, the alternative hypothesis that exhausts the parameter space is that they are unequal). Neyman and Pearson (1928, 1933) specified an all-or-none procedure whereby the null hypothesis is rejected or accepted based on how the test statistic compares to a critical value. Through it does not immunize the researcher against logical fallacy (Mayo & Spanos, 2006), it differs from NHST in that the inference is whether the evidence supports the null or alternative model, not whether the null hypothesis is sufficiently improbable. In the error-statistical approach, the test statistic (*t* score, *F* ratio, etc.) is not compared to a critical value. Instead, it quantifies the discrepancy between the null hypothesis and data. An important aspect of the error statistical approach is that the probative value of the test is tempered by its severity. A statistical test is severe when the data collected provide good evidence for or against the null hypothesis (Mayo & Spanos, 2011). Statistical significance is informative because it "[enables] the assessment of how well probed or how severely tested claims are" (Mayo & Spanos, 2006, p. 328), not the likelihood of the hypothesis.

Power analysis is an important component of Neyman-Pearson hypothesis testing, error-statistical analysis, and related inductive techniques for evaluating evidence with inferential tests. Using a priori power analysis to determine an appropriate sample size for an experiment does not guarantee a severe test of the hypothesis; as quantified by Mayo and Spanos (2006, 2011) severity can only be determined after data collection is complete. What it does is ensure that the test is optimally sensitive for detecting the effect size that is of greatest interest to the researcher. Using power analysis to determine sample size is an important step in research design no less because it increases the proportion of studies that yield conclusive results (Simmons, Nelson, & Simonsohn, 2013) than because it is required by the *American Psychological Association* (Wilkinson & the Task Force on Statistical Inference, 1999).

Power analysis was identified as a critical step in hypothesis testing 90 years ago (Neyman & Pearson, 1928) and relatively plain-language instructions on how to use power analysis in the design of psychological research have been available for more than half a century (Cohen, 1962).

Unfortunately, power analyses are not conventionally reported in behavior analytic research, perhaps owing to a skill and knowledge gap among behavior analysts as a group. Inferential statistics and power analyses are not necessarily covered in behavior analysis research methods classes. They are not mentioned in the accreditation standards of the Association for Behavior Analysis International Accreditation Board (2017) or the Behavior Analyst Certification Board's (2017) current task list. Most modern comprehensive statistics textbooks provide detailed treatments of power analysis, but they do not focus on the small-N, within-subject designs preferred by behavior analysts. The mathematical principles of power analysis are more or less the same regardless of design, but behavior analysts may be less likely to conduct power analyses for their own designs because they have not seen examples of power analysis that resemble the type of research that they do.

This tutorial is designed to remove the lack of appropriate models as one of the possible reasons behavior analysts continue to ignore statistical power when designing experiments and reporting results. It describes statistical power and the factors that determine statistical power and illustrates through case studies how behavior analysts can use G*Power (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007) in small-N research. G*Power is a free power calculator that can be used in power analysis for a wide range of research designs, including many of those popular with behavior analysts. (It is available for download from http://www.gpower.hhu.de/en.html.) Of course, not all analyses of behavior involve inferential tests and the power analyses that are possible using G*Power are not applicable to all of the research designs that are used by behavior analysts. Research designs that are amenable to power analyses in G*Power focus on group-average effects. They include both between-group and within-subject comparisons and can be categorical (i.e., involving analysis of variance) or continuous (involving regression). I am not trying to convince anyone who relies on other designs or analytic techniques to start evaluating hypotheses with significance tests. The aim here is to help those behavior analysts who sometimes compare group or condition means to maximize the probative value of their results.

Statistical Power

Statistical power, the ability of a test to detect an effect of a given size, is an important consideration in research design. Failing to detect a meaningful effect when one is present is a Type I error and falsely detecting a meaningful effect that is not there is a Type II error. In Neyman-Pearson hypothesis testing, the long-run probabilities of Type I and Type II errors are referred to as α and β, respectively. Power is 1- β, the long-run probability of *not* making a Type II error, that is, of correctly concluding that there is no meaningful effect. When testing hypotheses, failure to consider statistical power in the initial planning stages can produce sample sizes that are too small or large. A priori power analysis involves computing the sample size required to detect an effect of a given size with the desired power. Observed (or post-hoc) power is the power of an already-conducted statistical to detect as significant (i.e., to lead to a rejection of the null hypothesis in NHST) an effect size equal to the one obtained. The observed power of a test is directly (though nonlinearly) related to its *p*-value: when the *p*-value is high, power is always low and vice versa. As such, observed power reports the same information as a *p*-value, expressed a different way. It is no better or different from the *p*-value, and there is no reason to include it in a Results section that reports the *p*-value for the same test.

Nevertheless, many statistical analysis tools and applications report observed power, so it is important that researchers do not confuse a priori and observed power.

Underpowered tests have too few observations to identify effects that are clinically, theoretically, or practically meaningful (Peterson, 2009). For example, a new drug that offers a subtle improvement over the current treatment may be worth research and development even if the effect is small. Likewise, if the dependent measure is naturally highly variable, detecting a difference with a statistical test requires more evidence than one that varies little. When an effect is small or a dependent measure highly variable, a test that compares the two treatments in few patients (e.g., 10) will not produce a statistically significant result. Absence of evidence is not evidence of absence. In those cases, the absence of a statistically significant difference indicates that the amount of evidence is not sufficient to provide compelling support for either hypothesis, not because the two treatments are equally effective. Overpowered tests will detect effects that are trivially small. For instance, a difference of one-tenth of an IQ point is not meaningful in any practical sense of the word even if it is reliable, valid, consistently replicable, and otherwise real. However, a test comparing IQ scores of two groups of people that had a sufficiently large sample size (e.g., 500,000 per group) would detect a difference of 0.1 point as statistically significant and lead to rejection of the hypothesis that the two groups were equally intelligent. An experiment can have too few or too many observations to have genuine probative value.

A priori power is a conditional probability and subject to potential misinterpretation, like other *p*-values. A *p*-value in NHST is the probability of the data given the hypothesis, not the probability of the hypothesis given the data. Likewise, power is the probability of rejecting the null hypothesis, given the alternative hypothesis, not the probability the alternative hypothesis is true, given the null hypothesis was rejected. Designing a study to have high statistical power does not guarantee that the results obtained will accurately reflect the true state of affairs. Conducting a power analysis to determine an appropriate sample size maximizes the probative value of the test by ensuring it is neither underpowered nor overpowered for a selected critical effect size. Another advantage of thinking about power when designing behavior-analytic studies is that it quantifies the benefits of experimental control: better experimental control reduces behavioral variability, making effects easier to detect, so studies with a high degree of experimental control require fewer observations.

Factors that Affect Statistical Power

The factors that determine statistical power mathematically are α, β, effect size, and sample size. Knowing the value of any three factors makes it possible to solve for the fourth. In an a priori power analysis, researchers decide on the largest Type I and II error rates they are willing to tolerate and the smallest effect they would consider to be meaningful, then use those values to solve for the sample size required. The mathematical functions relating α, β, and effect size to sample size are described in detail elsewhere (Cohen, 1988 remains the gold standard, but its coverage of within-subject designs is limited), but researchers with a general conceptual understanding of power and subject-matter expertise can use G*Power effectively without direct manipulation of the mathematical equations.

Type of Test

Power calculations are test-dependent. All else being equal, the number of observations required to detect an effect of a given size is different for a chi-square than a *t*-test, and different for 2 x 3 analysis of variance (ANOVA) than a 2 x 4 ANOVA. Whether a test is parametric or nonparametric and whether it involves within-subject or between-group comparisons are also important considerations.

Parametric statistical tests typically assume that the parent distributions from which samples are drawn are all normal distributions with the same standard deviation. Nonparametric tests are not assumption free, but in general they do not require that parent distributions conform to a specific shape. A researcher who initially planned to use a parametric test might report a nonparametric test instead if one or more of their samples was nonnormal or there were large differences in sample standard deviations. A researcher might plan to use a nonparametric test because the dependent measure is ordinal (i.e., ranks rather than scores) or because prior research or quantitative theory suggests the assumptions of the parametric test will be violated. Davison (1999) recommended the use of nonparametric tests as a default for behavior analysts; however, when there is no reason to expect that their assumptions will be violated, one advantage to planning to use parametric tests is that power calculations are comparatively straightforward, both mathematically and in G*Power.

Within-subject designs are more powerful than comparable between-subject designs (Thompson & Campbell, 2004), but there are additional assumptions related to the covariance of repeated measures and the assumption of sphericity that must be addressed in power analysis in within-subject research. Along with potential order effects, concerns about sphericity have lead some scholars to suggest that within-subject designs ought to be avoided (Greenwald, 1976). Most behavior analysts appreciate the advantages of having each subject serve as its own control (Perone, 1999), but may not know when to anticipate violations of sphericity in their experimental designs or how to evaluate sphericity as part of their analyses. In simple (i.e., single-factor) repeated-measures designs, the assumption of sphericity is that the variances (denoted by s^2) of the difference scores are all equal. Difference scores are just the differences between scores from each pair of conditions for each subject. For example, in a within-subject design that measures response rate, R, in four conditions, there are six difference scores for each subject and the assumption of sphericity is that $s_{(R1\text{-}R2)}{}^2 = s_{(R1\text{-}R3)}{}^2 = s_{(R1\text{-}R4)}{}^2 = s_{(R2\text{-}R3)}{}^2 = s_{(R2\text{-}R4)}{}^2 = s_{(R3\text{-}R4)}{}^2$. Sphericity is not a concern when there are only two levels of the independent variable (there is only one difference score, so differences in the variances of difference scores are impossible). Violations of sphericity occur when scores from some conditions are more correlated than scores in other conditions (e.g., response rates from a baseline condition are correlated with response rates in treatment but uncorrelated with response rates in a follow-up).

Mauchly (1940) developed a test to examine repeated measures for violations of sphericity. For readers interested in learning more about sphericity, Lane's (2016) description of the assumption of sphericity and suggestions about what to do when it is violated is succinct, accessible, and also addresses more complex multifactor designs. The output of Mauchly's test is typically expressed as epsilon (ε), a measure of the degree to which sphericity is violated. Upper and lower bounds of ε are 1 (no violation) to 1/(*k*-1), where *k* is the number of measurements in the ANOVA. The sample size required to detect an effect with low ε (sphericity violated) may be higher than the sample size required to detect a smaller effect with $\varepsilon \approx 1$ (no violation of sphericity). For the purpose of a power analysis, a researcher might estimate ε based on previous research or pilot data. In some circumstances, researchers might be able to eliminate concerns about violations of sphericity through experimental control, which would mean they could power their experiment assuming $\varepsilon \approx 1$. As an alternative, a researcher might elect to be conservative and power their study assuming the largest possible violation of sphericity (and smallest value of ε).

There are broad differences between test families and more nuanced differences between specific tests within the same family. Tests from different families can be used to achieve the same goal, for example, regression, ANOVA, and *t*-tests could all be used to test whether difference between two independent samples was significant. Within the same test family, the parametric Student's *t*-test and the nonparametric Mann-Whitney *U* test both compare two independent samples, but they differ in some underlying assumptions. Detailing differences between types of tests and explaining when to use each type of test are best left to statistics textbooks, so I encourage readers seeking more information about selecting tests to consult their preferred statistics textbook.

Tolerable Rates of β and α

Power is 1- β, the long-run probability of not making a Type II error. Describing β as a factor that affects power (as statistics textbooks sometimes do) is a misnomer because they are two sides of the same coin, just as it would be uninformative to write that the number of incorrectly answered questions on an exam is a factor affecting exam score. By definition, whatever increases β decreases power and vice versa. Larger βs are associated with less statistical power. By contrast, larger αs are associated with greater statistical power. In hypothesis testing, tolerating a higher type I error rate (e.g., $\alpha = .10$ instead of $\alpha = .05$) means that the critical value for the test statistic is higher and the test is stricter overall. The hypothesis is less likely to be rejected whether it is true or not, so type II errors are less likely, β is lower and power is higher. Cohen (1992a, 1992b) noted that adopting an α of .05 is typical in psychological research and recommended that psychologists use a β of .20 (power = .80). For certain types of study designs, there are other, more principled ways of selecting error rates. For example, when the sample size of an experiment cannot be adjusted, Mudge, Baker, Edge, and Houlahan (2012) describe an approach for optimizing α to minimize the combination of Type I and Type II error rates at a critical effect size. Such techniques can mitigate some of the risks of misinterpreting results of low-power studies, but they are not applicable to sample size estimation.

Setting different error rates is a simple matter in G*Power but deciding what error rates are tolerable is highly dependent on the research question, the selected test and even the researcher's philosophy of science. Any attempt at prescriptive guidelines about tolerable error rates would be essentially useless. Nevertheless, examples of situations when power is more and less important than in your typical psychology experiment may be instructive. No one with any capacity for human compassion would accept power of 80% in criminal trials for capital offenses—the implication would be that they could tolerate putting to death 20% of the innocent people who happen to wind up on trial. By contrast, if the treatment for a fatal disease was very mild, presumably no one would object to very high rates of Type II errors (giving the treatment to healthy people) because it would minimize or eliminate Type I errors (failing to treat someone who is infected).

Meaningful Effect Sizes

An effect size is a descriptive statistic that estimates the magnitude of an effect. In research that compares means across different groups or conditions, some effect sizes (including Cohen's *d* and *f*) estimate the standardized difference between means. Others (e.g., eta-squared, η^2) estimate the proportion of variance in the dependent variable that can be explained by the different levels of an independent variable. Both types of effect size can be decomposed into the difference between population means and population variance. Larger differences between population means are easier to detect, so statistical tests have more power when differences between condition means are large. Likewise, consistent, reliable differences are easier to detect, so statistical tests have more power when population variance is low.

In psychophysics, the just noticeable difference (JND; Fechner, 1860/1912) is the smallest difference between two stimuli (e.g. brightness in lumens, volume in decibels, or pitch in Hz) that is perceptible to the subject. The objective of *a priori* power analysis is to determine the sample size required to detect a meaningful effect with the desired level of power, so in this article, the JND is the smallest effect that the researcher would consider to be practically, clinically, or theoretically significant.

Cohen (1988, 1992b) provided effect-size conventions that define small, medium, and large effects for several types of statistical tests. However, researchers investigating directly observable behavior and other similarly concrete dependent variables are advised to ignore conventions and consider the minimum absolute difference that would have impact or be meaningful in their research as the "unstandardized" JND to be used in power analysis (Thompson, 2002). For example, if five micrograms of lead per deciliter of blood is unsafe for children, medical tests of lead levels need to be able to detect that concentration of lead with high power (presumably >>>.80), even if it is only a tiny fraction of the standard deviation found in children in general and the standardized effect size is small. Conversely, if a child bites classmates on an almost daily basis, an intervention that reduces biting by 50% is probably insufficient to allow the child to return to the classroom even if the effect size is large. The case studies that follow illustrate the process of identifying an appropriate JND, converting it to a standardized effect size for computing the sample size needed to detect the JND with a desired level of power (given a specified α) in G*Power. Readers are encouraged to download G*Power and have it open as they read each case so that they can follow along with the calculations.

Case Studies

Case No. 1: Reducing Employee Absences

A large company suspects that employee absenteeism is having a significant negative impact on their bottom line, though lax record-keeping makes it difficult to know for sure. The records they do have indicate that on average, employees miss 8.0 (*SD* = 2.0) days of work per year in addition to annual leave and explained medical absences.

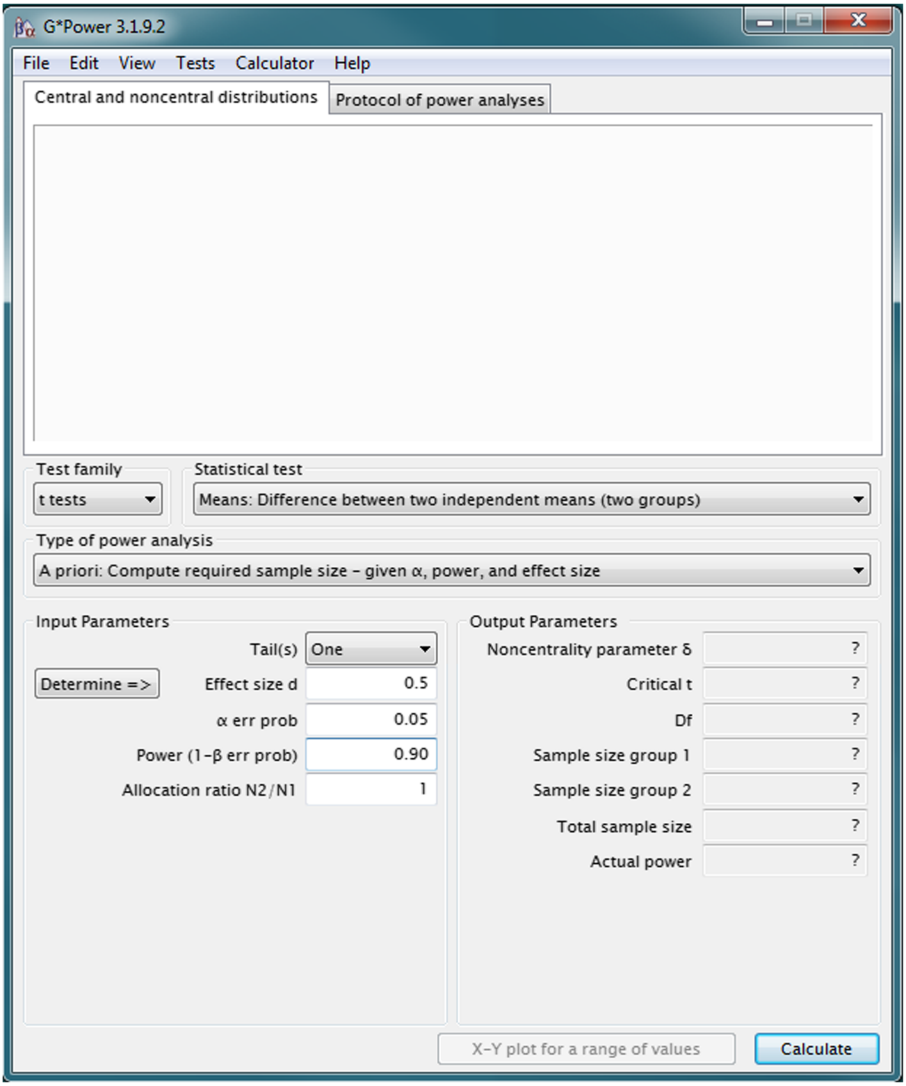
Someone has developed a strategy for reducing these unscheduled absences and the company has approved a request to run a month-long test to evaluate the efficacy and economic viability of this strategy.

The plan is to monitor absences in two groups of employees: a treatment group who will pilot the new strategy and a comparison group, with the same sample size in both groups. Based on the projected cost of implementing the new strategy throughout the company, it is estimated that the strategy needs to reduce absences by one day per employee per year to be cost neutral. If the board of directors is convinced that the strategy is likely to be effective enough to recoup the costs of implementation, they are prepared to adopt the new strategy.

Power analysis can determine how many employees to monitor during the test. The question of whether the strategy was effective *enough* in the pilot to be worth implementing throughout the company can be evaluated with an independent-samples *t* test. The test is one-tailed, because the company is specifically interested in reducing absences. The action taken if the treatment group has more absences than the comparison group would be the same as if both groups missed the same number of days: the company would not implement the strategy. To estimate the number of employees needed for the test, first determine the JND, convert it to a standardized effect size (Cohen's *d*), and compute an a priori power analysis in G*Power.

1. **Determine the JND.** The strategy needs to reduce absences by one day per employee per year to break even and the company will not adopt the strategy if it does not break even, so the JND is 1.0 day per year.
2. **Convert the JND to a standardized effect size.** The standardized effect size that G*Power uses to evaluate differences between two independent means is Cohen's *d.* Cohen's *d* expresses the difference between two group means in standard deviations. For example, a Cohen's *d* of 2.0 indicates that the two means are precisely two standard deviations apart. The standard deviations for the employee absence data that has not yet been collected are unknown, but they can be estimated using the simplifying assumption that they will be similar to the standard deviation of annual unscheduled absences in existing employee records. Those records indicate the distribution of unscheduled absences has a mean of 8 and a standard deviation of 2. The standardized effect size for the JND is Cohen's *d* = (break-even reduction in absences)/(standard deviation of absences) = (1 day per year)/(2 days per year) = 0.5. It does not matter that the pilot runs for one month and the effect size was estimated based on annual absences because the standardized effect is the same regardless of the unit of analysis.
3. **Compute an a priori power analysis in G*Power.** Figure 1 shows the settings to select in G*Power to run this analysis. The test family, *t tests*, must be selected first. The statistical test is *Means: Difference between two independent means (two groups)*. The type of power analysis is a priori because the objective is to determine the number of subjects needed to detect the JND with the desired power.

Figure 1 also shows the input parameters for this analysis. This is a one-tailed test with an effect size of 0.5. Setting the Type I error rate, α, equal to .05 is conventional in behavioral and social sciences rather than objectively correct. Using an α of .05 means


G*Power 3.1.9.2
File
Edit
View
Tests
Calculator
Help
Central and noncentral distributions
Protocol of power analyses
Determine =>
X-Y plot for a range of values
Calculate

Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s): One

Effect size d: 0.5

α err prob: 0.05

Power (1–β err prob): 0.90

Allocation ratio N2/N1: 1

Output Parameters

Noncentrality parameter δ: ?

Critical t: ?

Df: ?

Sample size group 1: ?

Sample size group 2: ?

Total sample size: ?

Actual power: ?

Fig. 1 Settings and input parameters for Case #1 (comparing means of two independent samples)

that if the true effect size is zero (the new strategy has absolutely no effect) and the samples are drawn randomly and representative of their respective populations, the probability that the sample means will be significantly different is .05. Setting power to .90 means that the Type II error rate (β) is equal to .10: if the true effect size is exactly equal to the JND and the samples are drawn randomly and representative of their respective populations, the probability that the sample means will be significantly different is .90. Using a higher value for power would decrease the chances that your pilot test fails to detect the "true" reduction in absences (if any) by increasing the sample size. An allocation ratio of 1 means that both groups will have the same sample size.

Clicking the *Calculate* button reveals that the test requires a total sample size of 140, 70 employees each in the treatment and comparison groups. With this sample size, the independent-samples *t* test has 90.3% power to detect the JND. Power analysis does not guarantee that if the strategy is effective, the statistical test will be significant. It also does not mean that if the result is statistically significant, the strategy is guaranteed to be cost-effective (Button et al., 2013). In this situation, power analyses are an assurance that the results of the inferential test provide useful information about the strength of evidence for the efficacy of the strategy and the appropriate course of action in light of that evidence.

This particular power analysis estimates the number of employees to monitor. It does not provide any insight about the optimal duration of the pilot test. On would assume that the duration of the test should be as long as it needs to be to ensure that results are reliable, but no longer. A pilot test that is too short might not yield a reliable estimate of the rate of absences, but beyond a certain point there are diminishing returns in continuing. The optimal duration is the variation in temporal distributions of absences. If the number of absences per day is relatively stable from one month to the next, it is advisable to run a shorter test than if absences varied dramatically. The JND in this example was estimated from annual records, but the planned test is only one month long, so the implicit assumption is that effect size (i.e., the difference between the treatment and comparison groups in standard deviations) will be comparable.

Case No. 2: Probability Discounting

Holt, Green, and Myerson (2003) assessed probability discounting in college students and found that college students who gambled discounted probabilistic rewards less steeply than nongamblers (they were more likely to choose the risky option). One way to replicate and extend these results with older adults is to measure probability discounting in older adults, in particular comparing discounting in senior citizens who belong to a local casino’s loyalty program with an equal number of age-matched participants who reported that they do not gamble. Many of the students in Holt et al.’s sample of gamblers reported only moderate rates of gambling. Problem gambling occurs at higher rates in older adults (Ladd, Molina, Kerins, & Petry, 2003), so it would be reasonable to aim to detect group differences that are *larger* than those reported by Holt et al. Determining the JND, converting it to Cohen’s *d*, and computing an *a priori* power analysis in G*Power will estimate the number of senior citizens to recruit for this replication and extension.

Determine the JND Holt et al. (2003) reported the area under the curve (AUC) for each participant for probabilistic amounts of $1,000 and $5,000. Area under the curve is a unitless measure that can take any value between 0 and 1. Smaller values indicate steeper discounting and arguably greater risk-aversion. For $1,000, the mean AUC was .23 (*SD* = .21) for gamblers and .10 (*SD* = .07) for nongamblers. For $5,000, the mean AUC was .17 (*SD* = .21) for gamblers and .09 (*SD* = .12) for nongamblers. The difference in college students’ probability discounting AUCs was larger for $1,000 than for $5,000. To detect larger differences than those observed in college students, one might select the largest difference in AUC that Holt et al. obtained as the JND. The larger difference was for the $1,000 reward giving a JND of .23 - .10 = .13.

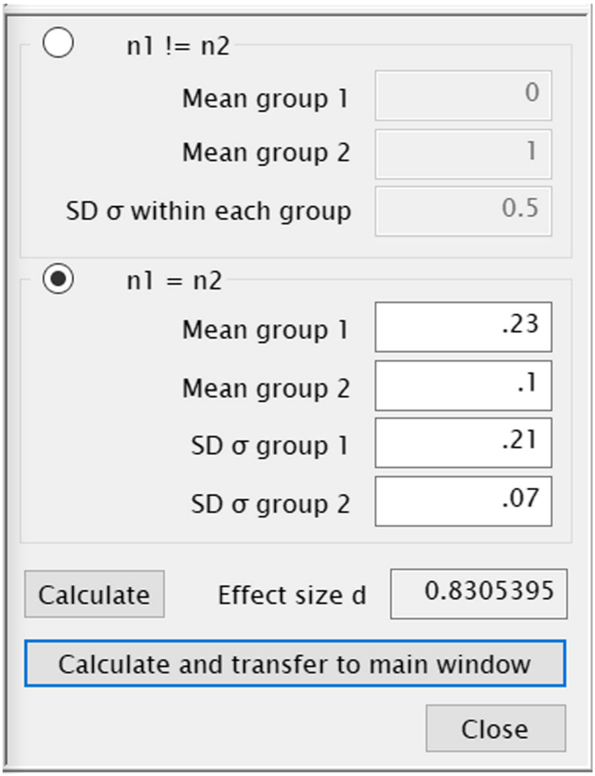
Convert the JND to a standardized effect size using G*Power Assume the standard deviations in the senior citizen samples will be similar to those that Holt et al. (2003) obtained for college students. The formula for estimating Cohen’s *d* based on two independent samples is $(M_1 - M_2)/s_{pooled}$, where M_1 and M_2 are sample means and s_{pooled} is the pooled standard deviation.[1] First, select settings for the type of power analysis you will run (in this case, test family, statistical test, and type of power analysis should be set to *t tests*, *Means: Difference between two independent means (two groups)*, and *a priori,* respectively). Next, clicking the "*Determine* =>" button under input parameters opens an effect-size calculator in G*Power that will compute Cohen's *d* based on sample means and standard deviations. Figure 2 shows the input parameters for this example (means and standard deviations from each sample). The standardized effect size for the JND is $d = 0.83$.

Compute an a priori power analysis in G*Power The settings and many of the input parameters for this example are the same as shown in Figure 1. This is a one-tailed test because the hypothesis is directional: gamblers will discount probabilistic rewards less steeply than nongamblers. The standardized effect size for the JND determined by G*Power's effect-size calculator based on means and standard deviations from Holt et al. (2003) is 0.83, a large effect according to Cohen's (1988, 1992b) conventions. Setting α and power to .05 and .80 is conventional in behavioral and social sciences. The allocation ratio of 1 means that both groups will have the same sample size.

Clicking the *Calculate* button reveals that this test requires a total sample size of 38, 19 gamblers and 19 nongamblers. With this sample size, the independent-samples *t* test has 80.7% power to detect the JND of 0.83 of a standard deviation. Although they did not mention power analysis explicitly, Holt et al. (2003) happened to include 19 participants in each group. Other research using a similar between-groups design to address similar research questions (e.g., Madden, Petry, & Johnson, 2009; Weller, Cook, Avsar, & Cox, 2008) has used the same sample size. For example, Madden et al. (2009) examined discounting in treatment-seeking pathological gamblers and demographically matched controls who did not gamble. They might have reasonably assumed that if differences between gamblers and nongamblers were detectable in Holt et al.'s sample of 38 college students, differences between *pathological* gamblers and nongamblers would be as large or larger, therefore the same sample size would be adequately powered for the desired comparison.

To evaluate whether there were differences between gamblers' and nongamblers' rates of discounting, Holt et al. (2003) compared AUCs using a parametric test, but they also reported Mann-Whitney U-tests to compare other dependent variables. The Mann-Whitney U test is a nonparametric equivalent to an independent-samples *t* test. It compares the ranks of scores rather than the scores directly and does not assume that the distributions take any particular shape (though it does assume that observations are independent and that the variances of the populations sampled are equal), so it can be used to compare ordinal or other nonnormal data. Computing the sample size needed to detect a JND of $d = 0.83$ with 80% power for a Mann-Whitney U test in G*Power requires some additional details about the shape of the parent distribution (under Input

[1] For two groups with standard deviations SD_1 and SD_2, and sample sizes N_1 and N_2, the pooled standard deviation is $\sqrt{\frac{(N_1-1)SD_1{}^2+(N_2-1)SD_2{}^2}{N_1+N_2-2}\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}$, which simplifies to $\sqrt{\frac{SD_1{}^2+SD_2{}^2}{2}}$ if sample sizes are equal.



○ n1 != n2

Mean group 1: 0

Mean group 2: 1

SD σ within each group: 0.5

● n1 = n2

Mean group 1: .23

Mean group 2: .1

SD σ group 1: .21

SD σ group 2: .07

Effect size d: 0.8305395

Fig. 2 Independent-samples *t* effect size calculator in G*Power with values set for Case #2, probability discounting

Parameters) and how the test will be calculated (under Options) that are beyond the scope of this tutorial.

Case No. 3: Increasing Physical Activity

Hayes and Van Camp (2015) increased children's physical activity during school recess with an intervention that involved reinforcement, self-monitoring, goal setting, and feedback. One way to extend their research might be to devise an alternative intervention and determine whether the alternative intervention is also effective at increasing physical activity. Like Hayes and Van Camp's experiment, the follow up might involve recording number of steps taken by elementary schoolchildren during 20-minute recess periods using accelerometers and evaluating the efficacy of the intervention using a withdrawal design. Data analysis for this experiment is likely to involve visual inspection of single-subject graphs showing number of steps as a function of session rather than a dependent-samples *t* test. Nonetheless, a power analysis for dependent means can be used to determine a sample size that is neither under- nor overpowered. To estimate the number of children needed as participants, determine the JND, convert it to Cohen's *d*, and compute an a priori power analysis in G*Power.

Determine the JND Hayes and Van Camp (2015) reported that their intervention successfully increased steps during recess by 47%, or M = 630 steps per student. If the objective were to determine whether the alternative intervention is at least as effective as Hayes and Van Camp's intervention, the unstandardized JND would be 630 steps. However, the objective is to determine whether the intervention is effective, without comparison to prior results. The results of the experiment might indicate that implementing the intervention is worthwhile even if the effect is smaller than the effect Hayes and Van Camp reported. One might reason that 249 additional steps during a single 20-minute recess period would have a negligible effect on distance traveled or minutes of moderate-to-vigorous physical activity, so a difference must be at least 250 steps to be considered meaningful. Of course, this decision is arbitrary because the difference between 249 steps and 250 steps is miniscule. The specific boundary of the JND is less important than whether the intended audience (e.g., an institutional review board, thesis committee, grant review panel, journal editors, the researchers themselves) is convinced by the rationale.

Convert the JND to a standardized effect size Assume the standard deviations will be similar to those obtained in previous research. Hayes and Van Camp (2015) did not report standard deviations, but Table 1 shows mean steps taken during baseline and the intervention for individual subjects from that study (C. M. Van Camp, personal communication, September 22, 2017). Cohen's d for dependent or related samples can be calculated either by dividing the mean of the difference scores for each subject by their standard deviation, or from means and standard deviations for each condition. Difference scores are $X_{Tx} - X_{Bl}$, mean steps taken by a subject in the intervention and baseline phases, respectively. From Table 1, the mean difference score for Hayes and Van Camp's six subjects was 630.17 steps (SD = 214.61 steps), giving an effect size d = 2.94. Replacing the obtained mean difference with the unstandardized JND gives a standardized JND of $d = 250/214.61 = 1.16$.

The effect-size calculator in G*Power can calculate Cohen's d either "from differences" as above or "from group parameters." Calculating d from group parameters requires the correlation between baseline and intervention scores (r = .071) in addition to means and standard deviations of both samples because correlation between baseline and intervention scores is an additional source of variance that must be accounted for in the power analysis. Select settings (in this case, test family, statistical test, and type of power analysis should be set to *t tests*, *Means: Difference between two dependent means (matched pairs)*, and *a priori,* respectively). Next, click the "*Determine =>*" button under input parameters to open the effect-size calculator in G*Power. Figure 3 shows the input parameters for this example. The mean for group 2 is the baseline plus the unstandardized JND, 1326 + 250 = 1576. Any pair of means that differ by 250 steps will produce the same effect size. Consistent with the calculation from difference scores, the standardized effect size for the JND is d = 1.16.

Compute an a priori power analysis in G*Power Figure 3 shows the settings and input parameters to select in G*Power for this analysis. This is a one-tailed test because it has a directional hypothesis (the intervention will increase steps). Setting α and power to .05 and .80 is conventional in behavioral and social

Table 1 Step counts for individual subjects from Hayes and Van Camp (2015)

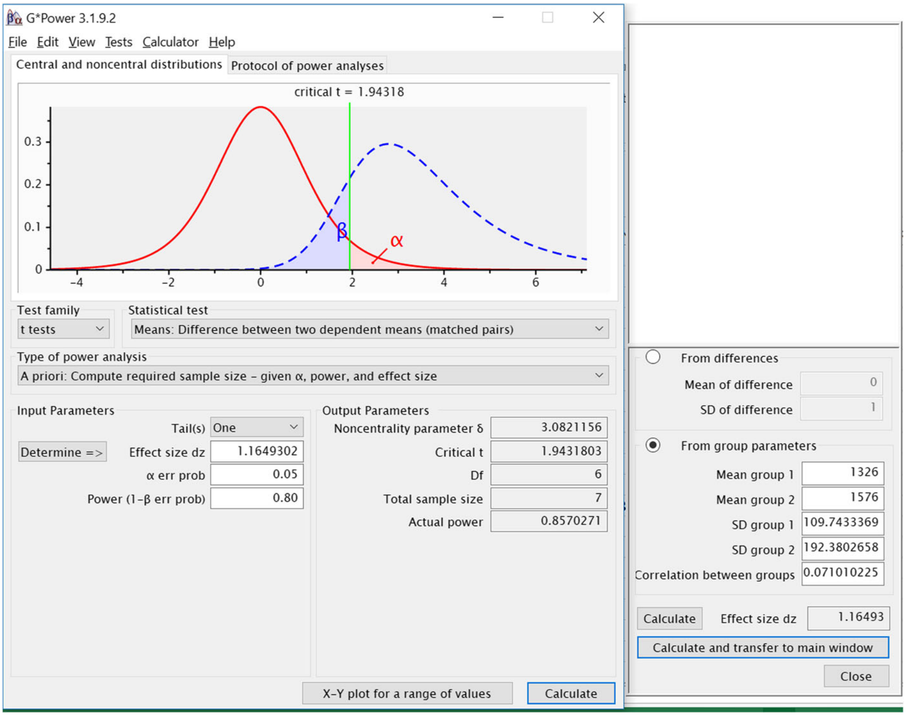
Subject	Steps		Difference Score
	Baseline	Intervention	
Ellen	1309	1640	331
Summer	1423	1960	537
Laura	1438	2085	647
Kate	1184	2177	993
Fallon	1210	1840	630
Sara	1392	2035	643
Mean	1326	1956.17	630.17
SD	109.74	192.38	214.61

sciences. There is no allocation ratio for this test, because it is within-subject: each subject will experience both baseline and intervention. Clicking the *Calculate* button reveals that for the specified input parameters, the dependent-samples test requires a total sample size of seven. With this sample size, a comparison of steps taken in baseline versus the intervention has 85.7% power to detect the JND of 1.16 standard deviations.

Case No. 4: Stimulus Discrimination

In three-key operant conditioning chambers, Kyonka, Rice, and Ward (2017) trained pigeons in a discrimination task that shares some features with slot machines. They compared several characteristics of responding across four trial types. Replicating results in chambers with different equipment (e.g., touch screens) might be a first step to conducting related follow-up experiments. An a priori power analysis in G*Power can be used to estimate the number of pigeons needed to confirm, with 80% power, that differences in responding to each trial type correspond to those reported in Kyonka et al.'s Table 1 (p. 35).

Determine the JND Partial eta squared ($\eta_p{}^2$) is the proportion of total variance in a dependent variable that is associated with the different treatments of the independent variable, with effects of other independent variables and interactions partialed out. It is calculated as $SS_{treatment}/(SS_{treatment} + SS_{error})$, where $SS_{treatment}$ is the sum of squared deviations of the treatment mean from the grand (overall) mean for each observation and SS_{error} is the sum of squared deviations of each observation from its treatment mean. In this experiment, the four different trial types are the four treatments. For the main effect of trial type on response proportion, conditional response rate, sample-phase response time, and "collect-phase" response latency, Kyonka et al. (2017) reported $\eta_p{}^2$s of .81, .66, .58 and .64. Trial type had the smallest effect on sample-phase response time, but violations of sphericity were observed for the effects on proportion and conditional response rate.


G*Power 3.1.9.2
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical t = 1.94318
0.3
0.2
0.1
0
β
α
-4
-2
0
2
4
6
Determine =>
X-Y plot for a range of values
Calculate
Calculate
Calculate and transfer to main window
Close

Test family: t tests

Statistical test: Means: Difference between two dependent means (matched pairs)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

- Tail(s): One
- Effect size dz: 1.1649302
- α err prob: 0.05
- Power (1-β err prob): 0.80

Output Parameters

- Noncentrality parameter δ: 3.0821156
- Critical t: 1.9431803
- Df: 6
- Total sample size: 7
- Actual power: 0.8570271

○ From differences

- Mean of difference: 0
- SD of difference: 1

● From group parameters

- Mean group 1: 1326
- Mean group 2: 1576
- SD group 1: 109.7433369
- SD group 2: 192.3802658
- Correlation between groups: 0.071010225

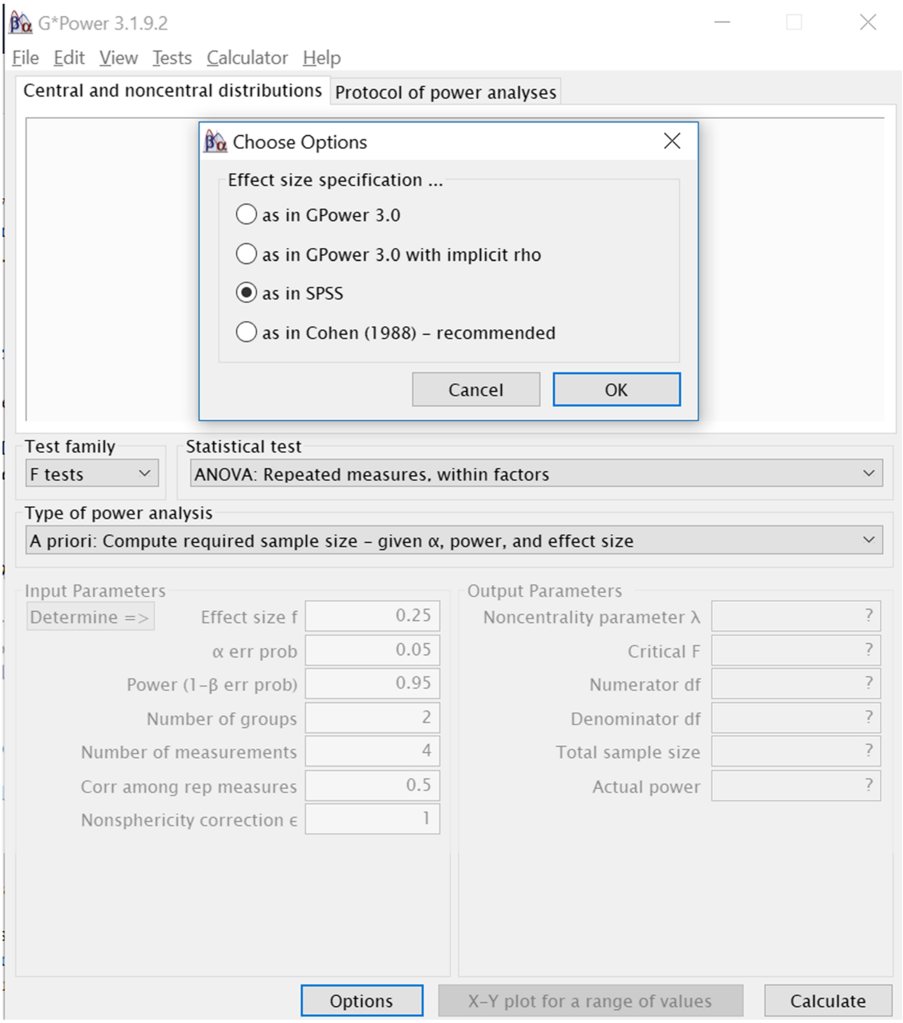
Effect size dz: 1.16493

Fig. 3 G*Power settings, inputs and outputs for Case #3 (comparing means of two dependent samples)

Sphericity is an assumption of repeated-measures ANOVA. Epsilon is a measure of the degree to which sphericity is violated, with upper and lower bounds of 1 (no violation) to 1/(*k*-1), where *k* is the number of measurements in the ANOVA. For an experiment with four trial types, the smallest ε that G*Power will accept is 0.34. Violations of sphericity do not affect the calculation of η_p^2 but can increase the sample size needed to achieve a certain power. To estimate the sample size needed to replicate an experiment with multiple dependent variables, a researcher might identify one critical dependent variable and power the experiment to detect that particular effect, or they might conduct separate analyses for multiple dependent variables and use the largest sample size indicated. This example estimates the number of pigeons needed to replicate the effect of conditional response rate, a measure of the relative conditioned reinforcing value of the stimuli presented (Kyonka et al., 2017).

Convert the JND to Cohen's *f* Partial eta squared is a standardized effect already, but G*Power uses Cohen's *f* in power analysis for repeated measures. Cohen's *f* is the standard deviation of the treatment means divided by their common standard deviation. It can be derived from partial eta squared as the square root of $[\eta_p^2/(1- \eta_p^2)]$ with the effect-size calculator in G*Power or any other calculator. The effect-size conditional response rate converts from η_p^2 = .66 to $f = \sqrt{1.94} = 1.39$.

Compute an a priori power analysis in G*Power Figure 4 shows the initial settings and options to select in G*Power. The statistical test that compares a dependent variable in four different trial types is a repeated-measures ANOVA. Select the test


G*Power 3.1.9.2
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
Cancel
OK
Determine =>
Options
X–Y plot for a range of values
Calculate

Choose Options

Effect size specification ...

○ as in GPower 3.0

○ as in GPower 3.0 with implicit rho

● as in SPSS

○ as in Cohen (1988) – recommended

Test family: F tests

Statistical test: ANOVA: Repeated measures, within factors

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size f	0.25
α err prob	0.05
Power (1–β err prob)	0.95
Number of groups	2
Number of measurements	4
Corr among rep measures	0.5
Nonsphericity correction ϵ	1

Output Parameters

Noncentrality parameter λ	?
Critical F	?
Numerator df	?
Denominator df	?
Total sample size	?
Actual power	?

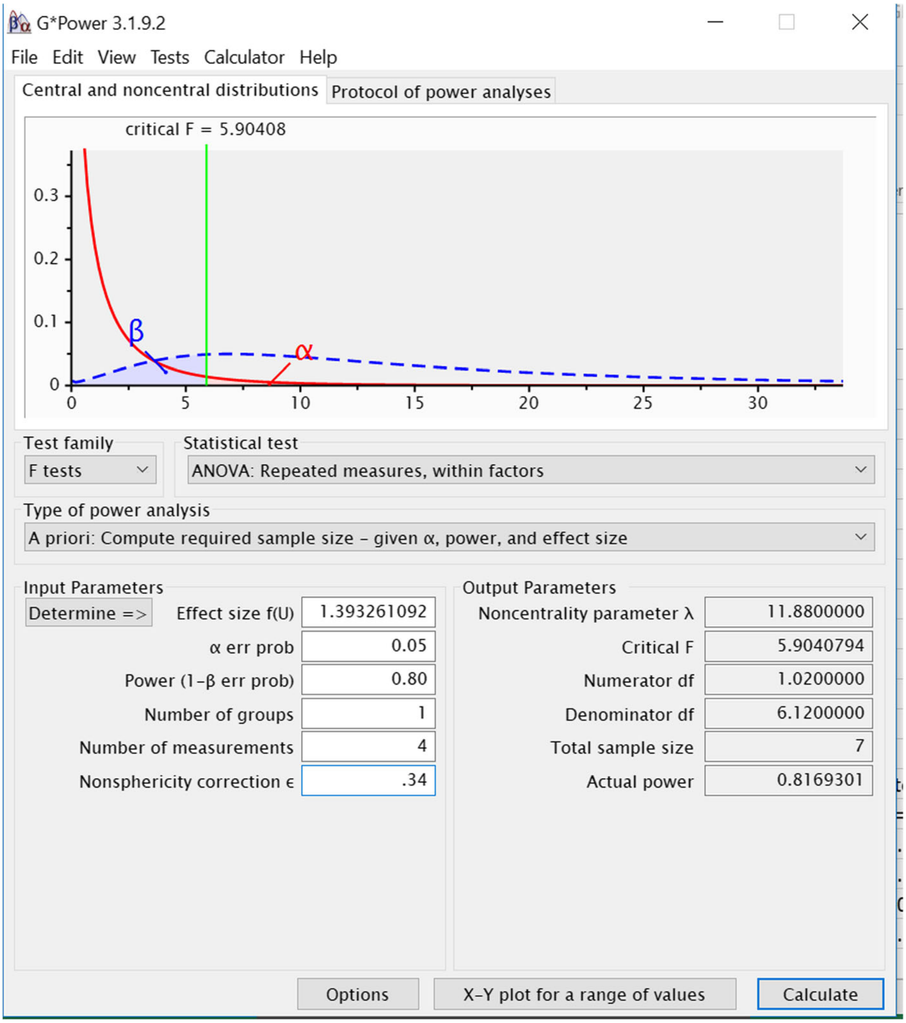
Fig. 4 Settings and options for Case #4 (comparing means from four within-subject conditions)

family, *F tests*, first. The statistical test is *ANOVA: Repeated measures, within factors* (rather than between or interaction). The type of power analysis is *a priori*.

G*Power 3 provides several different options for calculating the effect size of repeated measures. To change the effect-size calculator, click the *Options* button, select the radio button for the desired effect-size specification, and click *OK*. The option “as in SPSS” assumes that the effect size is calculated from $SS_{treatment}/(SS_{treatment} + SS_{error})$, so it is the appropriate option for this analysis regardless of the program used to calculate $\eta_p{}^2$. Using this option, G*Power can calculate Cohen’s f directly from $\eta_p{}^2$, or from treatment and error variances, sample sizes, and number of repeated measures in the within-subject factor, information that generally can be found in ANOVA source tables.

Figure 5 shows the input parameters for this power analysis. Setting α and power to .05 and .80 is conventional in behavioral and social sciences. This experimental design involves one group of subjects (i.e., there are no between-groups factors) and four measurements. The last input parameter is the nonsphericity correction factor, ε. Figure 5 also shows the output parameters for this power analysis. Clicking the *Calculate* button yields an estimated sample size of seven. By using seven pigeons in the experiment, the ANOVA has 81.7% power to detect a JND of *f* = 1.39 with the maximum possible violation of sphericity, ε = .34. Repeating the power analysis for different values of ε produces sample size estimates between four and seven.

Author Note The author thanks Don Hantula for his input on the utility of a tutorial on this topic, Carole Van Camp for providing additional data from a previously published experiment, Jarrett Byrnes for drawing attention to the Mudge et al. (2012) article on


G*Power 3.1.9.2
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical F = 5.90408
0.3
0.2
0.1
0
0
5
10
15
20
25
30
β
α
Determine =>
Options
X-Y plot for a range of values
Calculate

Test family: F tests

Statistical test: ANOVA: Repeated measures, within factors

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters
Effect size f(U) 1.393261092
α err prob 0.05
Power (1-β err prob) 0.80
Number of groups 1
Number of measurements 4
Nonsphericity correction ϵ .34

Output Parameters
Noncentrality parameter λ 11.8800000
Critical F 5.9040794
Numerator df 1.0200000
Denominator df 6.1200000
Total sample size 7
Actual power 0.8169301

Fig. 5 Settings, input and output parameters for Case #4 (comparing means from four within-subject conditions)

optimizing error rates, and Regina Carroll for valuable feedback on a portion of the manuscript.

References

Association for Behavior Analysis International Accreditation Board. (2017). *Accreditation handbook.* Portage, MI: Author.

Behavior Analyst Certification Board. (2017). *BCBA/BCaBA task list* (5th ed.). Littleton, CO: Author.

Branch, M. (2014). Malignant side effects of null-hypothesis significance testing. *Theory & Psychology, 24*, 256–277. https://doi.org/10.1177/0959354314525282.

Branch, M. N. (1999). Statistical inference in behavior analysis: some things significance testing does and does not do. *Behavior Analyst, 22*, 87–92.

Button, K. S., Ioannidis, J., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365–376.

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

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

Cohen, J. (1992a). Statistical power analysis. *Current Directions in Psychological Science, 1*, 98–101.

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

Cohen, L. L., Feinstein, A., Masuda, A., & Vowles, K. E. (2014). Single-case research design in pediatric psychology: considerations regarding data analysis. *Journal of Pediatric Psychology, 39*, 124–137.

Davison, M. (1999). Statistical inference in behavior analysis: having my cake and eating it? *Behavior Analyst, 22*, 99–103.

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.

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.

Fechner, G. T. (1912). Elements of psychophysics (H. S. Langfeld, Trans.). In B. Rand (Ed.), *The classical psychologists* (pp. 562–572). Retrieved from http://psychclassics.yorku.ca/Fechner/ (Original work published 1860).

Fisher, W. W., & Lerman, D. C. (2014). It has been said that, "There are three degrees of falsehoods: lies, damn lies, and statistics.". *Journal of School Psychology, 52*, 243–248.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33*, 587–606.

Greenwald, A. G. (1976). Within-subjects designs: to use or not to use? *Psychological Bulletin, 83*(2), 314–320.

Haig, B. D. (2017). Tests of statistical significance made sound. *Educational & Psychological Measurement, 77*, 489–506.

Hayes, L. B., & Van Camp, C. M. (2015). Increasing physical activity of children during school recess. *Journal of Applied Behavior Analysis, 48*, 690–695.

Holt, D. D., Green, L., & Myerson, J. (2003). Is discounting impulsive? Evidence from temporal and probability discounting in gambling and non-gambling college students. *Behavioural Processes, 64*, 355–367.

Kyonka, E. G., Rice, N., & Ward, A. A. (2017). Categorical discrimination of sequential stimuli: all SΔ are not created equal. *Psychological Record, 67*, 27–41.

Ladd, G. T., Molina, C. A., Kerins, G. J., & Petry, N. M. (2003). Gambling participation and problems among older adults. *Journal of Geriatric Psychiatry & Neurology, 16*, 172–177.

Lane, D. (2016). The assumption of sphericity in repeated-measures designs: what it means and what to do when it is violated. *Quantitative Methods for Psychology, 12*, 114–122.

Madden, G. J., Petry, N. M., & Johnson, P. S. (2009). Pathological gamblers discount probabilistic rewards less steeply than matched controls. *Experimental & Clinical Psychopharmacology, 17*, 283–290.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *Annals of Mathematical Statistics, 11*, 204–209.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *British Journal for the Philosophy of Science, 57*, 323–357.

Mayo, D. G., & Spanos, A. (2011). Error statistics. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Handbook of philosophy of science*, *Philosophy of statistics* (Vol. 7, pp. 153–198). Amsterdam, Netherlands: Elsevier.

Michael, J. (1974). Statistical inference for individual organism research: mixed blessing or curse? *Journal of Applied Behavior Analysis, 7*, 647–653. https://doi.org/10.1901/jaba.1974.7-647.

Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an optimal α that minimizes errors in null hypothesis significance tests. *PLoS ONE, 7*(2), e32734. https://doi.org/10.1371/journal.pone.0032734.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 20A*, 175–240 263–294.

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231*(694–706), 289–337.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531–536.

Perone, M. (1999). Statistical inference in behavior analysis: experimental control is better. *Behavior Analyst, 22*, 109–116.

Peterson, C. (2009). Minimally sufficient research. *Perspectives on Psychological Science, 4*, 7–9.

Sidman, M. (1960). *Tactics of scientific research: evaluating experimental data in psychology*. New York, NY: Basic Books.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). Life after p-hacking. Meeting of the Society for Personality and Social Psychology, New Orleans, LA, January 17–19, 2013. Available at SSRN: http://ssrn.com/abstract=2205186 or doi:https://doi.org/10.2139/ssrn.2205186.

Thompson, B. (2002). "Statistical," "practical," and "clinical": how many kinds of significance do counselors need to consider? *Journal of Counseling & Development, 80*, 64–71. https://doi.org/10.1002/j.1556-6678.2002.tb00167.x.

Thompson, V. A., & Campbell, J. I. (2004). A power struggle: between-vs. within-subjects designs in deductive reasoning research. *Psychologia, 47*, 277–296.

Trafimow, D., & Marks, M. (2015). Publishing models and article dates explained. *Basic & Applied Social Psychology, 37*, 1.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *American Statistician, 70*, 129–133. https://doi.org/10.1080/00031305.2016.1154108.

Weller, R. E., Cook, E. W., Avsar, K. B., & Cox, J. E. (2008). Obese women show greater delay discounting than healthy-weight women. *Appetite, 51*, 563–569.

Wilkinson, L., & The Task Force on Statistical Inference, American Psychological Association, Science Directorate. (1999). Statistical methods in psychology journals: guidelines and explanations. *American Psychologist, 54*, 594–604.

Zimmermann, Z. J., Watkins, E. E., & Poling, A. (2015). JEAB research over time: species used, experimental designs, statistical analyses, and sex of subjects. *Behavior Analyst, 38*, 203–218.