Understanding statistical power in the context of applied research

Dr Thom Baguley

Department of Human Sciences, Loughborough University

Email: t.s.baguley@lboro.ac.uk

Fax: 01509 223940

Abstract

Estimates of statistical power are widely used in applied research for purposes such as sample size calculations. This paper reviews the benefits of power and sample size estimation and considers several problems with the use of power calculations in applied research that result from misunderstandings or misapplications of statistical power. These problems include the use of retrospective power calculations and standardized measures of effect size. Methods of increasing the power of proposed research that do not involve merely increasing sample size (such as reduction in measurement error, increasing 'dose' of the independent variable and optimizing the design) are noted. It is concluded that applied researchers should consider a broader range of factors (other than sample size) that influence statistical power, and that the use of standardized measures of effect size should be avoided (except as intermediate stages in prospective power or sample size calculations).

Keywords: statistical power, applied research, experimental design

Statistical power in the context of applied research

1. Introduction

This paper aims to promote understanding of statistical power in the context of applied research. It outlines the conceptual basis of statistical power analyses, the case for statistical power and potential problems or misunderstandings in or arising from its application in applied research. These issues are important because recent years have seen an increase in the advocacy and application of statistical power, but not necessarily in the propagation of good understanding or good practice.

The statistical power of a null hypothesis test is the probability of that test reporting a statistically significant effect for a real effect of a given magnitude. In layman's terms, if an effect has a certain size, how likely are we to discover it? A more technical definition is that it is the probability of avoiding a Type II error.[1] An understanding of statistical power supports two main applications:

i) to estimate the *prospective* power of a study, and

ii) to estimate the parameters required to achieve a desired level of power for a proposed study.

This latter application is usually confined to sample size calculation, but in principle can also be used to estimate the effects of other parameters on sample size.

---

[1] Type I errors are statistical false positives (deciding there is an effect where there is no real effect). Type II errors are false negatives (deciding there is no effect when there is a real effect). Power = 1-β where β is the probability of a type II error.

To illustrate these applications consider the independent *t* test. Cohen (1988; 1992) has published power and sample size calculation methods for this and other common situations.[2] According to Cohen, statistical power for the independent *t* test is a function of three other parameters: sample size per group (*N*), significance criterion (α), and standardized population effect size (Cohen's **d**). The standardized population effect size will be discussed in more detail in section 3.2. Increasing any of these three parameters will increase the power of the study. Large effects are easier to detect than small ones. Large samples have smaller standard errors than small ones (making the sampling distribution of the means more tightly clustered around the population mean). Larger α levels mean that the threshold for declaring a difference significant is reduced (i.e., a smaller observed difference in the means will be considered significant). A more detailed account of how these parameters influence power can be found in Howell (2002). A worked example of a power calculation for independent *t* using Cohen's method is found in the Appendix. Table 1 illustrates how power of the independent *t* test is influenced by these three parameters for an arbitrary selection of values. While the details of the calculations differ, similar relationships exist for other forms of statistical inference.[3]

INSERT TABLE 1 ABOUT HERE

---

[2] Cohen's method is widely taught and widely used in fields such as Psychology, Medicine and Ergonomics. Independent *t* is chosen because the calculations are relatively simple and is therefore appropriate to illustrate the key concepts. Power calculations for more complex designs are also often done by reducing the main hypothesis or hypotheses to analogous tests of differences between means (e.g., linear contrasts).

[3] As an example, consider the relationship between independent *t* and Pearson's *r*. Calculating independent *t* is equivalent to calculating *r* between the group (dummy coding group membership as a 0 or 1) and the dependent variable (*r* itself is also a common measure of effect size for power calculations).

2. A case for statistical power

Many researchers now routinely use and report power or sample size estimates in proposals and published research. While there are good reasons to do so in many cases, there are also reasons to be cautious about the application of statistical power – particularly in applied research. Before considering some of the problems with statistical power, this article will first consider some of the advantages.

2.1 Avoiding low power. Many studies lack sufficient power to have a high probability of detecting the effects they are interested in. Understanding the role of sample size and other factors in contributing to statistical power allows researchers to design studies that have a satisfactory probability of detecting the effects of interest to the researchers.

2.2 Avoiding excessive power. In most applied research domains an over-powered study may be just as undesirable as an under-powered study. Increasing sample size, in particular, almost always has financial and other costs associated with it. Where the study involves exposing people to risk or discomfort the researcher has an ethical dimension to consider.[4] In such situations, both excessively low power and high power should be avoided. Participants should not be exposed to risk or discomfort such as extreme temperatures, noise or vibration if the study is unlikely to provide scientific or other benefits. Similarly, researchers should not recruit sample sizes greater than necessary for a reasonable level of power. As can be seen in Table 1, the relationship between sample size and power is not linear; larger samples provide diminishing returns in terms of increased power. For example, with $\alpha = 0.10$ and $\mathbf{d} = 0.8$

[4] Researchers may also have professional and legal restrictions in relation to studies of this type. It is increasingly common for those that evaluate research proposals (e.g., ethics committees, government agencies) to require sample size calculations for precisely the reasons outlined here. A full treatment of the ethical issues facing practitioners in relation to the power of a study is beyond the scope of this article, but would take into account the difficulties of conducting ethical, useful research while also satisfying the requirements of a client.

doubling sample size from 40 to 80 per group only increases power a couple of percentage points.[5]

2.3 Planning ahead. The technical and non-technical facets of power analysis require that researchers think carefully about their research prior to collecting data. A good researcher needs to define the main hypothesis or hypotheses that they wish to test, reflect on any ethical issues raised by the proposed experimental design and consider the costs of both Type I and Type II errors (e.g., Wickens, 1998). Careful consideration of these factors is an important aspect of research. It seems reasonable to argue that during statistical power estimation or sample size calculation is a suitable stage in the research process to review them. Furthermore, at least two aspects of a study must be considered in detail before a power calculation is possible.

The first aspect is the specific hypothesis to be tested. Although many studies have multiple hypotheses it is necessary to specify at least one principal hypothesis to set the sample size or estimate the power of the proposed study. The main hypotheses need to be specified in sufficient detail to determine the type of statistical procedure (and therefore the appropriate power calculations) to perform. For many studies a single power calculation for the main research hypothesis will suffice. If more than one hypothesis is essential to aims of the research then several power calculations are necessary. In any case, it is good practice to obtain a range of power estimates for plausible parameter estimates (perhaps laid out in a format similar to that of Table 1).

The second aspect is the standardized population effect size the researcher wishes to be able to detect. One common method for this is to run a pilot study (something that is also useful for other reasons). A successful pilot study (regardless of significance) should provide

---

[5] Note that the odds of a Type II error will still decrease dramatically in this situation from about 1 in 30 to about 1 in 300 (assuming **d** = 0.8), so it might still be acceptable to expose a further 40 participants to potential harm or discomfort if the consequences of a Type II error are sufficiently undesirable.

reasonable estimates of the necessary population parameters that contribute to the standardized effect size (the larger the pilot study the more accurate the estimates).

A pilot study is not always possible and an alternative route to effect size estimation is available. This route involves estimating not the *actual* effect size, but estimating the magnitude of effect required to be 'interesting' or 'important' in the context of the research (e.g., clinical importance, cost-efficiency and so forth). Such effects are often termed to have *practical significance* rather than *statistical significance*. Estimating an effect size that has practical importance in a field holds an obvious attraction in the case of applied research. For example, a researcher investigating the effect of the interior noise on a drivers ability to estimate the speed of a vehicle might decide that a difference of 6 km/h or greater will have practical importance (e.g., in terms of road safety). A statistical procedure that encourages researchers to plan ahead, specify clear hypotheses, select the statistical procedures they are likely to use and estimate the magnitude of effect that they are able to detect should have a positive impact on the conduct of research.

3. Misunderstandings and misapplication of statistical power

It should now be clear that there is a positive case for power calculations to form part of the research process. There are, however, several ways in which routine application of statistical power is problematic (in particular for ergonomics and related fields). In some cases these issues are due to straight-forward (but widespread) misapplications of statistical power and in other cases to subtle misunderstandings of aspects of the topic. This article will address these pitfalls under four main headings: retrospective power, standardized effect size, statistical power by numbers and design neglect.

3.1 Retrospective power.

The most widely reported misapplication of statistical power is in *retrospective* or *post hoc* power calculations (Hoenig and Heisey, 2001; Lenth, 2001; Zumbo and Hubley, 1998). A

retrospective power calculation attempts to determine the power of a study after data has been collected and analyzed. This practice is fundamentally flawed. The calculations are done by estimating the population effect size using the observed effect size among the sample data. Computer software such as SPSS readily perform these calculations under the guise of "observed power". Retrospective power and prospective power (estimating power or sample size as advocated in section 2) are not the same things (Zumbo and Hubley, 1998). In order to understand why retrospective power is a misapplication it is necessary to consider how retrospective power estimates are sometimes used.

A researcher might calculate retrospective power and arrive at a probability such as .85. This probability may then be interpreted as the ability of the study to detect significance (e.g., that the experiment would have detected the effect 85 times out of 100). In other words, retrospective power is frequently interpreted in much the same way as prospective power. This is particularly problematic when retrospective power calculations are used to 'enhance' the interpretation of a significance test. For example, low observed power may be used to argue that a non-significant result reflects the low sample size (rather than the absence of an important or interesting effect). Similarly, high observed power may interpreted as evidence that a significant effect is real or important. The main reason why such interpretations of retrospective power are flawed is that the observed power is a mere function of the observed effect size and hence of the observed *p* value (Hoenig and Heisey, 2001).[6] For test statistics with approximately symmetrical distributions marginal significance (where $p = \alpha$) will equate to observed power of approximately 0.5. In general, statistical significance will result in high observed power and non-significance will result in low observed power. Worse still, if the observed power for non-significant results is used as an indication of the strength of evidence for the null hypothesis, it

---

6 Note that this relationship only applies to retrospective power calculations where the other parameters that influence power (sample size and alpha) are fixed. For more complex statistical models observed power may no longer be a simple function of observed effect size – it may involve additional parameters (although, as long as these are estimated from the sample alone, even then observed power should be interpreted with considerable caution).

will erroneously suggest that the lower a *p* value (and therefore the larger an observed effect) the stronger the evidence is in favour of the null hypothesis (Frick, 1995; Hoenig and Heisey, 2001; Lenth, 2001). For significant results high observed power will act to (falsely) strengthen the conclusions that the researcher has drawn. In either case retrospective power calculations are highly undesirable.

One motivation for the use of retrospective power calculations is the desire to assess the strength of evidence for a null hypothesis – something that standard hypothesis tests are not designed for. Hoenig and Heisey (2001) and Lenth (2001) both suggest test of *equivalence* as alternatives to standard null hypothesis significance testing when researchers wish to show treatments are similar in their effects. A clear introduction to such equivalence tests can be found in Dixon (1998), and an inferential confidence interval approach that integrates equivalence with traditional null hypothesis significance tests has also been described (Tryon, 2001). The equivalence approach shares similarities with the approach proposed by Murphy and Myors (1999) which involves testing for a negligible effect size, rather than for a precisely zero effect (note, however, that the Murphy and Myors approach uses proportion of variance explained – a standardized measure of effect). In both cases the researcher needs to be able to provide a sensible estimate of the size of effect that has practical importance. Both approaches have their roots in applied work (e.g., pharmacology and applied psychology respectively).

3.2 Standardized effect sizes. Standardized effect size estimates are central to power and sample size calculations (though they are also widely used in other areas). An unstandardized effect size is simply the effect of interest expressed in terms of units chosen by the researcher. For example, a mean difference in the time to complete a task might be expressed in seconds. Unstandardized effect sizes have two obvious drawbacks:

i) that changing the units (e.g., from seconds to minutes) changes the value of the effect size, and

ii) that comparing different types of effects is not easy (e.g., how does a reduction of 23 seconds in task time compare to a reduction of 6.7% in error rate?).

The common solution to this problem in statistics is to re-express the effect size using 'standard' units (derived from an estimate of the variability of the population sampled). In the case of **d** and *r* the effect size is expressed in terms of an estimate of the population standard deviation. Cohen's **d** is the mean difference divided by the standard deviation, so a **d** of 0.5 represents a difference of one half a population standard deviation between two means. A value of 0.2 for Pearson's *r* represents an increase in the dependent variable of one fifth of a standard deviation for every s.d. increase in the predictor variable.

One well-known drawback of standardized effect sizes is that they are not particularly meaningful to non-statisticians. Ergonomists and other applied scientists typically prefer to use units which are meaningful to practitioners (Abelson, 1995). There is also a more technical objection to the use of standardized effect sizes. The standardized effect size is expressed in terms of population variability (which statisticians tend to call *error* in their models). This *error* can be thought of as having two components: the population variability itself and measurement error. Most of the time statistics can ignore the difference between these two components and treat them as a single source of "noise" in data. For standardized effect sizes measurement error is a bigger problem. Many applications of standardized effect size implicitly assume that studies with similar standardized effect sizes have effects of similar magnitudes. This need not be the case – even if exactly the same things are being measured (Lenth, 2001).

Consider two studies investigating ease of understanding of two versions of a computer manual. Assume the studies are identical in every respect except that one uses a manual stop watch to time people on sub-activities in the study, while the other uses the time stamp on a video recording. In both cases the mean difference in time to look up information in the two manuals is 23.4 seconds. Using the manual stop watch happens to introduce more measurement error than using the video time stamp (possibly because the video can be replayed to obtain more reliable times). The standard deviation of the difference is 51.2 seconds for the former study and 59.7 seconds for the latter study. The standardized effect size estimates (**d**) from the two studies are 0.39 and 0.46 respectively. In this case the difference in effect size is relatively modest (one effect is almost 20% larger than the other), but in principle comparing

effects from studies with low and high degrees of measurement error could produce extreme changes in standardized effect sizes. Schmidt and Hunter (1996) suggested that for many psychological measures error is "often in the neighbourhood of 50%" (p.200).

This is important, because the effectiveness of the computer manuals (or whatever is being compared) isn't influenced by the error in our measurements. In general the presence of measurement error leads us to underestimate the magnitude of effects if we use standardized effect sizes.[7] In the presence of measurement error, measures of explained variance such as r^2 are place are underestimates of the true proportion of variance accounted for by a variable. All other things being equal, r^2 can at best be thought of as a lower bound on the true proportion of variance explained by an effect (O'Grady, 1982). Unstandardized effect size isn't influenced by measurement error in this way. This means, for example, that if the mean number of accidents on a shift decreased from 5 to 4 after the introduction of new lighting, our best estimate of the reduction in accident rate would be 20% regardless of measurement error. As measurements become more accurate the influence of measurement error on standardized effect sizes becomes proportionately less important. Even so, it would be naïve to think that similar standardized effects based on measurements of different kinds represent magnitudes of a similar size. In extreme cases, identical standardized effect sizes could have real effects that differ by orders of magnitude.[8]

---

7 It can lead to overestimates in analyses incorporating several measurements with different reliabilities. For example, an unreliably measured covariate may lead to an overestimate of the effect of other variables in analysis of covariance.

8 It is worth noting that advocates of meta-analysis (which uses standardized effect sizes) recognize the importance of correcting standardized effect sizes for measurement error (Schmidt and Hunter, 1996). In practice, many researchers do not have appropriate estimates of the reliability of their measures available and therefore such corrections are unlikely to be performed. Work on meta-analysis also suggests that thinking of a single stable population effect size is also misleading. It is probably more realistic to think of effects being sampled from populations that have effects of differing magnitudes (Field, 2003).

3.3 Statistical power by numbers. Sample size and power calculations have begun to become routine aspects of research. As a consequence the focus of the calculations has increasingly fallen on the numbers themselves rather than the aims and context of the study. Cohen (1988) proposed values of standardized effect sizes for small, medium and large effects (**d** = 0.2, 0.5 and 0.8) which are widely used for statistical power calculations. The adoption of these "canned" effect sizes (Lenth, 2001) for calculations is far from good practice and, at best, condoned as a last resort. The use of canned effect sizes is particularly dangerous in applied research where there is no reason to believe that they correspond to the practical importance of an effect. In addition to the problems with measurement error outlined in the preceding section, the practical importance of an effect depends on factors other than the effect size. A very small effect can be very important if the effect in question is highly prevalent or easy to influence (Rosenthal and Rubin, 1979). In an industrial setting, a small reduction in the time to perform a frequent action (e.g., moving components from one location to another) can have a dramatic effect on productivity and profitability.

Many sample size calculations also aim for similar levels of power: usually about 0.80. This is undesirable because it is unlikely that this balances the risks of Type I and Type II errors appropriately for all types of research. As Type I and Type II errors have an important ethical dimension (as well as cost implications) the level of power should be considered individually for each study.[9] Power calculations are readily manipulated to meet external targets (e.g., by increasing the expected effect size until 0.80 power is achieved). This may be a consequence of the setting of naïve targets by journals or ethical advisory boards (Hoenig and Heisey, 2001), but it is also a by-product of the perceived difficulty of estimating effect sizes prior to collecting data.

---

9 It is possible to build the desired ratio of Type I and Type II errors into a power calculation. GPower (free-to-download power calculation software) terms this "compromise power" (Erdfelder et al., 1996).

3.4 Design neglect. While the use of power and sample size calculations is, in principle, an advance, there is a danger that routine use of statistical power leads to neglect of other important factors in the design of a study. The default assumption in statistical power appears to be that increased power is most easily achieved by manipulating sample size (Cohen, 1992).[10] Looking at Table 1 it should be immediately obvious that changes in α and standardized effect size are just as important in increasing or decreasing power. Increasing α is generally discouraged because it also increases the Type I error rate, yet (as previously noted) if the cost of Type I errors is low relative to Type II errors there is a strong case to increase α in return for increased power.

The effect size of a study is often assumed to be beyond the control of the researcher. A moment of reflection should convince one otherwise. First, the size of effect is often a function of the 'dose' of the independent variable. This is most obvious when drugs such as caffeine or alcohol are administered, but can also apply to other research situations. Consider the case of a researcher looking at the effects of heat stress on performance. In this case the 'dose' could be manipulated by the temperature of the thermal chamber in which a participant is working. Second, as noted above, the standardized effect size is also influenced by measurement error. Decreasing measurement error will increase the standardized effect size (though not the unstandardized effect) and therefore the power of the study. In some cases measurement error is out of the researcher's control (e.g., determined by the precision of equipment). In other situations the precision of the measurement is frequently over-looked, but relatively easy to influence.

---

10 This assumption has an element of truth for academic research in psychology and related disciplines where most studies rely on samples of undergraduate students, but this is rarely the case in Ergonomics. Even when a large potential pool of participants exists sample size may be hard to manipulate because participation is time-consuming, costly or because participants with appropriate characteristics (e.g., handedness or stature) are hard to find.

Consider a researcher who is looking at how the self-reported rate of minor household accidents varies with age. A common strategy would be to ask participants which age band they fall in to (e.g., "35-44", "45-54" and so forth). If the reported rate of accidents changes gradually with age (rather than being a step function that rises with age band) then testing the relationship in this way will have a higher measurement error (and therefore lower power) than using the numerical age of the participant.[11] Researchers often take continuous data and categorize it prior to analysis (e.g., using a median split). This artificially inflates measurement error, decreases standardized effect sizes and (amongst other things) reduces power (Irwin and McClelland, 2001; MacCallum et al., 2002). It is important to collect data and analyze in a way that minimizes measurement error.

Design neglect applies not only to the parameters of the power equations, but also to other aspects of study. For example, researchers often use omnibus tests in research (such as ANOVA or Chi-squared) rather than specific tests of the hypotheses they are interested in – such as focused contrasts (Judd et al., 1995; McClelland, 1997). In analysis of variance (ANOVA) conservative *post hoc* tests such as Tukey's HSD are often chosen by default, even when the costs of Type I errors might be negligible and when alternative procedures are readily available (Howell, 2002). Power can also be increased with repeated measures designs in situations where large individual differences are expected (Allison et al., 1997). Repeated measures analyses can also be used where participants have been matched (e.g., where each participant in one condition is paired with a participant in another condition on the basis of potentially influential confounding variables such as age, body mass index or social class). A repeated measures analysis controls for the correlation between pairs of observations (whether these correlations arise from measuring the same participant several times or by measuring

---

11 There may be other reasons for using age bands such as this, such as to facilitate confidentiality or to boost questionnaire return rates. Note also that measurement error won't be decreased by introducing spurious precision: a 7 point response scale might have more precision than a 2 or 3 point one, but there are probably few judgements that people make that can be precisely differentiated on a 30 or 40 point scale.

participants who have been carefully matched). The higher this correlation the more powerful the test (Howell, 2002). A similar outcome can be achieved by using analysis of covariance (ANCOVA) to control for confounding variables (Allison et al., 1997). These latter approaches are particularly useful for applied research.

Finally, the sampling strategy of a study can be designed to maximize power. For example, a researcher investigating a linear relationship between a predictor and a response variable should sample more cases from extreme levels of the predictor than from the mid-range. The extreme values have greater leverage in the analysis and provide more information about the slope of the regression line. Sampling such cases therefore produces more accurate parameter estimates and more powerful tests (McClelland, 1997). Similar "optimal" designs are available to detect non-linear effects (such as cubic or quartic trends) or for a compromise between two patterns (e.g., to maximize power to detect either a linear or cubic trend). McClelland (1997) notes that many studies reduce their chances of detecting the effects they are interested in by distributing participants equally between values of the independent variable or by dispersing participants between more values of an independent variable than necessary. It might appear that these design principles are difficult to implement outside laboratory research. This is not the case. Staged screening can also be used to sample optimal or near-optimal proportions of participants (Allison et al., 1997; McClelland, 1997).

4. Conclusions

Estimating statistical power or required sample size prior to data collection is good practice in research. It ceases to be good practice, though, when power is calculated retrospectively (either in place of a prospective power calculation or in attempt to add to the interpretation of a significant or non-significant result). Researchers need to consider a number of issues carefully when making use of power or sample size calculations:

1) It is important to derive meaningful estimates of effect sizes of practical significance or importance. Applied researchers are often in a position to use their knowledge of

their field or their relationship with practitioners to arrive at these estimates (e.g., international standards dealing with safety, or differences in efficiency that would impact on the profitability of a process).

2) Standardised effect sizes should normally be reserved for the intermediate stages in calculations and unstandardized effect sizes preferred where meaningful units can be communicated.

3) Applied researchers should use prospective power calculations where possible to make their findings more useful and their research more efficient. Such calculations need careful planning and should be based on a clear understanding of the range of design factors that influence power. Particular emphasis should be placed on reducing measurement error, the 'dose' of the independent variable and the proportions in which participants are sampled from or allocated to conditions (Allison et al., 1997; McClelland, 1997).

If power calculations are not thought through carefully the estimates obtained from them will be unreliable and the decisions they make may be flawed, unethical or both.

Acknowledgements

Dr. Thom Baguley, Human Sciences, Loughborough University, Loughborough, LE11 3LR, United Kingdom.

The author would like to thank Bruce Weaver, Jeremy Miles, Robin Hooper and Mark Lansdale and an anonymous reviewer for their helpful comments on earlier drafts of this work and to the numerous staff and students who have (directly or indirectly) assisted me in the preparation of the paper.

Correspondence concerning this article should be addressed to Dr. Thom Baguley at the Department of Human Sciences, Loughborough University or via electronic mail to t.s.baguley@lboro.ac.uk.

References

Abelson, R. P., 1995. *Statistics as principled argument*. Erlbaum, Hillsdale, N.J.

Allison, D. B., R. L. Allison, M. S. Faith, F. Paultre, and F. X. Pi-Sunyer, 1997. Power and money: Designing statistically powerful studies while minimizing financial costs. *Psychological Methods*, 2:20-33.

Cohen, J., 1988. *Statistical power analysis for behavioural sciences*. 2nd ed. Academic Press, New York.

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

Dixon, P., 1998. Assessing effect and no effect with equivalence tests. In M. C. Newman and C. L. Strojan (eds.), *Risk assessment: logic and measurement*, pp. 275-301. Ann Arbor Press, Chelsea, MI.

Erdfelder, E., F. Faul, and A. Buchner, 1996. GPower: a general power analysis program. *Behavior Research Methods, Instruments and Computers*, 28:1-11.

Field, A. P., 2003. The problems in using fixed-effect models of meta-analysis on real world data. *Understanding Statistics*, 2:77-96.

Frick, R. W., 1995. Accepting the null hypothesis. *Memory and Cognition*, 23:132-138.

Hoenig, J. M., and D. M. Heisey, 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55:19-23.

Howell, D. C., 2002. *Statistical methods for psychology*. 5th. ed. Duxberry, Pacific Grove, CA.

Irwin, J. R., and G. H. McClelland, 2001. Misleading heuristics and moderated multiple regression models. *Journal of Marketing Research*, 38:100-109.

Judd, C. M., G. H. McClelland, and S. E. Culhane, 1995. Data analysis: continuing issues in everyday analysis of psychological data. *Annual Review of Psychology*, 46:433-65.

Lenth, R. V., 2001. Some practical guidelines for effective sample size determination. *The American Statistician*, 55:187-193.

MacCallum, R. C., S. Zhang, K. J. Preacher, and D. D. Rucker, 2002. On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7:19-40.

McClelland, G. H., 1997. Optimal design in psychological research. *Psychological Methods*, 2:3-19.

Murphy, K. R., and B. Myors, 1999. Testing the hypothesis that treatments have negligible effects: minimum-effect tests in the general linear model. *Journal of Applied Psychology*, 84:234-248.

O'Grady, K. E., 1982. Measures of explained variance: cautions and limitations. *Psychological Bulletin*, 92:766-777.

Rosenthal, R., and D. B. Rubin, 1979. A note on percent variance explained as a measure of the importance of effects. *Journal of Applied Social Psychology*, 9:395-396.

Schmidt, F. L., and J. E. Hunter, 1996. Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, 1:199-223.

Tryon, W. W., 2001. Evaluating statistical difference, equivalence, and indeterminacy using confidence intervals: An integrated alternative method of conducting null hypothesis significance tests. *Psychological Methods*, 6:371-386.

Wickens, C. D., 1998. Commonsense statistics. *Ergonomics in Design*, 6:18-22.

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

Table 1

Statistical power as a function of two-tailed α, **d** and N per group for the independent *t* test

		α = 0.01	α = 0.05	α = 0.10
Effect size (**d**)	*N* per group			
0.2	10	.0165	.0708	.1313
	20	.0251	.0946	.1647
	40	.0448	.1431	.2299
	80	.0928	.2418	.3518
0.4	10	.0395	.1355	.2227
	20	.0862	.2343	.3456
	40	.2047	.4235	.5514
	80	.4711	.7104	.8090
0.8	10	.1717	.3951	.5308
	20	.4380	.6934	.7994
	40	.8226	.9422	.9714
	80	.9925	.9989	.9997

Note. All power calculations were made using GPower (Erdfelder et al., 1996)

Appendix

Sample size for independent *t*: a worked example

There are three main steps to the sample size calculation using Cohen's method (Cohen, 1988; Howell, 2002).

Step 1: Select values for α and power. These are required to look up the value of the non-centrality parameter δ. (Cohen's method simplifies calculation by using tables of non-centrality parameters common to a range of different statistical procedures.)

> Consider a study to compare the effects of a low dose of alcohol on passenger judgements of vehicle speed. As it is an exploratory study where the intervention is not deemed harmful and the cost of Type I error is considered relatively low, α is set leniently at 0.10 and the desired power is set high at 0.90. A two-tailed value of α is appropriate (as neither increases nor decreases in estimates can be discounted *a priori*). These values are used to look up tabulated values of δ (e.g., Howell, 2002, p. 743). Selecting the column corresponding to two-tailed $\alpha = 0.10$ it is possible to scan down until the value 0.90 for power (in the body of the table) is reached. This value lies on the row corresponding to $\delta = 2.90$.

Step 2: Determine the to-be-detected effect size: **d**.

> Assume that a difference between the group means of 6.0 km/h would have practical importance. It would therefore be desirable to detect an effect of this size. Using pilot data the population standard deviation is estimated as 16.0 km/h. This produces a target value of $\mathbf{d} = 6/16 = 0.375$. (Note that the use of the pilot study to estimate the mean difference is avoided).

Step 3: Calculate *N* per group using the formula for independent *t*.

$$N = 2(\delta/\mathbf{d})^2.$$

Substituting the values of δ and **d** from earlier steps produces the following calculation:

$$N \text{ per group} = 2(2.90/0.375)^2 = 2(7.433)^2 = 2 \text{ x } 59.80 = 119.61$$

Rounding up (to help avoid underestimating the required sample size) it is estimated that 120 participants per group (240 in total) are required to have a 90% chance of detecting a difference of 6km/h in the speed judgements.

This example illustrates the value of conducting sample size calculations early in the research planning stage. In this case the total sample size is substantial. The researcher might consider changes to the study that will reduce the required sample size such as an alternative design (e.g., repeated measures) or increasing **d** (e.g., by obtaining more reliable measurements).

Using software for power calculations (such as GPower) eliminates the need for tables and provides slightly more accurate sample size estimates. For example, GPower calculates δ as 2.9408 and total sample size as 246.