Justifying small-\(n\) research in scientifically amazing settings: challenging the notion that only “big-\(n\)” studies are worthwhile

Robert J. Ploutz-Snyder, James Fiedler, and Alan H. Feiveson
Universities Space Research Association, NASA Johnson Space Center, Houston, Texas
Submitted 13 December 2013; accepted in final form 2 January 2014

If you have ever competed for external research support, you know the drill. Every grant starts with some fantastic ideas that are honed, filtered, and shaped into a grant proposal. Primary aims lead to hypothesis statements, definitions of primary outcomes, and a statistical analysis section. At some point, sample size must be determined and justified. Perhaps more than ever, granting agencies are forced to make difficult choices about which studies to fund among many excellent proposals. Their funding decision is partly based on reviewers’ evaluations of the likelihood that the investigative team will arrive at scientifically valuable information. Often, reviewers consider rejecting the null hypothesis as the valuable outcome of a proposed study, and for that, sample size plays an important role.

Because the goal of most research is to reject a null hypothesis, thereby providing evidence that the alternative is true, sample size justifications are predominantly focused on power. Power is defined as the probability of rejecting a null hypothesis, given that the alternative hypothesis is true, and the minimum acceptable level is typically set to 80%. Traditionally, a “proper” sample size is determined by a power analysis. This requires estimates of an anticipated effect magnitude (how big of an effect will you discover?) and variability of the primary outcome(s) of interest. Also required, and usually set at a default, is the alpha level for the statistical test (typically \(\alpha = 0.05\)). Ideally, these parameters (anticipated effect, variance, power, and \(\alpha\)) are given to the statistician, the necessary calculations are made, and the sample size is the smallest \(n\) that delivers the desired power given all of the inputs. Studies are usually deemed “appropriately powered” if the sample size meets these criteria or “fundamentally flawed” if they do not (see Bacchetti (1) for a thoughtful discussion of this threshold myth).

But what happens if the power analyses suggest a sample size that far exceeds what would be feasible? Is there truly no value in pursuing research when a well-designed study cannot meet or exceed 80% power? Is \(P < 0.05\) really so fundamental to science? Shall we categorically and blindly follow these arbitrary defaults?

If your research involves very small or sparse populations (for example, those with a rare disease), has data collection costs tied to \(n\) that quickly exceed funding limitations, or for other reasons makes “big-\(n\)” sample sizes an irrational choice, then you probably already know that the usual sample-size justification methods are a real struggle. As an example, NASA researchers investigating effects of long-duration spaceflight on human systems understand this dilemma perfectly, given the very small number of astronauts (only 3–6 at any given time) available for research on the International Space Station (ISS), competing demands for their participation, the high costs of flying ISS research, and lengthy timelines associated with spaceflight research from concept to completion. And yet the ISS is an amazingly rich laboratory for scientific learning, even if only small-\(n\) studies are possible. This is a case where scientists must appreciate that there is value in small-\(n\) research, even if \(P \geq 0.05\) or if there is <80% likelihood of detecting significance.

If we limit ourselves to adhering strictly to defaults and traditions in terms of sample size justification and demand that any research plan that fails to achieve these default levels is “fundamentally flawed” and not worth pursuing, then spaceflight and other important fields of research may become seriously compromised.

Of course, there are several aspects of experimental design that can minimize sample size requirements, like using repeated-measures designs instead of independent groups, and considering one-tailed hypothesis testing where appropriate. Here, we outline a few additional approaches for justifying small-\(n\) research that we believe should also be considered. For discussion, we have categorized these into three strategies. The best strategy will involve a combination of approaches.

Balance your risks (scrutinize defaults). Some of the simplest ideas are to challenge default practices. These include setting critical \(\alpha = 0.05\), using 2-tailed tests in all cases, and requiring an 80% power threshold. These choices (alpha level, 1- vs 2-tailed tests, power threshold) are necessary for traditional sample size calculations, but they should be set with purpose, rather than given default values. Although we rarely consider it, these choices imply a relative valuation of Type 1 errors (rejecting a true null hypothesis) and Type 2 errors (failing to reject a false null). Typically, we set acceptable values of \(\alpha\) (the probability of making a Type 1 error) to 0.05, and \(\beta\) (the probability of making a Type 2 error) to 0.20 (by setting power = 1 − \(\beta\) to 0.80). The ratio of these two error probabilities is \(\beta/\alpha = 0.20/0.05 = 4\). Thus the defaults imply Type 1 error as being four times as costly as a Type 2 error. This risk posture is typically not considered by PIs, and we argue that such a universal standard is not appropriate for all of science. Mudge and colleagues (8) present a nice discussion of some alternative methods for sample size determination carrying over to the eventual data analysis phase.
Consider precision-based approaches. There are times when we see trends in our data—evidence leaning in the direction of our scientifically grounded theories—yet the $P$ value fails to fall below our critical cutoff (for example, 0.05). Rejection of a null hypothesis can be powerful support for a theory, but there can also be value in identifying and characterizing trends, even if $P \geq 0.05$. This is particularly true for novel areas of inquiry or where research is exceedingly difficult to conduct (for example, ISS).

Precision can be defined as the half-width of a 95% confidence interval for an effect. Choosing a sample size based on precision involves understanding the inherent variability in the outcome(s) of interest, but it does not require an anticipated effect magnitude or assuming that an effect will achieve statistical significance. In this way one can justify conducting research based on achieving a practical accuracy in measuring an effect, rather than trying to explain how the effect would be large enough to achieve statistical significance. Precision based sample size justification is nothing new (5, 7, 8) but it is seldom used. Indeed, for novel work, experimental methods courses and textbooks emphasize that early research steps are typically descriptive in nature, providing effect size estimates that may feed future studies with hypothesis-testing designs. Yet somehow these descriptive studies have inherited a reputation of “not being scientific enough” for many PIs and/or grant reviewers. We contend that there is huge value in these early glimpses into new ideas and that precision-based sample size methods can help justify these types of studies.

Consider the value of information, not just rejecting the null. A hypothesis test condenses a considerable amount of resources (data, costs, time) into a dichotomous decision to accept or reject a null hypothesis. Studies that reject the null are thought to have provided a valuable scientific contribution. Studies that fail to reject, unfortunately, are thought to have little or no value because of 1) a publishing bias against manuscripts that do not show statistically significant results, and 2) these studies are deemed inconclusive because the failure to reject does not imply the null was true. Value-of-information theory (6, 10) provides a framework for valuing research in terms of a ratio of sample size over (some measure of) study value. The definition of study value, however, eludes a common understanding. But there is compelling new work (1–4) that shows that some functions of sample size and cost can serve as surrogate measures of study value. Furthermore, these functions fit nicely with many currently accepted measures, including effect size, Shannon information criterion, Bayesian credible intervals, power, and others. This work is particularly relevant for extremely novel, expensive, or difficult-to-conduct research areas, where even a little knowledge is highly informative.

Such nontraditional approaches to communicating the value of small-$n$ research are appropriate when large-$n$ research is simply not feasible.

DISCLOSURES
No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS
Author contributions: R.J.P.-S. drafted manuscript; R.J.P.-S., J.F., and A.H.F. edited and revised manuscript; R.J.P.-S., J.F., and A.H.F. approved final version of manuscript.

REFERENCES