JBA
james bell associates

Serving the public and non-profit sectors through independent program evaluation, applied research, and technical assistance.

EVALUATION BRIEF

Working with Small Samples

July 2015

Introduction

In program evaluation and applied research, there are two distinct situations in which small sample sizes are encountered: a planned small sample study (e.g., pilot study) or an unplanned small sample that occurs as a result of unanticipated problems with participant enrollment or attrition. This brief provides recommendations for designing a small sample study as well as for working with unplanned small samples in the context of actual program implementation. Also of interest is how small sample sizes affect approaches to quantitative and qualitative analytical methods and the minimum requirements for each. The discussion is prefaced by the acknowledgement that there are no simple answers to most sampling problems, and ultimately the specific research questions and nature of the data that will be collected as part of a study determine how findings should be analyzed and reported. Guidance from the current literature and close collaboration with project and research partners will yield the best solutions to most sampling and methodological issues.

Designing a Small Sample Study

There are situations in which a small sample is necessary by design, for example, in a study involving a target population that is very small or difficult to access, or a program for which cost considerations necessarily limit the size of the population being served. A small sample may also be necessary when piloting an intervention or innovation to determine optimal logistics, preliminary effects, or the costs of future studies or large-scale program implementation. In addition, small samples are often a feature of some single case designs in which a limited number of cases (e.g., a person, school, or community) act as their own controls, often in the context of clinical studies of psychological and educational interventions (Kratochwill et al., 2010). In these situations, small samples have limitations that must be considered, including an increased likelihood of sampling error, strong data dependence, limited ability to detect effects, and restricted generalizability of findings. The primary challenge of a small sample study is to optimize the resources for data collection and analysis while maximizing the likelihood of detecting hypothesized effects. Peterson (2008) outlines the following steps when designing a small sample study.

This publication was developed by James Bell Associates on behalf of the Children's Bureau, Administration for Children and Families (ACF), HHS, under Contract Number GS10F0204K, Order Number 06Y010202. Its contents are the sole responsibility of the author and do not necessarily represent the official views of the Children's Bureau, ACF, or HHS.

Choosing a Sample

Knowing the population of interest is key to understanding how best to design a small sample study; the composition of the target population will shape the research questions and also provide insight into the optimal sampling strategy. When choosing a sample, controlling for specific population characteristics (homogeneity), ensuring equivalent sizes of the intervention and control/comparison groups, and assessing the characteristics of distributions (e.g., skewness and kurtosis) can strengthen the small sample by increasing the likelihood that the assumptions for certain statistical procedures are met (e.g., that the source population is normally distributed). Representativeness is also of concern; a sample can be small and still be representative depending on the target population, but a claim that it is representative must be logically justifiable. In other words, when a true random sample is not feasible, then the sampling method and the resulting sample should be reasonable and "make sense" given the context and constraints of the study.

Selecting a Research Design and Data Analysis Methods

In general, small sample studies employ the same research designs as studies involving larger samples; other factors, such as the purpose of the study (correlational or comparative analysis), the scale of the data (ratio, interval, ordinal, or nominal), and other data characteristics (e.g., independence, missing data, outliers) will affect the choice of statistical procedures. Also of importance is articulating the rationale for choosing a particular research design and how the possible disadvantages of a small sample are to be minimized. Explicitly acknowledging rival hypotheses and ruling them out when possible with additional tests strengthens the study and clarifies the results.

Assessing Statistical Significance, Statistical Power, and Effect Size

The size of the sample is an important component in understanding the relationship between the variables or outcomes of interest. Small sample sizes influence this relationship due to their impact on statistical significance, statistical power, and effect sizes. The significance (or alpha) level, which is typically set at .05 (5 percent), refers to the probability of rejecting the null hypothesis (i.e., that there is no effect from an intervention), given that it is true. The null hypothesis is rejected if the *p*-value (the probability of rejecting the null hypothesis given that it is true) is less than the alpha (α) level.

Statistical power (or sensitivity) refers to the ability of a particular test to detect an effect. Typically, the minimum desired level of statistical power is set at .8 (i.e., an 80-percent probability of rejecting the null hypothesis when it is false). If the hypothesized level of change in the outcomes of interest is known or can be estimated, determining the sample required to achieve the minimum level of statistical power can be done using a simple calculation. Finally, the level of strength of an observed relationship or outcome is referred to as its effect size. Sizes are often characterized in terms of small, medium, or large effects; for example, using Cohen's *d*, a common index of effect size, a "large" effect of .8 would \

indicate that the mean score of a treatment group on a given measure of change is at the 79th percentile of the untreated group on that score.[1]

Small samples usually require trade-offs between statistical significance and power. A small sample size always means less statistical power, holding the alpha level, effect size, and standard deviation constant. Statistical power can be augmented by increasing the alpha level (for example, to .10), but researchers are generally reluctant to increase power in this manner because it increases the likelihood of a Type I error (i.e., detecting false positives). Less power also has negative implications for effect size because small samples are only able to detect large effects; this is a serious limitation given that some small effects may be meaningful. In addition, power is influenced by the type of statistical test selected, with parametric tests having more power than non-parametric tests. Many free calculators are available online for estimating required sample sizes for achieving desired levels of statistical power and effect sizes (e.g., www.raosoft.com, www.surveysystem.com,), and statistical software packages such as SPSS and SAS include modules for estimating sample sizes.

Using Mixed Methods

Mixed method designs, including small sample studies, that employ both qualitative and quantitative research methods often produce the richest and most informative findings. Interviews, focus groups, and other qualitative methods provide a means for understanding a program or issue at a depth that cannot be achieved with quantitative methods alone. In addition, when effects cannot be detected through quantitative methods, qualitative tools can provide insights into the possible reasons (e.g., a flaw in the research design or problems with program implementation) while also elucidating a program's impact from the perspective of participants. Conducting an initial qualitative study may be more cost effective in order to understand the needs and characteristics of the study population, further refine the research questions, and identify the most relevant variables to measure.

Studies involving qualitative methods also have minimum recommended sample sizes. When using a qualitative analysis coding scheme, additional data collection is typically unnecessary when data saturation is achieved (i.e., no new information, themes, or codes emerge). According to one empirical study, 70 percent of all codes were identified after 6 interviews and almost completely after 12 interviews (Guest, Namey, & Mitchell, 2013). However, as with quantitative studies, the optimal size of the sample depends on the population and research questions of interest. If the main objective is to understand commonalities and patterns (homogeneity), a smaller sample is usually sufficient. On the other hand, a larger sample is needed to identify and understand the range of attributes of a population (heterogeneity). When implementing focus groups, a minimum of three groups per homogenous population is recommended with the group being the unit of analysis (Guest, Namey, & Mitchell, 2013).

[1] Lenth (2001) cautions against relying too heavily on rule-of-thumb categories like "small" and "large" when assessing effect size and the subsequent determination of an appropriate sample size. Rather, a variety of factors, including the type of research design, choice of instrumentation, and sample variance, all affect sample size selection.

Reporting Results

Study results should always be reported in a restrained and objective manner highlighting the qualified and provisional nature of any research findings. With studies involving small samples in particular, a detailed description of the context of the program or issues of interest (e.g., policies, legislation, cultural norms, current events, local conditions, community dynamics) helps to explain participants' behavior and provides a framework for future studies of the same population in similar environments. With detailed background information and an understanding of the limitations of the study methods and findings, other researchers and policy makers will have the requisite knowledge for studying comparable interventions and addressing similar methodological problems.

Unplanned Small Samples

Due to the uncertainties of implementing programs in real-world practice settings, evaluations and research studies rarely go exactly as planned. One issue that commonly arises is lower-than-expected enrollment or high program attrition, which results in a smaller sample available for analysis. Potential problems with enrollment or attrition are ideally addressed prior to program start-up or during the early phases of implementation; however, even the best planning cannot always prevent low numbers and researchers must adapt by working with the samples that are available. The following section provides suggestions for preventing small samples as well as possible statistical solutions when low enrollment cannot be avoided.

Monitoring Recruitment and Retention

Prevention is the best solution to almost every problem in the context of any applied research and evaluation endeavor, and can mitigate many issues with program enrollment or attrition. Of equal importance are flexibility and active problem solving during the implementation and data collection process. Both before and during the initial phases of research and project roll-out, an examination of four questions can prevent or shed light on the problem of low program participation and/or completion.

- What is the target population of interest and does the program address its needs? An assessment of the needs and characteristics of the target population is essential for designing and implementing an effective program. This assessment is ideally completed prior to implementation, but further assessments may be necessary if issues with enrollment or dropout arise. Issues to consider include whether critical attributes of the population have changed, whether the services offered to the population meet its needs, or whether the service needs of the population have changed since implementation.

- Is the population aware of the program? While it may seem self-evident that potential participants will know about a program once it has been established, many problems with low enrollment in fact result from a lack of knowledge about the program among both service providers and potential program participants. Questions to consider in this regard include whether information about the program has been adequately and effectively disseminated: what are the channels through which the target population is learning about the program, and what is the extent to which partnerships with potential recruiters and service providers (e.g., organizations active in the target population's community) have been established?

- Does the population have access to the program? Even when a target population needs and knows about the services available through a program, a variety of issues can impede access and participation. Issues to consider here include program eligibility requirements (e.g., are they overly restrictive or are they being applied inappropriately?) and concrete barriers (e.g., transportation, child care, other work and life commitments) that may limit participation.

- Does the population need or want the program? Lastly, investigating factors that contribute to high rates of program refusal or attrition may be necessary. Variables to examine here include whether people who decline or drop out of the program early are systematically different from those that accept and/or complete services; participant perspectives and attitudes towards the program (e.g., Do participants want or perceive a need for the service, did the service meet their expectations, how were they treated by service providers?); and practical impediments to ongoing participation (e.g. transportation) that may be similar to those that constrain initial program access.

Analytical Techniques

Even with the most careful planning and implementation, program enrollment and completion may fall short of expectations and require researchers to work with smaller than optimal samples. The critical decision at this juncture is whether the sample is large enough to enable the use of parametric tests or whether non-parametric tests must be considered. In general, parametric tests have more power to detect statistically significant differences than their nonparametric counterparts, but require that the variable or outcome of interest be intervally scaled (i.e., continuous) and that certain assumptions be satisfied (see below). Ultimately, the decision comes down to an understanding of the available data, how the variables have been operationalized (e.g., continuous or categorical), the assumptions underlying the statistical test options, and the trade-offs involved in choosing one type of test over others.

The assumptions for parametric tests are concerned with the parameters of the population from which the sample was selected. In order for a particular test to work properly (e.g., not overstate or understate the size of a relationship) certain parameters or characteristics must be present in the population. Assessing the structural characteristics of the sample (e.g., distribution, variance) provides a guide to the parameters of the population.

Parametric tests, such as independent samples t-test and ANOVA, have three assumptions that must be satisfied: independence, normality of the distributions for both groups, and equality of variance. Independence refers to the notion that the observations are independent; in other words, no one observation provides information about or otherwise affects the other observations. The question of independence is handled by the research design and is typically held to be true at the point at which the assumptions are assessed. Normality of distributions refers to the concept that the values of a given variable, when plotted, follow a bell-shaped curve (a normal distribution). A standard rule of thumb is that a sample size of 30 is sufficient to invoke the Central Limit Theorem, which posits that as sample size increases, the distribution of the sample mean more closely approximates a normal distribution regardless of the distribution in the population (Newton & Rudestam, 1999; Agresti & Finlay, 1997). However, distributions should be assessed by viewing plots of the data (e.g., stem and leaf plots, box plots) regardless of sample size. The equal variance assumption posits that a variable drawn from independent samples with different means will have the same variance. Equality of variance can be assessed using a variety of tests (e.g., Levene's test), which are available as part of many statistical software packages.[2]

Once these assumptions have been examined, a decision regarding the use of parametric or non-parametric tests can be made. It is important to note that some parametric tests are robust against violations of certain assumptions under certain conditions. Robustness refers to the extent to which a statistical test will give the correct answer even when the test's assumptions are violated. For example, one-way ANOVA is robust against non-normality when the data are not highly skewed and the sample sizes are balanced.

Table 1: Parametric Tests and Non-Parametric Equivalents, summarizes common parametric statistical tests and their non-parametric equivalents.

Table 1: Parametric Tests and Non-Parametric Equivalents

Research Design	Parametric	Non-Parametric	
		Continuous Data	Categorical Data
▪ Correlation	▪ Pearson	▪ Spearman ▪ Gamma[3]	▪ Fisher exact test ▪ Chi-square[4]
▪ Independent Measures (2 groups)	▪ Independent Samples t-test	▪ Mann-Whitney *U* test ▪ Kolmogorov-Smirnov test	▪ Chi-square
▪ Independent Measures (3+ groups)	▪ One-way ANOVA	▪ Kruskal-Wallis test	▪ Chi square
▪ Repeated Measures (2 conditions)	▪ Dependent Samples t-test	▪ Wilcoxon test ▪ Sign test[5]	▪ Mcnemar's test
▪ Repeated Measures (3+ conditions)	▪ Repeated Measures ANOVA	▪ Friedman test	▪ Cochran's Q test

[2] Statistical tests of equality should be used with caution in the case of very small samples because there is very little power to detect violations. In these instances, rules of thumb are often preferable, e.g., if the ratio of the larger to smaller standard deviation is greater than two, then tests that assume unequal variances should be used. See Keppel and Wickens (2004) for more information.

[3] The Gamma statistic is preferable when the data contain many tied observations, i.e., observations with the same value.

[4] The Fisher exact test is preferred with small samples, whereas chi-square is more appropriate with larger samples.

[5] The Wilcoxon test assumes that the magnitude of difference can be ordered in matched observations; if not, the Sign test is preferred.

Other Recommendations

In the case of both planned small sample studies and unplanned small samples, a number of additional recommendations are worth consideration.

- Check whether gaps or omissions in data collected from the sample are systematic or follow some observable pattern. Non-random patterns of missing data may uncover barriers to participant recruitment, enrollment, or retention (e.g., self-selection bias) that would otherwise not have been observed.
- Report effect sizes for future studies and possible meta-analysis. This strengthens the literature and further defines the findings' practical significance, in other words, a level of effect that demonstrates real change in the lives of children and families (McCartney & Rosenthal, 2000).
- Be aware that small samples may compromise the confidentiality and privacy of research subjects. Additional safeguards may be necessary to ensure that participant identities, and the information collected about them, are not inadvertently disclosed.
- Not all patterns are meaningful, especially when observed in small samples. Be careful not to overgeneralize findings.
- Incorporate qualitative research methods whenever feasible and methodologically justifiable. Qualitative data will produce a richer and well-rounded description of the program or population under study and may reveal previously unknown barriers to participant recruitment or retention.
- Identify and report implementation and evaluation barriers in detail. Careful documentation of these issues may provide insights into factors that contributed to small samples (e.g., program or research design, implementation difficulties, sample bias) and prevent similar problems from arising during future studies.

For more information about working with small samples, please contact a JBA team member.

James Bell Associates
3033 Wilson Boulevard, Suite 650
Arlington, Virginia 22201
703-528-3230 or 800-546-3230
www.jbassoc.com

References

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences*. Upper Saddle River, NJ: Prentice-Hall, Inc.

Guest, G., Namey, E. E., & Mitchell, M. L. (2013). *Collecting qualitative data: A field manual for applied research*. Thousand Oaks, CA: Sage.

Keppel, G., & Wickens, T. (2004). Design and analysis: A researcher's handbook (4th ed.). New York: Pearson.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M. & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician,* 55(3), 187-193.

McCartney, K. & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development*, 71(1), 173-180.

Newton, R.R., & Rudestam, K.E. (1999). *Your statistical consultant: Answers to your data analysis questions*. Thousand Oaks, CA: Sage Publications.

Peterson, N. J. (2008). Designing a rigorous small sample study. In J.W. Osborne (Ed.), *Best practices in quantitative methods*. Thousand Oaks, CA: Sage.