Published in final edited form as:

Prev Sci. 2015 October ; 16(7): 950–955. doi:10.1007/s11121-014-0542-7.

Maximizing the Yield of Small Samples in Prevention Research: A Review of General Strategies and Best Practices

Cameron R. Hopkin, **Rick H. Hoyle**, and **Nisha C. Gottfredson**

Duke University

Abstract

The goal of this manuscript is describe strategies for maximizing the yield of data from small samples in prevention research. We begin by discussing what "small" means as a description of sample size in prevention research. We then present a series of practical strategies for getting the most out of data when sample size is small and constrained. Our focus is the prototypic between-group test for intervention effects; however, we touch on the circumstance in which intervention effects are qualified by one or more moderators. We conclude by highlighting the potential usefulness of graphical methods when sample size is too small for inferential statistical methods.

Keywords

small samples; maximizing statistical power; graphical methods

Ideally, every prevention study would produce data from a sufficiently large sample that any research question of interest to the investigators could be informed by results from state-of-the-science analyses without concerns about meeting statistical assumptions or making errors of inference. In reality, many prevention studies, often for reasons beyond the control of the investigators (e.g., small, culturally distinct target population), result in data from samples which, because of their small size, are not suitable for some analytic methods. As a result, important research questions that could be addressed if the sample size were larger must be amended or abandoned altogether. The aim of this manuscript is to present a summary of general strategies and best practices within the existing literature to guide prevention researchers in maximizing the yield of analyses in prevention research when samples are relatively small.

When Is a Sample Small?

What do we mean when we describe a sample as "small"? Is an *N* of 50 small? How about 100? The answer, of course, is that it depends. An *N* of 1 is adequate for some study designs (Kratochwill & Levin, 2010), whereas an *N* of 200 or more may be considered a minimum for others (Hoyle & Gottfredson, in press; see Hoyle, 1999, for a fuller account). A related (and important) question is, how small is *too* small? Certainly samples that are small enough

Correspondence concerning this manuscript should be addressed to Rick Hoyle, Department of Psychology and Neuroscience, Duke University, Durham, NC 27708. Contact: rhoyle@duke.edu.

The authors declare that they have no conflict of interest.

that otherwise acceptable data from a single case can have disproportionate influence on parameter estimates and tests given the analytic method are too small (Fok et al., in press); however, samples that are large enough to minimize concerns about such influence may, for some analyses, still be considered small. In this manuscript, we use “small” to describe samples that are near the lower bound of the size required for satisfactory performance (including relative insensitivity to acceptable data from individual cases) of the particular statistical model chosen to address the questions that motivated the research. A sample is “too small” if its size falls below this lower bound. The evaluation of satisfactory performance can involve multiple dimensions. The most frequently cited dimension is statistical power, the likelihood of detecting an effect of a certain size if it is observed.[1] To that end, this manuscript focuses on coverage of strategies for maximizing statistical power when *N* is constrained. The results of prevention studies may have value beyond revealing a statistically significant effect (see, e.g., Bacchetti, Deeks, & McCune, 2011); thus, we offer suggestions that go beyond maximization of statistical power to touch on strategies for extracting value from a study when statistical power is inadequate for hypothesis testing (e.g., estimating effect sizes for planning future studies or inclusion in meta-analyses). Most of these strategies are recommended regardless of the adequacy of the sample size for a study given the design and research questions, but they are particularly useful for studies in which the available sample is (or will be) near the lower bound of the size required of the statistical model most appropriate to the research questions.

Practical Strategies for Contending with Small Samples

Hansen and Collins (1994) proposed strategies for increasing the statistical power of a study without increasing the sample size. In reality, two of their proposed strategies refer specifically to strategies for maximizing sample size. This seeming contradiction reflects a distinction between the number of cases (i.e., people, families, schools) sampled—the initial sample—and the number of cases from which data are analyzed—the *effective sample*. Although it is nearly always the case that at least some data are provided by all cases in the initial sample, it is frequently the case that some portion of the data are not provided by all cases, resulting in an effective sample with fewer cases than the study was expected to produce. Any strategy that can reduce the number of missing cases or make use of the incomplete information provided by some cases without the introduction of bias into the parameter estimates and standard errors will yield an increase in statistical power without additional sampling.

Attrition is commonplace in prevention trials, in which post-intervention data may be collected a year or more after individuals (or families, or schools) were initially assessed. In a meta-analysis of 85 longitudinal substance-abuse prevention studies, Hansen, Tobler, and Graham (1990) found that attrition ranged from an average of 19% for studies with three month follow-up to 34% for those with a three year follow-up. Attrition is particularly

[1] Our focus on statistical power assumes a traditional null hypothesis statistical testing (NHST) approach to data analysis. We recognize the shortcomings of this approach and its frequent misuse; however, because it remains the primary approach to the analysis of data from prevention trials, it is the approach on which our analysis and recommendations focus. For readers interested in concerns about NHST and potential alternatives, Nickerson (2000) and Harlow, Mulaik, and Steiger (1997) provide balanced, largely nontechnical presentations.

worrisome because, in addition to the loss of cases and, consequently, statistical power, bias may be introduced into the parameter estimates (e.g., means, correlation coefficients), raising questions about the meaningfulness of between-group comparisons. Hansen et al. found that the duration between waves accounted for little of the variance in proportion of attrition, pointing to the differential reactions to assessments and treatments as likely causes. A detailed analysis of attrition in a single study of inner-city middle school students in an alcohol, tobacco, and other drug prevention study found that students who dropped out prior to the eight-month follow-up were more likely than those who completed the study to belong to a family that had relocated between baseline and follow-up and reported higher levels of family conflict, less parental supervision, and greater perceived risk of alcohol and drug use (Zand, Thomson, Dugan, Braun, Holterman-Hommes, & Hunter, 2006). Despite impressive attempts to retain participants, the effective sample size of 104 was both significantly lower than the initial sample size of 127 and of a size that would be questionable for all but the simplest statistical models. Yet the efforts at retention likely made the difference between a study for which simple analyses could be conducted with adequate power and one for which power would be unacceptably low for even the simplest analyses. Investments in retention of participants narrow the gap between initial and effective sample sizes and, in so doing, improve statistical power and reduce bias without additional sampling.

It is possible for a longitudinal study to retain all members of the initial sample for the duration of the study yet nonetheless produce incomplete data. In the face of missing data due either to attrition or nonresponse, case-wise deletion discards valid data, thereby reducing power and biasing estimates and tests. Other strategies such as pairwise deletion (if correlations or covariances are to be analyzed) and replacement of missing data with imputed values retains data provided by research participants but introduces biases into estimates and tests. Fortunately, modern missing data methods allow researchers to take full advantage of the information provided by research participants without biasing estimates and tests by imputing values for missing data and treating them as legitimate values (e.g., Enders, 2010; Graham, 2009; Schafer, 1997).[2] In many cases, these methods can reduce the gap between the initial and effective sample sizes to zero, avoiding the loss in statistical power and bias in estimates and tests that result from traditional approaches to handling missing data such as case-wise deletion and mean substitution.

The remaining strategies suggested by Hansen and Collins (1994) concern increasing the size of the observed effect (i.e., difference between groups). Given the standard equation for computing effect size, which is a ratio of the effect of interest (e.g., difference between means, regression coefficient) and the population variance (expressed as standard deviation), there are two categories of approaches that, given a fixed sample size, would increase statistical power by increasing effect size: (1) increase the effect of interest, (2) decrease the population variance. We summarize each in turn.

Although effects can be reflected in a number of statistics, the focus of many, if not most, prevention studies is the difference between means; thus, we focus on practical measures for

[2] See von Hippel (2013) for potential problems and solutions for use of these methods with small samples.

increasing the difference between group means. To the extent that the intervention or manipulation can be modified by the researcher, it should be designed to target the primary mechanisms that would give rise to group differences. For example, if a manipulation is designed to increase resistance to peer influence and the exercise of such resistance requires self-efficacy, then the intervention or manipulation should focus squarely on the development of self-efficacy. The use of this commonsense strategy requires a clear understanding of the cognitive, affective, and motivational mechanisms that underlie prevention-relevant behaviors and the development of intervention components designed to change those mechanisms. The best designed intervention will not be effective if research participants do not receive full exposure to it. As such, an additional means of increasing the difference between groups given a well-grounded intervention is to invest in measures to ensure that the intervention is delivered with integrity (Dumas, Lynch, Laughlin, Phillips Smith, & Prinz, 2001). The intervention also should be delivered for a length of time necessary to change the targeted mechanisms and, thereby, produce behavior change. Relatedly, effects of the intervention should be assessed at a point in time when the effect is likely to be maximized. These considerations assume an understanding of how the intervention works in terms of exposure and timing (see Collins et al., 2011, for other considerations and strategies).

If sample size cannot be increased and the effect of interest is at its maximum, another means of increasing statistical power is to reduce variance other than variance attributable to the intervention or manipulation. Such variance arises from two sources: sample heterogeneity and unreliability of measurement (Hansen & Collins, 1994). The consideration of sample heterogeneity is one of balance—maintaining the representativeness of the sample while minimizing within-group variance that contributes to inflated test statistics. Although within-group variance can be reduced by including additional independent variables (e.g., ethnicity, gender), doing so leads to smaller *N*s per group and reduced power. An alternative, discussed below, is to account for the variance by including covariates in the analyses. Additional variance that decreases power by lowering effect sizes may arise from unreliability of measurement. For a given effect size and degree of true sample heterogeneity, an increase in reliability of measurement reduces variance not attributable to the intervention or manipulation and, in so doing, increases the likelihood of detecting an effect by reducing the confidence interval around estimates of means.

One simple way to reduce uncontrolled heterogeneity is to use within-subjects designs whenever possible. This is because for every participant, the score on the outcome variable can be attributed to three sources: 1) the effect of the intervention or predictor of theoretical interest, 2) measurement error due to the imperfection of any given measure's ability to tap the construct of interest, and 3) that person's extraneous personality and context variables that were not measured yet influence the score. In a within-subjects design, the same person participates in all possible conditions so that the third source of variability, which is potentially the largest of the three, can be eliminated. Interrupted time-series design with multiple baselines (described by Hawkins, Sanson-Fisher, Shakeshaft, D'Este, & Green, 2007) are a particularly efficacious type of within-subjects design for testing intervention effects within individuals or communities.

Despite their relative superiority in detecting effects compared to equivalent between-subjects designs, designs in which participants are exposed to all conditions are not always feasible or desirable. Powerful interventions, for instance, may lead to carryover effects; if a participant does not return to a reasonable baseline on the construct in question within the desired timeframe, his or her data in other conditions will be affected by the preceding intervention condition. When within-subjects designs are not appropriate, the power and precision of estimates in between-subjects designs may be increased by the inclusion of covariates that measure person-centered variables or account for individual differences in response to treatment. Raudenbush (1997) provides a detailed explanation of why including explanatory covariates may have a large impact on statistical power. Conceptually, it is clear that the more noise in the outcome variable that is explained by covariates, the easier it will be to detect meaningful predictor effects. In a slightly different context, Collins, Schafer, and Kam (2001) showed that using an “inclusive strategy” (i.e., including as many predictors as possible) decreased bias and increased efficiency of maximum likelihood estimates. Although Collins et al. (2001) were focusing on estimation in the presence of missing data, their work is relevant here. This is because random effects may be understood to be “missing” variables that must be estimated from all available information. Thus, inclusion of predictors may be particularly important for recovering variance component estimates.

When multiple predictors are tested simultaneously in an overall test of model significance, power is influenced by the number of predictors in the model (Cohen et al., 2003). Yet, it is a misconception that more covariates lead to lower power to detect an effect *for a single predictor of interest*. Rather, as shown by Raudenbush (1997), it is desirable to include covariates that explain a high degree of residual variance in the outcome when no inference is made regarding the effects of such covariates. Doing this essentially increases the effective reliability of the outcome variable. On the other hand, if an analyst makes multiple comparisons, then they should be compelled to make the appropriate corrections for them (e.g., Wang & Ware, 2013). Thus, it is wise for analysts with small samples to consider carefully which hypotheses they want most to test, and refrain from testing hypotheses of secondary importance.[3]

The flip side of minimizing error variance is maximizing the construct-relevant variance of measured variables. If the sample size is non-negotiable and small, a researcher might consider collecting a non-random sample in which individuals with high and low values of the independent variable are selected (Cohen, Cohen, West, & Aiken, 2003). This approach works because it maximizes the variability of the independent variable, thereby increasing the chance of detecting a significant effect of variation in the independent variable on variation in the dependent variable. Although such non-random sampling is not best practice, as it will tend to inflate the effect size as well as jeopardize the external validity of the study, an extremely small sample size might justify it so long as the sampling method is explicitly revealed and justified in the research report. This approach is used often in studies that seek to describe age-related effects on an outcome variable, for instance by sampling

[3] An informative discussion of the use of covariates to increase statistical power is provided by Dennis, Francis, Cirino, Schachar, Barnes, & Fletcher (2009).

from younger and older participants (with few in the middle). Similarly, a prevention scientist might sample very low-risk and very high-risk individuals to test the differential effect of an intervention on these groups, with the assumption that medium-risk individuals would fall in the middle.

Detecting Interaction Effects

Researchers working with small samples should think particularly carefully about testing interaction effects. In prevention research, intervention effects may only be effective for a range of individuals, or they may only be effective under certain conditions (Wang & Ware, 2013). This type of *moderated effect* is tested as a statistical interaction. In spite of their intrigue, interaction effects are doubly plagued by having both a relatively high Type I error rate compared to additive main effects (particularly when predictors contain measurement error; Embretson, 1996; Kang & Waller, 2005), as well as lower power than the main effects (Brown et al., 2011). For these reasons, tests of interaction effects when sample size is small should be approached thoughtfully. Collins, Dziak, and Li (2009) showed that reduced factorial designs are preferable to complete factorial designs. In other words, researchers should design studies that include only the experimental contrasts that are of specific theoretical interest; statistical models should conform to these specific hypotheses. Fractional factorial designs such as this come at the expense of being able to fully disentangle all possible interaction effects because not all variables are fully "crossed," but when multiple manipulations are planned and many of the higher-order interactions are assumed to be negligible in magnitude and of no theoretical interest, such designs can greatly reduce either the number of conditions or, more importantly, the number of participants required to achieve acceptable power. The target sample size should be dictated by the lowest number that is necessary for testing hypothesized statistical interactions with adequate power. A more complete discussion of strategies for detecting moderated effects can be found in a special issue of *Prevention Science* on the topic (Supplee, Kelly, MacKinnon & Barofsky, 2013).

The husbanding of research resources toward the variables and effects of greatest interest demonstrated in the fractional factorial design form the core of the multiphase optimization strategy (Collins et al., 2011), an overarching study design paradigm drawn primarily from engineering science in which possible intervention components are treated like candidates to be tested individually via small, focused trials that include as few comparisons as possible to test their efficacy before inclusion in larger intervention studies. Although many researchers prefer to think of their research in a more serial, independent fashion, approaching prevention studies in this programmatic fashion allows the careful research team to build a database of effective intervention components and be thriftier in the use of both research dollars and–relevant to our focus here–the number of participants required.

Even in the undesirable eventuality that a researcher's data from an individual study is hopelessly underpowered for traditional analyses, all is not lost. Increasingly, researchers are moving toward a model of collaborative science through meta-analysis and integrative data analysis across multiple independent studies (Brown et al., 2011; Curran & Hussong, 2009). This approach has gained traction particularly in the field of genetics because it would be impossible to detect miniscule effects of single genetic markers without pooling resources across multiple studies. In the event of an unworkably small sample size or otherwise unpublishable results, it is advisable to record in an easily-retrievable and readily-interpreted format certain data for easy inclusion in a future meta-analysis or integrative data analysis: sample size, primary variables involved, relevant measures of effect size for all outcomes, and confidence intervals for group means (indeed, consistent reporting of effect sizes and confidence intervals should be standard practice regardless of the "publishability" of the results). Such a post-mortem procedure is not time-intensive and can conceivably change a "wasted" study into a stepping stone for future findings. This practice can and should be encouraged in the field of prevention science, not least because doing so allows for an enhanced ability to detect moderated effects of prevention interventions (Brown et al., 2011). Because different studies invariably assess different subgroups of the population, there is more heterogeneity across studies than within (e.g., with respect to age, ethnicity, geography, or culture). If researchers take care to measure these characteristics, then this heterogeneity can be leveraged to test for moderation using meta-analytic methods.

When a Sample is too Small for Hypothesis Testing

Ideally, prevention scientists would always begin working with their data using data visualization methods (e.g., Friendly, 1995; Young & Bann, 1996). These methods can be particularly useful when sample size is too small for parameter estimation or hypothesis testing. Data visualization does not offer many options for increasing power, per se, but it may serve as an identifying end-point for situations in which statistical inference is not a reasonable possibility. When only a handful of cases are available, analysts should plot the within-group association between the predictor variable(s) of interest and the outcome variable. Plots should be used for in-depth data description and not for generalization to the population. Although this may seem to be the defeatist's option, it is in fact eminently practical in that the data are being employed to the maximum extent possible and serving as a springboard for more effective data collection instead of merely lining a file drawer in the back of an investigator's office. As an example, Carrig, Wirth, and Curran (2004) provide an easy-to-use SAS macro for visualizing person-specific growth trajectories with repeated measures data. A similar approach should be followed with data from individuals within groups.[4]

Summary and Conclusions

Many prevention researchers live with the unfortunate reality that limited availability of financial resources or limited access to, or size of, the population of interest results in samples that are smaller than they ideally would be given the requirements of the analytic strategy best suited to the research questions. We have suggested practical ways for prevention scientists to optimize statistical power and make good use of data when statistical power is inadequate for hypothesis testing using inferential statistics. At the same time, we have urged caution in generalizing too far beyond what is appropriate given study

[4] Information about approaches to data visualization can be found in Young (1996) and a collection of papers edited by Post, Nielson, and Bonneau (2003).

constraints. We would also caution that, when the costs of obtaining even a small sample are high (e.g., personnel costs, participant burden), the benefits might not be sufficient to warrant those costs. If, however, the costs associated with a study likely to yield a small sample can be justified, then the use of strategies we have described will serve to maximize the value of the study. Although the issues, strategies, and cautions vary from one study to the next, we offer these general suggestions for working with data from small samples.

Compensate for a small sample size by optimizing study features that you can control

There is more to power than just sample size. When planning studies, focus on study features that you can control, such as reliability of measurement. Measure as many theoretically-strong indicators as possible. Maximize the predictive power of your model by including covariates that are strongly related to the outcome of interest, and eliminate covariates that have no explanatory power. Finally, try to avoid censoring important variability in your outcome measures through coarse categorization (e.g., median splits), a practice that greatly reduces power.

Consider your research questions carefully; optimize resource allocation to maximize inference for the most important parameters

When dealing with complex models, not all model parameters are estimated with equal precision. Consider which parameters are trustworthy and focus on interpreting these, without placing much emphasis on the parameters that are not trustworthy. When interest centers on the effects of a particular predictor, aim to maximize variability across the full range of that predictor’s values. This can be achieved by over-sampling on the extreme ends of the variable distribution, for instance.

Visualize your data and use descriptive statistics liberally

With very small samples, it is usually best to limit statistical inference and to focus instead on describing the data with descriptive statistics and data visualization. Although the results of such analyses are not as persuasive as more rigorous analyses in which all relevant processes are considered simultaneously, they can move a research program forward, laying the groundwork for sharper focus and more efficient investment of resources in subsequent studies.

In short, the anticipation of a small sample for a prevention study should prompt an intense focus on other features of the study. These range from the choice of measures (including the number) to the inclusion of potentially useful covariates to the adjustment of research questions given the analytic options for which the sample size is appropriate. Optimizing these features of a study may make the difference between an acceptable and an ill-advised treatment of data produced by the study.

Acknowledgments

During the writing of this manuscript, the authors were supported by National Institute on Drug Abuse (NIDA) Grant P30 DA023026. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIDA.

References

Bacchetti P, Deeks SG, McCune JM. Breaking free of sample size dogma to perform innovative translational research. Science Translational Medicine. 2011; 3(87):87.

Brown CH, Sloboda Z, Faggiano F, Teasdale B, Keller F, Burkhart G, Vigna-Taglianti F, Howe G, Masyn K, Wang W, Muthén B, Stephens P, Grey S, Perrino T. Methods for synthesizing findings on moderation effects across multiple randomized trials. Prevention Science. 2011; 14:144–156. [PubMed: 21360061]

Carrig M, Wirth RJ, Curran PJ. A SAS Macro for estimating and visualizing individual growth curves. Structural Equation Modeling: An Interdisciplinary Journal. 2004; 11:132–149.

Cohen, J. Statistical power analysis for the behavioral sciences. 2nd ed.. New York: Academic Press; 1988.

Cohen J. A power primer. Psychological Bulletin. 1992; 112:155–159. [PubMed: 19565683]

Cohen, J.; Cohen, P.; West, S.; Aiken, L. Applied multiple regression/correlation analysis for the behavioral sciences. 3rd ed.. Mahwah, NJ: Erlbaum; 2003.

Collins LM, Baker TD, Mermelstein RJ, Piper ME, Jorenby DE, Smith SS, Christiansen BA, Schlam TR, Cook JW, Fiore MC. The multiphase optimization strategy for engineering effective tobacco use interventions. Annals of Behavioral Medicine. 2011; 41:208–226. [PubMed: 21132416]

Collins LM, Dziak JJ, Li R. Design of experiments with multiple independent variables: A resource management perspective on complete and reduced factorial designs. Psychological Methods. 2009; 14:202–224. [PubMed: 19719358]

Collins LM, Schafer JL, Kam C. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods. 2001; 6:330–351. [PubMed: 11778676]

Curran PJ, Hussong AM. Integrative data analysis: The simultaneous analysis of multiple data sets. Psychological Methods. 2009; 14:81–100. [PubMed: 19485623]

Dennis M, Francis DJ, Cirino PT, Schachar R, Barnes MA, Fletcher JM. Why IQ is not a covariate in cognitive studies of neurodevelopmental disorders. Journal of the International Neuropsychological Society. 2009; 15:331–343. [PubMed: 19402919]

Dumas JE, Lynch AM, Laughlin JE, Phillips Smith E, Prinz RJ. Promoting intervention fidelity. Conceptual issues, methods, and preliminary results from the EARLY ALLIANCE prevention trial. American Journal of Preventive Medicine. 2001; 20(1 Suppl):38–47. [PubMed: 11146259]

Embretson SE. Item response theory models and spurious interaction effect in factorial ANOVA designs. Applied Psychological Measurement. 1996; 20:201–212.

Enders, CK. Applied missing data analysis. New York: Guildford Press; 2010.

Fok CCT, et al. (in press)--insert reference to introductory article when it is known.

Friendly, M. Exploratory and graphical methods of data analysis. 1995. Retrieved from http://www.datavis.ca/courses/eda/

Graham JW. Missing data analysis: Making it work in the real world. Annual Review of Psychology. 2009; 60:549–576.

Hansen, WB.; Collins, LM. Seven ways to increase power without increasing *N. NIDA Research Monograph*. In: Collins, LM.; Seitz, LA., editors. Advances in data analysis for prevention intervention research (NIDA Research Monograph 142, NIH Publication No. 94-3599). Rockville, MD: National Institutes of Health; 1994. p. 184-195.

Hansen WB, Tobler NS, Graham JW. Attrition in substance abuse prevention research: A meta-analysis of 85 longitudinally followed cohorts. Evaluation Review. 1990; 14:677–685.

Harlow, LL.; Mulaik, SA.; Steiger, JH., editors. What if there were no significance tests?. Mahwah, NJ: Erlbaum; 1997.

Hawkins NG, Sanson-Fisher RW, Shakeshaft A, D'Este C, Green LW. The multiple baseline design for evaluating population-based research. American Journal of Preventive Medicine. 2007; 33:162–168. [PubMed: 17673105]

Hoyle, RH., editor. Statistical strategies for small sample research. Thousand Oaks, CA: Sage Publications; 1999.

Hoyle RH, Gottfredson NC. Sample size considerations in prevention research Applications of multilevel modeling and structural equation modeling. Prevention Science. (in press).

Kang S, Waller G. Moderated multiple regression, spurious interaction effects, and IRT. Applied Psychological Measurement. 2005; 29:87–105.

Kratochwill TR, Levin JR. Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. Psychological Methods. 2010; 15:124–144. [PubMed: 20515235]

Nickerson RS. Null hypothesis significance testing: A review of an old and continuing controversy. Psychological Methods. 2000; 5:241–301. [PubMed: 10937333]

Post, FH.; Nielson, GM.; Bonneau, G-P., editors. Data visualization: The state of the art. Boston: Kluwer Academic Publishers; 2003. Retrieved from http://www.springer.com/computer/image+processing/book/978-1-4020-7259-8?otherVersion=978-1-4613-5430-7

Raudenbush SW. Statistical analysis and optimal design for cluster randomized trials. Psychological Methods. 1997; 2:173–185.

Schafer, JL. Analysis of incomplete multivariate data. London: Chapman & Hall; 1997.

Supplee LH, Kelly BC, MacKinnon DM, Barofsky MY. Subgroup analysis in prevention and intervention science [Special issue]. Prevention Science. 2013; 14(2)

von Hippel PT. The bias and efficiency of incomplete-data estimators in small univariate normal samples. Sociological Methods & Research. 2013; 42:531–558.

Wang R, Ware JH. Detecting moderator effects using subgroup analyses. Prevention Science. 2013; 14:111–120. [PubMed: 21562742]

Young, FW. ViSta: The Visual Statistics System. Chapel Hill, NC: Thurstone Psychometric Laboratory Research Memorandum 94-1(c); 1996. Retrieved from http://forrest.psych.unc.edu/research/index.html

Young, FW.; Bann, CM. ViSta: A Visual Statistics System. In: Stine, RA.; Fox, J., editors. Statistical computing environments for social research. Thousand Oaks, CA: Sage Publications; 1996. p. 207-236.

Zand D, Thomson NR, Dugan M, Braun JA, Holterman-Hommes P, Hunter PL. Predictors of retention in an alcohol, tobacco, and other drug prevention study. Evaluation Review. 2006; 30:209–222. [PubMed: 16492999]