Seven Ways To Increase Power Without Increasing N

William B. Hansen and Linda M. Collins

ABSTRACT

Many readers of this monograph may wonder why a chapter on statistical power was included. After all, by now the issue of statistical power is in many respects mundane. Everyone knows that statistical power is a central research consideration, and certainly most National Institute on Drug Abuse grantees or prospective grantees understand the importance of including a power analysis in research proposals.

However, there is ample evidence that, in practice, prevention researchers are not paying sufficient attention to statistical power. If they were, the findings observed by Hansen (1992) in a recent review of the prevention literature would not have emerged. Hansen (1992) examined statistical power based on 46 cohorts followed longitudinally, using nonparametric assumptions given the subjects' age at posttest and the numbers of subjects. Results of this analysis indicated that, in order for a study to attain 80-percent power for detecting differences between treatment and control groups, the difference between groups at posttest would need to be at least 8 percent (in the best studies) and as much as 16 percent (in the weakest studies). In order for a study to attain 80-percent power for detecting group differences in pre-post change, 22 of the 46 cohorts would have needed relative pre-post reductions of greater than 100 percent. Thirty-three of the 46 cohorts had less than 50-percent power to detect a 50-percent relative reduction in substance use. These results are consistent with other review findings (e.g., Lipsey 1990) that have shown a similar lack of power in a broad range of research topics. Thus, it seems that, although researchers are aware of the importance of statistical power (particularly of the necessity for calculating it when proposing

research), they somehow are failing to end up with adequate power in their completed studies.

This chapter argues that the failure of many prevention studies to maintain adequate statistical power is due to an overemphasis on sample size (N) as the only, or even the best, way to increase statistical power. It is easy to see how this overemphasis has come about. Sample size is easy to manipulate, has the advantage of being related to power in a straightforward way, and usually is under the direct control of the researcher, except for limitations imposed by finances or subject availability. Another option for increasing power is to increase the alpha used for hypothesis-testing but, as very few researchers seriously consider significance levels much larger than the traditional .05, this strategy seldom is used.

Of course, sample size is important, and the authors of this chapter are not recommending that researchers cease choosing sample sizes carefully. Rather, they argue that researchers should not confine themselves to increasing N to enhance power. It is important to take *additional* measures to maintain and improve power over and above making sure the initial sample size is sufficient. The authors recommend two general strategies. One strategy involves attempting to maintain the effective initial sample size so that power is not lost needlessly. The other strategy is to take measures to maximize the third factor that determines statistical power: effect size.

MAINTAINING EFFECTIVE SAMPLE SIZE

Preventing Attrition

One of the best ways to increase power without increasing N is to avoid decreasing N through attrition. Of course, attrition has other consequences besides loss of power, such as internal and external validity problems. However, independent of these problems, a loss of subjects through attrition is accompanied by a loss of statistical power.

Many articles about attrition (Biglan et al. 1987; Ellickson et al. 1988; Hansen et al. 1990; Pirie et al. 1989) have helped to alert the research community to the potential causes of attrition so that measures can be taken to prevent it. Attrition has many and varied causes. Sometimes the causes are as simple as subjects moving out of the school district where the study is taking place. Usually, though, the causes are more complex and not totally unrelated to the study. In some studies where the treatment is aversive in some way, treatment group subjects drop out at a higher rate; in other studies where the treatment is a plum and nothing is done to compensate the control group, the opposite occurs. In substance use prevention studies, high-risk subjects are more likely to drop out (Hansen et al. 1985). Attrition can even reflect a political problem as, for example, when an institution like a school or a school district drops out of a study (Hansen et al. 1990). Researchers should become familiar with the studies that have examined retention of subjects (Ellickson et al. 1988) and political units (Goodman et al. 1991; O'Hara et al. 1991) to gain an understanding of how to manage attrition in practical terms. Every prevention effort should include funds in its budget for tracking and collecting data from subjects who have dropped out of the study.

Missing Data Analysis

Missing data analysis (Graham et al., this volume) is an exciting new data analysis strategy that recovers some (but not all) of the loss of power incurred through attrition. This is not a way of replacing missing data; rather, it is a way of making the most out of the remaining data. This methodology provides a way for the user to model the mechanisms behind attrition, allowing for estimation of what the results would have been if the full sample had been maintained. The chapter by Graham and colleagues (this volume) presents an in-depth look at this important topic.

MAXIMIZING EFFECT SIZE

Take a closer look at effect size:

$$\text{effect size} = \frac{\mu_A - \mu_B}{\sigma} \tag{1}$$

The numerator of equation (1) is the difference between the population mean for the treatment group (μ_A) and the population mean for the control group (μ_B). The denominator is the population variance (assuming homogeneity of variance, that is, the two populations have identical variances). The strategies suggested here are intended to increase effect size either by increasing the size of the numerator of equation (l), that is, increasing the difference between the mean of the treatment group and the mean of the control group, or decreasing the denominator of equation (l), that is, decreasing the population variance.

STRATEGIES TO INCREASE THE MAGNITUDE OF GROUP DIFFERENCES

Targeting (and Affecting) Appropriate Mediators

All prevention programs seek to change behavior by changing some mediating process. The choice of which mediating process to intervene on is the key to a powerful intervention. Only if the researchers developing a program understand the basic underlying processes that account for substance use behavior can they hope to identify the most appropriate mediators. Such an understanding is gained by examining very carefully and thoroughly existing theory and empirical evidence about the modifiable predictors and determinants of substance use behavior.

For example, in a series of studies conducted by Hansen and colleagues (1988, 1991), two mediating processes were targeted: the development of normative beliefs intolerant of alcohol and drug use and the

development of skills for resisting overt offers to use substances. Two programs were compared, each designed to address one mediator specifically and, to the extent possible, to not affect the mediator associated with the other program. The results consistently have shown success in achieving differential impacts on behavior.

Some program developers prefer a less systematic emphasis on which a mediator is targeted for change, basing program content and strategy on strongly held personal beliefs rather than on empirical evidence about which components offer potential for change in particular mediators. Such programs developed solely from instinct or good intentions will, over the long run, fail to have as much power as programs developed more scientifically.

Identifying the appropriate mediators is a necessary but not sufficient condition for increasing statistical power—the intervention must be strong enough to have an effect, ideally a large one, on the mediators. It is difficult to give advice on how to achieve this goal. It seems that, even at their best, researchers have little more than an intuitive understanding of what it takes programmatically to change mediating processes. Although the literature in this area can be of some help, the best methods for reaching school-age children change constantly. The impact of interventions probably could be increased, thereby increasing statistical power, by making better use of input from the people who know best how to teach youth, namely teachers, counselors, and youth workers.

Maintaining Program Integrity

Program integrity, the degree to which the program is adhered to in delivery, has predictable effects on outcome (Botvin et al. 1990; Hansen et al. 1991; Pentz et al. 1990); when program integrity is compromised, the treatment is less effective and differences between treatment and control shrink. Researchers have yet to develop a complete understanding of program integrity. For example, integrity to date has been defined by researcher standards rather than target audience-centered standards. Researchers may need to account for issues that they have not

considered when defining integrity, such as the need to tailor a program for specific audiences.

For some programs, there is a tradeoff between N and program integrity. In fact, Tobler (1993) found in a meta-analysis that effect size was reduced in prevention studies involving more than 400 subjects per condition. If the sample size is so large that a large staff must be hired to deliver the program and the researcher, therefore, cannot be highly selective about this staff and cannot supervise them closely, it is unlikely that the program will be delivered uniformly well. It is important for the researcher to be aware of this tradeoff, because there may be times when power is maximized in the long run by choosing a smaller N and a more manageable intervention.

Appropriate Timing of Longitudinal Followup

The magnitude of the difference between treatment and control groups partly is a function of the length of time between program implementation and followup. Hansen (1992) concluded that many prevention studies are conducted for too short a period of time. Prevention researchers sometimes argue that long-term impacts cannot be expected from prevention programs. The authors disagree for two reasons. First, the goal of prevention is to maintain existing nonbehavior. There is reason to be much more sanguine about the possibility of prevention to have long-term effects, especially if the forces that foster experimentation with alcohol and drugs have really been changed. Second, the outcome of interest in prevention studies is based, not only on the treatment group maintaining its level of use or nonuse, but on the control group changing its behavior. Since this change takes time, it makes sense to measure behavioral outcomes repeatedly over a long period of time in order to increase the potential for observing differences between treatment and control groups when they reach their peak. For more about timing of observations and its effects on results, see Cohen (1991) and Collins and Graham (1991).

STRATEGIES FOR REDUCING VARIANCE

Sampling Control

There often is some pressure on prevention researchers to make sure the studies they are planning involve heterogeneous samples. There are two reasons for this. One reason is the need to maximize external validity. The more representative the sample is of the population at large, the better the external validity of the study is. The second reason is political; for example, it is important to make sure that women and minority groups are not excluded from prevention studies.

These two reasons for using heterogeneous samples are very good ones. However, researchers should balance these considerations with the effects of heterogeneity on statistical power. When heterogeneity is enhanced and homogeneity is diminished, power is reduced. The reason for this is straightforward: All else being equal, a heterogeneous population has more variance than a homogeneous population. Consider two populations with identical variances, σ^2, but with different means. If these two populations are combined into one, the new variance, σ_*^2, will be:

$$\sigma_*^2 \;=\; \sigma^2 + \left(\frac{\mu_A - \mu_B}{2} \right)^2 \tag{2}$$

Thus, the larger the difference in means between the two populations is, the larger the variance of the combined population will be. This larger variance results directly in a decreased effect size (see equation [1]) and, therefore, decreased power.

The problem is compounded if analyses then are conducted separately on subgroups in the data because these analyses necessarily will be based on a smaller N and may have dramatically reduced power. Where appropriate, covariates can be used to model subgroup differences. This maintains degrees of freedom and, therefore, can reduce the threat that sampling from heterogeneous groups brings.

Using Reliable and Appropriate Measures

The disciplines of psychology and epidemiology have both greatly influenced the field of substance use prevention research. These fields have different, and at times opposing, methodological traditions, particularly with respect to measurement. Epidemiology has emphasized relatively straightforward measurement and the use of manifest, and often dichotomous, variables. In contrast, psychology has a long tradition of measurement theory, emphasizing scale development, multiple indicator models, latent variables, and continuous variables. Classical test theory, including reliability theory, came from psychology.

An immediate question that is raised by contrasting these two approaches is, "Which is more appropriate, using continuous measures of substance use or using dichotomous measures?" Of course, the answer depends partly upon the research question that is being posed. The ramifications of this question for statistical power are complex. Cohen (1983) showed that dichotomizing a normally distributed continuous variable essentially throws away information and leads to a considerable loss of power. The situation is less clear with the skewed distributions that are more the rule in substance use prevention research. In general, though, unless the distributions are severely nonnormal, a loss of power can be expected if continuous variables are dichotomized.

It also is worth noting the relationship between measurement reliability and statistical power. This relationship is more complex than it may appear at first glance. Recall that according to classical test theory, the total variance in a measure is made up of true score variance and error variance. Measurement reliability is defined as the proportion of total variance that is made up of true score variance. Zimmerman and Williams (1986) showed that the direction of the relationship between reliability and power depends upon which of the three components-total variance, true score variance, or error variance-is held constant while the others are varied. If a constant true score variance is assumed, it follows that the greater the reliability (that is, the less error variance there is in a measure), the greater the statistical power will be. This is true because,

because, under these conditions, when the error variance decreases, the total variance decreases, resulting in a decrease to the denominator in equation (1), However, if a constant error variance is assumed, when reliability is increased, the true score variance is increased and, therefore, the total score variance is increased, leading to a decrease in power.

Zimmerman and Williams (1986) pointed out that the answer to this seeming paradox lies in how reliability is increased in practice. If a measure is improved by, say, discarding a few items that do not belong in the instrument, then generally this improves reliability by decreasing error variance. This strategy can be expected to improve statistical power. On the other hand, if reliability is improved by changing the sample so that it is more heterogeneous and, therefore, there is more true score variance, this is likely to result in an overall increase in variance and, hence, a loss of power.

CONCLUSIONS

This chapter argues that, while obtaining a sufficiently large sample is important, it is not all there is to statistical power. Other strategies are important if statistical power is to be maintained over the course of a substance use prevention study. The authors made seven suggestions for ways to improve power without increasing N in prevention research. Except for missing data analysis, none of these suggestions are new. Most of them are based on common sense, and many of them will be recognized as recommendations often made to colleagues and students. It is ironic that scientists, researchers, and social advocates have largely failed to use these principles systematically to improve the power of research. They persist in thinking of statistical power only in terms of sample size but must adopt a wider view, as suggested here.

The suggestions made here do not translate directly into formulas that can be inserted "as is" into proposals or research designs. Instead, they represent principles that can be used to guide decision-making in practice. In the end, it is not the proposal or the research report that is the essence

of science, but increased understanding of the phenomenon of substance abuse and the procedures employed to prevent it. If researchers are ever to develop a thorough understanding of substance abuse and highly effective methods for preventing it, they must be aware of how research decisions affect statistical power.

REFERENCES

Biglan, A.; Severson, H.; Ary, D.; Faller, C.; Gallison, C.; Thompson, R.; Glasgow, R.; and Lichtenstein, E. Do smoking prevention programs really work: Attrition and the internal and external validity of an evaluation of a refusal skills training program. *J Behav Med* 10:613-628,1987.

Botvin, G.J.; Baker, E.; Dusenbury, L.; Tortu, S.; and Botvin, E.M. Preventing adolescent drug abuse through a multimodal cognitive-behavioral approach: Results of a three-year study. *J Consult Clin Psycho1* 58:437-446, 1990.

Cohen, J. The cost of dichotomization. *Appl Psychol Meas* 7:249-253, 1983.

Cohen, P. A source of bias in longitudinal investigations of change. In: Collins, L.M., and Horn, J.L., eds. *Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions, Future Directions.* Washington, DC: American Psychological Association, 1991.

Collins, L.M., and Graham, J.W. Comments on "A source of bias in longitudinal investigations of change." In: Collins, L.M., and Horn, J.L., eds. *Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions, Future Directions.* Washington, DC: American Psychological Association, 1991.

Ellickson, P.L.; Bianca, D.; and Shoeff, D.C. Containing attrition in school-based research: An innovative approach. *Eval Rev* 12:33l-351, 1988.

Goodman, R.M.; Smith, D.W.; Dawson, L.; and Steckler, A. Recruiting school districts into a dissemination study. *Health Educ Res* 6:373-385, 1991.

Hansen, W.B. School-based substance abuse prevention: A review of the state-of-the-art in curriculum. *Health Educ Res* 7:403-430, 1992.

Hansen, W.B.; Collins, L.M.; Malotte, C.K.; Johnson, C.A.; and Fielding, J.E. Attrition in prevention research. *J Behav Med* 8:261-275, 1985.

Hansen, W.B.; Graham, J.W.; Sobel, J.L.; Shelton, D.R.; Flay, B.R.; and Johnson, C.A. The consistency of peer and parent influences on tobacco, alcohol, and marijuana use among young adolescents. *J Behav Med* 10:559-579, 1987.

Hansen, W.B.; Graham, J. W.; Wolkenstein, B.H.; Lundy, B.Z.; Pearson, J.L.; Flay, B.R.; and Johnson, CA. Differential impact of three alcohol prevention curricula on hypothesized mediating variables. *J Drug Educ* 18:143-153, 1988.

Hansen, W.B.; Graham, J.W.; Wolkenstein, B.H.; and Rohrbach, L.A. Program integrity as a moderator of prevention program effectiveness: Results for fifth grade students in the adolescent alcohol prevention trial. *J Stud Alcohol* 52:568-579, 1991.

Hansen, W.B.; Tobler, N.S.; and Graham, J.W. Attrition in substance abuse prevention research: A meta-analysis of 85 longitudinally-followed cohorts. *Eval Rev* 14:677-685, 1990.

Lipsey, M.W. *Design Sensitivity: Statistical Power for Experimental Research.* Newbury Park, CA: Sage Publications, 1990.

O'Hara, N.M.; Brink, S.; Harvey, C.; Harrist, R.; Green, B.; and Parcel, G. Recruitment strategies for health promotion research. *Health Educ Res* 6:363-371, 1991.

Pentz, M.A.; Trebow, E.A.; Hansen, W.B.; MacKinnon, D.P.; Dwyer, J.H.; Johnson, C.A.; Flay, B.R.; Daniels, S.; and Cormack, C. Effects of program implementation on adolescent drug use behavior: The Midwestern Prevention Project (MPP). *Eval Rev* 14:264-289, 1990.

Pirie, P.; Murray, D.M.; Peterson, A.V.; Thomson, S.J.; Mann, S.L.; and Flay, B.R. Tracking and attrition in longitudinal school-based smoking prevention research. *Prev Med* 18:249-256, 1989.

Tobler, N.S. "Meta-Analysis of Adolescent Drug Abuse Prevention Programs." Paper presented at the National Institute on Drug Abuse Technical Review on Meta-Analysis of Drug Abuse Prevention Programs, Bethesda, MD, July 26-27, 1993.

Zimmerman, D.W., and Williams, R.H. Note on the reliability of experimental measures and the power of significance tests. *Psychol Bull* 100:123-124, 1986.

AUTHORS

William B. Hansen, Ph.D.
Associate Professor
Department of Public Health Sciences
Bowman Gray School of Medicine
Medical Boulevard
Winston-Salem, NC 27157-1063

Linda M. Collins, Ph.D.
Professor
Department of Human Development and Family Studies
The Pennsylvania State University
University Park, PA 16802-6504