INTRODUCTION

There are several purposes to this article, the first of which is an overview of power analysis: what it is, why it is important, and how to calculate it. The second purpose is the relative importance of power analysis to adequate survey return rates. While these two issues could be learned elsewhere (e.g., various research methods texts and journal articles), this article provides those readers who are less familiar with power analysis a summary of the key points as they relate to health education survey research. The third purpose of this article is to assess the use of power analysis in seven leading health education journals. This article is directed at readers unfamiliar with power analysis, as well as those who are better versed in its use, with the intent being to increase the appropriate use of power analysis in health education survey research.

Theory of Power Analysis

Anytime a researcher conducts a quantitative study, it is essential that the researcher calculate the statistical power of a study before any data are collected, with the possible exception of pilot studies. In fact, grant proposals to some federal agencies require that a power analysis be conducted before the proposal is submitted. A statistical power assessment tells us how likely it is that a statistical significance test (e.g., t-test, ANOVA, chi-square) will detect a significant difference between two or more groups, given that a difference actually exists. In other words, statistical tests attempt to disprove the null hypothesis that there is no difference or no association between or among various samples. Rejection of a null hypothesis means that a difference or an association may be inferred from the study sample to the population.

Using statistical significance tests to assess data from a study can result in several different outcomes (Figure 1). In the first cell (A), we see that the null hypothesis is
false in the population and if our study results find the null hypothesis to be false, we obtained a correct outcome. In this case, we find support for a hypothesis that says there is/are difference(s) between/among groups or an association between the variable(s) is/are difference(s) between/among groups.

In this case, we find the null hypothesis to be false, we will be correctly rejected. A higher power (e.g., 0.85, 0.90) would always be preferred, if possible. Both statistical significance and statistical power are influenced by the size of a sample. Under powered studies (e.g., too small sample size) are frequently the reason for not detecting differences between/among groups in a study. It is also possible to have the power of a study so high that very minor differences are detected as statistically significantly different, but in which the differences have no practical implications.

In the fourth cell (D), the example study results correctly support the population null hypothesis. Thus, there are two potentially correct, but different, outcomes when conducting a study (Figure 1): correct rejection or correct acceptance of the null hypothesis.

Most studies in the health education arena are more likely to be under-powered, rather than over-powered. In other words, because of time and costs, more health education researchers will use smaller samples (i.e., a few hundred subjects) rather than very large samples (i.e., 3,000 to 10,000 subjects). It should be noted that a case has been made in the professional literature to suggest that under-powered studies are unethical. This is, in part, due to research subjects being inadequately informed about the potentially limited value of being part of a study in which the research may not be able to detect important statistically significant effects.

**Forms of Power Analysis**

Statistical power is influenced by four factors: the level of statistical significance ($\alpha$); the effect size—the magnitude of the difference between the two sample groups being examined on a specific outcome variable; the variance of the responses to the outcome variable; and the size of the sample. The only factor that logically can be modified at the beginning of a study is the size of the sample. The second component, effect size (ES), is not known but needs to be estimated. Effect size often can be estimated from a review of the published literature, a pilot study can give an estimate, and one can use a “guesstimate” by using general effect sizes proposed by well known researchers in this field (e.g., Jacob Cohen). It is recommended that collaboration with a...
The second form of power analysis is when a researcher wants to be able to generalize the results of his/her sample to the population from which the sample was drawn. To determine this sample size, researchers need to know the following: how much sampling error, the size (n) of the population, how much variation there is in the population with respect to the outcome variable being studied; and the smallest subsample in the sample for which sample size estimates are needed. Table 1 provides sample sizes necessary to be able to generalize the sample results to the population given a variety of sampling errors, population sizes, and variation in the variable under study. For example, if one wanted to survey a community regarding firearm control and the researcher knew that the population had evenly split (50/50) perceptions regarding support for a ban on the sale of handguns to the general public, and the population of the community was 50,000 people, and one wanted the responses to the survey to have only a +/- 3% sampling error, then one would need a sample of 1,045 completed surveys. However, if the researcher was willing to have a larger sampling error, for example 5%, then one would need only 381 completed surveys. In other words, using the 5% sampling error column (and the 50,000 population row), this would mean that if the gun control survey found that 63% of the population supported eliminating the sale of handguns to the public, then one could be sure 95% of the time that, with a random sample of 381 individuals, the entire 50,000 adults believe the same results within a +/- 5% range (58% to 68%).

From Table 1, it can be seen that in very large populations (e.g., 100,000 or more) the samples needed are about the same size regardless of the size of the population. However, when a researcher is examining a population of 5,000 or less, then the sample size needed is a much larger portion of the total population. Also, it should be noted that the more diverse the beliefs in a population, the larger the sample size needed.

### Power Analysis Versus Survey Return Rates

The use of power analysis for determining sample size is needed for calculating statistical analyses and for appropriate generalization to the population. The latter of these, generalizing to the population (external validity), requires an additional consideration: the survey return rate. When the concern is the ability to generalize to the population, power analysis is important as an initial step to determine the number of completed and usable surveys needed. This needs to be taken a step further, however.

Suppose that power analysis was conducted to determine the number of usable surveys needed to be returned to generalize to a population of 5,000 (with 95% confidence, 50/50 split, and plus or minus 3% error). The number of completed surveys needed in this example is 880. If Survey A was sent to a sampling frame of 3,000 (of the 5,000) and 880 were returned, the needed number of surveys was achieved but with a return rate of 29.3% (880/3,000). In another example, Survey B was sent to a sampling frame of 1,500 (of the 5,000) and 880 were returned for a rate of 58.7% (880/1,500). Which situation is better? The answer depends on two issues: potential for sampling bias and potential for response bias.

Sampling bias occurs when the sample is obtained in such a manner that the sample is different from the population regarding characteristics important to the study. Sampling bias can be investigated if data are available from the population related to the subject matter being studied. In most cases in the health education arena, it may not be possible to have this information. Thus, the investigation of sampling bias is assessed based on the quality of the methods used to obtain a representative sample of the population.
Response bias occurs when the people responding to the survey are different from those not responding to the survey in regards to the subject of interest. In our previous handgun example, this could be a situation where members of the National Rifle Association (NRA), a conservative gun ownership support group, responded to the questionnaire more often than people who are not members of the NRA. This can be investigated by seeking out a sample of non-respondents and trying to collect the information originally sought. The extent to which those who responded were different from those who did not respond represents the magnitude of the response bias.

In the aforementioned examples, if both Survey A and Survey B were free from sampling bias and response bias, then the external validity of the responses of Survey A would be equal to the external validity of the responses of Survey B. Thus, the difference in the survey return rates would not be important when generalizing the results to the population (e.g., both have good external validity).

If both surveys contained sampling bias but were free from response bias, then Survey A would be better than Survey B. This is because the sampling frame of Survey A contained a larger portion of the entire population (3,000/5,000 [60%]) than Survey B (1,500/5,000 [30%]). A larger portion of the population included in the sampling frame increases the probability that the varied perceptions in the population are included in the responses of the sample. Because response bias does not exist in either survey in this example, the smaller sampling frame in Survey B is more likely to negatively impact the generalizability of the responses of the sample.

If both surveys were free from sampling bias (e.g., both were randomly selected) but they each had a response bias, then Survey B would be better than Survey A. Without sampling bias, the sampling frames for each survey were likely to be representative of the population. Thus, the ability to generalize the responses of the sample varies based on how well the people who respond to the survey represent the potential responses of the subjects composing the sampling frame. While both surveys have response bias, the magnitude of the impact from the response bias is greater in Survey A because two-thirds of the sampling frame did not respond. This is in contrast to Survey B where only one-third of the sampling frame did not respond. Thus, in this example, the survey return rate plays an important role in the ability to generalize the sample results to the population.

The importance of survey return rates already has been examined. However, of equal importance in assessing the quality of survey research is understanding the appropriate use of the size of samples (power analysis). Thus, another purpose of this manuscript is to examine the use of power analysis in health education research.

METHODS

Journals

Seven leading journals in the field of health education were studied to assess the reporting of power analysis. Criteria for journal selection included: health education orientation, a general nature instead of topic-specific (e.g., Journal of Drug Education), and availability in at least 25% of college and university libraries. The seven journals included in the sample were (in alphabetical order): American Journal of Health Behavior, American Journal of Health Education, American Journal of Health Promotion, Health Education & Behavior, Health Education Research, Journal of American College Health, and Journal of School Health. Power analysis deficiencies in articles in these journals potentially would have a major impact on health education research. Data were collected from the journals for the years 2000 through 2003, representing a span of four years.

Instrument

The selected journals were reviewed for articles meeting the criteria of a quantitative research article. These articles included Likert-type surveys, tallies, and other surveys containing data that could contain quantitative statistical analyses. Excluded articles included qualitative articles, review articles, editorials, and column articles that were not main articles (i.e., book reviews, letters from the editor, etc.).

The reviewers examined the methods sections of the selected articles, which were then recorded on a simple scoring sheet developed specifically for this project. The data recorded included: journal name and year, total number of main articles, total number of quantitative articles, and percentage of quantitative articles in which a power analysis was performed. Power analysis included any author self-reports of a priori power analysis to detect a statistical difference or to generalize the study findings to the population. In the event that the author of an article did not state that a power analysis was performed, the reviewers instead searched for key words and phrases indicating the potential use of a power analysis. These words included “sample size calculation,” “Cohen’s effect size,” and formulas and diagrams with power calculations. If the author of the article did not perform a power analysis prior to the study, but mentioned it in the limitations section, the article was not counted as containing a power analysis.

Analysis

Analysis of the data consisted of descriptive data, namely, frequencies, percents, and means. To assess accuracy of identifying reported survey return rates, a sample of two different journals was used and a Kappa coefficient was calculated to assess inter-rater reliability among the three journal reviewers. The Kappa coefficient was used to compensate for chance agreement of the “yes” or “no” assessments. The mean Kappa coefficient was 0.905.

RESULTS

Power analyses were rare in the seven health education journals (Table 2). Over the years 2000 through 2003, the average power analysis ranged from a high of 25%
of the quantitative research articles in the American Journal of Health Behavior to a low of 1% in the Journal of American College Health. Four (American Journal of Health Promotion, Health Education & Behavior, Journal of American College Health, and Journal of School Health) of the seven journals had power analyses of less than 5% of their quantitative research articles.

**DISCUSSION**

The current study has confirmed in
health education what has been found in other research fields, such as nursing and health psychology12,13; that few researchers are using a priori statistical power analysis. While it is not evident from this study why health education researchers, manuscript reviewers, and journal editors continue to discount this important attribute of quality research, it is likely that there are multiple reasons. One reason may be that many researchers are unfamiliar with the importance and appropriate use of power analysis in survey research. This would indicate a lack of training in health education programs pertaining to power analysis. Graduate programs in health education could help to remedy this issue by including units on power analysis in their research methods courses. Most health education researchers engage in research for altruistic reasons, such as to advance the field of health education and/or to advance the skills of graduate students. Thus, it is critically important to the quality of health education research that both graduate students (our future researchers) and our peers be better informed about power analysis.

Another reason for the lack of power analyses done in health education research could be that sample sizes based on appropriate power analysis would sometimes require larger samples than are seen in published health education research. This would require greater financial investment and/or time investment. These researchers may not consider power analysis to be essential when compared to tradeoffs for time and financial investment due to larger sample sizes. However, not to use power analysis can result in important hypotheses not being supported by underpowered research. For example, suppose a health education researcher investigated the effectiveness of a curriculum to increase the physical activity of students. In the evaluation, the researcher surveyed 150 students when 250 students would have been required, based on an appropriate power analysis calculation. The results of the evaluation conclude that there were no statistically significant differences between the intervention and control group. Because a power analysis was not conducted, one would be less confident in the findings. Due to the greater possibility of a Type II error, the curriculum may indeed be effective at increasing physical activity. By not conducting a power analysis and using the appropriate sample size, the evaluator/researcher may have wasted limited resources on an evaluation that has little to offer. Furthermore, the evaluator may be reporting a curriculum as ineffective when, in fact, it may have been very effective. In other words, underpowered studies can result in important research findings not being found. Effective interventions overlooked due to underpowered assessments could result in a serious problem for the health education field. To help reduce this problem in health education, researchers need to calculate power analysis before conducting studies or evaluations and then include the information on how sample size decisions were made when they report their findings.

Finally, the limitations of this study should be explored before accepting the results. First, it may have been that more published research studies than found in the current study actually were based on a priori power analysis, but the authors of the studies failed to report the analysis. Second, the authors of some studies may intuitively have used large enough samples such that power analysis would not have changed the sample size. However, guessing at adequate size samples could have led to overpowered studies and statistically significant trivial results. Third, the current analysis of statistical power simply examined whether a power analysis was reported; it did not attempt to assess if the power analysis was adequately conducted. Fourth, it may be that health education research published in journals with higher-impact factors may be reporting power analyses. Even if this were so, it would not appear to justify the limited use of power analysis in the majority of health education journals.

REFERENCES