Research in Nursing & Health, 2008, 31, 180–191

Considerations in Determining Sample Size for Pilot Studies

Melody A. Hertzog*

College of Nursing, University of Nebraska Medical Center, Lincoln Division, 1230 "O" Street, Suite 131, P.O. Box 880220, Lincoln, NE 68588-0220

Accepted 12 August 2007

Abstract: There is little published guidance concerning how large a pilot study should be. General guidelines, for example using 10% of the sample required for a full study, may be inadequate for aims such as assessment of the adequacy of instrumentation or providing statistical estimates for a larger study. This article illustrates how confidence intervals constructed around a desired or anticipated value can help determine the sample size needed. Samples ranging in size from 10 to 40 per group are evaluated for their adequacy in providing estimates precise enough to meet a variety of possible aims. General sample size guidelines by type of aim are offered. © 2008 Wiley Periodicals, Inc. Res Nurs Health 31:180–191, 2008

Keywords: sample size; pilot studies; effect size

Pilot studies can serve many purposes. Pilot study aims suggested by Prescott and Soeken (1989) based on a review of then-current nursing research textbooks included assessment of (a) feasibility, (b) adequacy of instrumentation, and (c) problems of data collection strategies and proposed methods. To these they added: (d) answering methodological questions, and (e) planning a larger study. In a more recent article, Jairath, Hogerney, and Parsons (2000) contributed a sixth potential use of a pilot study: (f) obtaining sufficient preliminary data to justify a grant award. Although some researchers might question whether this type of preliminary work is a true pilot study, it is referred to as such frequently enough in practice to be included for the purposes of this article.

Pilot studies thus range from relatively informal trying out of procedures on a handful of participants, to efficacy studies, or to small-scale clinical trials of interventions. How, then, does a researcher determine an appropriate sample size for a pilot study? Newer editions of several of the nursing textbooks cited by Prescott and Soeken (1989) offer limited guidance with respect to sample size for pilot studies. Burns and Grove (2005) and Polit and Beck (2004) make no specific recommendations. Others recommend obtaining approximately 10 participants (Nieswiadomy, 2002) or 10% of the final study size (Lackey & Wingate, 1998), the final decision to be guided by cost and time constraints as well as by size and variability of the population. Similar recommendations can be found in texts in other areas of clinical research, such as epidemiology (Hulley et al., 2001).

Informal guidelines based on the experience of seasoned researchers are probably sufficient for investigating the feasibility of procedures or methods. For example, implementing procedures with even a few cases is likely to be very informative with respect to difficulty of recruitment, acceptability of the intervention, logistics, and costs. When funding for the pilot study is being sought, review panels tend to expect further justification of its sample size. The researcher who turns to reference texts on power analysis (Cohen, 1977) or sample size determination in clinical studies (Chow, Shao, & Wang, 2003; Lwanga & Lemeshow, 1991; Machin, Campbell,

Correspondence to Melody A. Hertzog.
*Assistant Professor.
Published online 8 January 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/nur.20247

Fayers, & Pinol, 1997) will find that these sources focus exclusively on the design of fully powered studies. A search of periodicals on the topic of sample size and pilot studies will yield literature concerning internal pilot studies, also called sample size reestimation procedures. These use data from the first *n* cases of a clinical trial to more accurately estimate the total sample size needed for the study, techniques that are not applicable to situations in which the pilot and clinical trial are separate.

This lack of attention to pilot study size does not mean that we must always rely solely on the rough guidelines described above. Many of the aims listed involve estimation of a population value, but such estimates vary considerably across samples, becoming increasingly less precise as sample size decreases. The primary purpose of this article is to illustrate how a reasonable lower limit for sample size in a pilot study can be determined by using widely available methods to construct confidence intervals (CIs) around the estimates appropriate for a given aim. A secondary purpose is to encourage use of these methods in interpreting results from studies of any size. The discussion is divided into four major sections: feasibility, assessment of instrumentation, planning a larger study, and providing preliminary information for a grant proposal. As defined in Prescott and Soeken (1989), answering methodological questions could be considered an aspect of feasibility and will be discussed with that aim. Intervention efficacy sometimes is considered as a feasibility aim, but an efficacy study provides information more specifically related to planning a larger study and providing preliminary data for grant proposals and so will be included in those sections.

REFERENCE EXAMPLES

To keep computations simple and results easy to compare across a variety of statistics, the examples in this article that involve group comparisons use a between-subjects design having a single fixed factor with two levels. The discussions of effect size estimation and power then extend the results to factorial, single-factor within-subject, and split-plot designs. All reference examples use hypothetical sample sizes of 10–40 participants per group. The lower end of this range represents 10% of the typical size of a fully powered clinical trial comparing an intervention with a control group. The upper bound was chosen based on experience that a pilot study of more than 40 per group is likely to be unrealistic in terms of time and cost, and, in some cases, would not be an optimal use of a limited sample of participants available for a study. As an effect size measure, η^2 (eta-squared) or ω^2 (omega-squared) will be used because they can be generalized to more complex designs, although Cohen's *d*, a standardized mean difference, could have been used as an effect size measure in this simple two-group case. For details of the calculations of all effect size measures mentioned in this article and formulas for converting them, the reader is referred to Cohen (1992) or Fern and Monroe (1996). Given the size limitations and exploratory nature of a pilot study, conventional 95% confidence intervals may be unrealistically stringent, so 90% and 68% CIs will also be presented, with interpretation focusing on the most liberal level. All tabled confidence intervals are two-sided.

Feasibility

Feasibility encompasses a wide range of possible aims for identifying and resolving problems of implementing an intervention. Some problems are methodological and require estimation of population values important for planning the larger study. Examples include time to complete questionnaires, response rate for mailed surveys, rates of patient adherence, and percent attrition. For continuous variables (such as time to complete questionnaires), the desired CI around a mean value can be calculated as described in elementary statistics texts and produced by common statistical software. In the other examples cited, the estimate is a proportion, for which confidence intervals also may be calculated easily.

Unlike the standard error of a mean, the standard error of a proportion depends upon the value of the proportion itself, reaching its largest value when the proportion equals .50. This means that more extreme values are estimated with greater precision for a given sample size. Table 1 displays confidence intervals across the reference sample sizes for two values of a proportion: .15, which might arise in estimating attrition, for example, and .30, representing a plausible value for survey non-response. With an observed attrition rate of 15% in a pilot study with 20 participants, we can be 68% confident that our estimate is accurate within 8 percentage points. Doubling the sample size to 40 reduces the error in estimation to 6 percentage points. A further doubling to 80 results in a slightly smaller reduction in error to 5 points. Whether this gain in precision justifies the additional participants depends on the

Table 1. Limits (Lower, Upper) of Confidence Intervals for Proportions

N	p^a	CI_{95}	CI_{90}	CI_{68}
20 (10 per group)	.15	(.00, .31)	(.02, .28)	(.07, .23)
	.30	(.10, .50)	(.13, .47)	(.20, .40)
30 (15 per group)	.15	(.02,.28)	(.04, .26)	(.09, .22)
	.30	(.14, .46)	(.16, .44)	(.22, .38)
40 (20 per group)	.15	(.04, .26)	(.06, .24)	(.09, .21)
	.30	(.16, .44)	(.18, .42)	(.23, .37)
50 (25 per group)	.15	(.05, .25)	(.07, .23)	(.10, .20)
	.30	(.17, .43)	(.19, .41)	(.24, .36)
60 (30 per group)	.15	(.06, .24)	(.07, .23)	(.10, .20)
	.30	(.18, .42)	(.20, .40)	(.24, .36)
70 (35 per group)	.15	(.07, .23)	(.08, .22)	(.11, .19)
	.30	(.19, .41)	(.21, .39)	(.25, .35)
80 (40 per group)	.15	(.07, .23)	(.08, .22)	(.11, .19)
	.30	(.20, .40)	(.22, .38)	(.25, .35)

[a]Because the standard error of a proportion equals the square root of $[p(1-p)/n]$, it has the same value for p and for $1-p$. Therefore, the width of the confidence intervals tabled would also apply to proportions of .85 (i.e., 1–.15) and .70 (1–.3).

consequences of inaccuracy. For example, dropout rates may need to be estimated quite precisely because attrition will be explicitly taken into consideration in recruitment plans for a larger study. It is crucial that the final sample be adequately powered, but wise use of resources dictate that the sample be no larger than necessary. On the other hand, if adherence to a new protocol is being assessed, a value accurate within 10–15 percentage points, though very rough, might be sufficiently accurate to identify a problem that would justify modification of the intervention.

Adequacy of Instrumentation

For assessing clarity of instructions or item wording, acceptability of formatting, or ease of administration, a sample of 10 or even fewer may suffice. However, if the aims for a pilot are to estimate internal consistency or test–retest reliability or to assess item performance to evaluate or revise an instrument, such a small sample may be inadequate.

Evaluating item performance. The corrected item-total correlation can serve as an index of the ability of an item to represent performance on the total scale. Although the size of this index should be interpreted within the context of the breadth of the construct being measured (i.e., narrowly defined constructs tend to have higher item to total correlations than do broadly defined constructs) and compared to values of the index observed for other items on the same scale, .30 is often suggested as a minimum acceptable level (Nunnally & Bernstein, 1994). Table 2 shows CIs for correlations ranging from .30 to .60. Upper and lower limits were found by applying the standard procedure involving the Fisher z transformation outlined in many statistics textbooks and presented in a measurement context by Fan and Thompson (2001). An alternative strategy would be to use an on-line calculator such as SISA (Uitenbroek, 1997), which will calculate the limits for 95%, 90%, and 80% CIs using this same method. As can be seen from the width of the confidence intervals, the estimates are quite imprecise. Even with 40 participants per group and using the most liberal CI (68%), a discrimination index of .30 would have a lower limit of only .19. Goals of a pilot study generally would not include development of a new instrument, but quite commonly do include checking the performance of items on a previously developed instrument with a new population. Given the imprecision of these correlation estimates, it would not be advisable to make final decisions about item inclusion or deletion based on this criterion alone using pilot data.

Test–retest reliability. When estimating reliability by correlating scores from two administrations of a single instrument, the researcher usually seeks evidence that an observed correlation is at least .70, a lower bound to acceptable stability (Nunnally & Bernstein, 1994), and preferably that it exceeds .80 if it is an established instrument. These estimates are, of course, also somewhat dependent on the length of the interval between testing administrations. Table 3 contains limits of confidence intervals for the range of

Table 2. Limits (Lower, Upper) of Confidence Intervals for Pearson Correlation, r = .30–.60

N	*r*	CI_{95}	CI_{90}	CI_{68}
20 (10 per group)	.30	(−.16, .66)	(−.09, .61)	(.07, .50)
	.40	(−.05, .72)	(.03, .68)	(.18, .58)
	.50	(.07, .77)	(.15, .74)	(.30, .66)
	.60	(.21, .82)	(.29, .80)	(.42, .73)
30 (15 per group)	.30	(−.07, .60)	(−.01, .56)	(.12, .46)
	.40	(.05, .67)	(.11, .63)	(.23, .55)
	.50	(.17, .73)	(.23, .70)	(.34, .63)
	.60	(.31, .79)	(.36, .77)	(.46, .71)
40 (20 per group)	.30	(−.01, .56)	(.04, .52)	(.15, .44)
	.40	(.10, .63)	(.15, .60)	(.26, .53)
	.50	(.22, .70)	(.27, .68)	(.37, .61)
	.60	(.36, .77)	(.40, .75)	(.49, .69)
50 (25 per group)	.30	(.02, .53)	(.07, .50)	(.16, .43)
	.40	(.14, .61)	(.18, .58)	(.27, .51)
	.50	(.26, .68)	(.30, .66)	(.38, .60)
	.60	(.39, .75)	(.43, .73)	(.50, .68)
60 (30 per group)	.30	(.05, .51)	(.09, .48)	(.18, .41)
	.40	(.16, .59)	(.20, .57)	(.28, .50)
	.50	(.28, .67)	(.32, .65)	(.40, .59)
	.60	(.41, .74)	(.44, .72)	(.51, .68)
70 (35 per group)	.30	(.07, .50)	(.22, .47)	(.19, .41)
	.40	(.18, .58)	(.22, .55)	(.29, .50)
	.50	(.30, .66)	(.33, .64)	(.40, .59)
	.60	(.42, .73)	(.46, .71)	(.52, .67)
80 (40 per group)	.30	(.09, .49)	(.12, .46)	(.19, .40)
	.40	(.20, .57)	(.23, .54)	(.30, .49)
	.50	(.31, .65)	(.35, .63)	(.41, .58)
	.60	(.44, .72)	(.47, .71)	(.52, .67)

$r = .70–.80$. As is evident from the values for the most liberal confidence interval of those presented here, a test–retest correlation of .70 found in a small sample probably should be viewed as borderline. With 10 participants per group, the population value easily might be as high as .80 or as low as .56. Even with 40 per group, the lower bound is .64. With 35–40 per group and an observed value is .75, we could be reasonably confident that the population value would be .70 or above.

Cronbach's alpha. Internal consistency estimates of coefficient alpha are highly dependent upon item variances as well as upon their intercorrelations. If pilot data are used to check that reliability of a measurement tool is consistent with reported values or to support an instrument's use in a specific population, the researcher must consider whether the pilot sample exhibits variability representative of that in the new population. Too homogeneous a sample can result in low estimated alpha. Beyond this, the effect of sample size on precision of the estimate of alpha should be considered.

Using a method illustrated in Fan and Thompson (2001), upper and lower limits of confidence intervals for a selection of alpha levels were calculated and are presented in Table 4. There is some disagreement about the accuracy of these limits if the set of items does not follow a multivariate normal distribution (Bonett, 2002), but these intervals should be sufficiently precise for the purpose under discussion here.

The precision of coefficient alpha is also influenced by scale length (a twofold change in length changing interval limits by approximately .01), with the effect slightly larger for smaller sample sizes, but differences are so small that only values for a 20-item scale are reported here.[1] The values in Table 4 do not vary much across the range of sample sizes considered; for samples of 25–40 per group, observed alpha should probably be at least .75 in order to have reasonable

[1]Once data are collected, the researcher can use SPSS (2005) to compute limits of a confidence interval around alpha for that specific data by requesting computation of an intraclass correlation for a random effects model as an option when calculating reliability.

Table 3. Limits (Lower, Upper) of Confidence Intervals for Pearson Correlation, r = .70–.80

N	*r*	CI_{95}	CI_{90}	CI_{68}
20 (10 per group)	.70	(.37, .87)	(.44, .85)	(.56, .80)
	.75	(.46, .90)	(.52, .88)	(.62, 84)
	.80	(.55, .92)	(.60, .91)	(.70, .87)
30 (15 per group)	.70	(.45, .85)	(.50, .83)	(.59, .78)
	.75	(.53, .87)	(.58, .86)	(.65, .82)
	.80	(.62, .90)	(.65, .89)	(.72, .86)
40 (20 per group)	.70	(.50, .83)	(.54, .81)	(.61, .77)
	.75	(.57, .86)	(.61, .85)	(.67, .81)
	.80	(.65, .89)	(.68, .88)	(.73, .85)
50 (25 per group)	.70	(.52, .82)	(.56, .80)	(.62, .77)
	.75	(.60, .85)	(.62, .84)	(.68, .81)
	.80	(.67, .88)	(.70, .87)	(.74, .85)
60 (30 per group)	.70	(.54, .81)	(.57, .80)	(.63, .76)
	.75	(.61, .84)	(.64, .83)	(.69, .80)
	.80	(.69, .88)	(.71, .87)	(.75, .84)
70 (35 per group)	.70	(.56, .80)	(.58, .79)	(.63, .76)
	.75	(.63, .84)	(.65, .83)	(.69, .80)
	.80	(.70, .87)	(.72, .86)	(.75, .84)
80 (40 per group)	.70	(.57, .80)	(.59, .78)	(.64, .75)
	.75	(.63, .83)	(.66, .82)	(.70, .80)
	.80	(.70, .87)	(.72, .86)	(.76, .84)

Table 4. Limits (Lower, Upper) of Confidence Intervals for Cronbach's Alpha for a 20-Item Scale

N	Cronbach's α	CI_{95}	CI_{90}	CI_{68}
20 (10 per group)	.70	(.47, .86)	(.52, .84)	(.60, .80)
	.75	(.56, .88)	(.60, .87)	(.67, .83)
	.80	(.65, .91)	(.68, .90)	(.73, .86)
	.85	(.74, .93)	(.76, .92)	(.80, .90)
30 (15 per group)	.70	(.52, .84)	(.55, .82)	(.62, .78)
	.75	(.60, .86)	(.63, .85)	(.68, .82)
	.80	(.68, .89)	(.70, .88)	(.75, .85)
	.85	(.76, .92)	(.78, .91)	(.81, .89)
40 (20 per group)	.70	(.55, .82)	(.58, .80)	(.63, .77)
	.75	(.62, .85)	(.65, .85)	(.69, .81)
	.80	(.70, .88)	(.72, .87)	(.75, .85)
	.85	(.77, .91)	(.79, .90)	(.82, .88)
50 (25 per group)	.70	(.57, .81)	(.59, .79)	(.64, .76)
	.75	(.64, .84)	(.66, .83)	(.70, .80)
	.80	(.71, .87)	(.73, .86)	(.76, .84)
	.85	(.78, .90)	(.80, .86)	(.82, .88)
60 (30 per group)	.70	(.58, .80)	(.60, .79)	(.64, .76)
	.75	(.65, .83)	(.67, .82)	(.70, .80)
	.80	(.72, .87)	(.73, .86)	(.76, .84)
	.85	(.79, .90)	(.80, .89)	(.82, .88)
70 (35 per group)	.70	(.59, .79)	(.61, .78)	(.65, .75)
	.75	(.66, .83)	(.67, .82)	(.71, .79)
	.80	(.73, .86)	(.74, .85)	(.77, .83)
	.85	(.79, .90)	(.80, .89)	(.82, .88)
80 (40 per group)	.70	(.60, .79)	(.61, .78)	(.65, .75)
	.75	(.66, .82)	(.68, .81)	(.71, .79)
	.80	(.73, .86)	(.74, .85)	(.77, .83)
	.85	(.80, .89)	(.81, .89)	(.83, .87)

confidence that the population value is at least .70. Samples having fewer than 25 participants per group need observed alpha to be close to .80 to achieve this.

Planning a Larger Study

Estimating effect sizes. Ideally, judgments of clinical importance and effect size estimates reported in the literature should inform power analyses, but estimates from data collected using the exact design and instrumentation of a planned study are especially valuable. They help support an argument that a particular intervention is likely to be able to produce an effect of a given size. Unfortunately, estimates of population effect sizes based on very small samples are known to be positively biased (i.e., too large), and the problem is magnified for small effect sizes. Methods of bias correction vary, depending on the specific effect size measure used. For example, η^2 (eta-squared, the proportion of total variance explained by an effect) is a biased estimator, but ω^2 (omega-squared) attempts to correct for bias in the estimate by taking into account sample size and the number of factor levels. To illustrate the degree of bias that might be expected in an estimate based on pilot data, ω^2 was calculated for a two-group design having a total sum-of-squares redistributed to yield η^2 values corresponding to Cohen's often cited small (.01), medium (.06), and large (.14) effect sizes (Cohen, 1977). The values in Table 5 suggest that 15–20 participants per group may give reasonable bias-corrected estimates for medium to large effects, but even 40 per group is insufficient if the population effect size is small.

Not only are the effect size estimates from small samples biased, they are quite imprecise. Calculation of confidence intervals for effect sizes, although recommended (Thompson, 2002), involves use of noncentral distributions (Cumming & Finch, 2001) and is not implemented in most widely available statistical packages. Various methods have been proposed for calculating exact and approximate confident intervals for effect sizes. Interval limits for small, medium, and large values of η^2 shown in Table 6 were obtained using SPSS (2005) scripts that implement the method proposed by Smithson (2001a). This set of scripts, which will also calculate confidence limits for Cohen's *d*, can be downloaded from its author's website (Smithson, 2001b) at no cost. Table 6 illustrates that even a large observed effect size in a small sample is consistent with a very small to medium population effect size. With 40 participants per group, the lower limit of the 68% CI for a large effect size is only $\eta^2 = .07$, a medium effect. The lower limit of the confidence interval nevertheless could be used to help define a range of plausible values to use in power analyses for future studies, thus perhaps helping to avoid an underpowered study.

In practice, a researcher should estimate confidence limits around a bias-corrected effect size, so the confidence limits in Table 6 should be lowered by the appropriate correction for bias from Table 5 (for example, with 15 per group and η^2 of .14, the bias-corrected confidence interval would extend from .01 to .22). Results of each of these steps were presented separately here to illustrate the relative influence of each, but modifications of Smithson's scripts (Fidler & Thompson, 2001) may be used to produce confidence intervals already incorporating the bias correction. If the aim is to estimate effect sizes, none of the sample sizes examined here will give more than extremely rough estimates, even for large observed effect sizes. The improvement seen in doubling the pilot sample from 20 to 40 is minimal for moderate effect sizes (bias-corrected limits of .00–.11 for

Table 5. Effect Size (η^2) Corrected for Bias (ω^2)

	Corrected for bias (ω^2)		
N	$\eta^2 = .01$	$\eta^2 = .06$	$\eta^2 = .14$
20 (10 per group)	0[a]	.01	.09
30 (15 per group)	0[a]	.03	.11
40 (20 per group)	0[a]	.03	.12
50 (25 per group)	0[a]	.04	.12
60 (30 per group)	0[a]	.04	.12
70 (35 per group)	0[a]	.05	.13
80 (40 per group)	0[a]	.05	.13

[a]As a proportion of variance, η^2 is always positive, but applying the correction for bias can result in small negative values, which are customarily treated as zero.

Table 6. Limits (Lower, Upper) of Confidence Intervals for Effect Sizes (η^2)

N	η^2	CI_{95}	CI_{90}	CI_{68}
20 (10 per group)	.01	(.00, .21)	(.00, .17)	(.00, .07)
	.06	(.00, .32)	(.00, .27)	(.00, .17)
	.14	(.00, .41)	(.00, .37)	(.02, .27)
30 (15 per group)	.01	(.00, .17)	(.00, .13)	(.00, .06)
	.06	(.00, .27)	(.00, .23)	(.00, .15)
	.14	(.00, .37)	(.00, .33)	(.04, .25)
40 (20 per group)	.01	(.00, .14)	(.00, .11)	(.00, .06)
	.06	(.00, .24)	(.00, .21)	(.01, .14)
	.14	(.00, .34)	(.01, .30)	(.05, .23)
50 (25 per group)	.01	(.00, .12)	(.00, .10)	(.00, .05)
	.06	(.00, .22)	(.00, .19)	(.01, .13)
	.14	(.01, .32)	(.02, .29)	(.06, .23)
60 (30 per group)	.01	(.00, .11)	(.00, .09)	(.00, .05)
	.06	(.00, .20)	(.00, .18)	(.01, .12)
	.14	(.02, .30)	(.03, .27)	(.06, .22)
70 (35 per group)	.01	(.00, .10)	(.00, .08)	(.00, .04)
	.06	(.00, .19)	(.00, .17)	(.02, .12)
	.14	(.02, .29)	(.04, .27)	(.07, .21)
80 (40 per group)	.01	(.00, .09)	(.00, .07)	(.00, .04)
	.06	(.00, .18)	(.00, .16)	(.02, .12)
	.14	(.03, .28)	(.04, .26)	(.07, .21)

20 per group and .01–.11 for 40 per group), though somewhat greater when the effect size is large (the bias-corrected interval of .03–.21 narrows to .06–.20).

Extensions to other designs. A one-factor, between-subjects design was used in the simple example above, but in cross-sectional designs, a factorial design with additional between-subjects factors is common. Effect sizes estimated for use in power calculations should be calculated using a denominator that reflects the same sources of variability as those that will be later be used in the statistical test of that effect. In a factorial design, partial η^2 would replace the η^2 in the example and would be calculated removing the variability explained by other factors in the design. As with η^2, the partial η^2 should be corrected for bias by calculating partial ω^2 (see sources cited earlier for formulas). Partial η^2 or partial ω^2 are also appropriate for quantifying a treatment effect that has been adjusted for confounding variables, as in analysis of covariance, for example.

With longitudinal data, a commonly used design in nursing research is a split-plot or mixed design with intervention group as a between-subjects factor and repeated measurements over time as a within-subjects factor. In this design, the effect of intervention and the within-subject correlation are confounded, the distributions of effect sizes are complex, and tools to compute exact confidence intervals are not readily available (Kline, 2004). One solution is to use bootstrap methods to construct the desired confidence intervals (Steiger, 2004). A nonparametric bootstrap method involves resampling cases with replacement independently for each group in an obtained dataset, repeating the process a large number of times. The statistic of interest (here, an effect size) is calculated for each replication and its values accumulated across replications to create an empirical distribution of the statistic. Values in this distribution then are rank ordered and percentiles corresponding to the desired level of confidence used as estimates of the interval limits.

To illustrate this method, six datasets were generated, each having two equal-size independent groups of size 20, 30, or 40 with three repeated measurements. Observations for each group were drawn from a multivariate normal population with a correlation among repeated measures of .60. The sample data were transformed to produce observed effect sizes of approximately $\eta^2 = .06$ or $\eta^2 = .14$ for both the Group effect and the Group $\times$ Time effect, generally the two effects in this design of greatest interest in intervention effectiveness studies. Five hundred bootstrap samples were produced from these six generated datasets for each group using a SAS macro (Tonidandel, 2004), then a repeated-measures ANOVA was run on each bootstrap sample in order to calculate the observed effect sizes in each replication. Ideally, a larger number

Table 7. Limits (Lower, Upper) of Bootstrap Confidence Intervals for Group and Group × Time Effect Sizes (η^2) in RM-ANOVA Design

N	η^{2a}	Effect	CI_{95}	CI_{90}	CI_{68}
20 per group	.06	Group	(.00, .30)	(.00, .26)	(.01, .17)
		G × T	(.00, .23)	(.01, .21)	(.03, .15)
	.14	Group	(.01, .40)	(.02, .37)	(.07, .27)
		G × T	(.04, .34)	(.05, .31)	(.09, .25)
30 per group	.06	Group	(.00, .25)	(.01, .21)	(.02, .15)
		G × T	(.01, .19)	(.02, .16)	(.04, .12)
	.14	Group	(.02, .36)	(.04, .31)	(.08, .24)
		G × T	(.05, .29)	(.07, .26)	(.10, .22)
40 per group	.06	Group	(.00, .21)	(.00, .17)	(.02, .12)
		G × T	(.01, .16)	(.02, .14)	(.04, .11)
	.14	Group	(.02, .31)	(.04, .28)	(.08, .22)
		G × T	(.06, .26)	(.08, .24)	(.11, .20)

Note: Results are based on 500 replications.
[a]Group × Time effect size is partial η^2, removing the effect of time.

of replications (at least 1,000) would be performed, but a smaller sample was deemed sufficient to provide the rough estimates needed here.

The bootstrap confidence intervals are presented in Table 7. The confidence intervals are consistently narrower for the Group × Time interaction than for the Group effect. This would be expected, as the correlation among measurements increases power and precision more for within-subjects factors than for between-subjects factors. In addition, the lower limits of the confidence intervals for the effect of group are higher with this design than they were for the same effect size when samples were independent (Table 6), suggesting that pilot studies perhaps could be slightly smaller when this design is employed. Nevertheless, using the most liberal CI (68%), it is evident that, as with the independent-groups design, even group sizes of 40 will be barely adequate when the effect size is moderate.

A single-factor within-subjects design, such as might be used in an efficacy study with a single group being measured before and after treatment, is another common design in pilot studies. Ordinarily the aim of an efficacy study is to demonstrate change following implementation of an intervention as a preliminary step before comparing the intervention to an alternative. Although estimates of the mean difference and variance from an efficacy study may be useful in planning the two-group study, the effect size for the change within a single group may not correspond to the between-group effect that will be tested in the later study. Considering that direct effect size estimation is not the ultimate goal of an efficacy study, the confidence interval around the effect size will not be discussed here (as with the split-plot design discussed earlier, it is not easily calculated, but could be estimated using bootstrap methods). Most efficacy studies have additional aims such as assessing feasibility of methods, evaluating tools, estimating correlations, estimating variability, or providing preliminary information for a grant proposal. Sample size decisions for an efficacy study might be guided by results relevant to these other aims and are presented elsewhere in this article.

Estimating variances. Pilot data are sometimes used to estimate a within-group variance (or standard deviation) to use in power calculations. For example, the researcher might wish to plan a study with sufficient power to detect an effect size that corresponds to the smallest difference that can be justified as clinically important. In such a case, the means from the pilot study are irrelevant to the power calculations, but the size of the pilot sample is still important because it influences the accuracy of the variance estimate. As is true for all the statistics discussed in this article, the variability of the sample must be representative of the population variability, a condition that might be difficult to achieve with a very small pilot sample from a heterogeneous population. In addition, the standard deviation in a small sample tends to be negatively biased, underestimating the standard deviation in the population, and if used in sample size calculations for later studies, is likely to result in an underpowered study. In his review of 30 clinical trials published along with their pilot data in top medical journals, Vickers (2003) found this to be true 80% of the time. The degree of

underestimation was substantial, with half of the analyses needing at least 64% more patients than had been estimated in order to have adequate power based on the standard deviation found in the full study. This is consistent with findings in a simulation study by Browne (1995) that using the sample variance from a pilot study results in less than a 50% chance of having the planned power in the full study. Browne suggested that using the upper limit of the $100 \times (1-\gamma)$ percent confidence interval of the population variance in the formula for sample size calculation would guarantee the planned power with a probability of at least $1-\gamma$. In an article further investigating the theoretical basis for Browne's recommendation, Kieser and Wassmer (1996) concluded that for clinical trials of 80–250 participants, total pilot sample sizes of 20–40 will be adequate for applying Browne's method in a power analysis.

Estimating correlations. The aims of a pilot study may include estimation of correlations among predictors and/or outcomes. In addition to the situation where such relationships are of theoretical interest, estimates of within-group correlations might also be needed as input for power calculations if future analyses include repeated-measures ANOVAs, ANCOVAs, growth curve models, or various multivariate analyses. Results presented in Tables 2 and 3 can be used to evaluate the precision of correlations for these purposes.

Providing Preliminary Information for a Grant Proposal

Achieving the final aim considered in this article, providing preliminary information for a grant proposal, may involve more than estimation of variances or effect sizes needed for planning larger studies. When funding for a large-scale study such as an R01 is sought, proposals typically include analyses of pilot data with tests of statistical significance in the presentation of preliminary results. Even with little or no expectation that significance will be found at $\alpha = .05$, low p-values on at least some of the major outcomes are likely to be viewed as supporting the intervention's efficacy or effectiveness. Information from two sources was considered in evaluating sample sizes for this aim. The first is a power analysis for samples in the range of 10–40 per group. The second is a review of sample sizes in actual pilot studies that have been funded. The purpose of the latter was to assess whether sample sizes recommended in this article are within a range that review panels have previously found acceptable, while at the same time evaluating whether funded studies typically are large enough to produce adequate information for subsequent funding requests.

There may be instances when a pilot study's contribution to the literature is important enough to justify submission of the results for publication, so a review of sample sizes in published reports of pilot studies is also included. Some statistical issues related to the publication of small studies are discussed under limitations.

Power analyses. For a two-group cross-sectional design, analyses were performed to determine the smallest population effect size (η^2) that could be detected with a probability of at least .80 by a two-tailed independent t-test given the reference sample sizes. Using $\alpha = .05$, power will be adequate for medium-to-large population effect sizes of $\eta^2 = .14$, .12, .10, and .09 for group sizes of 25, 30, 35, or 40, respectively. For smaller samples, effect sizes would have to be very large ($\eta^2 = .31$, .22, .17 for group sizes of 10, 15, 20) to have power of at least .80. Using a more liberal alpha level of .10, which some researchers may find acceptable for exploratory or preliminary studies, decreases the needed sample sizes somewhat, with 20 instead of 25 per group needed to detect a large population effect size ($\eta^2 = .14$). For a split-plot design, results for the test of the main effect of the intervention would be very similar to those reported for the independent-groups design, as the gain in power from the repeated measurements is small for the between-subjects effect in this design.

In efficacy studies with a single-factor within-subjects design, power for a repeated-measures ANOVA will vary as a function not only of sample size, but number of repeated observations, average within-subject correlation across observations, and the extent to which the assumption of sphericity has been met. With three measurement times, a within-subject correlation of .60, an estimated epsilon of .80 (relatively mild departure from sphericity), and $\alpha = .05$, the RM-ANOVA will have estimated power of at least .80 only for fairly large effects with small samples of 10 ($\eta^2 = .17$) or 15 ($\eta^2 = .11$). However, power will be adequate for more moderate effect sizes with samples larger than 15 ($\eta^2 = .08$, .06, .05, .05, and .04 with 20, 25, 30, 35, or 40 participants, respectively). Increases in the number of measurement times will further increase power. If $\alpha = .10$ is used, power for a medium-to-large effect ($\eta^2 = .09$) would be adequate with only 15 participants. In terms of power, then, a much smaller total sample size can be considered for an

efficacy study than would be desirable if comparing interventions. Larger samples would be needed if there were greater violation of the sphericity assumption or if the within-subject correlation were less than .60. For example, with a within-subject correlation of only .40, 30 instead of 20 would be needed using $\alpha = .05$ to have power of .80 for an effect size of $\eta^2 = .08$.

Abstract Review

Abstracts of pilot studies (as defined in this article) funded by National Institutes of Health (NIH) National Institute of Nursing Research (NINR) R03 and R15 grants from 2002 to 2004 were obtained using the CRISP database (National Institutes of Health, 2005), and Medline was searched for articles on pilot studies published in 2004 and referenced in the category of nursing. Studies were eliminated if they were qualitative studies, reported an evaluation of a pilot program, or involved a clustered design. Studies whose abstracts gave no sample size information were also not included in the summary. During the defined period, NINR funded 30 studies that were identified in their abstracts as pilot studies; 14 studies met the criteria outlined above. Two of the 14 studies were multi-phase and contributed a second sample size to the review. Total sample sizes ranged from 24 to 400, with a median of 49. Of the 16 samples, 2 were purely psychometric studies ($n = 40$ and $n = 400$), 3 were single-group correlational or descriptive studies ($n = 40$, 100, 110), and 11 were quasi-experimental or small randomized clinical trials. This latter group included one study reporting results from two experiments ($n = 24$ and $n = 48$) in which all participants received multiple, randomly ordered treatments, 7 studies having two groups (median size of 24 per group) and two studies each having three groups ($n = 25$ and $n = 27$ per group).

The Medline search yielded 199 studies, 96 of which met criteria. Their total sample sizes ranged from 3 to 419, with a median of 34.5. Of those involving single groups, 13 were purely psychometric studies (median size of 84), 35 were correlational/descriptive (median size of 40), and 21 were feasibility or efficacy studies (median size of 18). Of the remaining 27 studies, 24 were two-group comparisons, with a median group size of 20.5. The remaining 3 studies each had 3 groups, with group sizes of 10, 15, or 30. Although the sample sizes recommended in this article are within the ranges found in the abstract reviews, it is clear that many of these studies would be too small to provide statistical support of promising results unless the interventions produced very large effects.

LIMITATIONS

The confidence limits constructed around a point estimate will be accurate only insofar as underlying statistical assumptions are met. For variables measured on a continuous scale, typically it is assumed that the variable is randomly selected from a population that is normally distributed and that groups being compared are drawn from populations having equal variances. To the extent that these assumptions are not met, the calculated limits may not be accurate; unfortunately, in very small samples, it is impossible to adequately evaluate whether violations of the assumptions are likely. In addition, many of the methods used in this article are based on large-sample statistical theory, and their accuracy in small samples has not been thoroughly investigated. The nonparametric bootstrap method used with the split-plot example has no distributional assumptions, but situations in which there are more groups or more measurement times, where population multisample sphericity is violated, or where the within-group correlation differs from the .60 used here may result in interval widths that differ from those reported. Given these limitations, the reader should view the approach presented in this article as producing only general guidelines.

The review of abstracts was not intended to support the publication of extremely small pilot studies, although examples of these were found, but to take a preliminary look at whether the sample sizes recommended in this article were within a range that review panels and editors were accustomed to considering. Final judgment of the adequacy of the sample size in any given study would require more information than was available in the abstract.

In the case of published studies located through Medline, it is important to consider the possibility that publication bias favors reports of statistically significant results. When an effect size is essentially zero in a population, even underpowered studies will sometimes show significant results entirely due to sampling variability. If only studies with significant findings tend to be published, the probability of Type I error in the literature as a whole will be much larger than the stated α. Even when the null hypothesis is not true, any publication bias makes it likely that reported effect sizes overestimate the true values (Kline, 2004).

Furthermore, a significant result does not mean that the result is likely to replicate; for a study reporting an isolated finding with $p = .05$, the probability of significance with an exact replication would be only .50 (Greenwald, Gonzalez, Harris, & Guthrie, 1996). If publication of results from a small study is warranted, reporting the a priori power and the confidence intervals for observed effect sizes is essential, as it helps emphasize the lack of precision of these estimates as well as the need for replication of results.

CONCLUSION

The aim of this article was to illustrate a method of determining an appropriate, justifiable sample size for a pilot study in which the aims include estimating a population value, evaluating selected feasibility issues, assessing the adequacy of instrumentation, calculating statistical estimates for use in planning a larger study, or obtaining preliminary data for a grant proposal. The examples illustrate that samples in the size range typical of pilot studies produce relatively imprecise and sometimes seriously biased estimates of relevant statistics, particularly at the smallest sample sizes considered here. Specific sample size recommendations for feasibility aims are not made. They will depend on the nature of the decision based on the estimate, with samples as small as 10–15 per group sometimes being sufficient. Twenty-five participants per group should probably be considered the lower threshold of sample size for aims related to instrumentation, although 35–40 per group would be preferable if estimating test–retest reliability or item discrimination. Decisions concerning instrument revision from pilots of this size should be treated as highly tentative. If the aim of a pilot study is to demonstrate intervention efficacy in a single group, a sample in the range of 20–25 will probably be adequate when population effect sizes are likely to be moderate or larger. For pilot studies involving group comparisons, the situation is somewhat different. If obtaining information for a power analysis when it is possible to specify meaningful group differences independently from the data itself (and therefore only the variance needs to be estimated), a smaller sample in the range of 10–20 participants per group would be sufficient to implement the method discussed in Kieser and Wassmer (1996). On the other hand, if direct estimation of a between-group effect size is desired or if estimates will be used as preliminary information to justify a grant application, 30–40 participants per group at a minimum will be needed in order to yield confidence intervals whose lower limits can help define the range of plausible values for a subsequent power analysis. Both efficacy and comparative effectiveness studies with sizes in the recommended ranges are represented among funded and published work. Even with these sample sizes, researchers are advised to consider the imprecision of their estimates as they interpret and report results from pilot studies.

REFERENCES

Bonett, D.G. (2002). Sample size requirements for testing and estimating coefficient alpha. Journal of Educational and Behavioral Statistics, 27, 335–340.

Browne, R.H. (1995). On the use of a pilot sample for sample size determination. Statistics in Medicine, 14, 1933–1940.

Burns, N., & Grove, S.K. (2005). The practice of nursing research: Conduct, critique & utilization (5th ed.). Philadelphia: W.B. Saunders Company.

Chow, S., Shao, J., & Wang, H. (Eds.). (2003). Sample size calculations in clinical research. New York: Marcel Dekker.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155–159.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on the central and noncentral distributions. Educational and Psychological Measurement, 61, 532–574.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. Educational and Psychological Measurement, 61, 517–531.

Fern, E.F., & Monroe, K.B. (1996). Effect-size estimates: Issues and problems in interpretation. The Journal of Consumer Research, 23, 89–105.

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. Educational and Psychological Measurement, 61, 575–604.

Greenwald, A.G., Gonzalez, R., Harris, R.J., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? Psychophysiology, 33, 175–183.

Hulley, S.B., Cummings, T.B., Browner, W.S., Cummings, S.R., Hulley, D.G., & Hearst, N. (2001). Designing clinical research: An epidemiological approach. Philadelphia: Lippincott, Williams, & Wilkins.

Jairath, N., Hogerney, M., & Parsons, C. (2000). The role of the pilot study: A case illustration from cardiac nursing research. Applied Nursing Research, 13, 92–96.

Kieser, M., & Wassmer, G. (1996). On the use of the upper confidence limit for the variance from a pilot sample for sample size determination. Biometrical Journal (Biometrische Zeitschrift), 38, 941–949.

Kline, R.B. (2004). Beyond significance testing: Reforming data analysis methods in behavioral research. Washington, DC: American Psychological Association.

Lackey, N.R., & Wingate, A.L. (1998). The pilot study: One key to research success. In P.J. Brink & M.J. Wood (Eds.), Advanced design in nursing research (2nd ed.). Thousand Oaks, CA: Sage.

Lwanga, S.K., & Lemeshow, S. (1991). Sample size determination in health studies: A practical manual. Geneva, Switzerland: World Health Organization.

Machin, D., Campbell, M.J., Fayers, P.M., & Pinol, A.D. (1997). Sample size tables for clinical studies (2nd ed.). London: Blackwell Science.

National Institutes of Health. (2005). CRISP (Computer Retrieval of Information on Scientific Projects). Retrieved 8/8/07, from http://crisp.cit.nih.gov.

Nieswiadomy, R.M. (2002). Foundations of nursing research (4th ed.). Upper Saddle River, NJ: Pearson Education.

Nunnally, J.C., & Bernstein, I.H. (1994). Psychometric theory. (3rd ed.). New York: McGraw-Hill.

Polit, D.F., & Beck, C.T. (2004). Nursing research: Principles and methods (7th ed.). Philadelphia: Lippincott Williams & Wilkins.

Prescott, P.A., & Soeken, K.L. (1989). The potential uses of pilot work. Nursing Research, 38, 60–62.

Smithson, M. (2001a). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. Educational and Psychological Measurement, 61, 605–632.

Smithson, M. (2001b). Scripts and software for noncentral confidence interval and power analysis calculations. Retrieved 8/08/07 from http://psychology.anu.edu.au/people/smithson/details/CIstuff/CI.html.

SPSS. (2005). Statistical package for the social sciences (Version 14.0). Chicago: SPSS.

Steiger, J.H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. Psychological Methods, 9, 164–182.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. Educational Researcher, 31(3), 25–32.

Tonidandel, S. (2004). Sampling with replacement. Retrieved 8/08/07 from http://www.davidson.edu/academic/psychology/Tonidandel/TonidandelResamp.htm.

Uitenbroek, D.G. (1997). Correlation: SISA home. Retrieved 8/08/07 from http://home.clara.net/sisa/correl.htm.

Vickers, A.J. (2003). Underpowering in randomized trials reporting a sample size calculation. Journal of Clinical Epidemiology, 56, 717–720.