Initial Scale Development: Sample Size for Pilot Studies

George A. Johanson and Gordon P. Brooks

Educational and Psychological Measurement 2010 70: 394 originally published online 18 December 2009
DOI: 10.1177/0013164409355692

Educational and Psychological Measurement
70(3) 394–400

DOI: 10.1177/0013164409355692
http://epm.sagepub.com
SAGE

Initial Scale Development: Sample Size for Pilot Studies

George A. Johanson[1] and Gordon P. Brooks[1]

Abstract

Pilot studies are often recommended by scholars and consultants to address a variety of issues, including preliminary scale or instrument development. Specific concerns such as item difficulty, item discrimination, internal consistency, response rates, and parameter estimation in general are all relevant. Unfortunately, there is little discussion in the extant literature of how to determine appropriate sample sizes for these types of pilot studies. This article investigates the choice of sample size for pilot studies from a perspective particularly related to instrument development. Specific recommendations are made for researchers regarding how many participants they should use in a pilot study for initial scale development.

Keywords

pilot study, sample size, instrument development

Whether constructing a new scale or revising an existing scale, researchers must confirm that the scale uses clear and appropriate language, has no obvious errors or omissions, and has at least adequate psychometric properties before it is used. A pilot study is often recommended to address these issues as well as to estimate response rate and investigate the feasibility of a study. If parameters are to be estimated or null hypotheses tested, then it is necessary to determine the sample size needed for adequate precision or statistical power, respectively, prior to data collection.

[1]Ohio University, Athens, OH, USA

Corresponding Author:
George A. Johanson, Ohio University, 201 McCracken Hall, Athens, OH 45701, USA
Email: johanson@ohio.edu

Perspectives From the Literature

Because pilot studies are so useful, a common question from students and researchers is "How many participants do I need for my pilot study?" This is a difficult question to answer. The number of participants recommended for a pilot study is influenced by many factors and is less straightforward than determining the sample size needed to detect a particular effect, given the level of significance and desired power for the statistical analysis.

Social science literature has surprisingly few sample size recommendations for pilot studies, given the popularity of the pilot. However, some relevant articles bring attention to the matter. For example, in a discussion of exploratory and pilot studies, Isaac and Michael (1995) suggested that "samples with N's between 10 and 30 have many practical advantages" (p. 101), including simplicity, easy calculation, and the ability to test hypotheses, yet "overlook weak treatment effects." For similar reasons, Hill (1998) suggested 10 to 30 participants for pilots in survey research. van Belle (2002) suggested that researchers "use at least 12 observations in constructing a confidence interval" (p. 11). In the medical field, Julious (2005) reiterated that "a minimum of 12 subjects per group be considered for pilot studies" (p. 291). Treece and Treece (1982), referring to piloting an instrument, noted that for a project with "100 people as the sample, a pilot study participation of 10 subjects should be a reasonable number" (p. 176) but were not clear whether this meant 10 cases or 10% of the project sample size.

Bootstrapped confidence intervals from pilot study data may be useful for a variety of purposes, particularly when more than a point estimate is required. Mooney and Duval (1993) noted that bootstrapped approximations of parameter estimates and confidence intervals are considered relatively high quality "when n reaches the range of 30-50, and when the sampling procedure is truly random" (p. 21). That is, $N = 30$ is recognized as a reasonable minimum sample size for bootstrapped confidence intervals.

Hertzog (2008) made several different recommendations for sample size depending on the purpose of the pilot study in her recent and comprehensive article. For a feasibility study, her recommendations were, "samples as small as 10-15 per group sometimes being sufficient" (p. 190). For instrument development, her recommendation was 25 to 40. Hertzog recommended 20 to 25 for intervention efficacy pilots, given reasonable effect sizes, but 30 to 40 per group for pilot studies comparing groups.

Because we want both accurate and precise parameter estimates from pilot studies, we need samples that are both representative of the population and sufficiently large, respectively. The implication is that we need to conduct pilot studies with a sufficient number of participants who serve as an accurate representation of our population of interest. Although the focus of the more current literature on pilot studies has been on sample sizes needed for precision, the nature of the sample, rather than its size, has the largest impact on accuracy of parameter estimates. For example, the accuracy of pilot study results becomes questionable when unrepresentative samples

are used. Pilot studies that use relatively homogeneous and convenient samples may not represent the population we wish to study and may lead to biased estimates. Such samples (e.g., college sophomores) are clearly suspect for use in effect size estimation but may also be unsuitable for instrument development when the instrument will be used with a very different audience. We should also have a representative sample for the more subjective feedback on instrument clarity, completeness, language, and so forth. Light, Singer, and Willett (1990) stated the following:

> One facet of a measurement pilot must not be compromised: the sample design. Be sure the sample in your pilot fully represents your chosen target population. You must evaluate your instruments in a context that makes the results of the pilot directly generalizable to your ultimate study. Reliability and validity coefficients must be portable between the pilot and future studies. (pp. 215-216)

Biased estimates may also arise when estimating effect sizes using published literature. One particular problem is publication bias in favor of studies showing statistical significance and, hence, the presence of overly large effect sizes (Hedges & Vevea, 1996). This is sometimes used as an argument for conducting a pilot study with a randomly selected portion of the sampling frame rather than relying on the literature for effect size estimation so as to better represent the population of interest for the larger study.

This purpose of this article is to address initial instrument or scale development. This would include preliminary item analyses, estimates of internal consistency, and proportions of persons responding to particular options. We will not address many of the common validity issues (such as dimensionality, group differences, and multitrait–multimethod analyses), because appropriate analyses for validity studies would clearly require larger samples than commonly used in pilot studies for initial instrument development. A comprehensive item analysis should be conducted with larger samples as well, perhaps $N = 100$ to 200 (Crocker & Algina, 1986).

Method

The primary approach we used was a cost–benefit analysis, similar to Julious (2005), where we identified that point at which a sample size increment produced a notably lesser effect in estimating relevant population parameters. Our criterion was relative efficiency. Note that all of our results are theoretical. In particular, graphs were constructed from formulas; data were neither collected nor analyzed.

For pilot item analyses, researchers might use (corrected) correlations between item responses and total scale scores as item discrimination indices. These discrimination indices are often simple Pearson correlations, whereas Cronbach's coefficient alpha is arguably the most commonly reported measure of internal consistency in survey research.

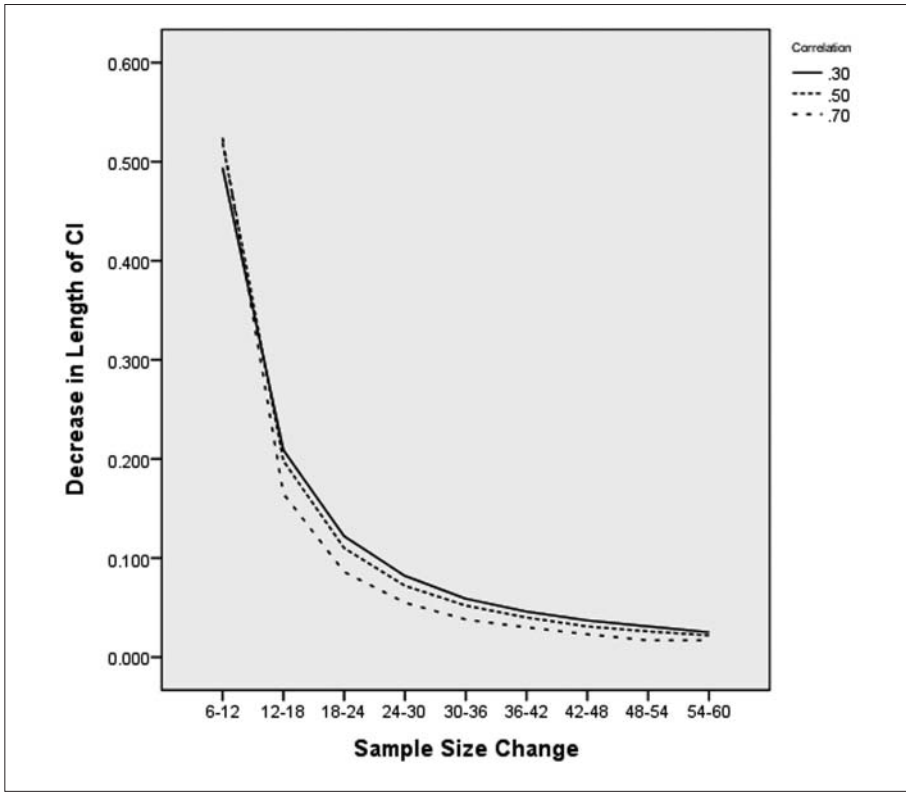

Figure 1. Decrease in length of confidence interval (CI) as sample size increases for a range of correlations

Results

Figure 1 was constructed using a program from Lowry (2008) and shows the impact of increasing sample size on the length of the confidence interval for Pearson correlations. From Figure 1, we see that as sample sizes increase from 24 to 30 and from 30 to 36, there is a decided flattening of the curve that suggests a loss of impact of sample size on the change in the length of the confidence interval, regardless of the magnitude of the correlation. If you have a predetermined level of precision, then you could use a program like Lowry's to choose your sample size, based on the desired standard error or confidence interval width. If you do not have a predetermined level of precision, then finding that point where the increase in precision is minimal (e.g., 24-36) may be a reasonable solution.

Figure 2 shows a similar pattern for confidence intervals about estimates of a proportion. Proportions are relevant when the purpose is to estimate response rates or the

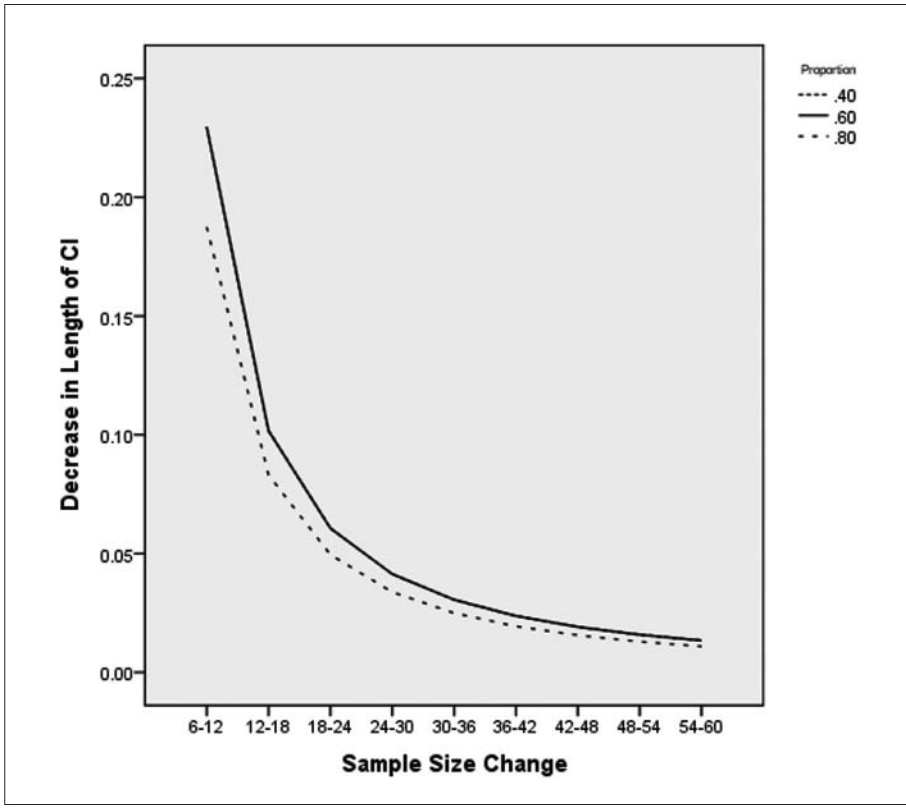

Figure 2. Decrease in length of confidence interval (CI) as sample size increases for a range of proportions

fraction of respondents choosing a particular option for an item. For binary response items, item difficulties or item means are proportions. As with correlations, the patterns do not differ very much with the magnitude of the proportion and show noticeable leveling in the intervals between $N = 24$ to 30 and $N = 30$ to 36.

Confidence intervals for the reliability coefficients were generated using formulas in Fan and Thompson (2001) and Feldt, Woodruff, and Salih (1987). Notice that Figure 3 also indicates transition points in the same range that we found for correlations and proportions, namely, $N = 24$ to 30 and $N = 30$ to 36, no matter the number of items. This observation is confirmed to some extent by Duhachek and Iacobucci (2004), who commented that standard errors for Cronbach's alpha

> are always larger for smaller sample sizes, as one might expect, though the differences between $n = 30$ and $n = 200$ are nominal for [mean interitem correlation] $r = 0.6$ or higher even when there are only two items. . . . (p. 796)

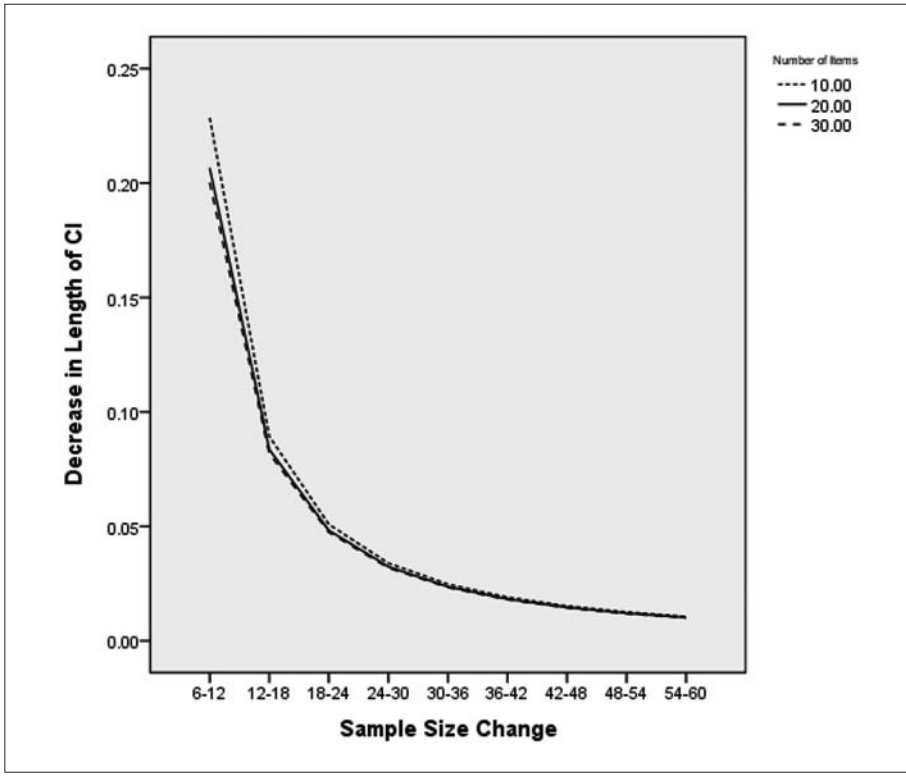

Figure 3. Decrease in length of confidence interval (CI) as sample size increases for a range of item lengths at Cronbach's α = .80

Only a reliability of .80 has been reported here, but additional reliability values show similar patterns.

Discussion

Pilot study sample size will depend on the particular purpose of the pilot study. What should be the sample size recommendation for pilot studies for initial scale development given a criterion of maximum information with minimum cost? Because the precision of our parameter estimates increases as sample size increases, all else being equal, larger samples are always better. The rate of increase in precision, however, is nonlinear, and we recommend that this information be used to help with this decision. If pressed for a single point estimate, we would suggest that 30 representative participants from the population of interest is a reasonable minimum recommendation for a pilot study where the purpose is preliminary survey or scale development. Both the

existing literature and our current investigation of confidence intervals converge nicely to this recommendation, and the redundancy in our plots serves to emphasize this consistency. An interval estimate of 24 to 36 is also supported by both our results and the existing literature in this area, where several scholars, for example, have recommended $N = 12$ per group in studies in studies where two or three groups might be expected.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Funding

The authors received no financial support for the research and/or authorship of this article.

References

Crocker, L., & Algina, J. (1986). *Classical and modern test theory.* New York: Holt, Rinehart & Winston.

Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology, 89*, 792-808.

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

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*, 93-103.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21*, 299-332.

Hertzog, M. A. (2008). Considerations in determining sample size for pilot studies. *Research in Nursing & Health, 31*, 180-191.

Hill, R. (1998). What sample size is "enough" in internet survey research? *Interpersonal Computing and Technology: An Electronic Journal for the 21st Century, 6*(3-4). Retrieved July 12, 2008, from http://www.emoderators.com/ipct-j/1998/n3-4/hill.html

Isaac, S., & Michael, W. B. (1995). *Handbook in research and evaluation*. San Diego, CA: Educational and Industrial Testing Services.

Julious, S. A. (2005). Sample size of 12 per group rule of thumb for a pilot study. *Pharmaceutical Statistics, 4*, 287-291.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education*. Cambridge, MA: Harvard University Press.

Lowry, R. (2008). *VassarStats: Website for statistical calculation*. Retrieved July 12, 2008, from http://faculty.vassar.edu/lowry/rho.html

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference.* Newbury Park, CA: Sage.

Treece, E. W., & Treece, J. W. (1982). *Elements of research in nursing* (3rd ed.). St. Louis, MO: Mosby.

van Belle, G. (2002). *Statistical rules of thumb.* New York: John Wiley.