Qual Life Res (2015) 24:147–151
DOI 10.1007/s11136-014-0752-2

BRIEF COMMUNICATION

# Sample size for pre-tests of questionnaires

**Thomas V. Perneger · Delphine S. Courvoisier · Patricia M. Hudelson · Angèle Gayet-Ageron**

Accepted: 1 July 2014 / Published online: 10 July 2014


**Abstract**
*Purpose* To provide guidance regarding the desirable size of pre-tests of psychometric questionnaires, when the purpose of the pre-test is to detect misunderstandings, ambiguities, or other difficulties participants may encounter with instrument items (called «problems»).
*Methods* We computed (a) the power to detect a problem for various levels of prevalence and various sample sizes, (b) the required sample size to detect problems for various levels of prevalence, and (c) upper confidence limits for problem prevalence in situations where no problems were detected.
*Results* As expected, power increased with problem prevalence and with sample size. If problem prevalence was 0.05, a sample of 10 participants had only a power of 40 % to detect the problem, and a sample of 20 achieved a power of 64 %. To achieve a power of 80 %, 32 participants were necessary if the prevalence of the problem was 0.05, 16 participants if prevalence was 0.10, and 8 if prevalence was 0.20. If no problems were observed in a given sample, the upper limit of a two-sided 90 % confidence interval reached 0.26 for a sample size of 10, 0.14 for a sample size of 20, and 0.10 for a sample of 30 participants.
*Conclusions* Small samples (5–15 participants) that are common in pre-tests of questionnaires may fail to uncover even common problems. A default sample size of 30 participants is recommended.

**Keywords** Questionnaires · Validity · Pre-tests · Power · Sample size · Cognitive interviewing

T. V. Perneger (✉) · D. S. Courvoisier · A. Gayet-Ageron
Division of Clinical Epidemiology, University Hospitals of Geneva, 6 rue Gabrielle-Perret-Gentil, 1211 Geneva, Switzerland
e-mail: thomas.perneger@hcuge.ch

T. V. Perneger · P. M. Hudelson · A. Gayet-Ageron
Department of Community Health and Medicine, Faculty of Medicine, University of Geneva, Geneva, Switzerland

D. S. Courvoisier · P. M. Hudelson
Division of Primary Care Medicine, University Hospitals of Geneva, Geneva, Switzerland

## Introduction

A qualitative pre-test is a key phase of the development, adaptation, or translation of any questionnaire or psychometric instrument [1–6]. The main purpose of the pre-test is to verify that the target audience understands the questions and proposed response options as intended by the researcher, and is indeed able to answer meaningfully. Typically, when pre-testing a self-report instrument, the pre-test participant will first fill in the questionnaire, and then debrief about each item in sequence. Many researchers also employ more or less extensive cognitive interviewing methods [1, 6]. Identification of problems—e.g., unclear question, unfamiliar word, ambiguous syntax, missing time-frame, lack of an appropriate answer—lead to a modification of the instrument and ideally another round of pre-tests. The process stops when no new issues arise.

Despite its importance for instrument validity, pre-testing is not well codified. Many experts acknowledge this—«the practice is intuitive and informal» [2], «the pretest is the most misunderstood and abused element of the survey process» [3]. Lack of guidance also applies to sample size. Several reference texts do not address the sample size of a pre-test at all [5, 7]. Others cite ranges of 5–8 participants [8], 5–15 [6], 7–10 [9], 8–15 [4], 10–15 [10], 10–30 [11], 30–40 [12], 25–75 [2] or «as many as you can» [2]. Little

rationale is provided for these numbers beyond the availability of resources; most authors describe what is usually done but do not actually take a stand. Streiner and Norman recommend «sampling to redundancy» by analogy with qualitative research, but without explaining how redundancy is to be assessed [4]. Other experts conclude that «additional standards are needed to determine optimal sample sizes» [6].

Regardless of the rationale for selecting a sample size, the ability of a pre-test to detect a problem that participants may encounter with a questionnaire is bound by basic laws of probability. If the sample size is too small, the probability can be large that no participant will report any given problem. A recent paper by Blair and Conrad explores this issue [13]. These authors show that sample size requirements are often much larger than the customary numbers cited above, for various values of problem prevalence, probability of detection, and power. Alongside useful formulas, Blair and Conrad reported sample sizes required for a limited set of problem prevalences (between 0.05 and 0.10). In this paper, we extend their work to help with the determination of sample size for pre-tests: (a) we compute the power to detect a problem for a wider range of prevalence values (from 0.01 to 0.30) given several plausible sample sizes, (b) estimate the required sample size to detect a problem for various levels of prevalence, for at least one occurrence of the problem and for a repeat occurrence, and (c) obtain upper confidence limits for problem prevalence in situations where no problems were detected in the pre-test.

## Methods

We computed power, required sample size, and upper bounds of confidence intervals for zero prevalence using basic probability calculus, for a range of reasonable scenarios.

To compute the power to detect a problem in at least one interview for a sample size $n$ and a prevalence of problem $p$, we used the equation proposed by Blair and Conrad:

$$\text{Power} = 1 - (1-p)^n \tag{1}$$

It is easy to see that $(1–p)$ is the probability of not finding the problem in one interview, $(1–p)^n$ the probability of finding no problem in $n$ independent interviews, and $1–(1–p)^n$ the probability of finding a problem in at least one interview. We computed power for 7 sample sizes (5, 7, 10, 15, 20, 30, 50) and 18 levels of prevalence (0.01–0.15 in steps of 0.01, 0.20, 0.25, 0.30). To compute the sample size needed to detect a problem in at least one participant with a given power, we resolved the power Eq. (1) numerically, finding the lowest n for which power exceeds 80 or 90 %. An alternative approach is to solve the equation for n as follows:

$$n = \frac{\ln(1-\text{Power})}{\ln(1-p)} \tag{2}$$

and to round the computed $n$ upward, to the next integer.

A single occurrence of a problem during a pre-test may not be enough to justify modification of the instrument. Therefore we also computed the sample size needed to observe a given problem in at least two individuals. We used for this the equation of the complement of binomial cumulative density function, where $r$, the number of occurrences of a problem in $n$ trials, was set to 1:

$$\begin{aligned}\text{Power} &= 1 - \sum_{k=0}^{r} \binom{n}{k} p^k (1-p)^{n-k} \\ &= 1 - (1-p)^n - np(1-p)^{n-1}\end{aligned} \tag{3}$$

Again, we resolved this equation for various values of $n$ and recorded the lowest $n$ for which power exceeds 80 or 90 %. Note that this equation is a general formulation of Eq. 1. Eq. 3 cannot be solved directly for n, because n appears both as a multiplier and as an exponent. However, we noticed that the numbers of observations required to observe 2 or more occurrences of a problem were almost directly proportional to the numbers required for 1 or more occurrences. The best empirically derived multiplier was 1.86 for a power of 80 %, and 1.70 for a power of 90 %. Thus a convenient formula for the number of observations needed to detect a problem at least twice in a sample is:

$$n = C \times \frac{\ln(1-\text{Power})}{\ln(1-p)} \tag{4}$$

where $C$ equals 1.86 if the desired power is 80 % or 1.70 if power is 90 %. Again, the computed n is rounded upward to the next integer. This method yields numbers that can be off by 1–2 observations when compared to the exact values obtained by numerical solution.

The probabilities were obtained form the program PASW statistics 18 [14].

The pre-test should stop when no (important) problem is detected. However, even when no problems are detected in the sample, there is no guarantee that the prevalence of a problem is zero in the population. We obtained confidence intervals for the proportion (or true prevalence of problems) when zero events are observed, for various sample sizes, using the Clopper-Pearson method [15]. We computed the upper bounds of 95, 90 and 80 % two-sided confidence intervals (the lower bound is always 0). Of note, when 0 events are observed, the upper limit of a two-sided $1–\alpha$ confidence interval corresponds to the upper limit of a one-sided $1 - \alpha/2$ onfidence interval. Because the confidence interval coverage is typically defined a priori, before the researcher knows how many events will be observed, the two-sided interpretation of coverage makes usually

**Table 1** Power (in percent) to discover a problem after *N* interviews in at least one interview, by prevalence of problem

| Prevalence | Number of interviews | | | | | | |
|---|---|---|---|---|---|---|---|
| | $N = 5$ | $N = 7$ | $N = 10$ | $N = 15$ | $N = 20$ | $N = 30$ | $N = 50$ |
| 0.01 | 5 | 7 | 10 | 14 | 18 | 26 | 39 |
| 0.02 | 10 | 13 | 18 | 26 | 33 | 45 | 64 |
| 0.03 | 14 | 19 | 26 | 37 | 46 | 60 | 78 |
| 0.04 | 18 | 25 | 34 | 46 | 56 | 71 | 87 |
| 0.05 | 23 | 30 | 40 | 54 | 64 | 79 | 92 |
| 0.06 | 27 | 35 | 46 | 60 | 71 | 84 | 95 |
| 0.07 | 30 | 40 | 52 | 66 | 77 | 89 | 97 |
| 0.08 | 34 | 44 | 57 | 71 | 81 | 92 | 98 |
| 0.09 | 38 | 48 | 61 | 76 | 85 | 94 | 99 |
| 0.10 | 41 | 52 | 65 | 79 | 88 | 96 | >99 |
| 0.11 | 44 | 56 | 69 | 83 | 90 | 97 | >99 |
| 0.12 | 47 | 59 | 72 | 85 | 92 | 98 | >99 |
| 0.13 | 50 | 62 | 75 | 88 | 94 | 98 | >99 |
| 0.14 | 53 | 65 | 78 | 90 | 95 | 99 | >99 |
| 0.15 | 56 | 68 | 80 | 91 | 96 | >99 | >99 |
| 0.20 | 67 | 79 | 89 | 96 | 99 | >99 | >99 |
| 0.25 | 76 | 87 | 94 | 99 | >99 | >99 | >99 |
| 0.30 | 83 | 92 | 97 | >99 | >99 | >99 | >99 |

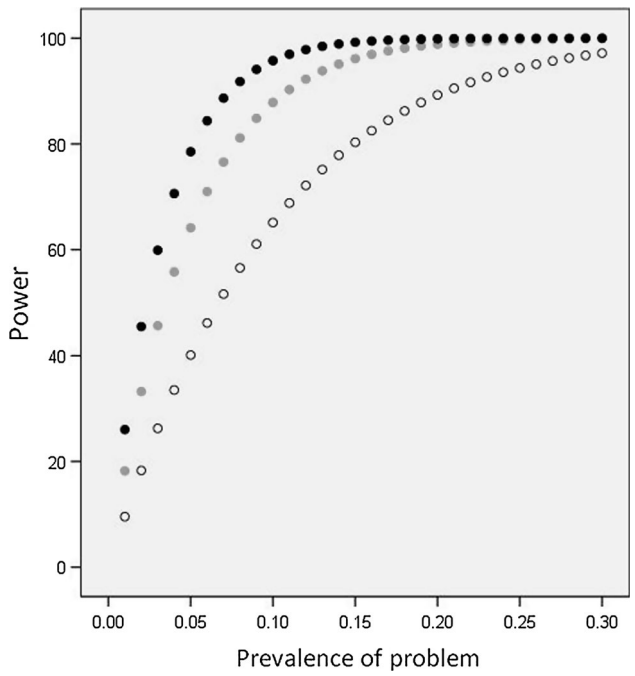


**Fig. 1** Power (in percent) to detect a problem in a pilot study, by prevalence of the problem, for sample sizes of 10 (*circles*), 20 (*grey dots*), and 30 (*black dots*)

**Table 2** Required sample size to detect with high probability (80 or 90 %) a problem with a questionnaire, by problem prevalence, for at least one or two occurrences of the problem in the sample

| Prevalence | Power >80 % | | Power >90 % | |
|---|---|---|---|---|
| | ≥1 occurrence | ≥2 occurrences | ≥1 occurrence | ≥2 occurrences |
| 0.01 | 161 | 299 | 230 | 388 |
| 0.02 | 80 | 149 | 114 | 194 |
| 0.03 | 53 | 99 | 76 | 129 |
| 0.04 | 40 | 74 | 57 | 96 |
| 0.05 | 32 | 59 | 45 | 77 |
| 0.06 | 27 | 49 | 38 | 64 |
| 0.07 | 23 | 42 | 32 | 55 |
| 0.08 | 20 | 37 | 28 | 48 |
| 0.09 | 18 | 33 | 25 | 42 |
| 0.10 | 16 | 29 | 22 | 38 |
| 0.11 | 14 | 27 | 20 | 34 |
| 0.12 | 13 | 24 | 19 | 31 |
| 0.13 | 12 | 23 | 17 | 29 |
| 0.14 | 11 | 21 | 16 | 27 |
| 0.15 | 10 | 19 | 15 | 25 |
| 0.20 | 8 | 14 | 11 | 18 |
| 0.25 | 6 | 11 | 9 | 15 |
| 0.30 | 5 | 9 | 7 | 12 |

sense. However, in this instance, we only compute confidence intervals if 0 events are observed, so the one-sided interpretation may be favored by some readers. For this reason we report both interpretations of the confidence bounds. The confidence limits were computed using StatXact 4.0 (Cytel software, Cambridge, MA) [16].

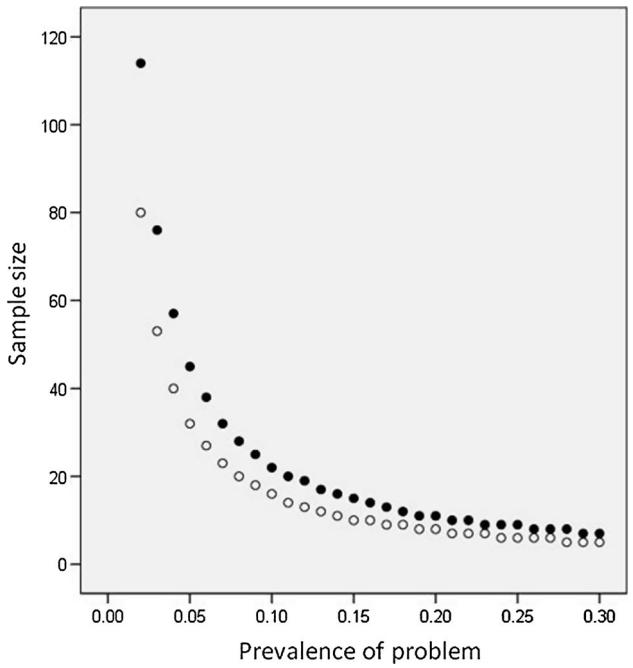


**Fig. 2** Sample size required to detect a problem in a pilot study, by prevalence of the problem, for power of 80 % (*circles*) and 90 % (*black dots*)

## Results

As expected, the power to detect a problem increased with the prevalence of the problem and with sample size (Table 1; Fig. 1). Power was insufficient for the detection of rare problems (prevalence 0.05) at small sample sizes, reaching 79 % only with a sample size of 30. In contrast, fairly frequent problems (prevalence 0.20) were detected with high power even when the sample size was 7.

We computed the minimal sample size that is required to achieve a reasonably high probability of detecting a problem, by prevalence of problem (Table 2; Fig. 2). To detect a low-prevalence problem (prevalence 0.05) at least once, 32 participants were required to achieve a power of 80 %, and 45 for a power of 90 %. If a repeat observation of a problem was required (i.e., 2 occurrences of a problem), the numbers increased to 58 (power 80 %) and 77 (power 90 %) individuals. Even for a problem prevalence of 0.10, the range of sample sizes went from 16 to 38, depending on the desired power and minimal number of occurrences.

How confident can a researcher be that a question works as intended if no pre-test participant reports any problem? This is reflected in the upper confidence bound for the true prevalence value (Table 3). With lower sample sizes, the absence of any problems during the pre-test does not rule out even a substantial prevalence of problems (up to 0.45 with 5 observations, and up to 0.10 with 30 observations, for a two-sided 90 % confidence interval or one-sided 95 % confidence interval).

**Table 3** Upper bound of exact confidence interval for the prevalence of a problem if no problems are identified, by sample size

| Sample size | One-sided coverage of confidence interval (two-sided coverage) | | |
|---|---|---|---|
| | 97.5 % (95 %) | 95 % (90 %) | 90 % (80 %) |
| 5 | 0.52 | 0.45 | 0.37 |
| 7 | 0.41 | 0.35 | 0.28 |
| 10 | 0.31 | 0.26 | 0.21 |
| 15 | 0.22 | 0.18 | 0.14 |
| 20 | 0.17 | 0.14 | 0.11 |
| 30 | 0.12 | 0.10 | 0.07 |
| 50 | 0.07 | 0.06 | 0.05 |

## Discussion

The tables in this paper provide realistic numbers regarding the ability of a small size pre-test to identify problems with a questionnaire or psychometric instrument, or the degree of reassurance provided by a problem-free pre-test. Small sample sizes (from 5 to 15 participants) are prone to missing even fairly common problems. One would not want to field a questionnaire that will cause a difficulty or produce unwanted results in 10 % of the participants, yet this cannot be ruled out if the sample size of the pre-test is 15 or less. To achieve a power of 90 % to detect a problem present for one out of ten respondents, 22 participants would be needed, or even 38 if the researcher required a confirmation in another participant before altering the instrument. Then, if a problem is identified and corrected, another fairly large pre-test is necessary (e.g., 30 participants), such that the absence of any detected problem would practically rule out actual difficulties.

What are the practical implications? Firstly, researchers should consider the actual power of the pre-tests that they conduct in order to avoid premature conclusions about the acceptability of their instruments. They should remain aware that an item may pose real problems to some respondents despite an uneventful small sample pre-test. Furthermore, the discovery of any given problem with a questionnaire item does not preclude the existence of other (presumably less prevalent) problems for the same item. A small-scale pre-test will only harvest the low-hanging fruit, so to speak. In order to identify less prevalent problems that have been missed during pre-tests, during the first large scale use of a new instrument, questions can be added to confirm the acceptability and clarity of the relevant items. Self-reported «questions about questions» have been used

to help select the most suitable satisfaction questionnaire [17], or to compare alternative versions of a health status instrument [18]. Such larger scale studies are the only feasible option for the detection of problems with a prevalence below 0.05.

Another consideration is to try and increase the yield of problems in the pre-test sample. This means that the researcher may select a purposeful sample of individuals who are likely to encounter problems with the proposed questions, such as cultural minorities, less educated persons, or people with low cognitive abilities. Related to this, the researcher may seek to increase the detectability of problems [13], through the use of various in-depth cognitive interviewing methods [1, 6], and also by selecting self-aware or introspective participants who have a greater capacity to reflect on their thinking patterns. If the detectable prevalence of the problem can be shifted from 5 %, which may be easily missed in a small sample, to 10 or 15 %, the pre-test will be considerably more productive.

Finally, sample sizes of 30 or more should be preferred for pre-tests whenever possible, to achieve a reasonable power to detect fairly prevalent problems (prevalence of 10 %). Indeed, if 30 or 40 % of a population has difficulty with an item, this suggests that basic rules of item writing were not adhered to, and that a critical review of the instrument should have identified the problem even before pre-tests. A sample size of 30 would achieve a reasonably high power (about 80 %) to detect a problem that occurs in 5 % of the population, and to detect a repeat occurrence of a problem that affects 10 % of the respondents. At the same time, if no problems are detected for a given question among 30 respondents, the upper 90 % two-sided confidence limit on the true prevalence of problems is 10 %. This suggests that 30 participants is a reasonable default value or starting point for pre-tests of questionnaires.

We do not suggest that the occurrence of a problem during a pre-test guarantees the success of the procedure, but rather that it is a necessary pre-condition for the improvement of the instrument. Once a problem has been identified, the instrument developers must decide if it is serious enough to warrant a modification, and if yes, what the modification should be. There is no guarantee that the new version will perform better than the older version. The questionnaire improvement process as a whole, and not only problem detection, may benefit from empirical research.

## References


1. Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., et al. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly, 68*, 109–130.
2. Converse, J. M., & Presser, S. (1986). *Survey questions. Handcrafting the standardized questionnaire*. Newbury Park: Sage Publications Inc.
3. Backstrom, C. H., & Hursch-César, G. (1981). *Survey Research* (2nd ed.). New York: Macmillan Publishing Company.
4. Streiner, D. L., & Norman, G. R. (2003). *Health Measurement Scales. A practical guide to their development and use* (3rd ed.). Oxford: Oxford University Press.
5. Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken NJ: Wiley.
6. Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*, 287–311.
7. DeVellis, R. F. (2012). *Scale development. Theory and applications* (3rd ed.). Los Angeles, Newbury Park: Sage Publications Inc.
8. Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., et al. (2005). Principles of good pratice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR task force for translation and cultural adaptation. *Value Health, 8*, 94–104.
9. Patrick, D. L., Burke, L. B., Gwaltney, C. J., Kline Leidy, N., Martin, L., Molsen, E., et al. (2011). Content validity—Establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: Part 2—Assessing respondent understanding. *Value Health, 14*, 978–988.
10. Sprangers, M. A., Cull, A., Groenvold, M., Bjordal, K., Blazeby, J., & Aaronson, N. K. (1998). The European Organization for Research and Treatment of Cancer approach to developing questionnaire modules: An update and overview. *Quality of Life Research, 7*, 291–300.
11. Fayers, P. M., & Machin, D. (2000). *Quality of life. Assessment, analysis and interpretation*. New York: Wiley.
12. Beaton, D. E., Bombardier, C., Guillemin, F., & Ferrsaz, M. C. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine, 25*, 3186–3191.
13. Blair, J., & Conrad, F. G. (2011). Sample size for cognitive interview pretesting. *Public Opinion Quarterly, 75*, 636–658.
14. PASW statistics, version 18, Chicago, IL.
15. Clopper, C., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika, 26*, 404–413.
16. StatXact version 4. Cytel software, Cambridge, MA.
17. Perneger, T. V., Kossovsky, M. P., Cathieni, F., di Florio, V., & Burnand, B. (2003). A randomized trial of four patient satisfaction questionnaires. *Medical Care, 41*, 1343–1352.
18. Cleopas, A., Kolly, V., & Perneger, T. V. (2006). Longer response scales improved the acceptability and performance of the Nottingham Health Profile. *Journal of Clinical Epidemiology, 59*, 1183–1190.