J Occup Health 1998; 40: 307–312

Journal of Occupational Health

Using Pilot Data to Estimate Sample Size and Compare Question Forms for a Crossover Study*

Maxia DONG, Martin R. PETERSEN and Mark J. MENDELL

National Institute for Occupational Safety and Health

Abstract: Using Pilot Data to Estimate Sample Size and Compare Question Forms for a Crossover Study: Maxia DONG, *et al.* National Institute for Occupational Safety and Health—A pilot study was conducted on thirty office workers to help determine if a scannable form of symptom severity questions would yield similar results as a nonscannable form. There were three goals of the pilot study: first, to observe if, in a questionnaire using two forms of Visual Analog Scales (VAS), questions using a scannable sequence of "boxes" for responses would elicit different mean responses than questions using unbroken "lines"; second, to observe if questions using a sequence of "boxes" would elicit different within person and week response variability than questions using unbroken "lines"; and third, to estimate the sample size needed for a crossover study, depending on the particular form of the question used, and the number of crossovers. The pilot study, consisting of three sequential weekly questionnaires, provided week, subject, and error variance components for each of three dependent variables from the two different VAS forms. Most of the calculations were performed with a log transformation of the data. For each VAS form, the number of subjects necessary for desired study power for each symptom was calculated. Based on this pilot study, neither the mean nor the within person and week variance component was consistently larger or smaller for the VAS_{box} form than for the VAS_{line} form. The linear models analysis showed that the two forms filled out by the same person on the same day had similar mean values and were highly correlated for all symptoms ($R^2 \geq 0.95$). Thus we chose the VAS_{box} form because of scanner compatibility and estimated the required number of subjects for our full-scale study based on this chosen form.
(*J Occup Health 1998; 40: 307–312*)

Received Mar 19, 1998; Accepted Sept 11, 1998
Correspondence to: M. Dong, National Institute for Occupational Safety and Health, 4676 Columbia Parkway, R-16, Cincinnati, OH 45226, USA
**This research was funded in part by the U.S. Environmental Protection Agency.*

Key words: Sample size, Pilot study, Crossover study, Intervention, Office worker, Indoor air, Questionnaire, Visual Analog Scale (VAS)

To study indoor environmental quality and office worker symptoms, an experimental intervention was planned. The goal of this intervention study was to assess if workers' symptoms would be decreased by an intervention that improved the indoor air quality (installing enhanced air filters in the ventilation system). In designing this study, Visual Analog Scale (VAS) questions were chosen to estimate the intensity of various occupant symptoms. A classical VAS is a horizontal line with two anchor points, one at each end[1]. Using VAS forms as a measurement for feelings and symptoms has been reported in many clinical trials as well as indoor air studies[2–6]. The majority of work on the use of VAS has been conducted relative to pain assessment. Sriwatanakul *et al*. concluded from their study that a graded linear, horizontal VAS was the most reliable and most preferred scale by their subjects[7]. These studies that have used the visual analog format for self-report of pain suggest that a similar approach might be successful for self assessment of symptom change in subjects for an intervention study.

For the intervention, it was hoped that a computer-linked scanner could be used to read questionnaire responses to reduce manual key entry error and save time and money. (Five 3-page questionnaires could be scanned in the time required for one to be manually keyed.) We considered using a conventional VAS form with an unbroken horizontal line (VAS_{line} form), on which respondents mark the level of their symptom severity with an intersecting vertical line (see Fig. 1, top). Our scanning software, however, could not interpret this form of question, but could read responses only from a sequence of boxes (VAS_{box} form, see Fig. 1, bottom). We therefore needed to determine if VAS questions designed with boxes would yield different average responses or if they would require substantially more subjects than the more commonly used one with lines.

The goal of the full-scale study was to test for an effect

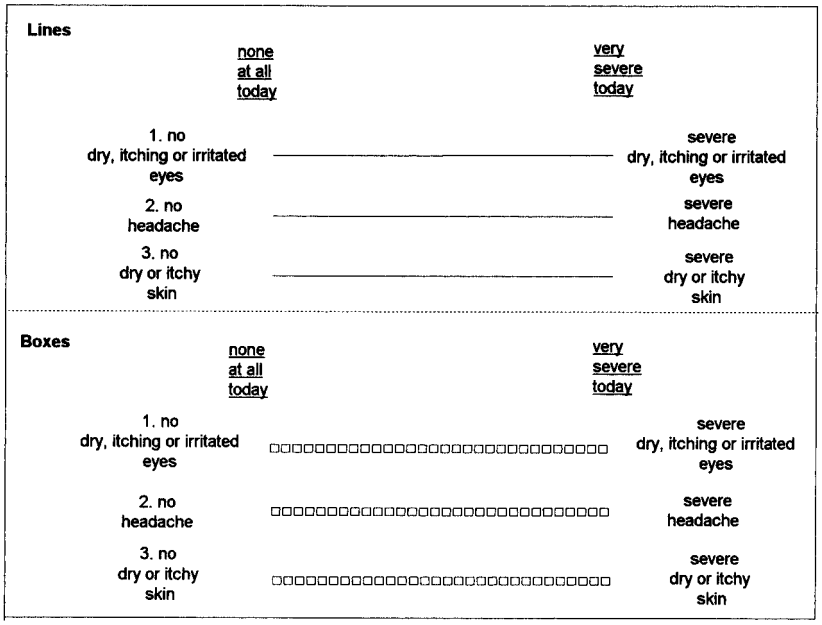

Fig. 1. Two forms of Visual Analog Scales questions used in the pilot study (The instruction on the top of two forms are not shown here).

of enhanced filtration using a crossover design involving two areas (different floors) and lasting either two weeks (single crossover) or 4 weeks (double crossover). Office workers in each area would complete one questionnaire containing both forms each week. This design removes week-to-week variability and subject-to-subject variability and uses the residual variability (within subject variance for a single week) as the error for testing the effect of the intervention.

There was no reasonable estimate of the within-person variance for a single week available from previous studies, and it was also not known if different VAS forms would result in different variances. To adopt an appropriate estimate of this variance, σ^2, for the two different forms of the question, a pilot study was conducted. The other purposes of the pilot study were to determine if the means and variances would be the same for both VAS forms, and to determine the sample size necessary for the main (crossover) study.

Methods

In order to estimate the within subject variance for a single week, it was not necessary to use a crossover design. Instead, the pilot study was conducted among thirty volunteers who were asked to complete the identical self-administered questionnaire three times, at weekly intervals. All subjects were office workers with professions of epidemiologist, statistician, or industrial hygienist. All of them had worked in the office located in the basement of a building for at least three months. Their jobs mostly involved data analysis by computer, reading and writing professional publications, and conducting scientific studies. Among them, 18 (60%) were females and 12 (40%) were males. The study was conducted from May 2 to May 17, 1996. During the three-week study period, the subjects were asked to complete the questionnaire on Thursday or Friday afternoon after 1 PM. On Thursday and Friday morning all subjects would receive an e-mail reminding them to complete and return the questionnaire. The overall response rate was 97%.

Three of the most important symptoms in the main (crossover) study were "dry, itching, or irritated eyes", "headache", and "dry, itchy, or irritated skin". Thus these were used in the pilot study questionnaire. This questionnaire asked about the current severity of those three symptoms using both forms of the VAS questions. As shown in Fig. 1, one form was a linear scale consisting of an unbroken line on which the respondent marked a response with a vertical line. The other form consisted of a row of 30 boxes, in one of which the respondent marked a response. A mark at the left border or in the leftmost box corresponded to the lack of a symptom, and a mark at the right border or in the rightmost box corresponded to a very severe symptom. The two forms of VAS questions were printed on separate sheets and were not assembled next to each other. Between these two forms there was one other set of questions asking about monthly and yearly symptom frequency, which was not included in the analysis here. By this procedure, the possibility that the subjects would give the same response could be diminished. For the pilot study, questionnaires

Table 1. Mean, median and correlation coefficients for three symptoms in two VAS forms

Outcome variables	Question forms	Number of observations	Minimum	Maximum	Mean	Median	correlation coefficient
Eye symptoms	Boxes	83	0	69.0	10.7	3.4	0.926**
	Lines	84	0	70.0	11.8	3	
Headache	Boxes	84	0	82.8	12.4	1.7	0.960**
	Line	85	0	79.0	12.8	1	
Skin symptoms	Boxes	84	0	48.3	6.0	0	0.906**
	Lines	85	0	49	5.5	1	

**Pearson coefficients, $p<0.01$.

were scored by hand.

The value obtained from the linear scale (y_{line}) was the distance of the marked response from the left end, divided by the total length of the scale, and then multiplied by 100. The boxes were effectively numbered from 0 to 29. The value obtained from the scale with boxes (y_{box}) was the number of the checked box, divided by 29 and then multiplied by 100. That is:

y_{line} = (marked length/line length) × 100

y_{box} = [box number/29] × 100.

The relationship between two VAS forms was calculated by linear models using the SAS GLM procedure[8]. The SAS CORR Procedure[8] was used to estimate means, distributions, and correlations between two forms.

Because the distribution of each dependent variable (the symptom outcomes) was skewed, with many responses of zero and few large responses, each dependent variable was log-transformed to achieve approximate normality. This was not necessary for estimating variance components, but it was thought that this transformation would be used in the main study so that a standard parametric analysis could be used. Since there were zero values for the dependent (y) variables, a constant (k, here=1) was added to y_i. Before log-transforming, the new dependent variable was t=y + 1. The actual dependent variable used was, then, r=ln(t)=ln(y + 1). The SAS VARCOMP Procedure[8] was used to estimate week, subject, and residual error variance components for each dependent variable. Further details are given in the Appendix.

Results

Within the pilot sample of 30 individuals over three weeks, there was a great deal of variability in the severity of symptoms reported. This can be seen in Table 1. The distributions of all three symptoms were so skewed that the estimated means were much larger than the medians (see Table 1). For eye symptoms, more than half of the observations were lower than 4. For headache and skin symptoms, more than half of the observations were lower than 2. Figure 2 depicts the distribution of the three symptom outcomes in two VAS forms. In the VAS_{box} form skin symptoms tended to have slightly more zero responses than in the VAS_{line} form.

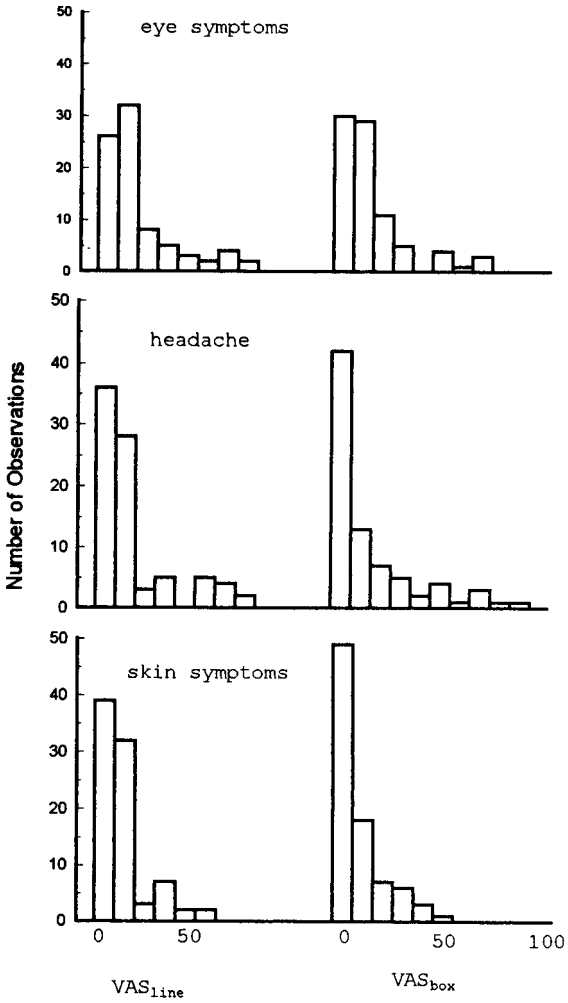

Fig. 2. Distribution of three symptom outcomes in two VAS forms.

Table 2 shows geometric means and variance

Table 2. Estimate of variance components of log-transformed data over three weeks

Outcome variable	Question form	Number of observations	Geometric mean*	Variance component for subjects	Variance component for weeks	Variance component for error, σ^2
Eye symptoms	Boxes	83	4.000	1.001	−0.0021**	0.940
	Lines	84	4.091	1.081	0.0181	0.920
Headache	Boxes	84	3.193	1.408	−0.0034**	1.150
	Lines	85	3.131	1.266	0.0092	1.085
Skin symptoms	Boxes	84	1.721	1.095	0.0475	0.519
	Lines	85	1.907	0.847	0.0569	0.677

*geometric mean of $t_{ij}=Y_{ij}+1$ (symptom severity plus one) over all weeks and all subjects. **Because the estimates are unbiased, they can be negative when the true variance component is close to zero.

Table 3. Number of subjects required for a single crossover design for different symptom outcomes and different forms of VAS at power 1-β=0.90 and 0.95 (α=0.05, c=1)

	Eye symptoms		Headache		Skin symptoms	
Δ*	Boxes	Lines	Boxes	Lines	Boxes	Lines
When 1-β=0.90						
1	406	397	321	303	57	74
2	125	120	101	95	20	26
3	65	63	55	52	12	15
4	42	41	37	35	9	11
5	31	30	27	26	7	9
6	24	24	22	21	6	7
7	20	20	18	17	5	6
When 1-β=0.95						
1	501	491	396	374	70	92
2	152	149	125	118	25	32
3	80	78	68	64	15	19
4	52	51	45	43	10	13
5	38	37	34	32	8	11
6	30	29	27	25	7	9
7	25	24	22	21	6	8

*change in symptom severity (mean value) from before to after treatment.

components for each dependent variable after log transformation. Neither question form had a consistently or substantially larger mean than the other. Thus the boxes appear to yield means which are similar to those yielded by lines. For each symptom, the lines form had a slightly larger variance component for week than the boxes form. Subject variability plays a large role in the overall variability of the study. Neither question form had a consistently or substantially larger variance component for subject or error. The error component (the variance of the residual error, σ^2) is the only one needed for sample size calculation in a crossover design.

Table 1 shows the relationship of two VAS forms. The correlation coefficients between the two VAS forms for three symptoms were larger than 0.90 ($p<0.001$) (See Table 1). The linear models analysis showed that the two VAS forms were not significantly different after adjustment for weeks, subjects, and their interaction ($p=0.1131$, 0.6517 and 0.2574 for eye symptoms, headache, and skin symptoms, respectively).

Table 3 shows the minimum number of subjects, N, for each VAS form and symptom, needed to obtain a power of 0.90 or 0.95 at the 0.05 significance level for a single crossover. For double crossovers, where c=2, the number of required subjects, N, would be halved. (See Appendix.) Different symptoms required different

numbers of subjects, which decreased with increasing Δ. For the skin symptoms, both forms required a very small number of subjects because of their lower means and error variance. For the other two symptoms (eye symptom and headache), the VAS_{box} form required slightly more subjects than the VAS_{line} form.

Discussion/Conclusion

Because of the high sensitivity and accuracy, using a VAS form could be more helpful for quantitative assessments of self-reported symptom changes than a verbal method. This has been reported in many studies[1, 4]. We designed a questionnaire response format with a large number of boxes that formed a linear-like scale, but could be read by our scanner software. The most commonly used VAS form is a classic line form[1, 3–7]. Using boxes is an economical approach to make computer software more applicable for data entry. The VAS form using boxes produced data with variability similar to that of the line version, while for each response variable the means from the two forms were fairly similar. The statistical tests for comparing the means of the two forms was not significant for any of the three symptoms. It is unlikely that a difference as large as two units (out of 100) was missed because the power for detecting a difference of two units was 0.78, 0.86, and 0.97 for eye symptoms, headache, and skin symptoms, respectively. Because the number of additional subjects needed when using the box form was small, and because the form simplified data entry, we chose it for our intervention study. Our pilot study found that a variable with a higher mean value may also have a higher error variance and therefore may require a larger number of subjects. This indicates that it is important in estimating sample size for a full scale study to include any variables with high mean values. It is obvious that the more boxes used the more alike the two forms should be. At some point, this will be restricted by the differentiating ability of a scanner.

In addition, in collecting such data it is desirable to use a fairly large sample size to improve the validity of parametric statistics, regardless of power, because the log transformation made only a slight improvement in the symmetry of the data. Thus, even with the transformation, the assumption of a normal distribution may be questionable with a small sample[4, 5, 9]. The log-transformation was used in this study, although others have reported using a different transformation[10].

Using the recommended numbers of subjects in our full-scale study should allow us to be 90–95% sure of detecting an effect of the intervention when in fact the effect is Δ. If the true effect is less than Δ, it may still be detected, but the probability of doing so will be less than 0.90–0.95.

In conclusion, the VAS_{box} form yielded means that were similar to those obtained from the VAS_{line} form, and they did not require a much larger sample size than did the line form. Since the VAS_{box} form could be interpreted by our scanner software, it was chosen for our full-scale study, and the sample size was based on this form.

Acknowledgment: We thank our volunteer office workers for their cooperation, and Charles Mueller, Dr. James Deddens, Dr. Richard Hornung, Dr. Avima Ruder, and Dr. Elizabeth Ward for helpful comments.

References

1) Aitken RCB. Measurement of feelings using visual analogue scales. Proc R Soc Med 1969; 62: 989–993.
2) Kildesø J, Tornvig L. The effect of improved cleaning methods on the composition of dust and the well-being of people. In: Proceedings of Indoor Air '96: The 7th International Conference on Indoor Air Quality and Climate, Nagoya, Japan, 1996, Vol. 2, 923–928.
3) Mador MJ, Kufel TJ. Reproducibility of visual analog scale measurements of dyspnea in patients with chronic obstructive pulmonary disease. Am Rev Respir Dis 1992; 146: 82–87.
4) Thomee R, Grimby G, Wright BD, Linacre JM. Rasch analysis of visual analog scale measurements before and after treatment of Patellofemoral Pain Syndrom in women. Scand J Rehabil Med 1995; 27: 145–151.
5) Quiding H, Okasala E, Happeonen RP, et al. The visual analog scale dose evaluations of analgesics. J Cli Pharmacol 1981; 21: 424–429.
6) Fritz G, Spirito A, Yeung A, Klein R, Freedman E. A pictorial visual analog scale for rating severity of childhood asthma episodes. J Asthma 1994; 31: 473–478.
7) Sriwatanakul K, Kelvie W, Lasagna L, et al. Studies with different types of visual analog scale for measurement of pain. Clin Pharmacol Ther 1983; 34: 234–239.
8) SAS Institute Inc. SAS/STAT User's Guide (vol. 2) Version 6, 6.04 Edition, NC, U.S.A. Cary. 1994.
9) Maxwell C. Sensitivity and accuracy of the visual analogue scale: a psychophysical classroom experiment. Br J Clin Pharmacol 1978; 6: 15.
10) Dexter F, Chestnut DH. Analysis of statistical tests to compare visual analog scale measurements among groups. Anesthesiology 1995; 82: 896–902.
11) Senn S. Cross-over Trials in Clinical Research, Chichester. England, John Wiley & Sons, 1993.

Appendix

For subject i and week j, let $r_{ij}=\ln(t_{ij})=\ln(y_{ij}+1)$, where y_{ij} is the response (between 0 and 100) for a given symptom on a given scale. Variance components were estimated using the model $r_{ij}=S_i+W_j+\varepsilon_{ij}$, where r_{ij} is the transformed symptom outcome for the ith subject on the jth week, S_i is the effect of the ith subject, W_j is the effect of the jth week, and ε_{ij} is the within subject within week (residual) error. The expected mean squares were equated

with the observed mean squares, and the system of equations was solved to yield unbiased estimates of the variance components for subject, week, and error.

A crossover experiment is similar to a matched pair experiment where the pairs are the *Nc* combinations of subjects (*N*) and crossovers (*c*), except that the variance is the residual error since a crossover analysis also removes the effect of week. The number of subjects required can be calculated using the following formula (modified from Senn[11]):

$$N = \frac{2\sigma^2}{c\,[\delta/(Z_{\alpha/2} - Z_{1-\beta})]^2}$$

where N is the number of subjects, σ^2 is the variance component for within subject within week error, c is the number of crossovers, δ is the assumed true mean difference in the r scale from before to after treatment, and Z_x (x=$\alpha/2$ or 1-β) is a value such that x is the probability of a standard normal random variable exceeding Z_x. Assuming μ_1 is the mean of t_{ij} after enhanced filtration and μ_0 is the mean of t_{ij} before enhanced filtration, then

$$\delta = \ln(\mu_0/\mu_1) = \ln(\mu_0) - \ln(\mu_1).$$

However, it is difficult to specify a meaningful δ directly, since it is the difference between two log values. Thus we specified $\Delta=\mu_0-\mu_1$, which is a difference in the non-log scale (y or t), and then calculated δ. Since the means were unknown for the intervention study, we used the average of the two geometric t_{ij} means (one for boxes and one for lines) for each variable in the pilot study to estimate μ_1. (We considered the building in which the pilot study was done to have a good ventilation filtration system and thus to be similar to an after-intervention environment in a building with a poorer filtration system originally.) We set α=0.05 and estimated the required sample size at powers of 0.90 and 0.95.