Repeatability of a questionnaire to assess respiratory symptoms in smokers

CELIA H WITHEY,¹ CHARLES E PRICE,¹ ANTHONY V SWAN,¹ ANNA O PAPACOSTA,¹ AND MICHAEL J HENSLEY²

From the Department of Community Medicine,¹ United Medical and Dental Schools of Guy's and St Thomas's Hospitals, St Thomas's Campus, London SE1 7EH and the Faculty of Medicine,² University of Newcastle, New South Wales, 2308 Australia.

SUMMARY  To evaluate the repeatability of a questionnaire designed to assess change in respiratory symptoms 90 smokers were interviewed on two occasions. The questionnaire included questions from the Medical Research Council questionnaire on respiratory symptoms, questions on acute chest illness and cough and phlegm production in the preceding two weeks, a modification of Field's card system for estimating frequency of cough, and an objective assessment of the presence of phlegm—the loose cough sign. The study was carried out in two parts. During the first part 30 male smokers were interviewed by one observer and then re-interviewed 1 to 2 hours later by a different observer. During the second part 60 subjects were interviewed and then after a period of 1 to 10 days re-interviewed by the same observer. The results showed that the within-subject variation representing the measurement error for Field's card system was 15.1% of the between-subject variation and was adequately Normal to justify the use of standard analytical techniques. Similar results were obtained from questions on cough and phlegm scored between 1 and 5, although the variation in this case was rather less Normal. In general, the between-observer, within-observer, and within-subject repeatability were satisfactory for all parts of the questionnaire with the exception of the loose cough sign which had a relatively low prevalence. There was no evidence of an observer order effect and there were no important systematic differences due to lapses in time or different observers. The findings indicate that techniques such as the cough scoring system may be used to permit studies of respiratory symptoms via questionnaire methods to be much smaller than those required to detect equivalent differences in prevalences.

There are a number of standard questionnaires for ascertaining the presence of particular symptoms and for identifying the presence of chronic cough and phlegm for epidemiological purposes.¹-³ In studies concerned with aetiological factors and various behavioural and environmental hazards they have served a useful function for a long time. They do have a serious limitation, however, in that they produce qualitative, often dichotomous, variables. A subject is classified as having or not having the symptoms, as having or not having chronic cough and phlegm. This means that analyses and thus the studies cannot be as sensitive as they would be with a measured outcome variable.

Field's card system⁴ goes a long way to remedy this situation in that it generates a 'measure' or score for cough and it could be extended to other symptoms such as phlegm. In a planned trial to assess changing respiratory symptoms over a period of modified smoking behaviour the need for sensitive measures was especially acute.⁵ For this reason a questionnaire was developed which included a modified form of the card system as well as new and standard respiratory symptoms questions. The study reported here was designed to assess the repeatability of the Field's scoring system and these symptoms questions within and between two observers.

Methods

The study was performed in two parts. The first part was designed to assess the effect of different observers applying the questionnaire to the same subjects; the second part was to assess differences arising when the same observer applied the questionnaire to the same subjects at different times.
Repeatability of a questionnaire to assess respiratory symptoms in smokers

The questionnaire constructed for this repeatability exercise was in five sections. The first included questions 1, 3, 5, 6, 8, 10, and 12 of the Medical Research Council questionnaire on respiratory symptoms with question 12 modified to cover a period of 12 months instead of three years. The second section asked about the occurrence of acute respiratory illness in the preceding two weeks, and whether and for how long this had kept the subject from his usual activities.

The third section consisted of two questions on cough and phlegm where the subject was shown two cards each containing five statements about cough and phlegm respectively and was asked to indicate the statement that best described his cough and phlegm over the preceding two weeks.

(a) For cough, the statements were: I have hardly coughed at all; I have coughed on some days; I have coughed on most days; I have coughed every day; I have coughed a lot every day.

(b) For phlegm, the statements were: I have not brought up any phlegm at all; I have brought up phlegm on some days; I have brought up phlegm on most days; I have brought up phlegm every day; I have brought up a lot of phlegm every day.

These produce a score of between 1 and 5 for each symptom.

The ‘loose cough sign’ has been proposed as an objective method for identifying subjects with persistent cough and phlegm. To investigate its repeatability and potential usefulness, it was included in the fourth section of this questionnaire. To obtain a response the observer asked the subject to take a deep breath and to cough. The response was taken as positive, and thus the cough classified as loose, if a rattling sound was heard.

In the final section a modified form of Field’s card system was used for estimating the frequency of cough. It required a set of cards each containing two statements about cough, one statement always being more severe than the other. The number of cards depends on the number of uniquely rankable statements used. For this study five statements were used:

I rarely ever cough
I cough occasionally
I cough a little on most days
I cough a little every day
I cough a lot every day.

These were grouped into seven specified pairs which were presented to the subject on cards, one pair to each card. By selecting the statement on each card which more closely described the subject’s cough over the preceding two weeks and marking this on a proforma, a cough frequency score was calculated. The minimum and maximum attainable cough scores were 1 and 8.

Subjects

The subjects selected for this study were male aged between 18 and 70 who were outpatients, inpatients or staff of a teaching hospital or were attending their general practitioner’s surgery. All subjects had smoked at least one cigarette per day or one cigar per week or one ounce of pipe tobacco per month for a minimum of 12 months during the preceding 18 months. There were 30 subjects in the first part of the study and 60 subjects in the second. In the first part 30 subjects attending an outpatient clinic were randomly allocated to one of two observers who interviewed them prior to their medical consultation. The second observer then repeated the interview after a period of 1 to 2 hours. In the second part, 60 patients and personnel of the hospital, or patients attending a general practitioner’s surgery, were randomly allocated to one of two observers. They were interviewed in the hospital or general practitioner’s surgery and again at their home or in the hospital after a period ranging from 1 to 10 days by the same observer.

Analysis

By design the observations were in pairs. Those from the first part of the study were repeated on the same subjects by different observers, and those from the second part were repeated on each subject by the same observer on different occasions. The problem of assessing deficiencies of a measurement technique or measurement error in pairs of repeated measurements on different subjects is a standard one. The most useful approach is to estimate the within-subject and between-subject variances. It is then possible to estimate what proportion of the total variation in a group of subjects is due to measurement error. The magnitude of this error and its effect on the sensitivity of any investigation can then be assessed. The measurement error obtained in this way is then a compound of subject variability and the effect due to observer variation or to the observers varying from one occasion to another.

Such an approach can be used for any measurement or score that can be treated as one but it is not obviously applicable to the qualitative, all or nothing, responses to symptom questions. They require another approach. The usual method is to quantify the degree of agreement between the pair of observations classifying an individual as positive or not. A useful
measure for this purpose is Cohen's kappa for categorical data, although it does have the limitation of requiring a reasonably even distribution of individuals in the various outcome categories. It is a measure of the degree of agreement beyond that predicted by chance. If the categorical variables are considered as measured outcomes of 0 and 1, it is in fact algebraically equivalent to the intra-class correlation coefficient which approximates the comparison of variance approach described above. A helpful categorisation for the interpretation of kappa scores given in Landis and Koch is:

<table>
<thead>
<tr>
<th>Kappa statistic</th>
<th>Strength of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0</td>
<td>Poor</td>
</tr>
<tr>
<td>0-0.20</td>
<td>Slight</td>
</tr>
<tr>
<td>0.21-0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41-0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61-0.80</td>
<td>Substantial</td>
</tr>
<tr>
<td>0.81-1.00</td>
<td>Almost perfect</td>
</tr>
</tbody>
</table>

And subsequent results will be discussed in terms of this categorisation.

**Results**

Although the aim of the study is to assess the repeatability of measures of respiratory symptoms it is necessary to have a sample with reasonable prevalence of these symptoms. Table 1 gives the prevalences of morning and day cough and phlegm, the prevalence of a loose cough, and the mean Field's cough score for both the groups in the study. The prevalence of morning cough was 23 and 30% in the first and second group respectively, whereas that of morning phlegm was 33 and 43%. At the interview only 5 out of the 30 subjects (17%) had a loose cough in the first group and 12 out of 60 (20%) in the second group. The group of subjects interviewed by different observers within 1 to 2 hours of each other gave a mean Field's cough score of 2.9 and those interviewed by the same observer on two different occasions, 1 to 10 days apart, gave a mean score of 3.3.

For Field's cough score, in the group re-interviewed after a lapse of from 1 to 10 days, the within-subject component of variance was obtained for subjects grouped according to the observer and the lapse of time between their interviews (Table 2). Results were obtained for only 59 out of the 60 subjects as the responses from one subject were repeatedly inconsistent and a final cough score could not be calculated. An analysis taking the components of variance as following a gamma distribution showed that there were significant differences between the “lapse of time” groups.

The within-subject component of variance estimated using all of the pairs re-interviewed after 1 to 10 days was 0.788. The between-subject variance component was 3.901 so the within-variance or measurement error was 16.8% of the total variation.

In the group interviewed after a period of 1 to 2 hours the within-subject variance with observer 1 first was 0.25 and with observer 2 first was 0.53. The differences were not significant (F19, 10 = 3.47; p<0.08) and the two were combined to give an estimate of 0.431.

**Table 2. Within-subject variances by observer and lapse of time between interviews for Field’s card system**

<table>
<thead>
<tr>
<th>Time (days) between interviews</th>
<th>1-2</th>
<th>3-4</th>
<th>5+</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observer 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance df</td>
<td>0.500</td>
<td>0.813</td>
<td>0.125</td>
<td>0.483</td>
</tr>
<tr>
<td><strong>Observer 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance df</td>
<td>0.778</td>
<td>5.667</td>
<td>0.167</td>
<td>1.083</td>
</tr>
<tr>
<td><strong>Observers 1 and 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance df</td>
<td>0.661</td>
<td>2.136</td>
<td>0.147</td>
<td>0.788</td>
</tr>
</tbody>
</table>

* The prevalences are of subjects responding positively at either or both interviews.
* The mean score was obtained using both observations for each subject.
Repeatability of a questionnaire to assess respiratory symptoms in smokers

All the data from the 90 pairs in both groups combined were then combined and gave an overall estimate of the within-subject variation which was 0-670 or 15-1% of the total variation between subjects. A Normal plot of the differences indicated that the distribution of scores obtained in this way was sufficiently close to Normality to justify the use of techniques such as the t test.

The two questions on cough and phlegm which produce scores of from 1 to 5 can also be treated as if they were measurements, although they are less likely to have Normal distributions. However, investigated in the same way as described above for Field’s cough score, with these data they did prove to be Normal enough for group comparisons using t tests, and the overall estimates of the within or error variations were 16-1% and 15-9% respectively of the total variations. There was no systematic evidence of the observer or time intervals affecting the measurement error.

To assess inter-observer agreement and within-subject variability for the different observers for the respiratory symptoms questions and the loose cough sign, Cohen’s kappa was calculated (table 3).

The results showed that agreement between the two observers interviewing subjects within 1 to 2 hours of each other was not affected by the order with which the two observers interviewed the subjects and was “moderate” to “almost perfect” (k = 0-57 – 0-93) for all the questions on respiratory symptoms and “moderate” for the loose cough sign (k = 0-53).

The agreement between the responses obtained from the same subjects interviewed by the same observer on two different occasions separated in time by 1 to 10 days was initially assessed separately for each observer. Agreement for each observer between first and second interviews was “moderate” to “almost perfect” (k = 0-49 – 1-00) for all the respiratory symptoms questions except for two questions on periods of increased cough and phlegm in the last 12 months and the question on phlegm production in the past 2 weeks for observer 1 where the agreement was “fair” (k = 0-21 – 0-33). Combining the results for both observers the agreement was found to be “moderate” to “almost perfect” (k = 0-44 – 0-83) for all questions. Finally, combining all the data thus using all 90 pairs of observations, the agreement for all the respiratory symptoms was found to be “substantial” to “almost perfect” (k = 0-61 – 0-82) except for the two questions on increased cough and phlegm in the last 12 months and the question on phlegm production in the past 2 weeks where the agreement was only “moderate” (k = 0-53 – 0-59).

The repeatability of the loose cough sign was “poor” (k = 0-03) for observer 1 and “substantial” (k = 0-67) for observer 2. Combining the results from both observers gave a kappa of 0-61, this agreement being “substantial”. For all 90 subjects agreement for the loose cough sign was “moderate” (k = 0-58). Such “poor” agreement found for one

Table 3  Repeatability by section of the respiratory symptoms questionnaire

<table>
<thead>
<tr>
<th>Questionnaire section</th>
<th>Between-observer repeatability in group re-interviewed after 1-2 hours by a different observer</th>
<th>Within-observer repeatability in group re-interviewed after 1-10 days by the same observer</th>
<th>Within-subject repeatability in all 90 pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kappa score (SE)</td>
<td>Kappa score (SE)</td>
<td>Kappa score (SE)</td>
</tr>
<tr>
<td></td>
<td>Observer 1</td>
<td>Observer 2</td>
<td></td>
</tr>
<tr>
<td>1. MRC Questions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morning cough in winter</td>
<td>0-67 (0-18)</td>
<td>0-84 (0-11)</td>
<td>0-79 (0-08)</td>
</tr>
<tr>
<td>Morning cough for more than 3 months of the year</td>
<td>0-57 (0-17)</td>
<td>0-86 (0-10)</td>
<td>0-72 (0-08)</td>
</tr>
<tr>
<td>Day cough in winter</td>
<td>0-71 (0-16)</td>
<td>0-64 (0-14)</td>
<td>0-61 (0-09)</td>
</tr>
<tr>
<td>Day cough for more than 3 months of the year</td>
<td>0-69 (0-17)</td>
<td>0-62 (0-13)</td>
<td>0-62 (0-08)</td>
</tr>
<tr>
<td>Morning phlegm in winter</td>
<td>0-75 (0-13)</td>
<td>0-59 (0-15)</td>
<td>0-65 (0-08)</td>
</tr>
<tr>
<td>Morning phlegm for more than 3 months of the year</td>
<td>0-70 (0-13)</td>
<td>0-64 (0-13)</td>
<td>0-65 (0-08)</td>
</tr>
<tr>
<td>Day phlegm in winter</td>
<td>0-89 (0-11)</td>
<td>0-49 (0-17)</td>
<td>0-77 (0-08)</td>
</tr>
<tr>
<td>Day phlegm for more than 3 months of the year</td>
<td>0-89 (0-10)</td>
<td>0-55 (0-15)</td>
<td>0-77 (0-07)</td>
</tr>
<tr>
<td>One or more periods of increased cough and phlegm</td>
<td>0-83 (0-11)</td>
<td>0-28 (0-22)</td>
<td>0-58 (0-10)</td>
</tr>
<tr>
<td>Two or more periods of increased cough and phlegm</td>
<td>0-69 (0-13)</td>
<td>0-21 (0-19)</td>
<td>0-53 (0-09)</td>
</tr>
<tr>
<td>2. Chest illness in past 2 weeks</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kept from usual activities</td>
<td>0-91 (0-07)</td>
<td>0-75 (0-13)</td>
<td>0-77 (0-07)</td>
</tr>
<tr>
<td>3. Cough in past 2 weeks</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kept from usual activities</td>
<td>0-93 (0-07)</td>
<td>0-76 (0-13)</td>
<td>0-82 (0-06)</td>
</tr>
<tr>
<td>4. Loose cough sign</td>
<td>0-53 (0-24)</td>
<td>-0-03 (0-02)</td>
<td>0-58 (0-12)</td>
</tr>
</tbody>
</table>
observer is explained by the fact that there were very few subjects with a loose cough (2 out of 30 (7%)) and in these circumstances the possibilities are limited and Cohen’s kappa is very imprecise. There was little change in the kappa scores when the results were stratified for the occurrence of a recent chest infection.

Discussion

This investigation arose from a need to assess how precisely symptoms could be measured or identified in smokers. The sample was therefore restricted to smokers. Even then the prevalences of the various symptoms were not always as high as necessary for accurate assessments of repeatability. This means that the results cannot automatically be generalised to non-smokers among whom the prevalences would be even lower.

To assess the repeatability of measures of respiratory symptoms, particularly that of Field’s scoring system for cough, it is necessary to estimate what the variation would be within subjects if the same observer could interview them repeatedly in a very short space of time with neither the observer nor the subject remembering anything about preceding interviews. The nearest it is possible to get to that is to have the same observers repeat interviews after a modest lapse of time. This was done with two observers for lapses in time of between 1 and 10 days. There was one substantial difference between observers due to one particular individual in the three subjects seen by observer 2 with a lapse of three or four days (table 2). This individual scored 6 on the first occasion and 1 on the second. It is believed that this was due to a response biased by the presence of the spouse on the second occasion. It is an example of how individuals may vary, and in that sense it should be included in the estimates of measurement error although it is not an observer effect.

The significant decrease in measurement error found between the shortest and longest ‘lapse of time’ groups (table 2) was somewhat surprising. It suggests that the longer the lapse in time between interviews the more repeatable they were. However there was no sign of a trend over the three groups so the difference was treated as chance, and an overall estimate of the within-subject variance using all the pairs of observations was obtained. This gave a value of 0.788 which was 16.8% of the total variation one would see in a sample like this with each individual observed once only. Thus the evidence is that the within-subject variation does not differ from one observer to another and is not a great proportion of the overall variation.

To assess the within-subject variability unaffected by a lapse of time the study design had included repeat interviews by different observers within 1 to 2 hours. This allows an assessment of not only the effect on the measurement error of different observers but also whether within-subject variability is affected by the order in which observers talked to the subjects. The within-subject variance with observer 2 was larger than that with observer 1 but not significantly so, and the two combined gave an estimate of 0.431. This estimate of the measurement error with very little lapse of time between interviews was smaller than that from the sample interviewed twice by the same observer (0.788), but again the difference was not significant. This supports the conclusion that neither the observer nor the lapses in time in this study have appreciably altered the within-subject variation, and all the data were combined to give an estimate of the within-subject variation of 0.670 or 15.1% of the estimated total variation between subjects. Thus the imprecision of the technique is such that sample sizes for studies using Field’s score as an outcome measure on subjects like these only need to be 15% larger than would be needed if the technique was perfect, with zero measurement error, in order to give equally sensitive studies.

Similar results were obtained from the two questions on cough and phlegm which result in scores of between 1 and 5. The overall estimates of measurement error were 16% of the total variation in both cases.

For the MRC questions the inter-observer agreement found in the group of subjects interviewed within 1 to 2 hours was “substantial” to “almost perfect”. The results for the second group were “moderate” to “almost perfect”. This good agreement for the MRC questions is not surprising as they have been extensively used and tested in the past.3 The repeatability of the loose cough sign was no more than “moderate” when all the data were combined, but the prevalence was quite low (17/90 or 19%). The reliability of the loose cough sign, its validity as an indicator of abnormal bronchial function, and the high degree of inter-observer agreement for this sign have been demonstrated.6

In conclusion, it appears that Field’s card system is a highly repeatable measure of cough frequency. Its repeatability is not significantly affected by different observers and its quantitative nature makes it considerably more sensitive than yes/no response questions in the detection of changes in symptoms. The technique requires that field workers administering it are given a small amount of training taking about 1 hour, and in the main study this has proved feasible and effective. The simple questions producing scores of between 1 and 5 are almost as repeatable as Field’s cough score although they may not quite as readily satisfy the assumption of an
Repeatability of a questionnaire to assess respiratory symptoms in smokers

underlying Normal distribution necessary for the use of techniques such as the t test.

The MRC questions have been widely used and our results confirm their repeatability. In contrast, Field's card system has not been used so extensively. Our results using this system and simpler questions producing scores have shown them to be highly repeatable. They present no practical difficulties and their use should be seriously considered in any epidemiological study concerned with the comparison of respiratory symptoms levels within subjects or between categories of subjects.

References

1 Medical Research Council. Questionnaire on respiratory symptoms 1966.

Accepted for publication October 1987
Repeatability of a questionnaire to assess respiratory symptoms in smokers.
C H Withey, C E Price, A V Swan, A O Papacosta and M J Hensley

*J Epidemiol Community Health* 1988 42: 54-59
doi: 10.1136/jech.42.1.54

Updated information and services can be found at:
http://jech.bmj.com/content/42/1/54

**These include:**

**Email alerting service**
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

**Notes**

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/