Brit. J. prev. soc. Med. (1959), **13**, 175–193

VARIABILITY IN ANSWERS TO A QUESTIONNAIRE ON RESPIRATORY SYMPTOMS

BY

A. S. FAIRBAIRN

Statistical Research Unit, Medical Research Council,

C. H. WOOD

Rockefeller Unit of Occupational Health, London School of Hygiene and Tropical Medicine,

and

C. M. FLETCHER

Postgraduate Medical School, London

Chronic bronchitis can be recognized, in its early stages, only from the patient's account of his symptoms. During the last few years the prevalence of respiratory symptoms has been surveyed in many populations in the United Kingdom. A more or less uniform questionnaire has been developed by certain workers in consultation; but until recently there has been no general agreement on its content or layout. The questionnaire is completed by one observer at one interview with each subject.

Cochrane, Chapman, and Oldham (1951) showed that observers reported widely different prevalence rates of respiratory symptoms in separate random samples of a population of miners. Schilling, Hughes, and Dingwall-Fordyce (1955) also found systematic differences in the extent to which two observers reported respiratory symptoms at interviews with the same group of cotton workers. In the surveys reported by Higgins (1959), the same observer conducted all the interviews and observer variability was thus avoided. If, however, the epidemiology of chronic bronchitis is to be more widely studied, the technique of symptomatic inquiry must be standardized so that the results of different investigators may be compared. Our purpose in this survey was to compare several observers in the use of a detailed standardized questionnaire. Differences in the frequency of symptoms reported by each observer were to be assessed statistically. Details of the interviews were to be investigated from tape-recordings, in order to find ways to improve both the questionnaire itself and the technique of the interview in future surveys.

METHODS

It was decided that a number of subjects were each to be interviewed twice, with a different allocation of subjects to observers on the second occasion. This enabled any discrepancies between the two answers given by the same subject to the same question to be investigated from the tape-recordings. The system of re-allocation enabled differences between observers, in the number of symptoms reported, to be tested for significance. At least 4 weeks elapsed between the two interviews. Kinsey, Pomeroy, and Martin (1948) considered that the answers given by an individual might influence his answers to the same question at a second interview unless at least 8 months intervened. Questions about repiratory symptoms, on the other hand, are not so emotionally charged as those put by Kinsey in investigating sexual behaviour. One of us (C.M.F.) found in a pilot study that hospital patients forgot their previous answers about respiratory symptoms even after one week. In the present survey no recollection of previous answers was apparent during the second series of interviews.

THE QUESTIONNAIRE.—This is shown in Appendix I. The questions concerned cough, phlegm, breathlessness, wheezing, and nasal catarrh, the effect on the chest of colds and the weather, past chest illnesses, and smoking habits. The majority of questions were designed to be answered either "Yes" or "No". For most of the remainder, alternative answers were laid down from which a choice had to be made. A few questions only were open-ended. Most questions (referred to as "compulsory") had to be asked of

every subject, but some "supplementary" questions were asked only when a particular answer had first been given; *e.g.*, when the answer to the question "Does the weather affect your chest?" was "Yes", the supplementary question "What sort of weather?" was asked.

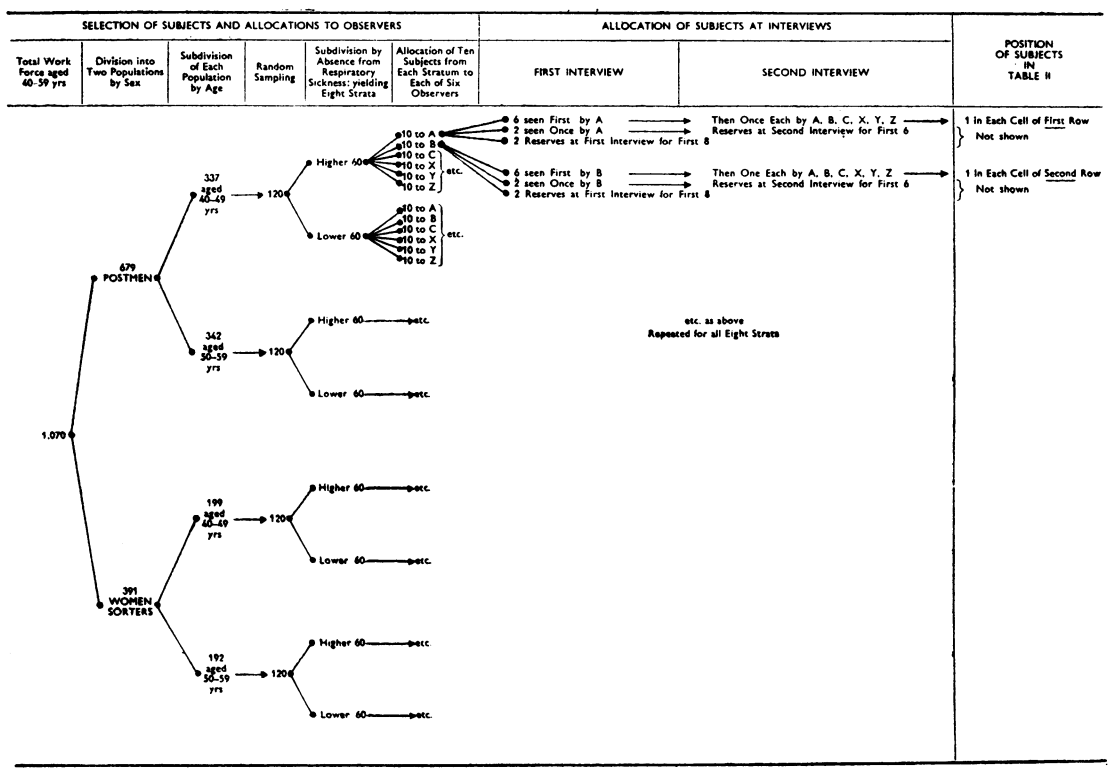
CHOICE OF OBSERVERS.—Since it is often convenient to employ medical auxiliaries in field surveys, we chose as observers three doctors and three health visitors. This allowed us to compare both the doctors with the health visitors as groups and also the individual observers within each group. Of the three doctors (A, B, and C), one was an epidemiologist, one a senior registrar, and one a consultant physician. The three health visitors (X, Y, and Z) were employed by the London County Council, one in a chest clinic and the others on general duties.

CHOICE OF POPULATION.—We wished to gauge the importance of observer differences under survey conditions, and looked for a readily accessible population of each sex with an appreciable incidence of bronchitis. We chose postmen from the London East Central District Office and women sorters from the Post Office Savings Bank in Hammersmith, London, between 40 and 59 years of age. The detailed sickness absence records of civil servants provided a means of checking the answers concerning past illnesses (Fletcher, Elmes, Fairbairn, and Wood, 1959).

SELECTION OF SUBJECTS AND ALLOCATION TO OBSERVERS.—Subjects were selected by random sampling. Every individual was asked to attend for interview, except for some reserves who were only approached if others failed to attend. The method of sampling and allocation to observers for the first interview is illustrated in Table I. The names of all postmen between 40–49 and 50–59 were found from the pay-roll and 120 were chosen at random from each age group. These two groups of 120 postmen were each listed in order according to the extent of each man's recorded sickness absence from respiratory causes. Division of each list into two equal parts with contrasting respiratory experience yielded four sub-groups of sixty postmen. The same procedure was then repeated for the women sorters. In this way eight sub-groups were obtained, each of sixty subjects, half the sub-groups consisting of postmen and half

TABLE I

SELECTION OF SUBJECTS AND ALLOCATION TO OBSERVERS



of women sorters. Six sets of ten subjects from each of the eight sub-groups were then allocated at random to the six observers. In this way a sample of eighty subjects stratified by age, sex, and respiratory sickness absence was allocated to each observer: the eight sub-groups formed by three successive halvings of the whole sample population are called "strata".

Table II shows the allocation to observers of subjects who were seen twice. Of the ten subjects in one stratum allocated to Observer A, six, chosen at random, were to be seen by A at the first interview; and these six subjects were then to be seen, one by each of the six observers, including A himself, at the second interview. If subjects are classified in a 6×6 Table (as in Table II) according to the observer at each interview, one of these six subjects appears in each of the six cells of the first row. The other four subjects in each set of ten were reserves, and are not shown in Table II. Two of these four reserves were seen once by A during the series of first interviews and could therefore be substituted for any of the first six subjects who were not available at the second interview. The remaining two reserves could be substituted for any of the first eight who failed to attend the first series of interviews.

The same allocation was repeated for all six observers and all eight strata. This completes Table II with eight subject in each of the 36 cells. It also provides 32 reserves, not shown in Table II, for the 48 subjects in each of the six rows, and half of these reserves were interviewed once. 288 subject were thus interviewed twice, and 96 subjects once only.

TABLE II

NUMBER OF SUBJECTS ALLOCATED TO OBSERVERS AT EACH OF TWO INTERVIEWS
(All 288 subjects interviewed twice)

		Observer at Second Interview					
		A	B	C	X	Y	Z
Observer at First Interview	A	8	8	8	8	8	8
	B	8	8	8	8	8	8
	C	8	8	8	8	8	8
	X	8	8	8	8	8	8
	Y	8	8	8	8	8	8
	Z	8	8	8	8	8	8

Of the 384 subjects chosen for at least one interview, only seven (1·8 per cent.) refused to be interviewed at all and eleven (2·9 per cent.) were not interviewed for other reasons (one through death, and the others through sickness, retirement, or transfer). Of the 288 subjects chosen for a second interview, only one refused (0·3 per cent.) and eleven lapsed for other reasons (3·8 per cent.). Each gap was filled by a reserve from the correct stratum.

BRIEFING.—The three doctors had all contributed to the design of the questionnaire and two of them had used a similar one in a pilot survey. The purpose of each question was explained to the health visitors. At a preliminary session each observer interviewed two hospital patients with chest illnesses. All six observers heard tape-recordings of these interviews and discussed together the interpretation of difficult answers. Written instructions, elaborated to cover any points of difficulty, were given to each observer (Appendix II).

CONDUCT OF THE SURVEY.—The survey began in November, 1956, and was completed in March, 1957. All interviews took place during working hours on Post Office premises, and postmen were seen on night duty when necessary. One of us (C.H.W.), who did no interviewing, recorded the sitting height and the weight of each subject and then made a series of slow and fast vital capacity tracings on a spirometer. The observer then interviewed the subject. The questionnaire was completed, and the observer then measured the peak expiratory flow rate using a Wright Expiratory Flow Meter (Wright and McKerrow, 1959).

After the first interview with the postmen, it was found that a number of "compulsory" questions had been left unanswered on the questionnaires. The percentage of these varied from 0·8 per cent. in the case of Observer Y to 4·8 per cent. in the case of Observer A (Table III). Since the answers to these questions were needed for the statistical analysis, they were obtained by playing back the tape-recording, where this sufficed, or by arranging a fresh interview. At the second interview with the postmen and at all interviews with the women sorters, the written answers were independently checked for completeness by C.H.W. before the subject left. The frequency of uncompleted questions might have been reduced by a better layout of the questionnaire to facilitate checking by the observer at the end of the interview.

TABLE III

UNCOMPLETED ANSWERS

Observer	Total Answers	Uncompleted Answers	
		No.	Percentage
A	896	46	5·1
B	896	32	3·6
C	896	25	2·8
X	896	52	5·8
Y	896	7	0·8
Z	896	14	1·6
Total	5,376	176	3·3

TABLE IV
DURATION OF RECORDED PART OF INTERVIEW (IN MINUTES)

Observer			A	B	C	X	Y	Z	All Observers
Interviews with	Postmen	Average	5·0	6·1	4·95	6·1	6·0	8·3	6·1
		Range	2–8	2–11·5	2·5–8·5	3–11·5	2·5–12·5	5–12	2–12·5
	Women Sorters	Average	4·75	7·6	4·9	5·75	4·5	8·5	6·0
		Range	2·5–8	3–11·5	2–8	4–10	2–9	4–24	2–24

LENGTH OF INTERVIEWS.—Table IV shows the duration of that part of the interview which was recorded on tape. The average duration for all observers was 6 minutes. This does not include the time spent on informal exchanges at the beginning and at the end of the interview, on taking the smoking history, or on taking the peak flow meter readings. The faster interviewers spent less time on these matters than the slower, thereby increasing the differences between them in the average duration of the whole interview. The average length of time that each subject was away from work was about 20 minutes.

PUNCH-CARDS.—The answers on the questionnaires were transferred on to 65-column punch-cards. It was found more convenient to have one card for each question, rather than to have one card for each subject as is more usual. Separate cards were punched for each of the two series of interviews. If an answer was positive, a hole was punched in the card for that question and for that series of interviews. The position of the hole was the same throughout for any one subject.

The total number of subjects answering "Yes" to a question at one of the series of interviews could be counted from the number of holes in the appropriate card. Cards were also punched to show individuals of different age or of different sex. The numbers in these different groups who answered "Yes" to one or more questions could be counted from the number of coincident holes when the corresponding cards were exactly superimposed. Additional cards could be punched to show individuals with any combination of characteristics. It was found useful to punch a number of negative cards in order to show those subjects who answered "No" to certain questions.

Cards of two contrasting colours were used for the two series of interviews, and enabled disagreements between the two corresponding answers by the same subject to be counted. The number of subjects who answered "Yes" at the first interview, and "No" at the second, could be found by placing the first interview card on top of the second interview card, and counting the number of dots where the colour of the second card showed through the holes in the first card. Those subjects who answered "No" at the first interview and "Yes" at the second could be counted in a similar way by reversing the order of the two cards.

METHOD OF STATISTICAL ANALYSIS.—The analysis was restricted to those subjects who were interviewed twice, and to questions which were answered by every subject in one of two alternative ways at each interview. The answers to twenty important compulsory questions, requiring the answer "Yes" or "No", were analysed. In four instances, the subject was classified in the same way by combining more than one answer. A subject who answered "No" to Question 31, "Does the weather affect your chest?" was then asked a "check" question, "Not even fog or cold?" (Question 31a). The answer "Yes" to either Question 31 or Question 31a was counted as positive, and the answer "No" to both questions as negative. The answers to Questions 43 and 43a were combined in the same was as those to Questions 31 and 31a. Subjects with breathlessness Grade 2 and over (Question 17) were considered as positive and those with Grade 1 as negative. An answer to part (*a*) of Question 29, "Do colds go to your chest?", was counted as negative, and an answer to part (*b*) as positive. At the first interviews with the postmen, Questions 4, 5, 9, and 10 (cough and phlegm during the day) were asked only if the answer to Question 2 (cough on rising) was positive. The data on these four questions were therefore incomplete and could not be analysed.

Systematic differences between observers in the number of positive answers reported to each question were tested for significance by analysis of variance. The quantity analysed was the difference in the total number of positive answers to one question given by the same subjects at the two interviews. The method of analysis is described in Appendix III.

RESULTS

The systematic tendencies of each observer are shown in Table V (opposite) and in Fig. 1 (overleaf).

TABLE V

PREVALENCE OF RESPIRATORY SYMPTOMS REPORTED BY THREE DOCTORS (A, B, C) AND THREE HEALTH VISITORS (X, Y, Z)

Percentage positive answers from a combined population of 288 subjects, 144 London postmen and 144 women sorters aged 40–59 years; adjusted

Question	Doctors (D)			Health Visitors (HV)			Significant Differences between Observers
	A	B	C	X	Y	Z	
(1) Cough usually	26·4	21·2	22·2	26·4	28·5	27·4	
(2) Cough on rising (Winter)	38·0	24·5	33·9	33·9	41·1	44·3	
(3) Cough on rising (Summer)	20·1	13·9	22·2	21·2	25·3	30·6	HV > D *
(6) Phlegm usually	30·2	29·2	28·1	32·3	20·8	34·4	
(7) Phlegm on rising (Winter)	30·7	24·5	24·5	30·7	35·9	38·0	HV > D *
(8) Phlegm on rising (Summer)	10·4	16·7	15·6	20·8	15·6	33·3	Z > X, Y *
(17: 2–5) Breathlessness	33·5	31·4	23·1	37·7	16·8	21·0	X > Y, Z **
(18) Breathing better in Summer	35·4	26·0	10·4	18·7	27·1	13·5	A > B, C ** A, B > C **
(20) Breathing varies day to day	9·9	8·9	9·9	6·8	12·0	18·2	Z > X, Y * Y, Z > X *
(21) Wheezing occasionally	63·0	54·7	51·6	57·8	50·5	56·8	
(23) Wheezing frequently	10·7	8·6	11·8	8·6	11·8	12·8	
(24) Symptoms of asthma	17·5	1·9	3·0	3·0	7·1	16·5	A > B, C ** Z > X, Y **
(29) Colds go to chest	40·8	36·6	40·8	33·5	34·5	27·3	D > HV *
(31) Weather affects chest	35·4	27·1	26·0	35·4	27·1	30·2	
or 31a: Not even fog?	53·1	47·9	31·2	42·7	36·5	44·8	A, B > C **
(41) Nasal catarrh (Winter)	42·9	35·6	30·4	35·6	30·4	51·2	Z > X, Y **
(42) Nasal catarrh (Summer)	23·1	25·2	20·0	33·5	13·7	23·1	X > Y, Z ** X, Z > Y **
(43) Chest illness last 3 years	39·2	37·2	38·2	48·6	31·9	40·3	
or 43a: Not even 'flu?	60·8	51·4	52·4	53·5	47·2	55·6	
(57) History of pneumonia	12·0	16·1	17·2	13·0	14·1	12·0	
(58) History of pleurisy	10·2	10·2	10·2	9·2	9·2	6·0	
(60) History of other chest illness	12·0	15·1	12·0	10·9	17·2	23·4	Z > X, Y *
(61) Heart trouble	6·8	1·6	4·7	2·6	7·8	4·7	
(62) Do you smoke?	61·2	60·2	60·2	59·2	60·2	60·2	

* = Differences significant at 5 per cent. ** = Differences significant at 1 per cent.

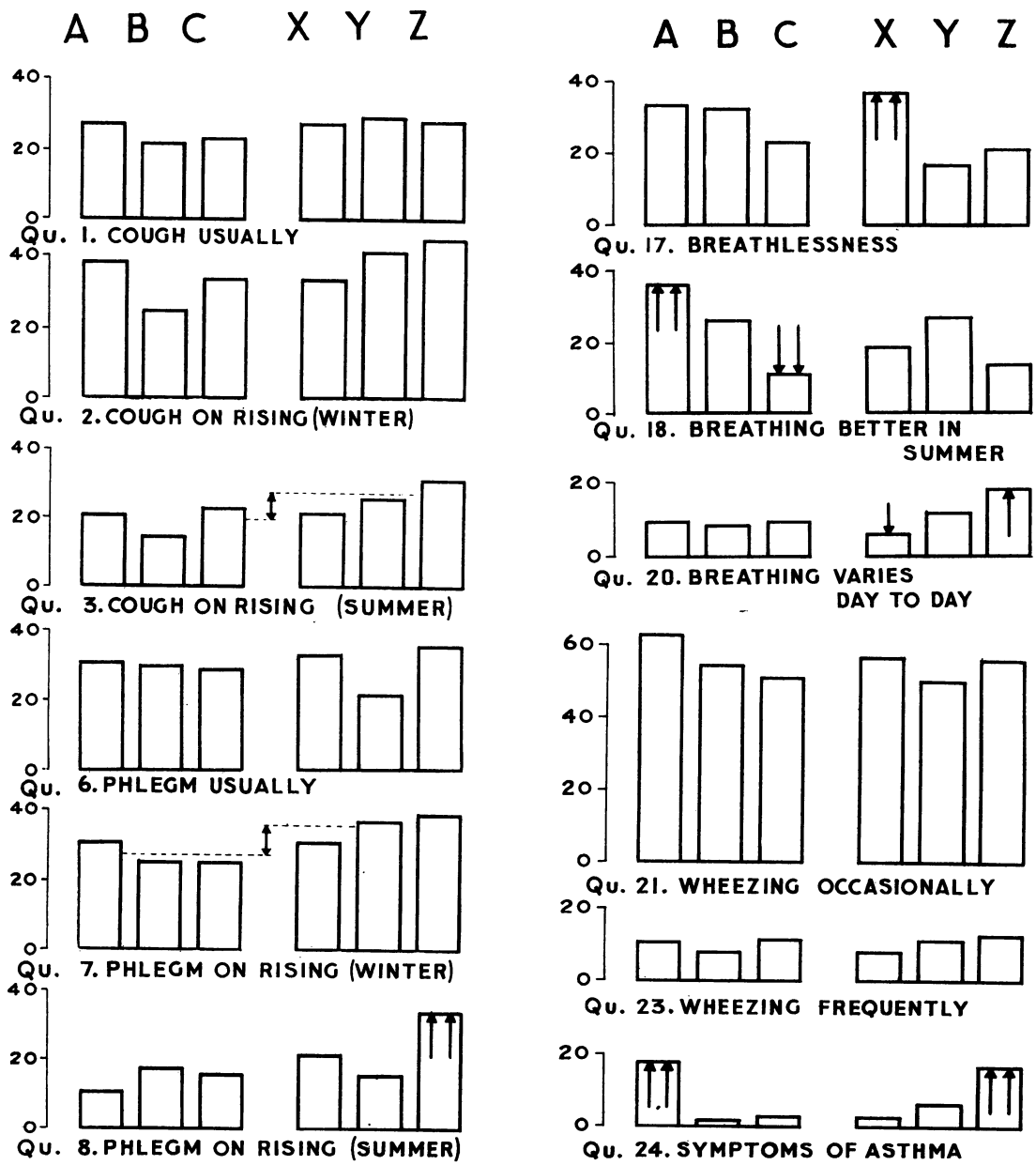


FIG. 1.—Prevalence of respiratory symptoms reported by three doctors (A, B, C) and three health visitors (X, Y, Z). Percentage positive answers from a combined population of 288 subjects, 144 London postmen and 144 women sorters aged 40–59 years; adjusted.

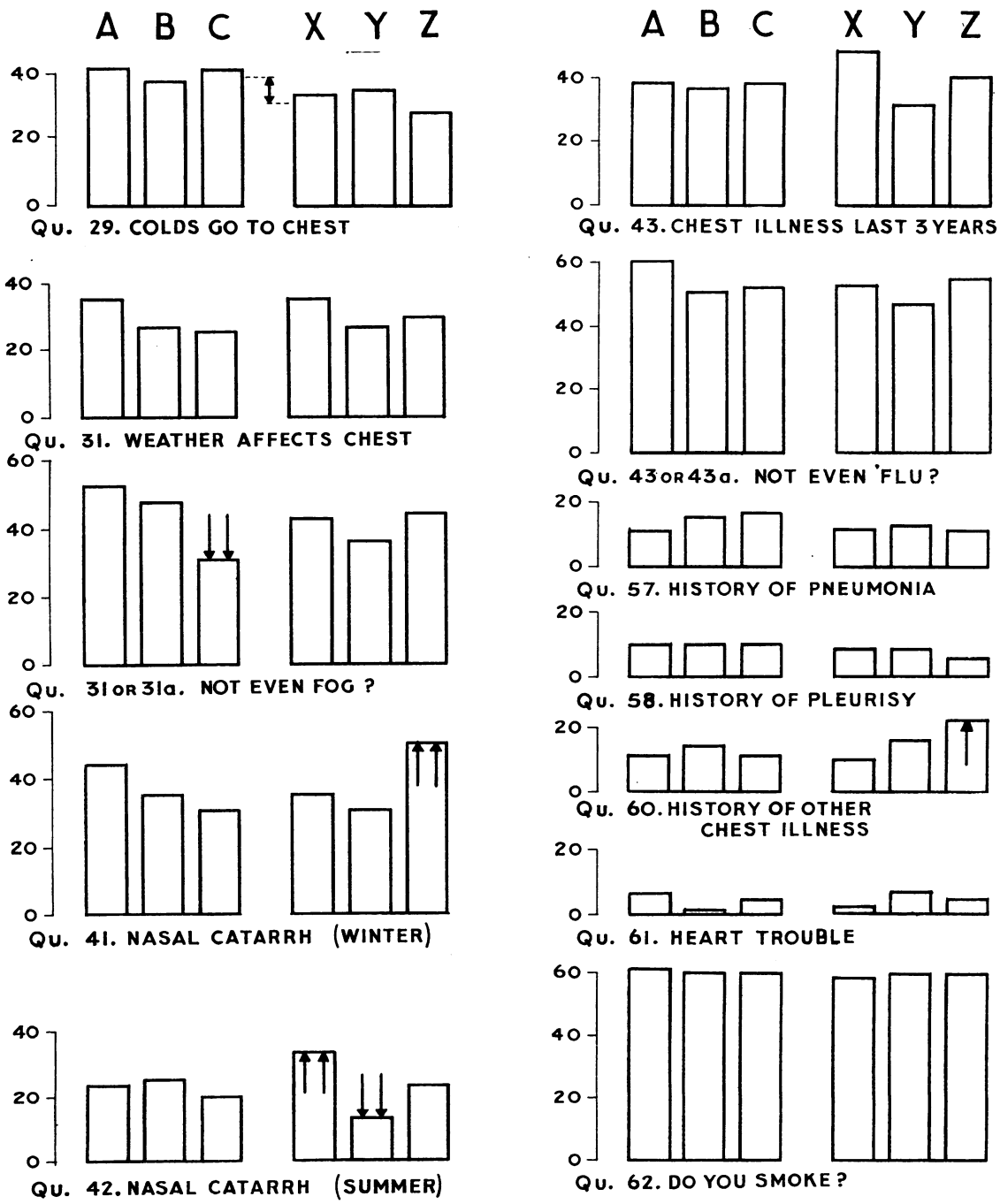

A B C X Y Z
40
20
0
Qu. 29. COLDS GO TO CHEST
Qu. 31. WEATHER AFFECTS CHEST
Qu. 31 OR 31a. NOT EVEN FOG ?
Qu. 41. NASAL CATARRH (WINTER)
Qu. 42. NASAL CATARRH (SUMMER)
A B C X Y Z
Qu. 43. CHEST ILLNESS LAST 3 YEARS
60
Qu. 43 OR 43a. NOT EVEN 'FLU ?
Qu. 57. HISTORY OF PNEUMONIA
Qu. 58. HISTORY OF PLEURISY
Qu. 60. HISTORY OF OTHER CHEST ILLNESS
Qu. 61. HEART TROUBLE
Qu. 62. DO YOU SMOKE ?

FIGURE 1.

Table V and the four parts of Fig. 1 show for each question the percentage of positive answers reported among all the subjects, both postmen and women sorters, interviewed by each observer. Comparison of the prevalence rates actually found by observers is often misleading, because each observer has interviewed different subjects. Observer Z, for instance, saw the 48 subjects in the last row of Table II at their first interview, and the 48 subjects in the last column of Table II at their second interview. At these 96 interviews she reported 52 positive answers to Question 2, a prevalence rate of 54·2 per cent. Now all six observers at both interviews reported 207 positive answers to the same question at 576 interviews, a rate of 35·9 per cent., so that Z appears to have over-reported the prevalence of this symptom by 18·3 per cent. But all six observers shared equally the second interviews with the subjects in the last row of Table II, and the first interviews with those in the last column. At these 96 interviews, 44 positive answers were reported, so that all six observers seeing the *same* subjects as Z reported a prevalence rate of 45·8 per cent., less by only 8·4 per cent. than that found by Z. The very high prevalence rate in Question 2 found by Z compared with the general average was in fact largely due to the chance allocation to her for interview of a group of subjects with a high prevalence of symptoms. In order, therefore, to show each observer's tendency to over- or under-report symptoms among the *same* subjects, every prevalence rate which is shown in Table V and Fig. 1 and which is discussed throughout the text has been corrected for these sampling variations between the subjects seen by different observers. The actual rate found by each observer has been adjusted by the difference between the rate found by all six observers at all 576 interviews, and that found by all six observers among the *same* subjects as that particular observer. Z's "adjusted" prevalence rate in Question 2 is therefore $54{\cdot}2 + (35{\cdot}9 - 45{\cdot}8) = 44{\cdot}3$ per cent. This method of adjustment was only made possible by the special design of this survey and is not of general application.

Significant differences between the prevalence of symptoms reported by observers are shown in Table V and Fig. 1. The same system of comparing observers was adopted throughout. The statistical significance of any group difference between doctors and health visitors, and of every individual difference between each observer and the other two in the same group was tested for each question. There are theoretical objections to this method of individual comparison (see Appendix III), but it gives a sufficient indication for practical purposes of the relative frequency and direction of real differences between the observers. Too much importance must not in any case be attached to one significant effect. A number of these may occur by the operation of chance, the choice of a significance level is itself arbitrary and alternative methods of analysis may not give quite the same results.

Significant group differences between doctors and health visitors occurred in four questions. (A technically significant group difference in Question 8 was due solely to the high prevalence reported by Z and is not shown in Fig. 1.) Health visitors reported more of the population as having morning cough in summer (Question 3) and more as having morning sputum in winter (Question 7), than did doctors. Doctors reported more colds going to the chest (Question 29). These three group differences were only significant at the 5 per cent. level. The difference in the prevalence of symptoms reported by each group amounted in each case to between 5 and 10 per cent. of the subjects interviewed. For instance, doctors found 27 per cent. of subjects to have morning sputum in winter, while health visitors found 35 per cent., a difference of 8 per cent. Health visitors also reported more positive replies to the three other questions on cough and sputum (1, 2, and 6) which were analysed, but the difference is not formally significant in any one question. Reasons for these differences are advanced later.

There were thirteen significant individual differences. Four of these occurred between the doctors. A reported a higher prevalence than B and C together in Questions 18 and 24. C reported a lower prevalence than A and B in Questions 18 and 31/31a. Nine significant differences occurred between health visitors. Z reported a higher prevalence than X and Y in five questions (8, 20, 24, 41, and 60). X over-reported as compared with Y and Z in Question 17 (grades 2–5) and in Question 42 and under-reported in Question 20. Y under-reported as compared with X and Z in Question 42.

Ten out of these thirteen individual differences were highly significant at the 1 per cent. level, and in these cases the difference in the prevalence of symptoms reported amounted to between 10 and 20 per cent. of the subjects interviewed. For instance, C found that the chest was affected by the weather in 31 per cent. of subjects (Question 31/31a) while A and B found this in 51 per cent. X found breathlessness Grade 2 or more (Question 17) in 38 per cent.

of subjects, while Y and Z found it in only 19 per cent. Significant over-reporting by one observer compared with the other two of the group (nine occasions) was commoner than under-reporting (four occasions).

These significant differences between individual observers suggest that A and Z were tending generally to report more symptoms, and C and Y less symptoms, than the other observers in their groups. This is borne out by a comparison, set out in Fig. 2, of the prevalence of symptoms reported by each observer, averaged for all 24 questions, among the postmen and among the women sorters. The systematic differences between the six observers in the average of the prevalence rates which they reported in each population are similar. The average prevalence rate found by A was in each case 4–5 per cent. greater than that found by C: A generally found 33·4 per cent. of the postmen to have symptoms, and C found 29 per cent.; A found 22·3 per cent. of the women sorters to have symptoms, while C found 17·4 per cent. In the same way Z generally found a higher prevalence of symptoms in each population than Y. This difference was about 3 per cent. in postmen (31·3 per cent. compared with 29·2 per cent.), but rather larger (5·6 per cent.) in the women sorters (23·9 per cent. compared with 18·3 per cent.). B and X were intermediate between the other observers in each group in their estimate of the level of symptoms in both populations.

In this statistical analysis, one effect of the differing techniques of observers has been studied, namely systematic differences between them in the number of positive symptoms reported. It must be remembered, however, that these systematic differences between observers do not show the full extent of the disagreements between their answers. If for example the number of subjects who answered "Yes" to the first observer and "No" to the second, in the same question, was equal to the number who answered "No" to the first and "Yes" to the second, each observer would report the same number of positive answers, yet the number of differing answers might be considerable. Nor do faults in observers' technique necessarily lead to systematic differences between the prevalence rates which they report. One observer may, for example, be given to different faults both of over- and under-reporting in the same question, the numerical effects of which cancel each other out; or two observers may be given to faults in the same direction which leave the differences between them unchanged.

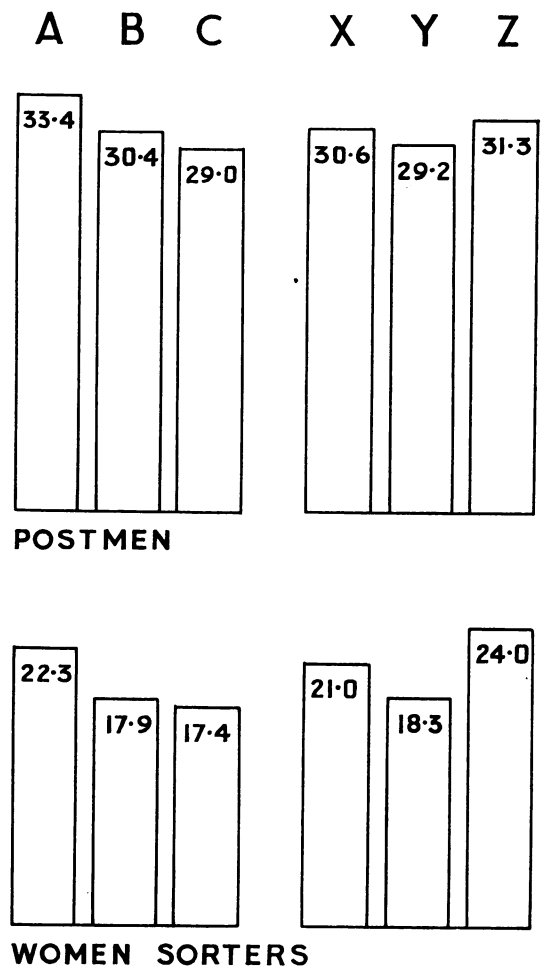


FIG. 2.—Prevalence of respiratory symptoms reported by three doctors (A, B, C) and three health visitors (X, Y, Z). Average of 25 different symptoms found in a combined population of 288 persons, 144 London postmen and 144 women sorters; adjusted.

TAPE-RECORDINGS.—The object of recording the interviews on tape was to find the reasons for disagreement between the answers obtained by two observers from the same subject to the same question. It was hoped in this way to explain the discrepancies between the symptom prevalence found by different observers, to assess the frequency of different causes of error, and to find ways to improve the phrasing of questions and the briefing of observers in future surveys.

Each subject was reassured that the interview was confidential and that the only purpose of the tape-recordings was to check the answers that the observers were writing down. All but two subjects consented to the interview being recorded, and in most cases no notice seemed to be taken of the machine.

160 disagreements were chosen at random from the answers to twelve questions in which the prevalence rates obtained by different observers differed significantly. The divergent answers were identified from the punch-cards. The recordings of the relevant parts of both interviews were played back from the tapes and transcribed in pairs on to master sheets, so that the reasons for the different answers might be discovered.

Observers had been briefed to use the exact wording of each question where possible and to record a definite answer as such. A "probing" question was required if the first answer was vague or on the borderline, but such questions were not to "force", *i.e.* to suggest the answer to the subject. If, after probing, the answer was still uncertain, it was then to be reported as negative. Examples of failure to observe these rules and of other reasons for disagreement are given below.

Question 3:

Do you cough at all when you get up, or first thing in the morning, in the summer? .. YES/NO

In Questions 1 and 6, the written instructions defined "usually" in borderline cases as "for most days for at least 3 months in the year". Observers had been instructed verbally that the same criterion of persistence should also apply to all questions about cough and phlegm. Symptoms of shorter duration, for instance with a cold, were to be excluded.

In one-third of the disagreements, the subject gave definite yet different answers to the correctly-asked question at the two interviews. Where the fault lay with the observer the commonest error was in assessing the persistence of the cough during the year. The health visitors, who as a group reported an excess of positive answers, interpreted vague answers such as "quite often" or "sometimes" as positive, without confirming that the cough was experienced "for most days for at least 3 months in the year". Coughs occurring only with colds were also incorrectly reported as positive.

Similar errors also gave rise to significant over-reporting by health visitors in Questions 7 and 8.

Question 17:

Grade 1: Are you ever troubled by breathlessness except on strenuous exertion?

Grade 2: (If yes) Are you short of breath when hurrying on the level or walking up a slight hill? .. .. ..

Grade 3: Etc. .. ..

If the first question was answered "No", the subject was put in Grade 1. If "Yes", the other questions were asked until a "No" was obtained and the grade was recorded as that opposite the last "Yes" received. Observer X overestimated the prevalence of breathlessness. Compared with the others, she placed too many subjects in grades other than 1 because, contrary to the rules, she often asked about Grades 2 and 3 and got anomalous positive answers when she had already had a negative answer to the first part of the question.

Question 18:

Is your breathing better in summer? .. YES/NO

There was no special briefing about this question. Disagreements mostly arose from radically different answers at each interview. There was no apparent reason for excessive prevalence reported by A.

Question 24:

In the past have you ever had attacks of wheezing and breathlessness, in between which your breathing was quite normal? .. YES/NO

This question was intended to discover those subjects who had suffered from spasmodic asthma. It was hoped that, by avoiding the word "asthma" and describing the symptoms, the question would mean the same thing to each subject. In fact, the question was so complex that many subjects seemed to answer "yes" without grasping its meaning. Some observers reported these answers as positive on the grounds that no supplementary questions were to be asked when a definite answer was given. Others went on to explain the meaning of the question to the subject and often got a negative answer. Observers A and Z, who over-reported compared with the others, also accepted as a positive answer a history of wheezing and breathlessness with a cold or with bronchitis. This was strictly correct according to the wording of the question, yet other observers realized that the question was not intended to include such cases and were reluctant to report them as positive. The fault here appeared to lie in a badly designed question.

Question 29:

When you have a cold in your head, does it go to your chest?

(*a*) *Never or only occasionally?* .. ..

(*b*) *Usually or always?* .. .. ..

A cross had to be put in the space opposite one part of the question only. Part (*a*) was treated as negative and part (*b*) as positive. According to the instructions, a cold "going to the chest" should have been followed by cough and phlegm. "Usually" implied that more than half the subjects' colds went to the chest, "occasionally" implied less than half.

The doctors, who reported significantly more colds as going to the chest than the health visitors, sometimes failed to offer the proper alternatives of "occasionally" or "usually", and so forced an answer. One observer, for instance, offered the alternatives "usually" or "always", both of which are in category (*b*). This sort of error arose partly from the fact that the question as written could not be asked in an easy conversational manner. One subject answered "Frankly, I'm fortunate that way. I very rarely get a cold in the head—if I do, it does end on the chest". This was wrongly interpreted as "never or occasionally". "Usually or always" was the correct answer because although colds were rare, more than half of them affected the chest.

Questions 31 and 31a:

(*31*) *Does the weather affect your chest?*.. YES/NO
(*31a*) (*If No, check: "Not even fog or cold?"*) YES/NO

No special instructions had been given about the interpretation of these questions, but the effect on the chest had subsequently to be determined in answer to Questions 35–39, and an effect on the nose, throat, or head should clearly not count as an effect on the chest.

The answers to Question 31 were analysed and no significant differences were found between the prevalence of positive answers found by different observers. The answers to Questions 31 and 31a were also combined. "Yes" in this case meant a positive answer to either question, and "No" a negative answer to both. A and B reported significantly more positive answers than C to both questions taken together. Their excess was due to the second "check" question (31a). They wrongly accepted replies such as "Well, I might catch a cold if it was foggy" as positive, without further inquiry as to the nature of the "cold". Some subjects said that their chest was affected, yet were unable to say in what way in answer to Questions 35–39. Observers often ignored the ruling that vague answers, not clarified even after probing, should be counted as negative. One striking example of "forcing" was recorded. The observer, a doctor, thought from the previous answers that the subject was bronchitic, and that his chest should be affected. In spite of the answer "No" to both Questions 31 and 31a, he persevered and obtained a reluctant "Yes" to "Not even a real pea-souper?" Clinical experience may lead an observer to form an opinion about the subject's condition as the interview proceeds. This episode suggests that it may also tempt him to deviate from the rules by forcing the subject to give the expected answer.

Questions 41 and 42:

Apart from colds do you usually have a stuffy nose or catarrh at the back of your nose?

(*41*) *In the winter?* .. .. .. YES/NO
(*42*) *In the summer?* .. .. .. YES/NO

Observer Z over-reported in Question 41, and Observer X in Question 42. The reason in each case was failure to exclude symptoms which only occurred with colds. Other observers found on probing that many indefinite positive answers related only to colds and correctly reported them as negative.

Questions 57–60:

Apart from the past 3 years have you ever had:

(*57*) *Pneumonia?* .. .. YES/NO
(*58*) *Pleurisy?* .. .. YES/NO
(*59*) *Tuberculosis?* .. .. .. YES/NO
(*60*) *Other Chest Disease?* .. .. YES/NO

Observer Z over-reported in Question 60. Her interviews were carried out more slowly and possibly gave more opportunity for recollection of minor illnesses. On some occasions of disagreement the subject volunteered a history of disease to Z and not to the other observer. On other occasions Z herself reminded the subject of a disease that had been mentioned earlier in the interview.

To assess the relative importance of different causes of disagreement, a further 149 disagreements were randomly selected from the answers to nine important questions and investigated in the same way. The results are shown in Table VI. 62 per cent. of the disagreements were classified as "due to the observer", *i.e.* at one of the two interviews the question was asked incorrectly or the answer misinterpreted. The remaining 38 per cent. of the disagreements seemed to be beyond the observer's control. About half of these were "due to the subject", *i.e.* radically different answers were given on each occasion to correctly asked questions. The rest were "due to the question", *i.e.* the question having been asked and answered correctly, the answer was on the borderline and could not be clarified by further probing. The answers were classified by the one of us (C.H.W.) who had not been an observer.

TABLE VI

CAUSES OF 149 DISAGREEMENTS IN NINE QUESTIONS

Disagreement due to	No.	Percentage
Observer	93	62
Subject	32	21
Question	24	16
Total	149	100

The three observers independently checked the results and found them to be substantially correct (Table VII, overleaf). The proportions of these causes of disagreement varied between questions, *e.g.* 80

per cent. of the causes of disagreement lay with the observers in Question 6, "Do you usually bring up phlegm from your chest?", whereas only 28 per cent. of causes lay with the observers in Question 17, "Do you have to walk slower than most people on the level?". These errors of technique by observers gave rise to positive answers four times more often than to negative answers.

TABLE VII

INDEPENDENT CLASSIFICATION OF CAUSES OF 102 DISAGREEMENTS IN FIVE QUESTIONS

Disagreement due to	Classification by One Observer	Check Classification by 3 Observers
Observer	78	74
Subject	16	23
Question	8	5
Total	102	102

In a small number of disagreements (about 3 per cent.), the technique of questioning was correct but the wrong answer was written down by mistake by one of the observers.

DISCUSSION

The purpose of a questionnaire on respiratory symptoms is to compare their prevalence and to follow their rate of progression in populations with different characteristics. The data are thus used for quantitative *comparison* between *groups*. This contrasts fundamentally with the usual objective of clinical medicine, a *decision* about an *individual*. Here an accurate diagnosis of the patient's condition is of first importance in order to determine the correct treatment or prognosis. In surveys of prevalence, on the other hand, valid comparisons can be made with less accurate data, provided that a sufficient population is studied, and that inaccuracies of reporting are randomly distributed between the groups being compared.

Accuracy of reporting is still desirable in prevalence surveys for several reasons. The most serious type of inaccuracy is caused by errors in the observer's own technique, because this may introduce bias and vitiate comparison of his results with those of other observers. Until the technique of interview can be improved it is important, therefore, to determine the likely extent of such systematic variation between observers, and to take precautions to avoid its effects.

Our results showed that two observers, A and C, differed from each other in the average of their estimate of the prevalence of 24 different symptoms in the whole population by about 5 per cent. (Fig. 2). In a single question, the greatest difference between the prevalence rates estimated by two different observers was 25 per cent.: in Question 18, A found 35·4 per cent. of subjects to have better breathing in summer, and C found only 10·4 per cent. If the three doctors are compared with the three health visitors in the 24 questions which we analysed, there is little difference in the average of their estimates of the prevalence of positive answers; doctors found 25·1 per cent. and health visitors 25·7 per cent. in the combined population. The largest difference between them in one question was 8·3 per cent. in Question 7: doctors found 26·9 per cent. of subjects to have morning phlegm in winter and health visitors found 35·2 per cent. The effects of these systematic differences between observers may thus be lessened by using several observers and grouping their results.

Much of the variation between the two answers to the same question appeared to be randomly distributed between different observers, and did not, therefore, lead to significant differences between the prevalence rates which they reported. Nonetheless, this random type of variation, though less serious than the systematic variation between two observers, should be reduced wherever possible, because it affects the *reproducibility* of the answers to the question, and hence both their *accuracy* and *discriminatory power*.

Though these several terms are closely related, it is important to keep in mind the distinction between them. The value of the answers to a question depends on their *discriminatory power*, *i.e.* their ability to discriminate between subjects with or without evidence of bronchitis. This has been discussed by Fletcher and others (1959). This *discriminatory power* is affected to an important degree by the *accuracy* with which the subjects describe their symptoms. Unfortunately, owing to the subject's fallibility, it is often impossible to determine the truth about his symptoms. In this case the *accuracy* of his answers can be determined only by their *reproducibility*, *i.e.* the extent to which the same answers are given to a question by the same subjects on two different occasions. Although reproducibility is a necessary condition for accuracy, it must be remembered that the accuracy of a question may be less than is implied by its reproducibility. A subject who swallowed his sputum might persistently deny its production to all questioners. Another, troubled by borborygmi, might think that the sound originated in the chest, and consistently, yet wrongly, admit to wheezing. These reproducible false negative and positive answers further reduce the discriminatory

power of the question. They are less important in the case of objective tests. It is fallacious, therefore, to argue that because symptoms are as reproducible as physical signs they are as accurate or valuable in diagnosis.

Reproducibility is usually calculated in inverse form as *percentage disagreement*, *i.e.* the number of subjects who answer differently on two occasions expressed as a percentage of all subjects. In the fourfold table shown in Table VIII this quality is equal to $\frac{100(b+c)}{n}$. In three out of the 24 questions which were analysed (57, 58, and 61), the small prevalence of positive answers reduced the percentage disagreement to a meaningless value. In the other 21 questions, the average percentage disagreement was 14·9 per cent. The greatest value was 24·3 per cent. in Question 21 (occasional wheezing) and the smallest was 2·4 per cent. in Question 62 ("Do you smoke?"). These figures imply substantial differences in the accuracy with which different answers were being reported. It is possible that differences in the discriminatory power of the questions found by Fletcher and others (1959) reflect to a certain extent the relative accuracy of the answers. If this accuracy could be improved, a different assessment of the value of the questions might result.

TABLE VIII

CLASSIFICATION OF *n* SUBJECTS ACCORDING TO THE ANSWER "YES" OR "NO" AT EACH OF TWO INTERVIEWS

		Number of Subjects answering at Second Interview		Total
		Yes	No	
Number of Subjects answering at First Interview	Yes	a	b	$a+b$
	No	c	d	$c+d$
Total		$a+c$	$b+d$	$a+b+c+d=n$

Much of the inaccuracy with which answers are reported is clearly irreducible by efforts on the part of the observer. Consistent false positive and false negative answers, and some of the random variation in the two answers to the same question cannot be avoided. There is the inevitable difficulty of classifying into discrete categories subjects with a continuous range of severity of symptoms, so that there are borderline answers which can justifiably be interpreted as positive or negative (*i.e.* "error due to the question"). There are also occasions when the subject quite unaccountably gives different replies (*i.e.* "error due to the subject").

On the other hand, the predominant cause of all disagreements between the answers to the same question at two interviews appears to be "error due to the observer", which accounts for about 60 per cent. of these discrepancies (Table VI). Many of these errors by observers appear not to lead to any systematic bias between their results, probably because all observers committed a certain number of errors which affected their reporting of answers to the same extent. For instance, the question in which the proportion of "errors due to the observer" was highest, Question 6, showed no significant difference between the prevalence rates found by different observers. Errors by the observers therefore cause not only systematic differences between them in the prevalence of symptoms reported, but also contribute to the "random" variation between the two answers to one question, and therefore to the general inaccuracy of the answers. It seems reasonable to suppose that a major improvement in the technique of the interview would not only eliminate any systematic bias between different observers, but also increase the accuracy and discriminatory power of the answers to the questions.

Apart from purely clerical errors, the disagreements "due to the observer" arose from failure to keep to the briefing. Observers were told to report as such a definite "Yes" or "No", but to try to clarify vague or borderline answers by "probing". Some forgot this rule and followed a definite answer with unwarranted probing which finally led to the opposite answer. Sometimes the opposite mistake was made, and when the answer was indefinite some observers did not try to clarify it by probing. In another group of cases the observer reworded the written question so as to suggest the answer. This mistake was easily made when, as in Question 29, the question was so worded that it could not be asked naturally. Some of the probing questions were also phrased so as to bias the subject's reply.

A further source of error became apparent in Questions 31 and 31a. The observer sensed from clinical experience that two answers by the same subject were inconsistent. He therefore rephrased the second question and asked it again in order to elicit the expected answer. The reply may possibly have been nearer the truth, but consistency with the replies obtained by other observers was lost. Observer A was tending in general to report more positive answers than the other two doctors (Fig. 2) and it is likely that the answers reported by him to other questions were biased in the same way. Memory is fallible, so that determined efforts to get at the real truth may only increase confusion. The aim in comparative surveys of prevalence should rather be to question in a simple and consistent way and scrupulously to observe the briefing instructions. Interviewers should abstain from unwarranted

probing even when the reply is unexpected. In this way variation in the subject's answers, though not avoided, is lessened, and is randomized between observers so as not to vitiate a comparison of their results. Health visitors or lay interviewers, starting with fewer fixed beliefs about the answers they will get, could quite possibly be trained to report more consistently than doctors. It is relevant that the difference in the number of symptoms reported in this survey by two doctors, A and C, was as great as that between two health visitors, Y and Z (Fig. 2).

The consistency of an observer's results does not necessarily depend on the time taken over the interview. The observer with the fewest disagreements with the other observers was a health visitor. She interviewed at a brisk pace. She asked the first question as the subject sat down and succeeding ones almost before the subject had finished answering. She adopted a matter of fact tone and was sometimes impatient at irrelevant answers. This contrasts with another health visitor who had the most disagreements with the other observers. She took time to put subjects at their ease but in doing so appeared to suggest positive answers. A business-like approach may thus be successful while well-meant efforts to encourage the subject may bias his replies.

Our observers were deliberately given no more and no less briefing than is commonly given on such occasions. Two doctors had had previous experience with a similar questionnaire. One of them had been largely responsible for its construction, and he recorded the fewest positive answers of any of the observers. Errors of observers' technique appeared usually to give rise to positive answers. Training and experience would therefore probably tend to reduce the number of symptoms recorded.

Some of the questions which we used are open to criticism, and contributed to errors by the observers or confusion on the part of the subjects. Some questions were not conversationally worded and the observer therefore rephrased them, but did so in such a way as to suggest the answer. Some questions were too complex and the subject was being asked in effect more than one question at the same time. When interpreting the answers about cough and phlegm, for example, observers readily forgot that the symptoms should persist for "most days for at least 3 months in the year". This error has been avoided in later versions of the questionnaire by the inclusion of a separate question. The inquiry about symptoms of asthma (Question 24) confused both subjects and observers. This question should be split into two or three parts.

The instruction of the new interviewers, the preliminary interviews with hospital patients, and the discussion between all the interviewers were insufficient in our survey to avoid bias due to errors of technique. Moreover, the two most experienced interviewers, A and C, differed as much as any two other interviewers in the prevalence rates which they reported (Fig. 2). We do not consider that the briefing of our interviewers was seriously wrong, but insufficient time was spent both in preliminary interviews with patients and in subsequent discussion of difficult points. Most important of all, a single test interview with a patient was insufficient to make certain that the observers were putting the rules of briefing into practice.

The degree of competence required of interviewers must be judged in the light of the particular circumstances of the survey. If, for example, the results of a single interviewer are to be compared with those of others, avoidance of systematic bias is more important than if the results of several interviewers are first grouped together.

After a full study of the questionnaire and the briefing instructions and discussion with experienced colleagues of any difficulties, the new interviewer should conduct a pilot series of at least ten interviews with subjects likely to show at least some chest symptoms. These interviews can be attended by one or more colleagues who can afterwards discuss critically the technique employed. The interviews can with advantage be recorded on tape so that doubtful points can be studied at leisure by playing back any part of the recording, or it may be more convenient and less embarrassing for a nervous interviewer to see the subject alone and for criticism to be confined to the tape recording. Another method of instruction which could usefully supplement the essential pilot interviews is for the new interviewer to study word for word transcripts of questions and answers where known faults of technique have occurred. Although a written transcript cannot convey important aspects of the interview such as emphasis or tone of voice, yet this method is convenient and obviates the need of a tape-recorder, the transcripts can be studied at leisure and can cover a comprehensive variety of faults. The interviewer might be asked to comment on the transcripts and it should be ensured that he can reliably detect any errors of technique. The transcripts should include about an equal number of questions and answers where the technique has been correct and incorrect.

In this survey, we have been concerned chiefly with the problem of the systematic errors introduced by observers into estimates of symptom prevalence.

A related problem, still largely unsolved, is that of the relative importance of the different questions. The inclusion of some questions must depend, amongst other things, on the circumstances and the purposes of the investigation: the shortest possible questionnaire would sometimes be needed, sometimes a longer and more detailed one would be suitable. We consider, however, that in order to maintain comparability between the results of different workers, some questions should always be included. The wording of the questionnaire, and the method of reporting of replies should be standardised, at least for a period. We have therefore consulted with others concerned in surveys of respiratory symptoms, and we have all agreed upon a questionnaire and instructions which appear to us to be the best in the light of present knowledge.* Modifications will obviously be necessary at a later date as experience increases.

SUMMARY

Each person in a random sample of 144 London postmen and 144 women post-office sorters aged 40 to 59 was interviewed twice at an interval of 6 weeks. At each interview the same questions about respiratory symptoms were asked by one of three doctors or three health visitors.

The prevalence of symptoms reported by the six observers, both individually and as two groups, was compared, and a number of significant differences were found. The frequency and extent of these differences are estimated. The differences between the number of symptoms reported in the same subjects by two doctors was as great as that between two health visitors.

The causes of disagreement in the answers at the two interviews were investigated from tape-recordings. In 62 per cent. of cases of disagreement, one of the observers failed to abide by the briefing instructions; about half of the remainder were due to radically different replies by the subject and half to the difficulties of interpreting vague or borderline answers.

Failure to observe the rules of briefing tended to increase the frequency of positive answers.

The significance of these results is discussed, and suggestions are made for lessening the effects of observer variability in future surveys.

* Copies of this questionnaire, which is shortly to be published, are available from Dr. C. M. Fletcher, Postgraduate Medical School, Ducane Road, London W.12.

We are indebted to the Post Office authorities for permitting the investigation and for their efficient help in arranging the interviews; and to Dr. W. E. Chiesman and Dr. M. C. W. Long for their continuing encouragement of such surveys among civil servants. We are also happy to acknowledge the goodwill of the Union of Post Office Workers and the Civil Service Clerical Association.

We are grateful to Dr. Richard Doll, Dr. Peter Armitage, and Dr. Ian Sutherland for statistical guidance and criticism, and to Dr. Peter Elmes and the three London County Council health visitors, who participated in the interviews. The method of card-punching is a modification of that used by Mr. E. G. Brisch.

Miss Doreen Bobby did the computing and Mrs. B. M. Hunt drew the diagrams.

REFERENCES

Cochrane, A. L., Chapman, P. J., and Oldham, P. D. (1951). *Lancet*, **1**, 1007.
Fletcher, C. M., Elmes, P. C., Fairbairn, A. S., and Wood, C. H. (1959). *Brit. med. J.*, **2**, 257.
Higgins, I. T. T. (1959). *Ibid.*, **1**, 325.
Kinsey, A. C., Pomeroy, W. B., and Martin, C. E. (1948). "Sexual Behaviour in the Human Male." Saunders, Philadelphia.
Schilling, R. S. F., Hughes, J. P. W., and Dingwall-Fordyce, I. (1955). *Brit. med. J.*, **1**, 65.
Tukey, J. W. (1949). *Biometrics*, **5**, 99.
Wright, B. M., and McKerrow, C. B. (1959). *Brit. med. J.*, **2**, 1041.

The three following appendices will be found overleaf:

Appendix I. Respiratory Symptoms Questionnaire

Appendix II. Notes on Use of Questionnaire

Appendix III. Method of Statistical Analysis

APPENDIX I

RESPIRATORY SYMPTOMS QUESTIONNAIRE

INTERVIEWER : A B C X Y Z INTERVIEW: 1st 2nd

NAME ..... SERIAL NO. ..... SEX .....

ADDRESS ..... BIRTH DATE ..... AGE .....

..... WEIGHT .....kilos

DEPARTMENT ..... STEM HEIGHT .....cms.

OCCUPATION ..... DATE .....

TAPE RECORDER Tape No..... Start at.....

COUGH

Do you usually have a cough? ... Yes/No 1.

Do you cough at all when you get up or first thing in the morning:
(*a*) in the winter? ... Yes/No 2.
(*b*) in the summer? ... Yes/No 3.

Do you go on coughing during the day:
(*a*) in the winter? ... Yes/No 4.
(*b*) in the summer? ... Yes/No 5.

PHLEGM

Do you usually bring up phlegm from your chest? (not from back of nose) ... Yes/No 6.

Do you bring up any phlegm at all when you get up first thing in the morning:
(*a*) in the winter? ... Yes/No 7.
(*b*) in the summer? ... Yes/No 8.

Do you go on bringing it up during the day:
(*a*) in the winter? ... Yes/No 9.
(*b*) in the summer? ... Yes/No 10.

Is this phlegm:

(*a*) always clear white, grey, or blackish? ... [ ] 11.

(*b*) occasionally yellow or green, at least in parts? ... [ ] 12.

(*c*) usually yellow or green? ... [ ] 13.

ONSET AND DURATION OF COUGH AND PHLEGM

How long have you had this morning cough/phlegm in the winter:
(*a*) less than 1 year? ... [ ] 14.

(*b*) 1 to 3 years? ... [ ] 15.

(*c*) more than 3 years? ... [ ] 16.

BREATHLESSNESS (average in last winter) Grade No. 17.

Grade 1. Are you ever troubled by breathlessness except on strenuous exertion? (No disability) ... [ ]

(If yes) Grade 2. Are you short of breath when hurrying on the level or walking up a slight hill (Slight disability)... [ ]

(If yes) Grade 3. Do you have to walk slower than most people on the level? Do you have to stop after a mile or so (or after ½ hour) on the level at your own pace? (Moderate disability)... [ ]

(If yes to either) Grade 4. Do you have to stop for breath after walking about 100 yards (or after a few minutes) on the level? (Severe disability) ... [ ]

(If yes) Grade 5. Are you too breathless to leave the house, or breathless after undressing? (Total disability) ... [ ]

Is your breathing better in summer? ... Yes/No 18.
(If yes) Grade in summer—after checking ... 19.
Does your breathing vary from day to day? ... Yes/No 20.

WHEEZING

Does you breathing ever sound wheezy or whistling? ... Yes/No 21.

(if yes) (*a*) occasionally (for example when you have a cold)? ... 22.

(*b*) most days (or nights)? ... 23.

In the past have you ever had attacks of wheezing and breathlessness, in between which your breathing was quite normal? Yes/No 24.
(If yes) How old were you when the attacks began? ... 25.
Do you still have them? ... Yes/No 26.
(If no) How old were you when they stopped? ... 27.

COLDS

When you have a cold in the head, does it go to your chest? (*i.e.* is it followed by increased cough or phelgm?)

(*a*) never or only occasionally? ... 29.

(*b*) usually or always? ...

WEATHER

Does the weather affect your chest? ... Yes/No 31.
(If no, check "not even fog or cold?") ... Yes/No 31(*a*).
(If yes) What sort? Fog ... Yes/No 32.
Cold ... Yes/No 33.
Other ... Yes/No 34.
Does it make you wheeze? ... Yes/No 35.
Does it make you cough? ... Yes/No 36.
Does it make you bring up phlegm? ... Yes/No 37.
Does it make you breathless? ... Yes/No 38.
Any other effect? ... Yes/No 39.
(If yes to 36 or 37) Does the cough/phlegm clear up entirely when the weather is not (as in 32–34) ... Yes/No 40.

NASAL CATARRH

Apart from colds do you usually have a stuffy nose or catarrh at the back of your nose?
(*a*) in the winter? ... Yes/No 41.
(*b*) in the summer? ... Yes/No 42

CHEST ILLNESSES

During the last 3 years have you had a chest illness which has kept you in bed, off work, or indoors? ... Yes/No 43.
(If no) check "not even 'flu?" ... Yes/No 43(*a*).
(If yes to 43 or 43(*a*))
(*a*) Did you have increased cough with the illness(es)? ... Yes/No 44.
(*b*) Did you have increased phlegm with the illness(es)? ... Yes/No 45.

(*c*) Did you have only one such illness? ... 46.

(*d*) Did you have more than one illness? 2–5, 6–10, 10+ (Ring) ... 47.

(e) Estimated average duration ... 48.

Doctor's diagnoses in the illness(es)	Number of times diagnosed	
Doctor not seen	.....	49.
No diagnosis	.....	50.
Bronchitis	.....	51.
Influenza	.....	52.
Pneumonia	.....	53.
Pleurisy	.....	54.
Asthma	.....	55.
Other (specify)	.....	56.

(Ring appropriate number and put the number of illnesses after each alternative)

Apart from the past 3 years have you ever had:
Pneumonia? Age..... Yes/No 57.
Pleurisy? Age..... Yes/No 58.
Tuberculosis? Age..... Yes/No 59.
Other chest disease?..... Age..... Yes/No 60.
Have you ever had heart trouble? ... Yes/No 61.
(If yes) Doctor's diagnosis.....

(STOP RECORDER)

SMOKING

Do you smoke? ... Yes/No 62.
(If no) Have you ever smoked? (Record "No" if less than 1 per day for a year) ... Yes/No 63.
(If yes) How much?

	Now	Changes during past 15 years
Cigarettes per week ...	.....	.....
Oz. tobacco/week (hand-rolled) ...	.....	.....
Oz. tobacco/week (pipe) ...	.....	.....
Cigars (number/week) ...	.....	.....

N.B. Weekend

Age of starting regular cigarette smoking .....
Age of stopping regular cigarette smoking .....

Tape Recorder end at.....

APPENDIX II

Notes on Use of Respiratory Symptoms Questionnaire

Start by introducing yourself as a doctor or nurse. Say that you are going to ask some questions about chest symptoms and that the replies are entirely confidential.

Each section may be prefaced with a few words such as "Now, Mr. ....., I want to ask you about a cough".

Stick to the ***actual*** *wording of each question as far as possible.* If the answer to the question is doubtful ask any supplementary questions you like to clarify what the subject means.

Where squares are provided for X to be inserted, the X must only be put in one square, since the answers are alternative and if one is positive the other must be negative.

Mention the subject's name when you switch on recorder.

Individual Questions

Cough

The aim is to provide a classification of severity according to those who cough first thing in the morning and those who have a repetitive cough during the day and whether this cough occurs only in the winter or is perennial. Clearing the throat does not count as cough.

Questions 1 and 6.—"Usually" means most days for at least part of the day. Transient cough with a cold should be excluded, the cough must last for at least 3 months during the year. The point of these questions is that they have been included in previous epidemiological studies and if they are excluded comparison might be difficult.

Question 2.—Even if the answer to 1 is "No", 2 should be asked.

Questions 2 and 3.—The words *at all* should be stressed.

Question 4.—Only asked if the answer to 2 is positive. The words "go on coughing" imply a frequency of about once an hour or more.

Phlegm

Any suitable word can be used instead of phlegm according to local custom, but "phlegm" should come up from the chest and post-nasal discharge is excluded. Some subjects admit to bringing up sputum without admitting to cough. This claim should be accepted and it is not necessary to change the answers under cough if they are negative but the answers to phlegm are positive. If phlegm is coughed from the chest and is swallowed, this counts as positive.

Question 7.—To be asked even if the answer to 6 is negative.

Question 9.—To be asked only if answer to 7 is positive.

Questions 11–13.—If the phlegm is stated to be "grey, etc.", the alternative of 12 must be asked to be sure that the phlegm is *always* mucoid.

Questions 14–16.—Remember that the coronation was three years ago.

Breathlessness

Question 17.—*This refers to average condition this winter.* If Grade 1 is answered "No", this is the grade. After that, go on asking till you get a "No". The grade is the last one to which the answer is "Yes" (In the case of Grade 3 *either* question may be answered "Yes").

Write the number of the grade in the appropriate space. Write the grade in summer in the space provided.

Questions 18–20.—To be asked in *all* cases. The word "breathing" is preferable to "breathlessness".

Wheezing

Question 21.—To be asked as written. Whistling is not a separate question.

If subjects are uncertain about the meaning of "wheezing", it may be worth asking whether their relatives, *e.g.* husbands or wives, notice wheezing during the night.

Questions 22 and 23.—These are meant to distinguish intermittent wheezing with long periods free from persistent wheezing.

Questions 24 to 27.—These are meant to discover subjects who have had occasional asthmatic attacks in the past. In some cases the symptoms will have stopped completely, but in others they may have become persistent. In the latter case, 27 should give the age at which they became persistent.

Colds

Question 29.—This should be "usually *or always*".

Questions 30 and 40.—Only check Questions 2–16 if all of Questions 1–13 have been answered "No". Remember to note the Question (30 or 40) on account of which these questions are checked.

Questions 35–39.—If the subject has admitted to cough/phlegm/breathlessness, ask if the weather *makes these worse.*

Chest Illnesses

Question 43.—Add "at home" after "indoors". Remember these are illnesses since the coronation.

Question 45.—Refers only to the illnesses with increased cough/phlegm.

Question 47.—Ring the number of illnesses as well as putting the X in the square.

Question 48.—Give duration in weeks. If there has been a wide variation give the range, *e.g.* "1 to 13 weeks".

Question 49 to 56.—If necessary, offer these diagnoses, *i.e.* say "Was it called bronchitis, 'flu, pneumonia, pleurisy, asthma?"

Questions 57 to 60.—If any of these illnesses have been experienced more than once, record all occasions.

APPENDIX III

METHOD OF STATISTICAL ANALYSIS

The answers given at two interviews by n subjects to a question requiring the answer "Yes" or "No" can be classified as in Table VIII by a fourfold table. Such a table can be constructed for any number of the 288 subjects in Table II, all of whom were interviewed twice. The smallest possible value of n is 1, and the largest possible value is 288. The difference in the number of positive answers reported among these n subjects at the two interviews is $b-c$; this quantity is positive if more positive answers were recorded at the first interview than at the second. For one subject, $b-c$ can only take the three alternative values $+1$, 0, or -1. The variance of $b-c$ can be resolved into its component parts; the sources of variation and degrees of freedom are shown in Table IX. There is no replication in the design, and therefore no term in the analysis of variance corresponding to residual variation. The second order interaction is used as an estimate of residual variance.

TABLE IX

ANALYSIS OF VARIANCE

Source of Variation	Degrees of Freedom
Main Effects	
Between First Observers (Rows of Table II) ..	5
Between Second Observers (Columns of Table II)	5
Between Strata	7
Interactions	
Strata × First Observers	35
Strata × Second Observers	35
First × Second Observers	25
Residual (Second Order Interaction)	175
Total	287

The total value of $b-c$ for the 48 subjects in the first row of Table II is equal to the number of positive answers reported by A at the first set of 48 interviews, less those reported *among the same subjects* at the second set of 48 interviews, by all six observers, including A himself, who shared these interviews equally between them. This quantity will be positive if A reports a higher prevalence than all the observers, negative if he reports a lower prevalence.

Comparison of the totals of these rows will therefore indicate how each observer over-reports or under-reports symptoms compared with all the observers at interviews with the *same* subjects. Each column total of $b-c$ in Table II affords a similar comparative measure of the tendency of an observer to over- or under-report symptoms, but in this case a *negative* value of $b-c$ represents over-reporting by the single observer and a positive value represents under-reporting.

In any one question, estimates of the variance of $b-c$ derived from the rows or columns of Table II, divided by the residual variance, will yield values of F which can be used to test the significance of differences between observers in the number of positive answers reported.

The two interviewer terms in Table IX, each with five degrees of freedom, can be further subdivided, as shown in Table X. Within each group of three observers there are three alternative ways of comparing one observer with the other two. It was generally found that differences between the mean values of $b-c$ for rows of Table II were of opposite sign to the differences between the corresponding columns. Observers therefore who reported more symptoms than others, as judged by the first interviews, also generally did so at the second. The individual terms in Table X, each with one degree of freedom, can in such cases be recombined, as shown in the last column, by adding the corresponding sums of squares and obtaining a fresh variance estimate based on two degrees of freedom.

TABLE X

SUBDIVISION AND RE-COMBINATION OF OBSERVER TERMS

Source of Variation		Degrees of Freedom		
		First Observers	Second Observers	Re-combined
Between Doctors and Health Visitors		1	1	2
Within Doctors	Between A and B, C	1	1	2
	Between B and C	1	1	2
Within Health Visitors	Between X and Y, Z	1	1	2
	Between Y and Z	1	1	2
Total		5	5	10

The significance of a group difference between doctors and health visitors, and of each of the six individual differences between one observer and the other two in the same group, were tested in each of the 24 questions. The significant results shown in Fig. 1 and Table V were in every case based on a value of F found by combining the difference between the corresponding rows and columns of Table II. This re-combination yielded many significant differences which did not appear from the comparison of either rows or columns alone. In other instances the degree of significance of the differences was increased from the 5 per cent. to the 1 per cent. level. The significance of all first order interactions was tested, but none was found to be significant.

Tukey (1949) has discussed the difficulties of comparing individual means in the analysis of variance. The method of comparing individual observers, used here, somewhat increases the percentage of positive results which would be expected purely by chance, in a series of significance tests, above that of the particular significance level chosen. No satisfactory yet simple method of individual comparison, however, is known. Comparison of one observer with two others leads to simplicity of computing and is satisfactory in practice (see page 178).