Thorax (1969), **24**, 4.

Repeatability of physical signs in airways obstruction

S. GODFREY, R. H. T. EDWARDS, E. J. M. CAMPBELL, P. ARMITAGE, AND E. A. OPPENHEIMER[1]

From the Department of Medicine, Royal Postgraduate Medical School and the School of Hygiene and Tropical Medicine, London

We have examined the observer variation in the detection of physical signs in airways obstruction. In the first study, 10 relatively experienced physicians examined 11 patients for the presence of six familiar signs and seven less well-known signs. Six experienced observers then underwent a training period after which they examined another group of 21 patients. The repeatability of all the signs fell about midway between that expected by chance and the maximum possible. There was no difference between the familiar and unfamiliar signs. The training period resulted in a slight but not significant improvement in reliability.

We have examined the observer error in the elicitation of some familiar and some less well-known physical signs in patients with airways obstruction. To provide additional information on the importance of skill and training, the study was undertaken in two stages. In the initial study a group of physicians were briefly instructed in the elicitation of the signs, some of which were relatively unfamiliar, and they then examined a series of patients. In the second stage, a smaller group of the more experienced members of the team were trained together for a time and they then examined another series of patients.

METHODS

In the initial study, 10 observers examined 11 patients. These observers were all physicians at the Hammersmith Hospital. Seven were members of the Royal College of Physicians, one was of equivalent American standard, and two were post-registration house physicians. In the second study, six observers examined 21 patients. These observers were all members of the Royal College of Physicians and ranged in experience from consultant to registrar. The patients for both studies were selected from the wards and outpatient clinics, and their co-operation was sought after the nature of the study had been explained to them.

PHYSICAL SIGNS All patients were examined while reclining on a couch with the back rest raised to 45°. The following list of physical signs is based on those described by Campbell (1969), and can be divided into two groups.

(A) Classical signs familiar to all observers

1. Filling of the external jugular veins during expiration
2. Increased resonance to percussion
3. Diminished intensity of breath sounds
4. Wheezes (continuous rales) at the bases
5. Crepitations (discontinuous rales) at the bases
6. Contraction of the sternomastoid muscles during inspiration.

(B) Signs unfamiliar to most observers at beginning of study

7. Excavation of the supraclavicular fossae during inspiration
8. Excessive contraction of the scalene muscles during inspiration
9. Tracheal tug—a downward movement of the trachea during inspiration
10. Costal paradox—inward or biphasic movement of the costal margin during inspiration instead of the normal outward movement
11. Exaggerated 'pump-handle' movement of the upper chest—an exaggeration of the upward and forward movement of the anterior ends of the ribs instead of the normal outward and upward rotation of the mid-point of the ribs ('bucket-handle' movement)
12. Tracheal length; measured in finger-breadths between the lower border of the cricoid cartilage and the upper border of the sternal notch at the end of expiration
13. Forced expiratory time—the time in seconds taken to expel the vital capacity forcibly (Lal, Ferguson, and Campbell, 1964). The expiration was stopped if it lasted longer than 7 seconds in order not to

[1]Present address: Department of Medicine, University of Chicago, Illinois

distress the patients, and it was not measured in all patients in the initial study for the same reason.

The observer was instructed to decide whether or not the abnormality was present for signs 1 to 11 inclusive; in order to simplify the analysis, grading was not allowed. If he was in doubt, he was told to regard the sign as absent. If the sign in question was not present during quiet breathing, the patient was asked to breathe more deeply. The presence of the sign during either quiet or deep breathing was recorded as positive. Signs 12 and 13 were expressed in finger-breadths or seconds.

CALCULATION OF RESULTS There are several possible ways of calculating observer error (Armitage, Blendis, and Smyllie, 1966). We have chosen to use the standard deviation agreement index (S.D.A.I.), to conform with previous studies, and an experience agreement index (E.A.I.). The meaning of these indices is briefly explained.

S.D.A.I. Suppose n observers study a given sign in k patients and let the total number of positive findings for all observers in all the patients be r. This number expressed as a proportion of the total possible will then be $r/nk=p$ (say). Now, in the individual patient there will be a number of positive findings $=r_i$ (say) representing a proportion p_i of the total possible, where $p_i=r_i/n$. The greater the agreement between observers in any one patient, the nearer will r_i approach n or zero and hence the nearer will p_i approach 1 or 0. In a group of patients with an overall proportion of positive signs p, complete agreement will occur when the individual values of p_i are either 1 or 0. In other words, for a given value of p, the wider the scatter of p_i, the closer the agreement. The S.D.A.I. measures this scatter as follows:

$$\text{S.D.A.I.} = \sqrt{\frac{\Sigma r_i^2 - (\Sigma r_i)^2/k}{k-1}} \qquad (1)$$

It can be shown that the maximum possible value of this index for any given value of p is given by

$$\text{S.D.A.I. (max)} = n\sqrt{p(1-p)} \qquad (2)$$

and the chance value of the index based on a binomial distribution is given by

$$\text{S.D.A.I. (chance)} = \sqrt{np(1-p)} \qquad (3)$$

E.A.I. While the S.D.A.I. measures the overall variation, it does not take any account of the possibility that some observers are more 'correct' than others, and it is not applicable to quantitative signs such as 12 and 13. For these reasons we have also calculated an index to compare the findings of the group of observers with those of its most experienced member (E. J. M. C.). In any one patient, for any one sign, there will be a number of agreements a_i with the opinion of E. J. M. C. In the whole group of a_i, and, since the maximum number of agreements Σa_i, and, since the maximum number of agreements possible will be nk, we have

$$\text{E.A.I.} \frac{\Sigma a_i}{nk} \times 100\% \qquad (4)$$

It was also possible to apply this index to the two quantitative signs by accepting agreement as being $\pm 1/2$ finger-breadth and ± 1 second respectively.

PLAN OF STUDIES

First stage In the initial study, nine observers were individually instructed by the tenth and most experienced observer in the elicitation of signs in three patients. The 10 observers then examined a further group of 11 patients at a single session lasting 1½ hours. The observers rotated around the patients and did not communicate with one another. The forced expiratory volume in the first second ($F.E.V._{\cdot 1\cdot 0}$) and the relaxed vital capacity (V.C.) were measured for each patient before and after the session by the method of Freedman and Prowse (1966).

Second stage Six observers, five of whom (including E. J. M. C.) had participated in the initial study, then underwent a period of training together for one month. During this time they examined patients individually and then compared their results. When differences of opinion arose, they re-examined the patient together. The six observers then examined a further group of 21 patients over the course of the next three months. The form of the examinations was similar to that in the initial study. Most of the patients were seen by all six observers within about 30 minutes, but in two or three cases the examinations were spread over five or six hours. $F.E.V._{\cdot 1\cdot 0}$ and V.C. were recorded on all the patients before examination. No patient in whom the spirometry had been varying over the preceding three to four weeks was included in either study.

PATIENTS There were 11 patients aged between 27 and 77 years, with a mean age of 54 years in the first study. Most of them were suffering from chronic bronchitis, emphysema or asthma, but one patient had pulmonary tuberculosis and one had sarcoidosis. Neither of these two patients had evidence of airways obstruction, and they were included as 'blind controls'. There was a reasonably wide distribution of spirometry among the patients (Fig. 1A). The mean $F.E.V._{\cdot 1\cdot 0}$ fell from 1·50 l. before the study to 1·23 l. afterwards, and the mean V.C. fell from 2·77 l. to 2·42 l., but the significance of the fall in the individual patient was low ($0{\cdot}1<P<0{\cdot}5$).

In the second study there were 21 patients aged 28 to 73, with a mean age of 53 years. All had chronic bronchitis, emphysema, or asthma, but in two cases it was very mild. There was a similar scatter of $F.E.V._{\cdot 1\cdot 0}$ and V.C. as in the first study (Fig. 1B). The mean $F.E.V._{\cdot 1\cdot 0}$ (1·10 l.) and V.C. (2·28 l.) did not differ significantly from the first study. Spirometry was not measured after the examinations. The frequency with which the most experienced observer reported positive findings was approximately the same in the two groups.

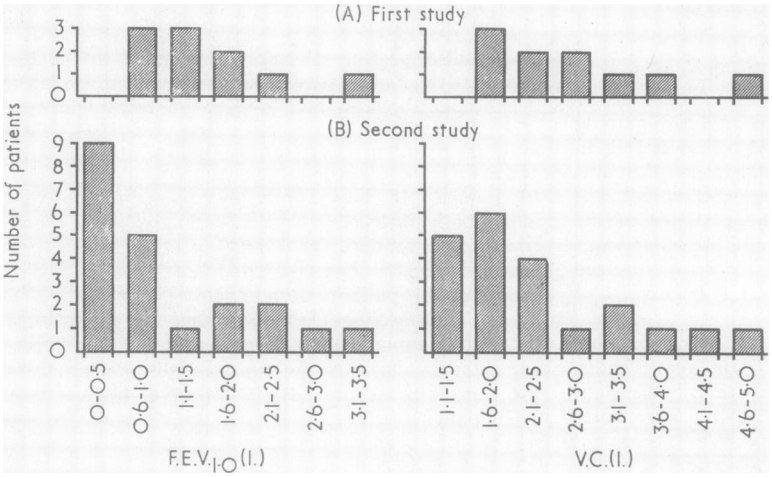

FIG. 1. *Distribution of spirometry in patients in the two studies.*

RESULTS

In order to show that the signs that were studied were indeed related to airways obstruction, we have compared the frequency of their presence in relation to the F.E.V.$_{1\cdot0}$ (Table I). All the signs were thought to be present by the majority of observers far more frequently in patients with an F.E.V.$_{1\cdot0}$ of 1 l. or less compared with those with an F.E.V.$_{1\cdot0}$ of more than 1 litre. We did not attempt any further analysis of this relationship in the present study.

FIRST STAGE There was considerable observer variation during the initial study as judged by either the S.D.A.I. or the E.A.I. (Table II). Part of the difference in S.D.A.I. between the signs was due to the frequency with which the signs were recorded as positive, *i.e.*, the differing values of P (see Calculation of Results). This has been allowed for by calculating the maximum and chance values for the S.D.A.I. for each value of P from 0 to 1. The observed values of the S.D.A.I. for each sign have then been plotted on the same axes (Fig. 2A). This shows that the indices for all the signs were scattered between the chance and maximum possible values. To compare one sign with another, we have calculated the S.D.A.I. as

TABLE I

PERCENTAGE OF PATIENTS WITH SIGN PRESENT[1]

Sign	F.E.V.$_{1\cdot0}$= 1·0 l. or less (17 patients)	F.E.V.$_{1\cdot0}$= 1·1 l. or more (14 patients)
1. External jugular filling	53	0
2. Resonance to percussion	44	14
3. Diminished breath sounds	62	21
4. Wheezes	70	36
5. Crepitations	38	21
6. Use of sternomastoids	74	29
7. Excavation of supraclavicular fossae	91	14
8. Use of scaleni	100	39
9. Tracheal tug	91	43
10. Costal margin paradox	74	43
11. Upper chest movement	65	36
Mean value for group		
12. Tracheal length (finger-breadths)	0·8	1·5
13. Forced expiratory time (sec.)	6·9	6·2

[1] Percentage of patients in whom the majority of observers found positive signs is given in relation to their F.E.V.$_{1}$. The mean value for tracheal length and forced expired time in each group is also given.

TABLE II

S.D.A.I. AND E.A.I. FOR EACH SIGN IN THE TWO STUDIES

	S.D.A.I.		E.A.I.		S.D.A.I. as % Max.[1]	
	First Study	Second Study	First Study	Second Study	First Study	Second Study
1. External jugular filling	1·68	2·35	62	86	43	80
2. Resonance to percussion	2·93	2·13	82	75	63	72
3. Diminished breath sounds	2·07	2·06	67	84	40	68
4. Wheezes	3·36	2·24	87	88	68	75
5. Crepitations	3·41	1·75	78	87	65	64
6. Use of sternomastoids	3·46	2·08	73	79	70	68
7. Excavation of supraclavicular fossae	3·90	2·22	76	94	75	79
8. Use of scaleni	2·05	2·08	84	93	46	77
9. Tracheal tug	2·59	1·27	71	88	55	52
10. Costal margin paradox	2·51	1·70	69	86	48	60
11. Upper chest movement	3·23	2·30	60	93	62	75
12. Tracheal length	—	—	66	76	—	—
13. Forced expiratory time	—	—	80	92	—	—
Mean	2·83	2·02	73	86	58	70

[1] In the last two columns, the S.D.A.I. is given as a percentage of the maximum possible for the particular sign.

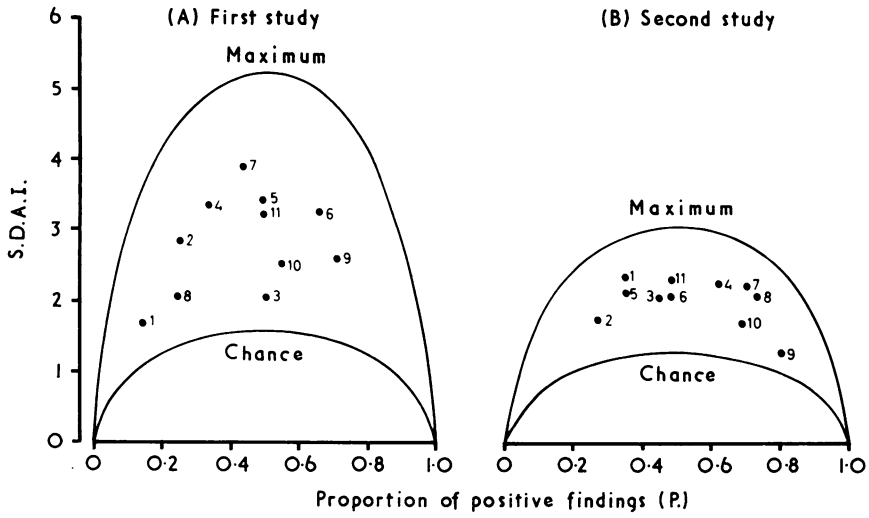

FIG. 2. *Relationship between the S.D.A.I. for each sign and frequency with which it was recorded as positive. The theoretical chance and maximum values are shown. The values differ in the two studies because of the number of observers and patients (see text). The numbering of the signs corresponds with the text.*

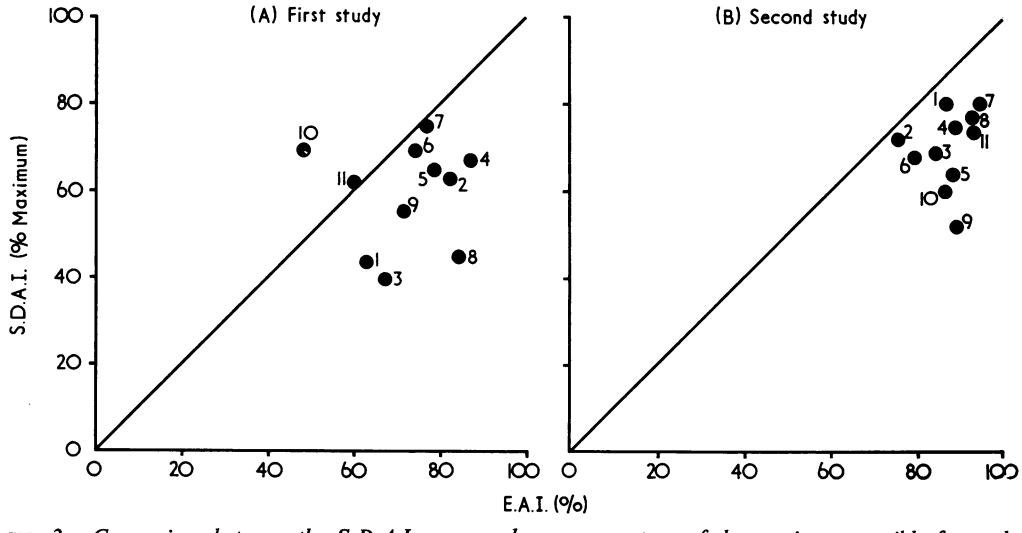

FIG. 3. *Comparison between the S.D.A.I. expressed as a percentage of the maximum possible for each sign and the E.A.I. The numbering of the signs corresponds to the text.*

a percentage of the maximum possible value for each sign from the data contained in Figure 2. These percentages appear in the last column of Table II. On this basis excavation of the supraclavicular fossae and contraction of the sternomastoids were the easiest signs to agree on, while external jugular filling and diminished breath sounds were the least reliable. There was poor agreement between the S.D.A.I. as a percentage of maximum and the E.A.I. (already a percentage of maximum) for the various signs (Fig. 3A).

SECOND STAGE The observer variation has been expressed in the same form for the second study (Table II and Fig. 2B). Because of the different number of observers and patients, the chance

and maximum values of the S.D.A.I. are different, but the general distribution of the indices was similar. Expressed as a percentage of the maximum S.D.A.I., the highest indices and most reliable signs were excavation of the supraclavicular fossae, upper chest movement, and wheezes; the least reliable were tracheal tug and costal paradox.

There was no close agreement between the S.D.A.I. and E.A.I. (Fig. 3B). The E.A.I. was consistently higher than the S.D.A.I. expressed as a percentage of maximum.

In both studies the reliability of tracheal length and forced expiratory time as measured by the E.A.I. was similar to the reliability of the other signs.

The effect of the training period between the two studies was further examined by comparing the performances of the five observers common to both studies. Patients were selected from the two groups and matched for age and spirometry. The mean values for the 10 patients selected from the initial study (55 years, $F.E.V._{\cdot 1\cdot 0}=1\cdot 47$ l., V.C. $=2\cdot 68$ l.) did not differ significantly from the values for the 10 patients from the second study (51 years, $F.E.V._{\cdot 1\cdot 0}=1\cdot 34$ l., V.C. $=2\cdot 58$ l.). The S.D.A.I. expressed as a percentage of the maximum possible for each sign in the first study has been plotted against the value for the second study (Fig. 4). Most of the indices were higher in the second study, indicating some improvement, but this was only significant in the case of external jugular filling. An analysis of improvement based on the E.A.I. gave similar results for all the signs including tracheal length. It was impossible to compare forced expiratory time in the two studies because it was not measured in enough patients in the first study.

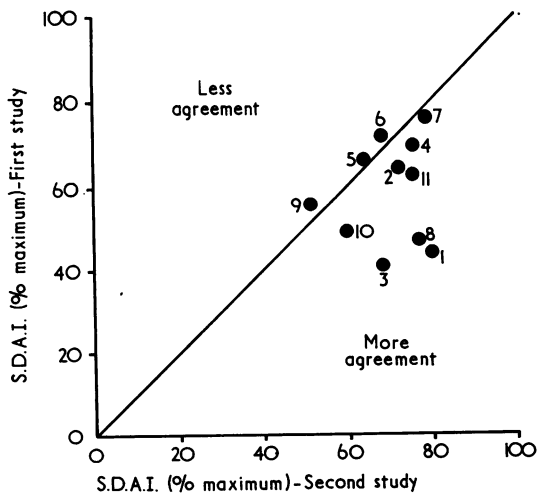

FIG. 4. *Comparison of results of five observers common to both studies on matched pairs of patients. The numbering of the signs corresponds to the text.*

DISCUSSION

Observer variation has now been demonstrated in a wide variety of medical techniques (Garland, 1960; Fletcher, 1964). One of the earliest demonstrations was the finding of observer variation in the detection of the clinical signs of 'emphysema' (Fletcher, 1952). Most of the signs in this study were actually related to increased lung volume.

In a more recent study, Smyllie, Blendis, and Armitage (1965) also found considerable observer variation in the detection of a wide variety of signs in the chest, but only included four signs related to lung volume or airways obstruction. We have shown (Table I) that the signs we have chosen to study were related to the presence of airways obstruction. Some are probably related to the secondary over-inflation of the chest and also to the age of the patient and the duration of symptoms (Campbell, 1958; Campbell, 1969).

The design of the study provided a severe test of the signs because some of them were unfamiliar to most of the observers, although in fact these signs fared no worse than the more familiar signs. The 'all or none' method of recording also has its disadvantages because the disagreement between observers will increase if the sign is close to the boundary between present and absent. Oldham (1968) has recently discussed this problem and he pointed out that discussion between observers might be fruitful if one could suggest a reason for their previous differences. We have assumed that these differences were likely to be due to technique and we hoped that the period of training between the two studies, when observers did discuss their findings, would eliminate some of the error. The fact that we found little difference between our two studies and between our quantitative and qualitative signs suggests that boundary errors were not the major source of disagreement.

The statistical methods for measuring observer variation have been fully discussed previously (Armitage *et al.*, 1966). The standard deviation agreement index (S.D.A.I.) is an excellent measure of group scatter, but it fails to allow for the possibility that the majority of a group of observers might in fact be giving the wrong answer; if five out of six observers record a sign as absent, the contribution to the calculated value of the

S.D.A.I. will be the same as if five out of six had recorded it as present. For this reason we also calculated the experience agreement index (E.A.I.) based on agreement with the decision of the most practised observer. We realize that this approach is also open to criticism, but we felt that the extra information was worth having. The E.A.I. increased more from the first to the second study than did the S.D.A.I. (Fig. 3). The fact that the majority was more often correct in the second study would have relatively little effect on the S.D.A.I. (as explained above) but would have a much greater effect on the E.A.I. provided that the experienced observer was consistent in his reporting.

When using the S.D.A.I. or similar indices, the most reliable results will be obtained by having as large a sample of observers and patients as possible. On the other hand, practical and clinical considerations make a large number of observers undesirable and, as we have shown (see Patients), repeated examination can cause a deterioration in the patient's spirometry. We have attempted to compromise on these requirements by the different formats of the two studies and by choosing a reasonably wide scatter of abnormality among the patients. The least frequently reported sign in the first study was recorded in 15% of possible cases and the most frequently reported sign in the second study in 80% (Fig. 2); the other signs in both studies were reported within these limits.

The results of our study agree quite well with the previous studies of signs in chest disease (Fletcher, 1952; Smyllie *et al.*, 1965). All the studies have demonstrated that no sign is infallible and that most of them have a repeatability about midway between that due to chance and the maximum possible. It has been possible to compare some of the signs in the present two studies with the same signs reported in the two earlier studies quoted above. The results for the S.D.A.I. as a percentage of its maximum value are given in Table III. The general trend of the results is similar, and the higher values in our present studies probably reflect a closer awareness of the problem and greater attention to detail.

We know of no studies of the effect of training on the observer variation in physical signs. It is tacitly assumed that experience always increases reliability (Abrahams, 1932). The period of training certainly improved the agreement of most of the physical signs, but the only significant improvement was in the detection of external jugular filling. These results rather suggest that while attention to detail may result in improvement in even familiar signs, a certain amount of observer variation will always remain.

TABLE III

COMPARISON OF RESULTS FOR EQUIVALENT SIGNS IN TWO STUDIES FROM THE LITERATURE AND THE PRESENT TWO STUDIES

	Study			
	Fletcher (1952)	Smyllie *et al.* (1965)	First Study	Second Study
No. of observers	8	9	10	6
No. of patients	20	20	11	21
Signs	*S.D.A.I. as % Maximum*			
Increased resonance	47	40	63	72
Impaired breath sounds	52	65	40	68
Use of scaleni[1]	54	—	46	77
Use of sternomastoids[1]	54	—	70	68
Movement en bloc[2]	53	—	62	75
Crepitations	—	59	65	64
Wheezes	—	46	68	75

[1] Fletcher only quotes use of accessory muscles without specifying muscle group.
[2] Movement en bloc is taken as equivalent to 'pump handle' movement of upper chest (sign 11).

We wish to express our thanks to our patients and our colleagues for taking part in these studies, and to Miss M. Chandler for help with the computations.

REFERENCES

Abrahams, A. (1932). Errors in diagnosis. *Lancet*, **2**, 661.
Armitage, P., Blendis, L. M., and Smyllie, H. C. (1966). The measurement of observer disagreement in the recording of signs. *J. roy. statist. Soc.*, *A*, **129**, 98.
Campbell, E. J. M. (1958). *The Respiratory Muscles and the Mechanics of Breathing*. Lloyd-Luke, London.
—— (1969). Physical signs of diffuse airways obstruction and lung distension. *Thorax*, **24**, 1.
Fletcher, C. M. (1952). The clinical diagnosis of pulmonary emphysema—an experimental study. *Proc. roy. Soc. Med.*, **45**, 577.
—— (1964). The problem of observer variation in medical diagnosis with special reference to chest diseases. *Meth. Inform. Med.*, **3**, 98.
Freedman, S., and Prowse K. (1966). How many blows make an $F.E.V._{1.0}$? *Lancet*. **2**, 618.
Garland, L. H. (1960). The problem of observer error. *Bull. N.Y. Acad. Med.*, **36**, 570.
Lal, S., Ferguson, A. D., and Campbell, E. J. M. (1964). Forced expiratory time: a simple test for airways obstruction. *Brit. med. J.*, **1**, 814.
Oldham, P. D. (1968). Observer error in medicine. *Proc. roy. Soc. Med.*, **61**, 447.
Smyllie, H. C., Blendis, L. M., and Armitage, P. (1965). Observer disagreement in physical signs of the respiratory system. *Lancet*, **2**, 412.