Comparative rating of consultation performance: a preliminary study and proposal for collaborative research

I.M. STANLEY, MRCP, FRCGP
C.A. WEBSTER, BA
J. WEBSTER, FSS, MBCS

SUMMARY. Factors governing the appropriateness, reliability and validity of rating scales in the measurement of professional performance are reviewed. The origin and preliminary testing among undergraduates and general practitioners of a brief consultation rating schedule is described.

Statistical criteria are proposed for the analysis of ratings, by groups, in the comparison of consultation performance. Using these criteria the capacity of the 10 rating schedule items to discriminate between two contrasting consultations was examined. Each of the items was used at some time by students or doctors to express significant preference for the same consultation; and on this basis all the items are considered to merit inclusion. One item showed highly significant intra- and inter-observer reliability.

The schedule is reproduced in full, together with a data-collection document and significance chart, with the aim of encouraging groups of doctors to test the validity of the items in the comparison of other pairs of consultations. It is proposed that future versions of the schedule should reflect the experience of such groups in testing existing items and in defining additional items which satisfy the proposed criteria.

Introduction

THE rating scale has become familiar in research and educational assessment as a method of quantifying judgement.[1] The measurement of professional performance by rating is appropriate in situations where more objective measures do not exist or are not feasible. Here, limited objectivity is both a weakness and a strength; rating cannot provide absolute statements by one individual about the performance of another because in the process it reveals something of the rater. The method, then, is better suited to use by groups, where collective wisdom can amount to objective statements and the bias of individuals may be seen in atypical ratings.[2]

Although widely used, rating has attracted a good deal of criticism. Recently McNamara and McNamara have reiterated the need for reliability and validity to be established before conclusions are drawn:

> 'Before using rating schedules to judge the general practitioner's performance we must have a clear notion of what constitutes effective performance and how it is to be measured.'[3]

I.M. Stanley, Lecturer and C.A. Webster, Research Assistant, Department of Community Medicine and General Practice, University of Leeds; J. Webster, Principal Lecturer, Department of Mathematics and Computing, Leeds Polytechnic.

© *Journal of the Royal College of General Practitioners*, 1985, 35, 375-380.

This 'clear notion' must be translated by the schedule constructor into clear statements (or questions) about each item of performance to be rated; as far as possible the items should be distinct from one another. When used with an unvarying source (such as a consultation video recording) inappropriate, poorly-defined or overlapping items will generate ratings which vary widely between observers or between different occasions with the same observer.

Reliability alone however is not sufficient. What McNamara and McNamara call low inference items — matters of fact rather than judgement — can be rated reliably but are likely to be uninteresting and uninformative. Validity involves decisions about what is interesting and informative; and in relation to consultation performance, requires the schedule constructor to make value judgements in the selection of significant behaviours.

What is to be measured?

'Effective performance' begs the question, effective for whom? Consultation is an interaction in which doctor and patient have similar though not necessarily identical priorities. Evidence from patients and from observers of this interaction suggests that a number of aspects of the performance of doctors are identifiable and of interest.[4-9] The priorities of the doctor properly extend beyond meeting the immediate expectations of patients;[10] and for a rating schedule to represent consultation analysis validly, the ratings when taken together should reflect both patient expectations and professional concerns. Thus to be informative in education such a schedule must examine the consultation as an interview — 'conversation to gain insight' — and the capacity of the doctor to promote effective communication.[11]

How is it to be measured?

The choice of rating scale is itself an important issue and a controversial one. Types in use include nominal, ordinal, visual analogue and interval scales. Whereas analysis of the first two types is limited to counting, both visual analogue and interval scales permit measurement.[12] In situations where ratings are used to make comparisons of performance, interval scales can provide discrimination without the need to assume that equal intervals are of equal value or that single ratings have any absolute quantitative meaning. Furthermore, when used in groups an important requirement is that the rating given by individuals should be capable of immediate comparison. These practical considerations lead us to prefer interval scales.

The number of intervals provided must reflect the nature of the judgement required of the rater — too many will imply a disproportionate capacity to resolve issues, too few will encourage middle-of-the-road judgements.[13]

The use of multiple items each with its own scale introduces further difficulties through the possibility of interaction. The rater may base his judgement of one item on external standards of performance while with another item make a relative judgement. Thus the order and grouping of items could be important.

Finally, we must recognize that the total number of items capable of being rated reliably is limited. To make judgements the rater must recall the evidence under each item; the larger the number of items the more likely it becomes that the rater fails to observe or cannot recall such evidence.

The scope of testing

At a superficial level testing is concerned with the acceptability of the schedule and its component items; thus compliance by observers with the rating process is a useful indicator.

Reliability testing is fundamental. It examines the capacity of items to provide similar data from different observers and from the same observer on different occasions.

Clearly much that is subsumed by the term validity is a matter of opinion: 'A perfectly valid measure is one that correlates completely with "God's opinion" concerning the attribute'.[14] However, unless an item is reliable it cannot be valid — discreteness of an item may be assumed where reliable differences are shown between it and other items in a single consultation; for the same item, reliable differences between consultations suggest that the item is detecting true differences in performance. In these circumstances testing should define the number of observers required to detect such differences.

Testing a rating schedule is a protracted process. Useful findings may emerge from a preliminary study of the schedule in selected consultations. However, before validation is achieved these preliminary conclusions must be confirmed in a wide variety of consultations and by significant numbers of observers.

The present study

In this paper we describe the design and preliminary testing of a schedule for consultation analysis, the University of Leeds Consultation Rating Schedule (ULCRS); and suggest ways in which it might be used by groups of observers to detect and define significant differences in performance between doctors in consultation.

Method

Background

In 1980, owing to an increase in the number of students attached at any one time to the Division of General Practice at the University of Leeds it became necessary to provide a programme of seminars to alternate with practice experience. Two seminars on interview and communication form part of this programme.

In the first of these seminars, groups of fourth year undergraduates view and comment on consultation video tapes comprising material recorded in practice, and student consultations recorded in the Division with nurses acting the role of patient.

In the second seminar students analyse their own performance in consultation and thus need to identify and quantify aspects of consultation behaviour; a schedule of rateable items was considered to be a prerequisite.

In planning the seminars the teachers reviewed a number of published rating scales, but concluded that they contained too many items, lacked clear definition of certain items or did not address both interview strategies and communication skills. For these reasons a rating scale was devised consisting of 10 items; alongside each item was an interval scale numbered from zero to five.

Data collection

The first seminar began in the same way for each group of about 12 students, two recorded consultations being viewed and rated without comment from the teacher or other members of the group. The consultations comprised a young general practitioner with an elderly patient in his own surgery (consultation X) and an undergraduate with a middle-aged nurse complaining of tiredness and sleep disturbance (consultation Y). The two recordings had been chosen to reflect important differences in consultation. In the judgement of the teachers the performance of the student in consultation Y is exceptional in its maturity whereas consultation X appears to be ineffective.

Subsequently, and without prior warning, certain groups of students were shown the same consultations and asked to rate them on a second occasion two weeks after the first rating.

During 1982 the order of items on the rating scale was altered from the initial to an alternative format (Figure 1); its use by students remaining unchanged.

From time to time the opportunity arose to seek the rating of the same consultations by similar-sized groups of principals in general practice.

Completed rating scales were collected before discussion of the consultations. They form the basis of our analysis.

Results

Numbers and categories of rater

The two consultations were rated, on one occasion, by 143 undergraduates and by 33 principals in general practice. Of the undergraduates, 107 used the initial format of the schedule and 36 the alternative format. All the principals used the initial format. The same consultations, shown in the same order were rated again, after an interval of two weeks, by 100 of the 143 undergraduates, 83 using the initial and 17 the alternative format. On this basis the study population can be divided into five categories:

1. Undergraduates, first rating, initial format of schedule.
2. Undergraduates, second rating, initial format of schedule.
3. Undergraduates, first rating, alternative format of schedule.
4. Undergraduates, second rating, alternative format of schedule.
5. Principals, initial format of schedule.

Acceptability

The schedule proved acceptable to students and principals in either format. In use, 176 individuals provided 5504 item ratings out of a possible 5520, representing a non-compliance rate of 0.29%. There was no difference between the compliance rate of students and principals.

Analysis of ratings

Inter-observer variability. The schedules were used by our raters to compare the two consultations (X and Y). Table 1 shows the proportion of raters in a typical group of students preferring one consultation to another under each item of the schedule; and those who were uncertain, that is gave ratings to both consultations not differing by at least one whole mark. The items

Table 1. Analysis of the item ratings given by one group of 11 students to consultations X and Y, showing the number of raters in each of the four categories.

Schedule item	Preference of at least one mark		Uncertain	No rating
	for Y over X	for X over Y		
A	9	0	2	0
B	11	0	0	0
C	7	1	3	0
D	11	0	0	0
E	9	0	1	1
F	11	0	0	0
G	11	0	0	0
H	10	0	1	0
J	7	0	4	0
K	11	0	0	0

In establishing a relationship: did the doctor exhibit			
involvement?	0	5	A
warmth?	0	5	B
In gathering information: were the doctor's questions appropriately selective/sufficiently wide ranging regarding			
the patient's primary problem?	0	5	C
other problems raised?	0	5	D
potential problems not presented?	0	5	E
In creating understanding of the nature of the problem(s), were opportunities provided for the patient to			
define/clarify them?	0	5	F
In giving the patient guidance: did the doctor provide opportunity for the patient to			
participate in decision making?	0	5	G
In the use of communication skills: in general did the doctor			
listen enough?	0	5	H
use appropriate language?	0	5	J
recognise cues?	0	5	K
.....	0	5	
.....	0	5	
.....	0	5	

Figure 1. *University of Leeds Consultation Rating Schedule.*

are referred to by their alphabetical order in the alternative format of the schedule (Figure 1).

In our analysis, then, the first requirement of preference is defined as a difference between consultations of at least one whole mark on a scale of zero to five in the rating of an item. Where raters marked scores between whole numbers, inspection was used to allocate a fractional score (0.25, 0.5, 0.75) above the lower value; the same order of difference between consultations being required for preference.

Using pooled data and in terms of crude numbers consultation Y was preferred, under all items, by all categories of rater. However the differences in favour of consultation Y vary between items and categories of rater; and not all of them are, in our terms, statistically significant, since our second requirement is that, in any group, 65% of individuals must prefer one consultation to another under one or more items. Table 2 shows the items used by different categories of the study population to express preferences which are significant (using the binominal distribution where $P = 0.65$). It can be seen that all the items on the schedule were used by at least one category to express significant preference. The alternative format of the schedule was used to express such preference under a larger number of items than the initial format.

Table 2. Items for which significant preference was shown by different groups of observers. * = significant preference.

Categories of study population (see text)	Schedule item A	B	C	D	E	F	G	H	J	K	Number of raters
1.		*					*				107
2.	*	*		*			*				83
3.	*	*		*	*	*	*	*		*	36
4.	*	*		*	*	*	*	*	*	*	17
5.			*				*	*			33

On the basis of these findings it is possible to construct a hierarchy of items which, for the chosen consultations and using our two criteria of preference, operated with decreasing reliability between categories of observer. This is represented in Figure 2. Thus rating schedule item G discriminated between the two consultations in the hands of all raters; items A, B, D, H in at least three of our five categories; and items C, E, F, J, K in at least one category.

Intra-observer variability. It was possible to study intra-observer variability for undergraduates only; 100 of 143 students rating both consultations on a second occasion. Table 3 is an analysis of these second ratings of consultations X and Y, contrasted with the first ratings given by the same individuals, under each of the schedule items. Using our first requirement of significant preference, the analysis is concerned with three general categories of rating behaviour — preference expressed on both occasions; uncertainty on the first occasion but preference expressed on the second; and uncertainty on the second or on both occasions. In the first two categories the direction of any change is shown.

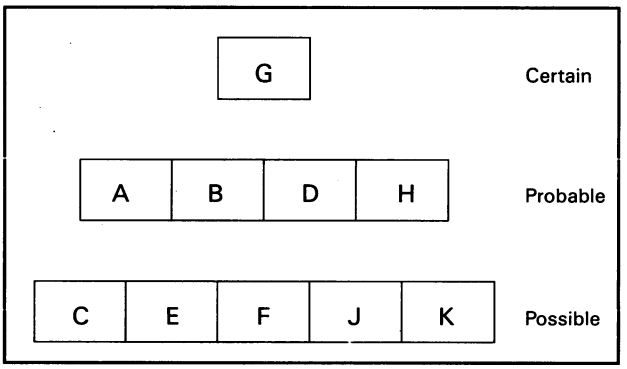

Figure 2. *The effectiveness of items on the rating schedule.*

The analysis shows that among those expressing clear preference on both occasions the number of raters favouring Y over X has increased under all items. However, for some items the number of raters expressing uncertainty has also increased. In general there is a shift in the population exhibiting uncertainty, newcomers to the ranks in part replacing those who now express preference.

However, we are concerned only with the number of raters who expressed preference for consultation Y on both occasions (first column of Table 3); as might be expected, the number (equivalent to percentage) varies between items. From these numbers a descending rank order of intra-observer reliability of items can be created: G, B, A, K, F, J, E, D, H and C.

Once again we suggest that 65% is an appropriate requirement and on this basis only items G and B can be considered to show significant intra-observer reliability among students, for the chosen consultations.

Discussion

To some doctors the very idea of comparing the performance of colleagues in consultation may be unacceptable. Certainly we agree with Pendleton and colleagues that this area of professional performance must be approached with sensitivity.[15] Our aim in proposing statistical criteria is to create confidence

Table 3. Analysis of the first and second ratings by 100 undergraduates of consultations X and Y.

Schedule items	Preference expressed at both ratings: consistent for Y over X	consistent for X over Y	changing from X to Y	changing from Y to X	Uncertain at 1st rating, preference at 2nd rating: for Y over X	for X over Y	Uncertain at 2nd rating: previously uncertain	previously showed preference	Number of raters
A	62	2	5	2	18	2	4	5	100
B	72	0	5	2	9	2	6	4	100
C	40	3	7	2	22	1	11	14	100
D	54	1	8	1	21	0	6	8	99
E	56	1	9	1	14	0	6	11	98
F	60	1	6	2	12	2	6	11	100
G	92	0	1	0	6	0	0	1	100
H	54	0	4	1	19	0	7	15	100
J	56	1	3	0	18	1	9	11	99
K	61	3	3	1	14	0	4	13	99

Comparison 1 or 2 (please circle) Date .....

(N.B. Data from a second comparison is acceptable only if the same individuals rate on both occasions)

Let X be the first and Y the second consultation rated by a group of size *n*. To determine the significance of ratings, first enter, in the boxes under each of the rating schedule items:

1. The number of raters who expressed a preference for consultation X or Y, of one mark or more.
2. The total number who rated in any way, both consultations.

	A	B	C	D	E	F	G	H	J	K
No. of raters preferring X to Y										
No. of raters preferring Y to X										
Total no. of raters (*n*)										

Then, using the chart below determine the significance of preference, under each item, for *n* raters.

n	Definite evidence: over 65% show preference							Some evidence: over 50% show preference								No evidence of preference				
36	36	35	34	33	32	31	30		29	28	27	26	25				24	23	22	21
35	35	34	33	32	31	30			29	28	27	26	25			24	23	22	21	
34	34	33	32	31	30	29			28	27	26	25	24			23	22	21		
33	33	32	31	30	29	28			27	26	25	24	23			22	21	20		
32	32	31	30	29	28	27		26	25	24	23	22			21	20	19	18		
31	31	30	29	28	27			26	25	24	23	22			21	20	19			
30	30	29	28	27	26			25	24	23	22	21		20	19	18	17			
29	29	28	27	26	25			24	23	22	21			20	19	18				
28	28	27	26	25	24			23	22	21	20		19	18	17					
27	27	26	25	24				23	22	21	20		19	18	17					
26	26	25	24	23			22	21	20	19		18	17	16						
25	25	24	23	22			21	20	19	18		17	16	15						
24	24	23	22	21			20	19	18		17	16	15							
23	23	22	21			20	19	18	17		16	15	14							
22	22	21	20			19	18	17			16	15	14							
21	21	20	19			18	17	16		15	14	13								
20	20	19	18		17	16	15			14	13	12								
19	19	18			17	16	15		14	13	12									
18	18	17			16	15	14		13	12	11									
17	17	16		15	14	13			12	11	10									
16	16	15		14	13			12	11	10										
15	15	14		13	12		11	10	9											
14	14			13	12		11	10	9											
13	13			12	11		10	9	8											
12	12		11	10			9	8	7											
11	11		10				9	8	7											
10	10		9			8	7	6												

Figure 3. *Significance of a preference in a group of n observers (P <0.05). Completed forms for submission should be posted to: ULCRS, Division of General Practice, Clinical Sciences Building, St James's Hospital, Leeds LS9 7TF.*

in peer ratings where these are soundly based, or to minimize their impact where they are not.

In certain circumstances such comparisons are likely to be more acceptable. Group rating may be seen as useful for comparing, for example, a trainer and his trainee or a trainee before and after vocational training. Similarly, a doctor (principal or trainee) who feels that one of his consultations went badly and another well, may wish to test these feelings by showing video recordings of the consultations to a group of his peers.

Techniques of video recording have recently been reviewed;[15] their increasing use in undergraduate education and vocational training is likely to lead to their acceptance in continuing medical education. In time the comparison, through rating, of consultation material submitted by established practitioners may become a widely accepted method of learning from a group of colleagues, provided that reliable and valid rating methods are available.

This work is no more than a preliminary stage in the testing of the rating schedule. The present conclusions rest on two assumptions in the comparison between consultations — that differences in ratings of an item must amount to one whole mark (on a scale of zero to five) to indicate a difference between consultations, and that at least 65% of the group should show such a difference. Separately these assumptions may be challenged; taken together we suggest that they form a reasonable yardstick.

On this basis we can find no firm evidence to reject any of the chosen items in the rating schedule since in use by different categories of rater each of them satisfied these criteria; in its alternative format the schedule appears to be more discriminating. It is interesting that there is substantial common ground between these 10 items and the seven tasks of consultation described independently by Pendleton.[15]

However, agreement between all categories of observer was demonstrated with only one item (G) on the schedule. Since this item also demonstrated a satisfactory level of intra-observer reliability it is almost certainly valid and reflects a true difference in performance between the chosen consultations.

This then is a framework for further testing of the schedule; each of the items has some justification for inclusion and may, when used to make other comparisons of performance in consultation, behave like item G in this preliminary study. How then is the validity of the remaining items to be confirmed or refuted? We believe that this task is an appropriate one for collaborative research between an academic department and our colleagues in practice and in vocational training. We have reproduced the rating schedule in full (Figure 1) so that photocopies may be taken for use by groups in the comparison of two consultations. In addition there is a chart (Figure 3) which shows the significance of preferences in terms of numbers of raters within groups of varying size. We would like groups using the schedule in this way to report their findings to us, using the data-collection form incorporated in Figure 3. Wherever possible the comparative rating should be repeated by the group after an interval of not less than two weeks. In this way experience of the operation of items used in rating the performance of colleagues in consultation can be accumulated. For our part we undertake to analyse the results of this collaboration and report directly to those who contribute data and from time to time by publication.

We recognize that groups may wish to amend the schedule to incorporate other items. We suggest that they do so by adding new items in the spaces provided and deleting existing items to keep the total number at 10 items. New items may prove to be more valid than those we have chosen; where groups have shown significant differences between consultations with a new item we will consider incorporating it in future versions of the schedule.

References

1. Moser CA, Kalton G. *Survey methods in social investigation.* London: Heinemann, 1971.
2. Topping J. *Errors of observation and their treatment.* London: Chapman and Hall, 1978.
3. McNamara DR, McNamara HI. Judging general practice by numbers: another educational fad? *J R Coll Gen Pract* 1983; **33:** 117-120.
4. Cartwright A. *Patients and their doctors.* London: Routledge and Kegan Paul, 1967.
5. Congalton AA. Public evaluation of medical care. *Med J Aust* 1969; **2:** 1165-1171.
6. Davis A, Horobin G (eds). *Medical encounters: the experience of illness and treatment.* London: Croom Helm, 1977.
7. Korsch BM, Negrette VF. Doctor–patient communication. *Sci Am* 1972; **227:** 66-71.
8. Stimpson GV. Obeying doctor's orders: a view from the other side. *Soc Sci Med* 1974; **8:** 97-104.
9. Jaspars J, King J, Pendleton D. The consultation: a social psychological analysis. In: *Doctor–patient communication.* Pendleton D, Hasler J, (eds). London: Academic Press, 1983.
10. Stott NCH, Davis RH. The exceptional potential in every primary care consultation. *J R Coll Gen Pract* 1979; **29:** 201-205.
11. Froelich RE, Bishop FM. *Medical interviewing. A programmed manual.* St Louis: Mosby, 1969.
12. Wright HJ, MacAdam DB. *Clinical thinking and practice.* Edinburgh: Churchill Livingstone, 1979.
13. Bennett AE, Ritchie K. *Questionnaires in medicine. A guide to their design and use.* London: Oxford University Press, 1975.
14. Abramson JH. *Survey methods in community medicine.* Edinburgh: Churchill Livingstone, 1979. (The author attributes this concept to AL Cochrane.)
15. Pendleton D, Schofield T, Tate P, Havelock P. *The consultation. An approach to learning and teaching.* Oxford University Press, 1984.

Acknowledgements

The authors wish to thank Drs Jamie Bahrami, Takis Christou, John Wright and Arnold Zermansky who contributed to the design of the rating schedule; and Sarah Stanley, who helped with the analysis of ratings.

Address for correspondence

Professor I.M. Stanley, Department of General Practice, New Medical School, Ashton Street, Liverpool L69 3BX.

Doctors on strike

Experience accumulated during the doctors' strike of 1983 in Israel was analysed to compare the effect on the work of the family doctor of direct-charge as against pre-paid insurance arrangements, in three different settings—suburban, rural and working-class small town. The imposition of direct charges greatly reduced the consultation rate; more of the patients consulting received prescriptions, especially for antibiotics; laboratory investigation, referral and admission to hospital were unchanged, but referral for specialist consultation was reduced; the most frequently seen diagnostic categories remained respiratory diseases and undefined signs or symptoms, but pneumonia was seen much more frequently; there was no change in the proportion of follow-up visits, but house calls were more frequent. These trends were stable over the four-month period of the strike, and partial reimbursement of the fee did not change the picture significantly. The evidence did not conclusively support the hypothesis that direct charges selectively deter trivial complainers.

Source: Weingarten MA, Monnickendam MS. The effect of direct charges on consultations in family practice: a study of a doctors' strike. *Family Practice* 1985; **2:** 35-41.