

THE 'WESSEX' BEHAVIOUR RATING SYSTEM FOR MENTALLY HANDICAPPED PEOPLE: RELIABILITY STUDY

JOHN PALMER and JUDITH JENKINS
Health Care Evaluation Research Team,¹ Winchester, U.K.

INTRODUCTION

The Wessex behaviour schedule (Kushlick, Blunden & Cox 1973), originally designed for the 1963 Wessex survey of mental handicap (Kushlick & Cox 1973), enables an individual to be rated on various aspects of skills and incapacity. Since the original Wessex survey, the behaviour schedule has been regularly used in collecting information for the Wessex Mental Handicap Register. The schedule covers aspects of self-care, walking, continence, literacy, behaviour problems, vision and hearing. It is completed by someone who is familiar with the mentally handicapped person being rated, usually a member of the care staff in a residential institution, or a teacher or instructor in a school or training centre. The rater is generally given no instructions on how to conduct the rating except those on the back of the schedule. From the ratings given for each of the items, the mentally handicapped person can be categorised either according to Social and Physical Incapacity (SPI) and Speech, Self-help and Literacy (SSL) (Dawes 1977), or according to a newer fivefold classification which combines the two scales (Figure 1). Further, the Wessex plan for mental handicap services (Joint Planning Group 1973) introduced a rule for determining which agency (National Health Service or Local Authority Social Services) was to be responsible for providing residential care for mentally handicapped clients. The behaviour schedule provides sufficient information to categorise a person as the responsibility of NHS or of Social Services (Figure 1).

Figure 1

SPI	SSL	New Classification	Abbrev.	Wessex Joint Plan rule:	
				children	adults
Nonambulant	Any value	Nonambulant	NA	NHS	NHS
Severe behaviour disorder with severe incontinence, or Severe behaviour disorder alone	Any value	Severe behaviour disorder	SB	NHS	NHS
Severe incontinence	Any value	Severe incontinence	SI	Social Services	NHS
Mild handicap or No handicap	No SSL or Speech only	Continent, ambulant, no severe behaviour disorder	CAN	Social Services	NHS
Mild handicap or No handicap	Self-help only, or Speech and self-help, or Literate	Continent, ambulant, no severe behaviour disorder, feeds, washes and dresses self	CAN FWD	Social Services	Social Services

1. The Research Team is supported by Department of Health and Social Security, Medical Research Council, University of Southampton and Wessex Regional Health Authority, the MRC grant number being G971/145.

Besides research functions the Register has two important service functions. One is to produce data for service planners, enabling them to estimate the scale of provision needed for various kinds of service. The second is to provide lists of mentally handicapped clients who may, according to their age, home area and disability level, be eligible for a particular type of service. The behaviour schedule is central to both functions. It has also been widely used or recommended elsewhere, in its original or a modified form (Wing 1971, Department of Health and Social Security 1972, Martindale 1977, Wing and Gould 1978, Development Team for the Mentally Handicapped 1978, National Development Group 1978).

Given these functions, it is important to have estimates of the behaviour schedule's reliability as a measurement tool, as it allows one to estimate confidence limits for numbers found from the Register. Similarly, knowing the inter-rater agreement level is important to users who ask for lists of people who are eligible for a specific service because of their degree or kind of handicap. If such lists are to be of practical use, they must not exclude people who are in fact eligible for the service, but at the same time not contain many people who are in fact ineligible.

Kushlick, Blunden & Cox (1973) published some inter-rater agreement findings using the 1963 survey data. However these findings are now less applicable because in 1963 raters received more instruction in the completion of the schedule than is now the practice, and the sample in that study was not controlled to give adequate representation to all levels of disability; further, the two ratings for each person were made in different settings (e.g. residential unit and training centre, or residential unit and school), so some of the apparent unreliability may in fact represent real differences in behaviour between the two settings.

The present study therefore investigated the inter-rater reliability of the behaviour schedule, when completed according to present-day Wessex practice, using pairs of raters in the same setting, with adequate representation of different ages, disability levels and settings.

METHODS

Sample selection

It was proposed to obtain 50 people in each of the following categories: CHILDREN: nonambulant, SB, SI, CAN (whether or not FWD); ADULTS: nonambulant, SB, SI, CAN and CAN FWD. Each category was to be evenly divided between residential and non-residential clients. The Wessex Mental Handicap Register was used to identify institutions at which previously collected data showed that subjects in the desired categories would be found. Nevertheless the number of people in certain categories (notably SI) is not large at any establishment, and for reasons of time the collection of data was stopped before the intended numbers had been reached for these categories, while in other categories they were exceeded. The actual composition of the achieved sample is given in Table 1. Since there were two schedules for each person, the number of schedules in each category has been counted and halved.

Data collection

The first author visited each establishment and arranged for two or more members of staff to act as raters. Raters were members of the daytime residential care or teaching staff and were personally well acquainted with the people they were to rate. Where possible the two raters for each person were taken from opposite shifts to reduce the risk of collusion. Table 2 gives the numbers of establishments and of raters. All raters except 5 (who made between them less than 3% of the ratings) were personally briefed by the first author who stressed that they must not compare ratings nor ask advice from each other. All raters received written instructions that included the warning against collusion and specified when they should make their ratings, so that the two ratings made of each person were never separated in time by more than 7 days.

Table 1
*Numbers of mentally handicapped
 people* in the completed sample*

		<i>residential</i>	<i>nonresidential</i>
Under 16	nonambulant	26	40½
	severe behaviour disorder	28	13
	severely incontinent	19	11
	CAN or CAN FWD	32½	48
	Disability category undetermined**	1½	2½
Total children		107	115
16 and over	nonambulant	35	8½
	severe behaviour disorder	58	14
	severely incontinent	23	7½
	CAN	63½	36
	CAN FWD	50½	75½
	Disability category undetermined	8	9½
Total adults		238	151
All ages		345	266

*Calculated by dividing the number of schedules in each category by two.

In total, 611 children and adults were included in the sample.

**Because the schedule was not fully completed.

Table 2

		<i>Establishments</i>	
<i>Residential</i>		<i>Non-residential</i>	
11 Adult hospital wards		8 ESN(S) schools	
3 Children's hospital wards			
5 Children's locally-based hospital units		5 Adult training centres	
1 Adult locally-based hospital unit		1 Adult day centre (social services)	
		<i>Raters</i>	
29 Raters	rated	1—5 subjects each	
31 Raters	rated	6—10 subjects each	
4 Raters	rated	11—15 subjects each	
24 Raters	rated	16—20 subjects each	
5 Raters	rated	21—25 subjects each	
2 Raters	rated	29 (2 charge nurses at an adult locally-based hospital unit)	
2 Raters	rated	52 (Nursing officer and charge nurse at an adult hospital ward)	
2 Raters	rated	59 (Head and deputy head at an ESN(S) school)	
1 Rater	rated	73 (Manager of an adult training centre: the other ratings for these subjects were shared among the other staff)	

Analysis

All ratings were scored by the PL/1 computer program given by Dawes (1977). The reliability statistic used is the coefficient kappa (Cohen 1960). Because kappa is relatively insensitive to variation in the actual distribution in the sample of the variable under investigation, no weighting has been applied to the sample.

Table 3
Reliability (measured by kappa) of the fivefold classification

	<i>Residential children</i>	<i>adults</i>	<i>Nonresidential children</i>	<i>adults</i>	<i>Whole sample</i>
kappa	.55	.54	.65	.72	.62
95% confidence limits	±.10	±.07	±.09	±.08	±.04
No. of people rated	104	224	110	135	573

RESULTS

The overall classification

Table 3 shows the reliability coefficients obtained for the new five-fold classification, for the four main divisions of the sample and the whole. Mentally handicapped people for whom either rating was incomplete are omitted from this analysis. The confidence limits are calculated as described by Cohen (1960) and represent the limits within which kappa will fall 19 times out of 20. Kappa has the value zero when the frequency of agreement is equal to the frequency expected by chance. It can therefore be seen that the reliabilities found are in all cases well above the chance level, though they are not particularly high. Kappa was also used to investigate whether the raters made more reliable discrimination between certain categories of the classification than between others. Table 4 gives the results. It shows that 'nonambulant' is reliably distinguishable by the raters from all other categories, though least reliably from 'severe incontinence'. 'CAN FWD' is very reliably distinguished from 'nonambulant' and 'severe incontinence' but less so from 'severe behaviour disorder' or 'CAN'. 'Severe behaviour disorder', 'severe incontinence' and 'CAN' are rather unreliably discriminated from each other. Similar results were obtained for each of the four main divisions of the sample.

Table 4
Kappa values for the reliability of discrimination between pairs of categories

	<i>NA</i>	<i>SB</i>	<i>SI</i>	<i>CAN</i>	<i>CAN FWD</i>
NA		.95	.81	.94	.99
SB			.67	.50	.76
SI				.57	.96
CAN					.73
CAN FWD					
95% confidence limits for the above values:					
		±.03	±.07	±.04	±.01
			±.10	±.09	±.06
				±.10	±.03
					±.06

Table 5

If one rating is:-

the probability per cent. of another rating made at the same time in the same environment being as below is:-

			<i>SAME</i>	<i>NA</i>	<i>SB</i>	<i>SI</i>	<i>CAN</i>	<i>CAN FWD</i>	<i>no of cases*</i>
NA	ch	res.	92		0	6	2	0	52
		nres.	93		0	5	3	0	80
	ad	res.	86		6	4	3	1	70
		nres.	94		0	0	6	0	17
SB	ch	res.	61	0		9	20	11	56
		nres.	62	0		8	27	4	26
	ad	res.	52	4		8	23	13	114
		nres.	69	0		4	19	8	26
SI	ch	res.	54	8	14		24	0	37
		nres.	36	18	9		32	5	22
	ad	res.	55	7	20		16	2	44
		nres.	71	0	7		21	0	14
CAN	ch	res.	37	3	30	24		5	37
		nres.	60	4	13	13		9	53
	ad	res.	61	2	21	6		10	123
		nres.	68	2	8	5		18	65
CAN FWD	ch	res.	69	0	23	0	8		26
		nres.	82	0	3	3	13		39
	ad	res.	70	1	15	1	12		97
		nres.	90	0	1	0	8		148

*Since each person has two ratings, the number of *persons* is on average half of the number in this column.

The user of this classification may therefore have considerable confidence in a rating of 'nonambulant'. In the case of a rating of 'CAN FWD', however, he should bear in mind the possibility that another observer may rate the same person as 'CAN' simply, or even 'severe behaviour disorder', while for a rating in any of the categories 'SB', 'SI' and 'CAN', there is considerable probability that another rater would place the subject in a different one of the three categories. Table 5 shows what this probability was, for each of the four divisions of the present sample.

From this table it can be seen that the probability of a rating of 'CAN FWD' being reclassified as 'SB' is much higher for people rated in residential care than for those rated in schools or training centres. This is probably due to greater prevalence of severe behaviour problems among the residential clients. The probability of a 'SB' or 'SI' rating being reclassified as 'CAN', or of a 'CAN' rating being reclassified as 'SB', are of the order of 20%.

The reader will observe that the 'Wessex rule' for responsibility for residential care depends heavily, for children, on the distinction between 'severe behaviour disorder' on the one hand and 'severe incontinence' and 'CAN' on the other, and for adults, equally heavily on the somewhat less disputable distinction between 'CAN' and 'CAN FWD'. For example, a child rated 'SB' is NHS responsibility by the Wessex rule; but table 5 shows that the probability of a second rating putting that child in the category of Social Services' responsibility is about 40%. Similarly an adult rated 'CAN FWD' is Social Services' responsibility; but in the present sample, 30% of adults so rated in residential care were assigned to an NHS responsibility group by the other rater. Clearly there will be severe problems in using the behaviour rating as the basis on which to apply the Wessex rule to individual clients.

The individual questions

It remains to consider the reliability of answers to individual questions on the schedule, with a view to identifying parts of the questionnaire that could be improved. There are 20 questions in all (excluding that on clarity of speech, which is not used at all in the derivation of the fivefold classification, nor in that of SPI or SSL).

If raters have difficulty in answering the questions, that may be shown either by missing answers or by unreliability in the answers given. Only three questions were unanswered in more than 2% of cases. Of these, the questions on incontinence at night are optional for staff of non-residential establishments and a classification can still be made if they are unanswered. Likewise the question on 'walking with help' is unimportant in many cases provided that 'walking by self' is unanswered.

Table 6
Values of kappa with 95% confidence limits

	Residential		Nonresidential	
	Adults	Children	Adults	Children
speech	.71	.73	.78	.78
counting	.62	.67	.89	.77
walk-by-self	.77	.72	.84	.61
writing	.52	.63	.83	.78
dressng	.62	.69	.69	.74
washing	.63	.72	.52	.67
day-wetting	.53	.66	.69	.67
walk-with-aid	.61	.59	.80	.57
night-wetting	.43	.64	.76	.72 ± .21
reading	.59	.40	.73	.75
self-injury	.56	.56	.59	.59
vision	.77	.56	.56	.48
day-soiling	.49	.56	.55	.74
feeding	.52	.55	.79	.54
hitting people	.51	.54	.37	.59
hearing	.50	.56	.42	.48
over-activity	.49	.41	.40	.57
night-soiling	.40	.52	.48	.52 ± .27
damaging things	.39	.51	.34	.58
attention-seeking	.34	.34	.33	.61
Confidence limits:	± .09 to .13	± .06 to .09	± .06 to .15	± .07 to .11 except where shown

Table 6 shows the reliability of the separate questions, with the most reliable at the top and the least reliable at the bottom. It can easily be seen that the most unreliable group of questions is the group on behaviour problems. All these are in the lower half of the table and between them they account for all the kappa values below .4. This emphasises the unreliable discrimination of 'severe behaviour disorder' noted earlier. The two soiling questions are also in the lower half, though the wetting questions rank higher; this is consistent with the slightly better (though still poor) discrimination of 'severe incontinence' from other categories.

The remainder of the lower half comprises 'vision' and 'hearing' (which are not used in the fivefold classification though they are of interest as individual items), and 'feeding'. The other two self-help questions, 'wash' and 'dress', rank relatively high; thus it is likely that the uncertainty in discriminating 'CAN FWD' from 'CAN' is due more to the unreliability of 'feeding' than the other two questions.

DISCUSSION

It appears that the 'Wessex classification' is an instrument of modest but not negligible reliability. It is based on a simple questionnaire which is easily and quickly completed and so it is still a useful device in large-scale surveys, the purpose for which its authors intended it.

Whether this rating system should be used in service management, for selecting individual clients as potential recipients of a particular form of service, is much more questionable. Relatively few errors will be made in identifying the 'nonambulant'; rather more will be made in identifying the 'CAN and FWD' — the most able — particularly, it would seem, in residential settings. For persons placed in the other three categories of the fivefold classification, the usefulness of the rating is much less clear, and managers of services, doctors, social workers and all direct-care staff would be well advised to obtain further information about anyone who is not clearly in one of the extreme categories mentioned. It is notable that the Wessex policy rules for residential care (Joint Planning Group 1973) require decisions on individual clients to be taken upon distinctions which are unreliably made using this system, and have not been more precisely defined using any other standard instrument.

The results indicate that the greatest improvements in the reliability of the instrument are likely to be obtained by increasing agreement on the answers to the incontinence questions and the behaviour problem questions. To achieve this will require some experimentation. One possible step would be to replace the vague expressions of frequency used in these items by more precise expressions, based on the explanatory notes that at present appear on the back of the form. Thus, for the incontinence questions, 'frequently, occasionally, never' might be replaced by 'more than once a week, once a week or less, never'; and for the behaviour problems, 'marked, lesser, no' might be replaced by 'in the last month and still a management problem, not in the last month or no management problem, does not occur'.

Still such changes may not be enough to increase the frequency of agreement, since these two sets of questions ask the rater to report on events of low frequency (especially some of the behaviour problems) which the rater may not have been personally present to observe. Wetting and soiling is often observed only by the night staff and, unless these events are systematically recorded, the day staff may have little knowledge of them. Likewise, the frequency of day wetting and soiling may be misreported by staff who work shifts and are not continuously present throughout the day. Behaviour problems are even more likely to be misreported, since the time span which the rater is asked to review is a month rather than a week, and he is also asked to make a judgment about the existence of a 'management problem', this being undefined.

Other possibilities might involve a more radical change in the way data are collected. For example, schedules might be completed at a staff meeting at which differences of opinion could be discussed and resolved. (This procedure would make difficult the measurement of reliability). Alternatively, if it were possible to ask staff to maintain careful diaries of the occurrence of incontinent incidents and problem behaviour over suitable periods of time, it might be possible to make the reporting of these events significantly more reliable. On the other hand the behaviour rating schedule is designed to be completed quickly by a direct-care staff member with the minimum of trouble, and thus it achieves acceptability with informants in large-scale surveys and registers. To introduce more elaborate procedures of data collection could spoil that acceptability and would certainly increase the cost of collecting the information; some balance must be struck between the aims of obtaining highly reliable data and of obtaining any data at all.

None of the above suggestions would overcome the difficulty that each of the five behaviour problem questions asks about a rather imprecisely defined class of problem behaviour. Careful training of raters might help them to identify each sort of behaviour correctly, but again this would greatly increase the cost and difficulty of obtaining data.

The vision, hearing and feeding questions were also low in reliability. It seems unlikely that any easy change to the questionnaire could improve the situation for vision or hearing, since the problem appears to be one of discrimination of steps in a scale of impairment that is actually continuous. Some sort of formal sight and hearing tests might be necessary to raise reliability appreciably. However these two questions do not affect the composite ratings.

Feeding however is important in distinguishing 'CAN FWD' from the other categories. It is possible that the main difficulty is the definition of 'feeding with help'. The explanation on the back of the schedule is open to a range of interpretation, and refers to a 'central kitchen' which is not a feature of all environments. However, before attempting to revise this explanation, it would be advisable first to investigate how often raters refer to the back of the form at all.

One other question with which informants clearly have some difficulty is that on walking with help. The difficulty is indicated mainly by the frequency with which it is not answered. In practice this has little consequence for reliability because the assessment of ability to walk is made principally by means of the question on walking by oneself. However the form could be made less confusing by reversing the order of the two walking questions.

The foregoing suggestions for improvements must be regarded as tentative and subject to test. It would be undesirable to introduce changes to the schedule or scoring procedure unless it were shown that they do actually increase reliability and that the revised version is comparable with the old.

SUMMARY

This study investigates inter-observer agreement in classifying mentally handicapped people using the Wessex behaviour rating schedule. 611 children and adults of varying levels of handicap, some in residential establishments and some in schools and training centres, were each rated independently by two staff members who knew them well. The study concludes that the rating system is usable for large scale surveys but that in assessment of individual clients it must be treated with great caution. Some possible improvements are discussed.

ACKNOWLEDGEMENT

The authors were greatly assisted by discussion with David Felce. They would also like to thank all the raters and the other staff of various agencies who made this study possible.

References

- COHEN, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 1, 37-46.
- DAWES, B. J. (1977) A method of rating behaviour characteristics: flow chart of classification procedure. Winchester: *Health Care Evaluation Research Team Report No. 139*.
- D.H.S.S. (1972) *Census of mentally handicapped patients in hospital in England and Wales at the end of 1970*. H.M.S.O.: London.
- DEVELOPMENT TEAM FOR THE MENTALLY HANDICAPPED (1978) *First Annual Report*. D.H.S.S.: London.
- JOINT PLANNING GROUP (1970) *Forward Planning of Services for the mentally handicapped in Wessex*. Winchester: Wessex Regional Hospital Board.
- KUSHLICK, A., BLUNDEN, R. and COX, G. R. (1973) A method of rating behaviour characteristics for use in large scale surveys of mental handicap. *Psychological Medicine*, 3, 4, 446-478.
- KUSHLICK, A. and COX, G. R. (1973) The epidemiology of mental handicap. *Developmental Medicine and Child Neurology*, 15, 748-759.
- MARTINDALE, A. (1977) A case register as an information system in a development project for the mentally handicapped. *Brit. J. Mental Subnormality*, 22, 2, 1-7.
- NATIONAL DEVELOPMENT GROUP FOR THE MENTALLY HANDICAPPED (1978) *Helping mentally handicapped people in hospital*. D.H.S.S.: London.
- WING, L. (1971) Severely retarded children in the London area: prevalence and provision of service. *Psychological Medicine*, 1, 5, 405-415.
- WING, L., and GOULD, J. (1978) Systematic recording of behaviours and skills of retarded and psychotic children. *J. Autism and Childhood Schizophrenia*, 8, 79-97.