British Journal of Psychiatry (1988), **152**, 670–673

AIMS Ratings – Repeatability

J. A. BERGEN, N. B. CARTER, J. CRAIG, D. MACFARLANE, E. F. SMITH and P. J. V. BEUMONT

In the present study, two raters, a psychologist and a nurse, each made five independent ratings of 30 different video-recorded patient-examinations. Having thus excluded patient fluctuation, individual-rater consistency and between-rater agreement over the 6 weeks of the study are examined. While between-rater agreement was apparently being maintained, mean AIMS scores steadily increased. In the hands of these raters, AIMS items 2 and 4 emerged as very reliable, while items 1, 6, and 7 showed high variability. Some patients appeared to be hard to rate. Differences between the study raters and the author JB highlight the issue: how reproducible is an AIMS rating?

How repeatable is an AIMS rating? A rater's ability to repeat a rating in response to the same stimulus on a separate occasion will obviously influence both between-rater agreement and test–retest agreement, using the scale. Our interest in rerating reliability arose because we wished to attempt to assess change using the AIMS (Abnormal Involuntary Movement Scale; Guy, 1976). The key question was: what change in an AIMS score represents meaningful change in a patient's tardive dyskinesia (TD)? We have previously explored the magnitude of fluctuations in the signs of TD in a pilot study (Bergen *et al*, 1984). Three psychiatrists made multiple ratings of multiple video-recorded AIMS examinations of four patients. The unexpected finding was that, for each of the patients, measurement variability attributable to patient fluctuation was dominated by variability attributable to reratings. It was thus decided to explore rerating variability over a broader range of patients; whether some of the rater variability could be explained in terms of rater change with time or in terms of the distribution or severity of the patient's signs; whether some patients are simply hard to rate. The study was designed in such a way as to make optimal use of available facilities. A major limitation would be the use of only two raters.

Method

There was one video-recorded AIMS examination of each of 30 patients. Two raters, J and D, made five independent ratings of each examination.

The 30 patients were 16 males and 14 females, mean age 50.3 ± 2.3 years. When the video examinations were rated by JB, mean AIMS sum 1–7 was 7.4 ± 0.8, and 20 patients (67%) met Schooler & Kane's (1982) AIMS criteria for TD (one '3' or two '2' item scores). The 30 patient-examinations were selected from an available pool of 45 by firstly subgrouping on the basis of global severity and distribution of signs, using ratings by JB, then making random selections from within subgroups.

The AIMS examinations had been video-recorded according to a set format which closely resembled that recommended by the NIMH. The 10–12-min examinations included 1.5–2 min of unobtrusive observation. The 30 patient-examinations were randomly ordered for presentation at the first rating, then randomly reordered for each of the four reratings. Eight examinations were rated in each rating session. Sessions were 2-h duration, always at the same time of day, and evenly spaced, with three sessions per week. Raters rated concurrently but independently.

Practical considerations precluded the use of psychiatrists as raters. We selected as raters a psychologist and a registered nurse. Both were experienced in research and patient observation; however, neither had worked with long-term psychiatric patients. They had no prior knowledge of TD and were given no information concerning the study subjects or the purpose or design of the study. Before commencing the study, nine 2-h training sessions were conducted by JB. Training material consisted of: 1. the AIMS training tape (borrowed from NIMH); 2. a film depicting the AIMS examination and a wide range of signs (courtesy of Sandoz); and 3. the remaining 15 video-recorded examinations. Raters were provided with standard descriptions of abnormal involuntary movements in different body areas. They were told that a rating was not a diagnosis. Difficulties that JB and psychiatrist co-raters had experienced in using the scale were described, but no solutions were offered. Raters were informed that the purpose of training was to promote consensus between them, and that they should attempt to concur with the AIMS training tape as the international standard.

Examinations had been recorded so that the examiner rarely appeared on screen, but was in a predictable location in relation to the patient at all times. The method of filming was explained to the raters. This knowledge, together with the soundtrack of the recording would be important for decisions about the significance of certain patient movements.

https://doi.org/10.1192/bjp.152.5.670 Published online by Cambridge University Press

Results

Before detailing the results, it is important to point out that these raters, whose training included repeated exposure to the AIMS training tape, rated differently from the author JB. They rated with greater sensitivity to abnormality. Individual rater's scores, represented by the sum of AIMS items 1–7 (AIMS sum 1–7), are displayed against the corresponding scores of JB in Fig. 1; the diagonal line represents points of complete agreement.

There were 19 evenly spaced 2-h rating sessions and eight ratings were made in each session. A regression analysis of the AIMS sum 1–7 scores was used to explore the effects of sessions and order of presentation. There was no effect of the order of presentation within the sessions but there was a positive linear trend over the sessions ($P<0.001$) that was similar for the two raters. This can also be seen in Table I, where the mean scores over 30 patients for each rater's five ratings are presented.

The proportion of the 30 patients for whom the two raters gave the same score, and the corresponding unweighted kappa coefficients, are tabulated in Table II for each AIMS item 1–8 at each of the five ratings. There are two sources of correlation between these proportions: that between items on the same patient and that between reratings. In the analysis of these dependent proportions, the correlations were estimated from the data and comparisons between items and between reratings were tested using contrasts. There was greater agreement between J and D for items 2, 3, and 4 (lips and perioral, jaw and tongue) than for item 1 (face) ($P<0.01$), for item 5 (upper extremity) than item 6 (lower extremity) ($P<0.05$), for AIMS sum 1–4 than AIMS sum 5–7 ($P<0.05$), and for item 8 (global severity) than AIMS sum 1–7 ($P<0.05$). There was no linear trend with rerating in the proportions of agreement between J and D. These analyses cannot be done with kappa coefficients because the adjustments for chance agreements vary from rating to rating. Differences between raters were also explored by testing the null hypothesis that, given that J and D differed, it was in a random direction. J rated lower than D on AIMS items 1 and 5 ($P<0.001$) and tended to do so on items 6 ($P=0.06$) and 8 ($P=0.07$). Note that, in these analyses, J and D are matched at each rating, i.e. they are compared at equal stages of their rating experience.

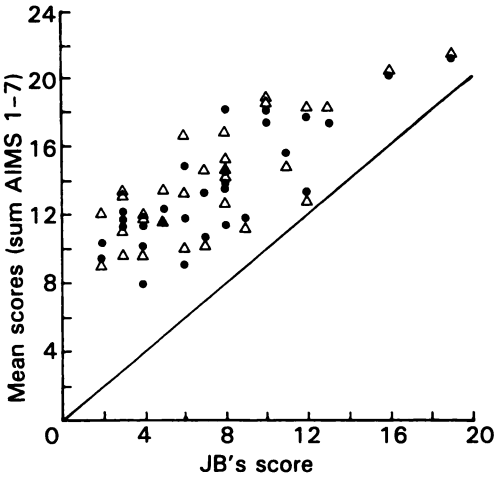


FIG. 1 Comparison of raters' with JB's scores. Raters: ●, J; △, D.

We wanted an estimate of the number of AIMS points that represents a real change in the patient when two AIMS evaluations by one of these raters are compared. A variance components model and analyses of variance for three variables, AIMS sum 1–4, AIMS sum 5–7 and AIMS sum 1–7, gave estimates 3.09, 2.79, and 4.47 respectively for the minimum changes, with a 95% level of confidence. The mean values for these variables are tabulated in Table I.

In order to compare the rating difficulty of different AIMS items, we also calculated the following indices of rater agreement and disagreement for AIMS items 1–8 respectively: 1. within-rater variability (the standard deviation for each item, pooled over patients and over two raters), 0.49, 0.36, 0.44, 0.38, 0.41, 0.52, 0.54, and 0.36; 2. the proportion of patients for which raters scored the items the same on each of the five ratings, averaged over the two raters, 0.27, 0.57, 0.33, 0.43, 0.42, 0.33, 0.23, and 0.48; and 3. the proportion of patients for which individual raters showed variation of more than 1 point over repeated ratings, 0.22, 0.08, 0.07, 0.02, 0.05, 0.17, 0.28, and 0.00. Because individual raters' results were similar, these are averaged. These raters showed considerable differences between items on these various indices of reliability;

TABLE I

Mean ± s.e.m. for AIMS sum 1–4, AIMS sum 5–7 and AIMS sum 1–7 over 30 patients for the two raters J and D

Rating	*Sum 1–4*		*Sum 5–7*		*Sum 1–7*	
	J	*D*	*J*	*D*	*J*	*D*
1	7.2±0.5	7.8±0.5	4.9±0.4	5.2±0.4	12.1±0.7	13.0±0.8
2	8.0±0.5	8.2±0.4	5.3±0.3	5.5±0.3	13.3±0.7	13.7±0.6
3	8.3±0.4	8.7±0.3	5.2±0.3	5.7±0.4	13.5±0.6	14.3±0.6
4	8.3±0.4	8.7±0.3	5.5±0.3	5.4±0.4	13.8±0.6	14.1±0.6
5	8.5±0.5	8.7±0.3	6.0±0.3	5.9±0.4	14.4±0.7	14.6±0.6
Overall	8.0±0.2	8.4±0.2	5.4±0.1	5.6±0.2	13.4±0.3	14.0±0.3

https://doi.org/10.1192/bjp.152.5.670 Published online by Cambridge University Press

TABLE II
Agreements between raters

Rating	*Item*							
	Face	*Lips and perioral*	*Jaw*	*Tongue*	*Upper extremity*	*Lower extremity*	*Trunk*	*Global judgement*
1								
p[1]	0.57	0.70	0.50	0.60	0.63	0.57	0.53	0.87
κ[2]	0.37	0.49	0.29	0.34	0.41	0.43	0.33	0.81
2								
p	0.67	0.80	0.80	0.80	0.77	0.60	0.73	0.70
κ	0.47	0.66	0.69	0.60	0.48	0.47	0.58	0.52
3								
p	0.60	0.87	0.73	0.77	0.70	0.53	0.57	0.83
κ	0.25	0.76	0.58	0.56	0.44	0.37	0.38	0.71
4								
p	0.50	0.83	0.60	0.80	0.53	0.40	0.67	0.77
κ	0.14	0.67	0.38	0.56	0.17	0.20	0.51	0.64
5								
p	0.53	0.73	0.60	0.77	0.70	0.57	0.63	0.63
κ	0.29	0.45	0.34	0.46	0.40	0.37	0.48	0.46
Mean								
p	0.57	0.79	0.65	0.75	0.67	0.53	0.63	0.76
κ	0.30	0.61	0.46	0.50	0.38	0.37	0.46	0.63

1. p is the proportion of 30 patients where raters J and D gave the same score.
2. κ is the unweighted kappa coefficient. Standard errors for the κ, assuming that κ was not zero, varied between 0.09 and 0.17 and were typically 0.12.

items 4 (tongue) and 2 (lips and perioral region) were scored most reliably, while items 1 (face) and 7 (trunk) were scored least reliably. In order to assess whether patients were equally hard to rate, the proportion of agreements between raters was calculated over the seven AIMS item scores and five reratings for each patient. Testing whether this proportion was the same for all patients gave $\chi^2 = 42.73$ on 29 degrees of freedom. Thus some patients were harder to rate than others ($P < 0.05$). Two patients had very low proportions of agreement but the rater disagreements were random in direction.

Discussion

Repeatability of ratings is clearly relevant to the clinical monitoring of individual patients and in the design of any clinical study needing to evaluate change using the AIMS. How do you know that you are a good enough rater to begin a study? How do you establish that you remain a good enough rater with the passage of time during a protracted study? For the two raters in this study, 18 h of training was insufficient to prevent a steady increase in the score values they ascribed to the same stimulus. One assumes there will be a ceiling to this increase, but it was not obviously reached during this study, which involved 38 h of rating in 6 weeks. Between-rater agreement remained roughly steady, concealing this shift in scoring. No distinction can be made in this study between the effects of practice, repeated exposure to the same patients, and repeated exposure to the same examinations.

In evaluating between-rater agreement we have used a simple statistic: the proportion of complete agreement. It must be for the reader to decide if these levels of agreement are sufficient and comparable to their own experience of rating with the AIMS. In considering disagreement between these raters, the major point to emerge is that they disagreed more frequently on some scale items. In some instances, this was in a non-random manner. Moreover, for some items, the raters' disagreement with themselves exceeded 1 point in a substantial number of patients.

For items 2 (lips and perioral area) and 4 (tongue) the proportion of complete agreement was high, differences were random in direction, individual rater's standard deviations on the items were least, raters responded with the same score at each rerating in approximately 50% of patients, and differences of greater than 1 AIMS point were infrequent. This was also true for item 8 (global judgement) although there was a tendency for J to rate lower than D. Such reliability in a subjective measurement intuitively suggests that the measurement is valid. Item 5 (upper limb) was handled generally well, but differently, by these raters. In contrast, items 1 (face), 6 (lower extremity), and 7 (trunk) showed high variability.

https://doi.org/10.1192/bjp.152.5.670 Published online by Cambridge University Press

In the hands of these raters, the meaning of a score on these items must certainly be questioned. These conclusions are supported by inspection of the kappa coefficients if those less than 0.4 are interpreted as indicating poor agreement and those greater than 0.6 as indicating good-to-excellent agreement (Flack, 1987).

The finding that some patients emerged as harder to rate than others concurs with our clinical experience with both live and recorded examinations. We could not identify any particular features of these patients in terms of total AIMS score or distribution and severity of signs.

Finally, note that the study raters, who trained from the AIMS training tape, practising and comparing between themselves and without reference to JB's scoring, rated differently to JB. Their lower scoring threshold would have major implications in terms of diagnostic cut-off point in a prevalence study. It will be clear that these raters were not clinicians; however, it is not uncommon to use non-psychiatrist raters. This prompts the question: how comparable is our use of the scale to that of other authors? If training together is a critical issue in achieving between-rater agreement, then an international pool of training tapes, together with standards for acceptable inter- and intra-rater variability, would be a worthwhile objective.

Acknowledgements

This work was undertaken while the principal author was a Research Fellow of the N.S.W. Institute of Psychiatry. Our thanks to the National Institute of Mental Health, USA and Sandoz, Australia for the loan of training material.

References

BERGEN, J. A., GRIFFITHS, D. A., REY, J. M. & BEUMONT, P. J. V. (1984) Tardive dyskinesia: fluctuating patient or fluctuating rater. *British Journal of Psychiatry*, **144**, 498–502.

FLACK, V. F. (1987) Confidence intervals for the interrater agreement measure kappa. *Communications in Statistics – Theory and Methods*, **16**, 953–968.

GUY, W. (1976) *ECDU Assessment Manual for Psychopharmacology*, pp. 534–537. Washington, DC: US Department of Health, Education and Welfare.

SCHOOLER, N. R. & KANE, J. M. (1982) Research diagnoses for tardive dyskinesia. *Archives of General Psychiatry*, **39**, 486–487.

*J. A. Bergen, MB, BS, FRANZCP, *Research Fellow, NSW Institute of Psychiatry, Dept of Psychiatry, University of Sydney*; N. B. Carter, BSc, *Senior Experimental Scientist, CSIRO, Division of Mathematics and Statistics, Sydney*; J. Craig, BA Registered General Nurse, *Research Assistant, Department of Psychiatry, University of Sydney*; D. MacFarlane, BSc (Hons), *Research Assistant, Department of Psychiatry, University of Sydney*; E. F. Smith, BA, *Experimental Scientist, CSIRO, Division of Mathematics and Statistics, Sydney*; P. J. V. Beumont, MB, ChB, MSc, MPhil, MRCPE, FRACP, FRANZCP, FRCPsych, *Professor of Psychiatry, University of Sydney*

*Correspondence: *Department of Community Medicine, Royal Prince Alfred Hospital, Missenden Road, Camperdown, 2050*

https://doi.org/10.1192/bjp.152.5.670 Published online by Cambridge University Press