Use of the Global Assessment of Function scale in learning disability*

PATRICIA OLIVER, SHERVA COORAY, PETER TYRER and DOMENIC CICCHETTI on behalf of the Parkside Learning Disability Research Initative (PLDRI) Group

Background The Global Assessment of Function (GAF) scale is widely used in adult psychiatric practice and research but it has not often been used in learning disability, which is inherently more complex.

Aims To evaluate the reliability of GAF in the assessment of learning disability.

Method GAF reliability was tested by simultaneous multiple rating of unselected case vignettes (*n*=19–25) from health professionals of different disciplines, under controlled conditions. Analysis of reliability was made with the intraclass correlation coefficient (R_I) with separate assessments to determine rater bias and individual performance of raters.

Results The results of three data-sets showed generally poor overall levels of agreement, with R_I levels of 0.35 and 0.28 and somewhat better levels for current GAF scores (R_I=0.49). However, a subset of raters was identified that achieved much higher levels (R_I=0.54 to 0.74).

Conclusions The GAF, in its current format, is not reliable enough to be used in the routine assessment of learning disability. A subgroup of raters, however, have ratings that are, by current biostatistical criteria, sufficiently reliable.

Declaration of interest None.

The assessment of function in learning disability is a necessary clinical skill. Function is, however, more difficult to describe and standardise in learning disability than in other forms of psychiatric disorder, because function is relative to the intellectual level of the individual as well as to any problems created by mental illness. Routine global assessments of function are becoming more common in general psychiatry and increasingly are likely to be used in ordinary clinical work, as evidence-based medicine develops and quality standards become necessary to monitor performance. One of the earliest published global rating scales was the Health Sickness Rating Scale (HSRS) developed by Luborsky (1962). This was revised by Endicott *et al* (1976) as the Global Assessment Scale (GAS), the aim of which was to address the shortcomings of the HSRS. The GAS was subsequently modified as the Global Assessment of Function (GAF) scale which, since 1987 has been Axis V of the DSM–III–R multi-axial classificatory system (American Psychiatric Association, 1987). The GAF score is frequently recorded in routine clinical practice, but no such general instrument exists for learning disability. As such we thought it would be valuable to examine the reliability of the GAF in this population group and, in particular, to determine whether the elements of personality disorder and intellectual disability, combined in this axis of classification, might complicate assessment.

METHOD

The intention of the investigation was to replicate as nearly as possible the assessment of clinical data in ordinary practice. The approach used was the measure of agreement between raters who scored case vignettes. An example of a case vignette is shown in the Appendix. To determine whether levels of agreement were robust a large number of assessors were used, not all of whom were involved in clinical practice with patients with learning disability. The case vignette approach is a measure of inter-judgement agreement rather than inter-observer agreement, as the element of observation has been removed (Bech *et al*, 1986; Hjortso *et al*, 1989); however it was appropriate for this enquiry since the major difficulty in recording scores comes from the judgement of behaviour and symptoms.

Each phase of the study included the following stages: the selection of vignettes; explanation of the scoring system and of the completion of ratings; and analysis of data.

In a first phase, preliminary testing of a modified form of the GAF scale with more tightly defined anchor points (Hall, 1995) was carried out on 48 vignettes of clients with mild to moderate learning disability by 19 raters. In a second preliminary phase, the original GAF scale was used and training given to all 25 raters. The second data-set included 38 case vignettes of clients with severe learning disability. Although the 38 case vignettes were prepared to specific World Health Organization (2002) guidelines, not all provided information on the clients' current clinical presentation so that only the worst symptomatology scores were recorded for this data-set.

Selection of vignettes

Case vignettes were selected from the caseload of 12 senior psychiatrists to represent the heterogeneous psychopathology in people with learning disability. This process ensured that there was a representative selection of case material that was heterogeneous in nature but which correctly reflected current practice and documentation in the catchment area. The psychiatrists were asked to include a summary of the presenting problem, history findings and course and treatment–response information, although the last of these was optional.

Scoring procedure

The vignettes were assessed independently and simultaneously by 19 professionals in a first phase (Table 1) and 25 in a second phase (Table 2). In the first phase, all participants received written course material and 2 hours' common introduction to scoring the Modified GAF scale. In the second

*Paper presented at the second conference of the British and Irish Group for the Study of Personality Disorders (BIGSPD), University of Leicester, UK, 31 January to 3 February 2001.

https://doi.org/10.1192/bjp.182.44.s32 Published online by Cambridge University Press

Table 1 First-phase interrater reliability of modified Global Assessment of Function (GAF) (worst/current scores)[1]

Variable	Reliable assessors' scores (n=8)	Unreliable assessors' scores (n=11)	Overall level of agreement (n=19)
Mean level of agreement (R_I) (worst scores)	0.63 (good)	0.26 (poor)	0.35 (poor)
Mean level of agreement (R_I) (current scores)	0.74 (very good)	0.36 (poor)	0.49 (fair)
Assessors aged <45 (worst score assessors), %	12.5	18.1	15.7
Psychiatrists (worst score assessors), %	75	63.6	68.4
Excellent levels of agreement (>0.75) (worst score assessors), %	47.2	2.2	N/A

[1] Distribution of scores and reliability for GAF ratings at worst and current level of function in 48 cases rated by 19 raters.

Table 2 Second-phase interrater reliability of original Global Assessment of Function (GAF) (worst scores)[1]

Variable	Reliable assessors' scores (n=12)	Unreliable assessors' scores (n=13)	Overall level of agreement (n=25)
Mean level of agreement (R_I)	0.54 (fair)	0.15 (poor)	0.28 (poor)
Assessors aged <45, %	25	15.4	20
Psychiatrists, %	66.6	46.2	56
Good or excellent levels of agreement (>0.75), %	5.1	0	N/A
Assessor ratings with significant rater bias, mean (s.d.)	0.77 (11)	3.42 (3.4)	N/A

[1] Distribution of scores and reliability for GAF ratings at worst level of function in 38 cases rated by 25 raters.

phase, they received written course material and 2 hours' common introduction to the scoring of the original GAF. The training emphasised that both scales were continuous and the anchor points were only guides; and that although all forms of disability and symptomatology should be assessed, some allowances should normally be made for the intellectual level of the subject concerned when scoring her/his function. For each vignette, during the first phase the assessor was asked to record the GAF score both currently and at the time of greatest dysfunction or worst score (the choice about this time being left to the assessor). During the second phase, the assessor was asked to record only the worst score.

Analysis of data

All data were analysed for interrater reliability using the intraclass correlation coefficient (Bartko, 1966). This is appropriate for the assessment of continuous data and allowance is made for chance association in calculating agreement. Using a computer program BigRi (Cicchetti & Showalter, 1988), both overall levels of agreement and rater bias were assessed for the raters. We also applied a new reliability statistic that assesses examiner agreement and bias in ratings on a case-by-case basis (Cicchetti *et al*, 1997, 1999; Cicchetti & Showalter, 1997; Baca-Garcia *et al*, 2001). The step-by-step method for data analysis is described in Table 3.

RESULTS

The results are shown in Tables 1 and 2 for the two phases of the study. There was a greater than twofold difference between the mean GAF scores of the raters and this was associated with significant rater bias during the second phase of the study, most markedly for those with poor reliability. Examination of those with good and poor reliability showed no marked differences in terms of the raters' age, experience, discipline, gender or practice in learning disability. The reliable and unreliable raters were similar with regard to worst and best GAF

Table 3 Step-by-step methods for interrater analysis

(a) We obtained an overall intraclass reliability coefficient (R_I) among all the raters in a given data-set, using the BigRi program (Cicchetti & Showalter, 1988).
(b) We obtained a separate R_I for each rater with every other rater.
(c) Applying the clinical or practical criteria of Cicchetti & Sparrow (1981), we classified each of the rater R_I coefficients into one of four categories, such that: R_I <0.40=poor, R_I between 0.40 and 0.59=fair, R_I between 0.60 and 0.74=good and R_I between 0.75 and 1.00=excellent.
(d) We assigned a weight to each of the four categories of clinical significance, as follows: poor=0; fair=1; good=2; and excellent=3.
(e) We obtained a total Clinical Level Score (CLS) for each of the raters.
(f) We rank-ordered the CLSs from lowest to highest.
(g) We located the median CLS score across all of the raters.
(h) We classified those raters whose CLS was above the median value as the reliable examiners; and we classified the remaining raters, those at or below the median, as the unreliable examiners.
(i) We recalculated separate R_I values for the reliable and unreliable raters.

scores in the first study, with 75% and 82% concordance for reliable and unreliable rater groups, respectively.

DISCUSSION

The findings demonstrate the positive and negative aspects of the GAF. The ease with which it can be applied to the wide range of patients with learning disability on the basis of clinical vignettes alone, some of which are vague and not particularly conducive to quantitative assessment, shows the versatility of the instrument. The staff involved had a wide range of professional expertise, and no difficulties were experienced in understanding the instrument despite only minimum training. However, the level of agreement was relatively low for both current and worst-case scenarios. It is clear from the large range of scores that there is considerable difficulty in rating global function across the domains of personality, intellectual level and symptomatology of mental state disorder.

There was considerable rater bias in the assessments of GAF scores, with a wide variation between mean scores for each rater. The variation was associated with poorer agreement. The fact that there was concordance between reliable and unreliable raters suggests that the achieving of good and poor reliability is not a chance event and is probably accounted for by different perceptions of the GAF scale in its current form.

The findings are similar to those of Loevdahl & Friis (1996), who estimated the level of GAF agreement with 104 raters from 6 therapeutic centres in their assessment of 5 clinical case vignettes. Systematic differences between centres were up to 6 points, and the authors concluded that GAF reliability was unsatisfactory in routine clinical settings. However, Rey *et al* (1995), using well-trained raters, reported interrater reliability ranging from 0.83 to 0.87 for the GAF of general psychiatric patients in a clinical setting. The reliability and the validity of the GAF was also tested by Jones *et al* (1995) with psychiatric patients, and their trained raters had an interrater reliability score of 0.72 for the GAF in total.

Several methods could improve agreement in learning disability. These include:

(a) splitting the scale into clinical and social function sections (Tyrer *et al*, 1998);

(b) better standardisation of case vignettes (but excessive rigidity could improve reliability spuriously);

(c) formally stating that intellectual function level should (or should not) be taken into account in making a rating;

(d) more extensive training of raters;

(e) changing the examples given in the scale from those derived from general psychiatry to those from learning disability practice;

(f) alternatively, a major modification of the scale could be used for learning disability, but this would not be comparable with the original GAF scale.

We conclude that, although in its present form the GAF scale is not suitable for general learning disability use, it is none the less possible to identify from among a larger pool of independent examiners those whose ratings are, by current biostatistical criteria, sufficiently reliable for both clinical and research applications. Specifically, we have been able to find and cross-validate subsets of reliable raters (R_I values between 0.53 and 0.74) from among a larger pool of clinical examiners.

ACKNOWLEDGEMENTS

We thank Parkside Health NHS Trust for their funding and support of the Parkside Learning Disability Research Initiative (PLDRI) and Helen Bond, Senior Library Assistant, Hertfordshire Partnership NHS Trust, for her invaluable assistance throughout the project. We also appreciate the work of Donald Showalter, Senior Computer Programmer, VA Northeast Program Evaluation Center,West Haven, CT, USA, who wrote the reliability assessment computer programs used in this investigation. The Parkside Learning Disability Research Initiative (PLDRI) Group involves seven NHS trusts and health authorities. *Parkside Health NHS Trust*: Mary Antony, Michael Attwood, Alina Bakala, Angela Brady, Yang Chang, Cathy Claydon, Fred Cowperthwaite, Kofi Krafona, Zenobia Nadirshaw, Nihal Ranasinghe, Vijaya Sharma and Heather Shaw; *Barnet Healthcare NHS Trust*: Shridhar Mahadeshwar; *Brent and Harrow Health Authority*: Nandha Balan; *Harrow and Hillingdon NHS Trust*: Adrienne Regan and Iqbal Singh; *Hertfordshire Partnership NHS Trust*: Marius Cooray, Nimal Marker, George Matthew, Jack Piachaud, Renuga Rasaratnam, Poppy Sebaratnam and Shyamala Thalayasingham; *Hounslow and Spelthorne NHS Trust*: Stephanie De Silva, Venkat Murthy and Manga Sabaratnam; *Leicestershire and Rutland Healthcare NHS Trust*: Regi Alexander.

APPENDIX

Sample case vignette

C is a 35-year-old, single African–Caribbean man institutionalised since the age of 4 years.

Problems include:

(a) unprovoked, unpredictable, opportunistic aggression against others, several of these incidents resulting in grievous bodily harm;

(b) property destruction;

(c) sexual attacks on vulnerable persons of both genders;

(d) self-injurious behaviour including biting, slapping, poking causing tissue damage;

(e) sexual over-arousal and masturbation;

(f) antisocial behaviour, inclusive of faecal smearing, screaming, overactivity;

(g) poor sleep pattern.

The above problems have been present over most of his life since adolescence. Longitudinal monitoring of his behaviour indicates that there is a definite waxing and waning of the intensity, and the pattern appears to be cyclical regardless of environmental and other variables. Functional analysis demonstrates that there is also a clear relationship to attention-seeking and staff changes.

History

C comes from a close-knit but disorganised, large family. Very little is known about his natural father who left home when C was an infant. Early history is sparse, except that his mother had a prolonged labour. He was described as slow and difficult from childhood. Speech was limited to the odd word and noises. At the long-stay institution he continued to be disruptive and aggressive towards other people. From the age of 12 he was sexually active and needed constant supervision in the mixed children's ward to prevent attacks on both male and female children. He was admitted to a community children's unit for people with severe learning disability (National Health Service) and subsequently to an assessment–treatment facility where he has remained in view of his complex needs. Intensive work within the unit has resulted in considerable improvement of his activities of daily living and communication.

Findings

On examination, C is a well-built man who is likely to be intimidating to strangers or, alternatively, overfriendly. He has no dysmorphic features. He has limited eye contact and is able to communicate his basic needs using single words or very short sentences in conjunction with Makaton signs. Attention span is limited. He likes repetitive movements and flicking as well as ritualistic tapping and slapping. Likes playing with his bodily fluids. Does not like changes in routine, repeats the same words and sounds. He enjoys music, especially rhythms with a strong beat. Periodically he becomes persistently

https://doi.org/10.1192/bjp.182.44.s32 Published online by Cambridge University Press

over-excited, when meaningful communication is replaced by increased episodes of hooting, screaming and constant slapping as well as sexual over-arousal. At such times his sleep pattern becomes even more disrupted, reducing from about 3–5 hours at night to sometime less than 1 hour. Despite this he does not appear to be tired. Since his speech improved, staff have commented that he goes through his whole repertoire of language parrot-fashion repeatedly. Self-injurious behaviour is common and he appears to have a very high pain threshold.

Course

Management has particular emphasis on social-skills training. The behaviour problems have responded in a limited way as a result of the specialist input, structure and discipline, within the unit. Nevertheless, he continues to need intensive supervision at all times and has been detained under Section 3 of the Mental Health Act since 1990, following a serious physical attack on a fellow resident. The cyclicity of his hyperactivity inclusive of escalation of behaviour problems and sleep disorder has been much reduced by the current regimen of medication.

CLINICAL IMPLICATIONS

- Ratings of global function using the Global Assessment of Function (GAF) scale in learning disability are not reliable for ordinary clinical practice.
- Reliability is better for current function than for a description of worst lifetime function.
- The interaction between intellectual disability level, personality, behavioural status and mental symptomatology may need to be acknowledged in scoring instructions.

LIMITATIONS

- Ratings of global function were compared using the case vignette method only.
- Most of the raters were not familiar with the GAF scale before the study.
- The quality of the case vignettes was variable and, even though this reflected ordinary clinical practice, it could have influenced levels of agreement.

PATRICIA OLIVER, PhD, Faculty of Medicine, Imperial College, Paterson Centre, London, UK; SHERVA COORAY, FRCPsych, Parkside Health NHS Trust, Kingsbury Community Unit, Brent, London, UK; PETER TYRER, FRCPsych, DOMENIC CICCHETTI, PhD, Department of Psychological Medicine, Imperial College, London, UK

Correspondence: Sherva Cooray, Consultant Psychiatrist, Parkside Health NHS Trust, Kingsbury Community Unit, Honeypot Lane, London NW9 9QY, UK

REFERENCES

American Psychiatric Association (1987) *Diagnostic and Statistical Manual of Mental Disorders* (3rd edn, revised) (DSM–III–R). Washington, DC: American Psychiatric Association.

Baca-Garcia, E., Blanco, C., Saiz-Ruiz, J., *et al* (2001) Assessment of reliability in the clinical evaluation among investigators in a multi-center clinical trial. *Psychiatry Research*, **102**, 163–173.

Bartko, J. J. (1966) The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, **19**, 3–11.

Bech, P., Haaber, A., Joyce, C. B., *et al* (1986) Experiments on clinical observation and judgement in the assessment of depression: profiled videotapes and judgement analysis. *Psychological Medicine*, **16**, 873–883.

Cicchetti, D. V. & Sparrow, S. S. (1981) Developing criteria for the rating of specific items in a given inventory. *American Journal of Mental Deficiency*, **86**, 127–137.

__ & Showalter, D. (1988) A computer program for determining the reliability of dimensionally scaled data when the numbers and specific sets of examiners may vary at each assessment. *Educational and Psychological Measurement*, **48**, 717–720.

__ & __ (1997) A computer program for assessing inter-examiner agreement when multiple ratings are made on a single subject. *Psychiatry Research*, **72**, 65–68.

__, __ & Rosenheck, R. (1997) A new method for assessing inter-examiner agreement when multiple ratings are made on a single subject: applications to the assessment of neuropsychiatric symptomatology. *Psychiatry Research*, **72**, 51–63.

__, Rosenheck, R., Showalter, D., *et al* (1999) Inter-rater reliability levels of multiple clinical examiners in the evaluation of a schizophrenic patient. Quality of life: level of functioning and neuropsychological symptomatology. *Clinical Neuropsychologist*, **13**, 157–170.

Endicott, J., Spitzer, R. L., Fleiss, J. L., *et al* (1976) The Global Assessment Scale. *Archives of General Psychiatry*, **33**, 766–771.

Hall, R. C. (1995) Global Assessment of Functioning – a modified scale. *Psychosomatics*, **36**, 267–275.

Hjortso, S., Butler, B., Clemmesen, L., *et al* (1989) The use of case vignettes in studies of inter-rater reliability of psychiatric target syndromes and diagnoses – a comparison of ICD–8, ICD–10 and DSM–III. *Acta Psychiatrica Scandinavica*, **80**, 632–638.

Jones, S. H., Thornicroft, G., Coffey, M., *et al* (1995) A brief mental health outcome scale-reliability and validity of the Global Assessment of Functioning (GAF). *British Journal of Psychiatry*, **166**, 654–659.

Loevdahl, H. & Friis, S. (1996) Routine evaluation of mental health: reliable information or worthless 'guesstimates'? *Acta Psychiatrica Scandinavica*, **93**, 125–128.

Luborsky, L. (1962) Clinicians' judgements of mental health. A proposed scale. *Archives of General Psychiatry*, **7**, 407–417.

Rey, J. M., Starling, J., Wever, C., *et al* (1995) Inter-rater reliability of global assessment of functioning in a clinical setting. *Journal of Child Psychology & Psychiatry & Allied Disciplines*, **36**, 787–792.

Tyrer, P., Evans, K., Gandhi, N., *et al* (1998) Randomised controlled trial of two models of care for discharged psychotic patients. *BMJ*, **316**, 106–109.

World Health Organization (2002) *The International Classification of Mental and Behavoural Disorders (ICD–10 Chapter V) Educational Kit* (App. 2). http://www.who.int/msa/ems/icd10/icd10ekit/intro.htm#contents.

https://doi.org/10.1192/bjp.182.44.s32 Published online by Cambridge University Press