British Journal of Psychiatry (1995), **166**, 654–659

A Brief Mental Health Outcome Scale

Reliability and Validity of the Global Assessment of Functioning (GAF)

STEVEN H. JONES, GRAHAM THORNICROFT, MICHAEL COFFEY and GRAHAM DUNN

Background. The Global Assessment of Functioning (GAF) is a quick and simple measure of overall psychological disturbance. However, there is little research on the reliability and validity of this measure in severely mentally ill populations.

Method. Multidisciplinary keyworkers assessed 103 patients at monthly intervals over a 6-month period. Overall GAF scores were obtained, with additional separate ratings for symptoms and disability. These were compared with changes in antipsychotic medication and support needs over the same period.

Results. Satisfactory reliability was obtained for total GAF score and for symptom and disability measures, in spite of raters having only one brief training session. All GAF scores were associated with current support needs of patients. Symptom and disability scores were associated with changes in antipsychotic medication in the previous month. Only symptom score was associated with increases in antipsychotic medication at time of rating.

Conclusion. GAF proved to be a reliable and, within the limits of the indicators used, a valid measure of psychiatric disturbance in our sample of the severely mentally ill. Differences in relationships between the three GAF scores and medication/support needs indicate the usefulness of obtaining all three scores for monitoring levels and type of psychiatric disturbance in this population.

As changes continue to occur apace in the provision of mental health services, the need for quick and accurate indicators of severity of mental illness is increasing. Such measures are potentially useful in tracking the mental health of individual clients and also for initial evaluations of characteristics of client populations attending particular services. DSM–III–R (American Psychiatric Association, 1987) provides a rating scale, the Global Assessment of Functioning (GAF), for measurement of overall psychiatric disturbance. Like the earlier Global Assessment Scale (GAS) (Endicott *et al*, 1976), the GAF has a number of ranked sentences descriptive of psychiatric disturbance, associated with numerical ratings.

Dworkin *et al* (1990) used the GAS with multiple raters in a large sample of chronically mentally ill patients. They found that this scale was reliable both within and between raters over the course of their 18-month study. There have been an increasing number of studies that have employed the GAS as an outcome/severity measure for a range of patient groups. Use of GAF has been less common, although there are exceptions (e.g. Lyness *et al*, 1993; Shanks, 1994). Furthermore, there are few reports that address the issue of the reliability and validity of GAF for a chronically mentally ill population. Shanks (1994) reported that clinician's ratings of case vignettes using GAF achieved acceptable levels of reliability and validity, but felt that the implications of such results for the use of GAF in clinical practice were limited. A further issue with respect to GAF is that in its original form it confounds two areas of functioning: symptomatology and social functioning. As Goldman *et al* (1992) indicate, it has been suggested that these two areas be separated in ratings.

In the present study the GAF was used to assess severity of disturbance in a sample of chronically mentally ill patients, assessed over 6 months. In addition to original GAF scores, separate ratings are taken for symptoms and social functioning. As noted above it is important to have some indication of the validity of these measures: patients were rated therefore on a three-point measure of support needs used routinely within the clinical team and on changes in antipsychotic medication. If valid, then GAF scores would be expected to be associated in a logical manner with changes in support needs, which reflect the practical clinical response to perceived changes in levels of disturbance among clients. Association with medication changes would be expected to be lower, as the team's aim is to maintain patients on minimum effective doses; hence non-pharmacological responses will normally form the first interventions to initial indications of increasing disturbance.

Method

Patients

A total of 103 patients were assessed over a 6-month period. These were all patients attending the District Services Centre of the Maudsley Hospital. They had the following diagnoses: schizophrenia ($n=75$); manic–depressive psychosis ($n=13$); major depression ($n=6$); dysthymia ($n=1$); personality disorder ($n=6$); and obsessional compulsive disorder ($n=2$).

Raters

A total of 12 raters took part in this study. They were all members of the multidisciplinary mental health team treating the above patients. This group comprised five nurses, four psychiatrists, two occupational therapists and one psychologist.

Instruments

Support needs (SUPP) were allocated according to a three-point scale. High support needs (scored 3) were indicated by daily attendance, poor social support, complex family circumstance/no family, poor physical state, high risk to self or others. Medium support needs (scored 2) were indicated by weekly attendance, medium social support, medium family circumstances, medium physical state, medium risk to self or others. Low support needs (scored 1) were indicated by monthly attendance, good social support, good family circumstance, good physical state, low risk to self or others.

Antipsychotic medication (MED) was also rated according to a three-point scale: increased (3 points), remained the same (2 points), decreased (1 point) over the preceding month. Changes in other types of medication were not scored.

GAF consists of nine behavioural descriptors ranging from "absent or minimal symptoms (e.g. mild anxiety before an exam) . . . no more than everyday problems" to "persistent danger of severely hurting self or others . . . or persistent inability to maintain minimal personal hygiene or serious suicidal act with clear expectation of death". Patients are rated between 0 (most severe) and 90 (least severe), each descriptor having a nine-point range. Therefore raters have two decisions to make: they have to decide which of the descriptors is the best summary of the particular client's problems and then for that descriptor they use the nine-point scale to indicate the severity of the problems indicated. Final score was coded as GAF Total (GAFT).

Two further GAF ratings were also made for each patient: symptoms (GAFSYM) and disability (GAFDIS), following Goldman *et al* (1992). Since each of the above nine descriptors confounds symptoms and disability, the symptom scale is formed by taking the part of each descriptor that deals only with symptoms; the remaining aspects of disturbance are then described in the disability scale.

Procedure

Instruments were described to raters in a brief group training session of approximately 30 minutes. Any raters added after this first session were given individual instruction in the use of these measures. Ratings took place on the first Friday of each month, over a 6-month period. Ratings were made with respect to the patient's lowest level of functioning at the time of rating. Each patient was rated by their keyworker (clinician from any discipline responsible for co-ordination of care for that patient).

It was not possible to obtain complete rating sets for any of the six ratings recorded. Numbers varied from a maximum of 92 in month 5 to a minimum of 56 in month 6 out of a possible 103 ratings.

Statistical analysis

Data were analysed by means of mixed effects analysis of variance models using the REML program (Scottish Agricultural Statistical Service, University of Edinburgh). The use of REML in the analysis of reliability studies is described in Dunn (1989, 1992). This approach allows for the estimation of both variance components and fixed effects in repeated data in which not all patients have a complete set of observations. Patient number was included in all models as a random effect. Rater identity was also included as a random effect for the estimation of reliability coefficients. Time of rating (e.g. months 1–6) was included as a fixed factor in all of the following analyses, although in no case was its effect statistically significant.

Results

Table 1 presents mean scores for GAFT, GAFSYM and GAFDIS over the 6-month period of the study. This indicates apparently similar levels of overall disturbance, symptoms and disabilities. These scores indicate at least moderate levels of disturbance in the majority of patients, which would be consistent with the predominance of psychotic illnesses in the group studied.

Table 1
Mean (s.d.) for GAF scores averaged over the six (monthly) rating periods

GAFT	52.4 (14.6)
GAFSYM	53.8 (15.2)
GAFDIS	55.2 (15.2)

GAFT, total Global Assessment of Functioning (GAF) score; GAFSYM, GAF symptom score; GAFDIS, GAF disability score.

Reliability

Separate analyses were carried out for GAFT, GAFSYM and GAFDIS. Models included Patient (1–103) and Rater (1–12) as random effects, with time (1–6) as a fixed factor. For GAFT the estimated variance components for Rater, Patient and 'Error' were 52.27 (s.e. 27.43), 127.90 (s.e. 20.45) and 49.34 (s.e. 3.77), respectively. The reliability (generalisability) of a rating in which each assessment for each patient is made by a randomly chosen rater is therefore given by the ratio 127.90/(52.27 + 127.90 + 49.34) = 0.56. If a single rater were to rate all patients the rater component could be dropped from this expression, yielding a reliability (generalisability) coefficient of 127.90/(127.90 + 49.34) = 0.72.

The corresponding variance components for GAFSYM were: Rater = 32.76 (s.e. 19.65), Patient = 139.60 (s.e. 22.69) and 'Error' = 63.75 (s.e. 4.88). The reliability for an assessment made by a randomly selected rater is 139.60/(139.60 + 32.76 + 63.75) = 0.59; that for ratings when all made by the same rater is 139.60/(139.60 + 63.75) = 0.69. The variance components for GAFDIS were: Rater = 39.93 (s.e. 22.19), Patient = 127.00 (s.e. 20.46) and 'Error' = 53.19 (s.e. 4.08). The reliability coefficients for randomly selected raters for each assessment, or for a single rater to make all assessments, are 127.00/(127.00 + 39.93 + 53.19) = 0.58 and 127.00/(127.00 + 53.19) = 0.70, respectively.

Table 2
Rater effects on GAFT, GAFSYM and GAFDIS (BLUPs, measured as deviations from an overall mean)

Rater	GAFT	GAFSYM	GAFDIS
1	−2.74	−1.89	−4.33
2	2.44	1.82	2.59
3	−0.04	2.92	1.30
4	1.17	1.11	0.95
5	−8.07	−2.69	−7.99
6	12.80	9.48	10.14
7	2.48	2.44	0.87
8	−10.80	−7.44	−8.68
9	4.22	−0.88	4.70
10	1.33	0.82	0.50
11	5.22	2.84	5.30
12	−8.02	−9.53	−5.35
s.e.[1]	4.37	4.34	4.28

1. Average standard error of differences between any pair of raters. GAFT, total Global Assessment of Functioning (GAF) score; GAFSYM, GAF symptom score; GAFDIS, GAF disability score.

The above results indicate that rater effects are a significant source of variation in the assessments. This was confirmed by the following. Individual Rater effects (BLUPs, best unbiased linear predictors) were also produced by the REML analyses, together with the standard errors for the differences between pairs of raters (see Table 2). The difference between any two raters represents their constant bias, relative to each other. In the case of GAFT, the maximum relative bias was about 23.5 points (s.e. about 4.5). For GAFSYM, the maximum relative bias was about 18 (s.e. about 4.5). Finally, for GAFDIS the maximum relative bias was about 18.5 (s.e. about 4.5).

These apparently large rater effects are likely to have arisen from the fact that the study was not designed as a formal generalisability study in which a selection of raters rated each of the 103 patients; the values given in Table 2 probably reflect changes in patient means rather than biases. In fact most of the patients (84%) will have been rated by only one rater (their keyworker). These patients are essentially nested within raters. The rater effects and patient differences will, therefore, be partially confounded. On the assumption that the rater effects were negligible the above analyses were repeated after dropping Rater as a random effect. If, indeed, the dropped rater effects were confounded with patient differences, rather than inflating measurement error, we would expect the new patient variance components to be higher. This was the case. For GAFT the variance components were 167.20 (s.e. 25.39) and 53.68 (s.e. 4.07) for Patient and 'Error', respectively. The corresponding reliability coefficient is 167.20/(167.20 + 53.68) = 0.76. For GAFSYM the variance components were 160.50 (s.e. 24.96) and 66.85 (s.e. 5.08) for Patient and 'Error', respectively. The reliability of GAFSYM is therefore 160.50/(160.50 + 66.85) = 0.71. Similarly, the variance components of GAFDIS were 161.80 (s.e. 24.76) and 56.34 (s.e. 4.29) for Patient and 'Error' respectively, with a corresponding reliability of 161.80/(161.80 + 56.34) = 0.74. These reliabilities are quite satisfactory. A similar analysis for SUPP gave a reliability of 0.55 (equivalent to a weighted kappa).

Relationship between GAF and support needs

Separate analyses were carried out for each of the three GAF measures in relation to current support

https://doi.org/10.1192/bjp.166.5.654 Published online by Cambridge University Press

Table 3
Estimated effects[1] of SUPP1, SUPP2, MED1 and MED2 on current GAF scores (GAFT2, GAFSYM2 and GAFDIS2), each effect assessed in separate models

	SUPP1				SUPP2				MED1				MED2			
	1	2	3	s.e.[2]	1	2	3	s.e.[2]	1	2	3	s.e.[2]	1	2	3	s.e.[2]
GAFT2	0.00	−2.08	−6.32	1.66	0.00	−4.48	−16.50	1.51	0.00	−1.27	−2.34	2.34	0.00	1.24	−4.88	2.21
GAFSYM2	0.00	−1.97	−5.27	1.80	0.00	−4.09	−14.96	1.68	0.00	−1.25	0.19	2.62	0.00	−1.76	−9.35	2.41
GAFDIS2	0.00	−1.09	−5.63	1.77	0.00	−4.86	−17.28	1.63	0.00	2.02	4.95	2.56	0.00	0.95	−7.81	2.34

1. Measured as contrasts from level 1 of SUPP2 and MED2, as appropriate.
2. Average standard error of difference between pairs of effects.
GAFT2, current total Global Assessment of Functioning (GAF) score; GAFSYM2, current GAF symptom score; GAFDIS2, current GAF disability score; GAFT1, total GAF score in preceding month; GAFSYM1, GAF symptom score in preceding month; GAFDIS1, GAF disability score in preceding month; SUPP2, current level of support needs; MED2, current needs for antipsychotic medication; SUPP1, level of support needs in preceding month; MED1, needs for antipsychotic medication in preceding month.

Table 4
Estimated effects[1] of medication and support on current GAF scores after allowing for effects of the previous months' GAF scores[2]

	SUPP1				SUPP2				MED1				MED2			
	1	2	3	s.e.[2]	1	2	3	s.e.[2]	1	2	3	s.e.[2]	1	2	3	s.e.[2]
GAFT	0.00	0.58	6.44	1.73	0.00	−3.66	−15.36	1.66	0.00	1.42	2.85	2.55	0.00	2.77	−0.25	2.56
GAFSYM	0.00	−0.22	5.91	1.92	0.00	−2.36	−13.03	1.95	0.00	−0.09	6.30	2.78	0.00	−0.59	−4.88	2.78
GAFDIS	0.00	1.35	5.81	1.78	0.00	−3.53	−15.92	1.80	0.00	2.70	11.41	2.54	0.00	3.16	−1.16	2.55

1. Measured as contrasts from level 1 of SUPP2 and MED2, as appropriate.
2. Average standard error of difference between pairs of effects.
GAFT, total Global Assessment of Functioning (GAF) score; GAFSYM, GAF symptom score; GAFDIS, GAF disability score; SUPP2, current level of support needs; MED2, current needs for antipsychotic medication; SUPP1, level of support needs in preceding month; MED1, needs for antipsychotic medication in preceding month.

Table 5
Estimated joint effects[1] of SUPP2 and MED2 on GAF scores in preceding month (GAFT1, GAFSYM1 and GAFDIS1)

	SUPP2				MED2			
	1	2	3	s.e.[2]	1	2	3	s.e.[2]
GAFT2	0.00	-3.59	-10.60	1.88	0.00	-1.47	4.49	2.58
GAFSYM2	0.00	-5.26	-12.57	2.00	0.00	-0.18	3.47	2.77
GAFDIS2	0.00	-4.05	-6.44	1.82	0.00	-1.41	1.44	2.47

1. Measured as contrasts from level 1 of SUPP2 and MED2, as appropriate.
2. Average standard error of difference between pairs of effects.
GAFT2, current total Global Assessment of Functioning (GAF) score; GAFSYM2, current GAF symptom score; GAFDIS2, current GAF disability score; GAFT1, total score in preceding month; GAFSYM1, GAF symptom score in preceding month; GAFDIS1, GAF disability score in preceding month; SUPP2, current level of support needs; MED2, current needs for antipsychotic medication.

(SUPP2), current medication (MED2), support in the preceding month (SUPP1) and medication in the preceding month (MED1). In this, and all the following analyses, time is fitted as a fixed effect, and Patient and Rater as random effects. The remaining effects are fixed. The estimated effects are shown in Table 3. Large negative associations were noted between SUPP2 and all GAF measures. Smaller associations, but of the same sign, were noted between SUPP1 and GAF measures. MED1 did not have a consistent association with GAF scores. MED2, when analysed as a separate factor, was found to be negatively associated with GAF scores, particularly GAFSYM.

Current total GAF score (GAFT2) was next used as the dependent variable, with factors SUPP1, SUPP2, MED1 and MED2 simultaneously entered as fixed effects, covarying for GAF score (GAFT1) in preceding month. The results are shown in Table 4. This indicated no significant effects for time, MED1 or MED2. Increases in support in the preceding month were associated with increased current GAF total (improved functioning). The clearest effect was that of lower GAF total (GAFT2) being associated with higher SUPP2.

Current GAFSYM and GAFDIS scores (GAFSYM2 and GAFDIS2, respectively) were also entered as dependent variables (in separate analyses) with SUPP1, SUPP2, MED1 and MED2 as simultaneous factors and the appropriate prior GAFSYM and GAFDIS scores (GAFSYM1 and GAFDIS1, respectively) entered as covariates. As above, SUPP1 was positively associated with current GAF scores of both types, while the relationship for SUPP2 was larger and of the opposite sign. Similarly, MED1 was positively associated with current GAF (particularly symptoms), while MED2 was modestly negatively associated. These results (Table 4) suggest that increased support and medication in the previous month leads to decreased psychological disturbance as measured by GAF. Increases in disturbance lead to immediate changes in support and, to a lesser extent, to increases in medication.

Finally, we examined the relationship between preceding GAF scores (GAFT1, GAFSYM1 and GAFDIS1) and support level at the point of rating (SUPP2) and antipsychotic medication, also at the point of rating (MED2). The GAF scores were used as the dependent variables in the REML runs, with SUPP2 and MED2 as predictive factors (fixed effects). The estimated effects for SUPP2 and MED2 are shown in Table 5. These estimates indicate a substantial relationship between SUPP2 and all GAF scores, this being largest for GAFSYM and smallest for GAFDIS. In all cases a decreased GAF score was associated with an increase in the clinically identified support needs of the client. There was no significant relationship between GAF scores and MED2 for these comparisons.

Discussion

The current data indicate that GAF is a reliable measure of disturbance of psychological functioning in a cohort of long-term mentally ill patients. Following Goldman *et al* (1992), the GAF was administered as both an overall scale and as two separate measures assessing symptoms and disability. Reliability of ratings was satisfactory in all cases, indicating the viability of these more specific scales for use in clinical practice. It should be emphasised that these reliability figures were obtained for raters given only limited training in the use of the above measures. It is therefore possible to be relatively confident, in contrast to Shanks (1994), that these figures represent a realistic estimate of reliability figures in general clinical practice.

It is clear that a reliable measure can derive its reliability from a number of sources, some having no association with validity of the measurement being taken. In the case of global measures it is

necessary to be alert to the possibility that reliability figures may derive from insensitivity of measurement. Therefore, in the current study, GAF scores were assessed in relation to two indicators of clinical need routinely used in clinical practice by the multidisciplinary team. These were the level of support required by a given patient and the changes made to antipsychotic medication.

When considered as separate factors it was found that support levels in both current and preceding month were negatively associated with all three GAF scores. Thus greater psychological disturbance, as indicated by symptoms, disability or total GAF score, was associated with higher levels of clinical support offered by the clinical team. Current medication levels showed a similar negative association with GAF scores, this being clearest in relation to increased symptomatology rather than disability.

When all factors were considered simultaneously, the strongest single association found was that between current support level and all three GAF scores. When REML estimates control for preceding month's GAF score level the relationship between current GAF and prior support becomes positive, suggesting that early increases in support level are followed by improved functioning in subsequent months. Current medication is no longer associated with total GAF score in this analysis, suggesting it may be an artefact of the large association with current support. When symptom and disability scores are analysed in a similar manner, the relationship with support levels reflects those for total GAF score. However, for both there is a positive relationship with medication in the previous month, when preceding GAF scores are controlled for. This indicates that increases in antipsychotic medication in the previous month are followed by improved levels of current functioning, as measured by both symptom and disability scores. Association between current medication and GAF remains negative for symptom score, but is minimal for disability.

The GAF has been shown to be a reliable, quick measure of disturbance in functioning, which can be readily used by multidisciplinary raters, without the need for extensive training. Whereas more detailed measures of symptomatology and social functioning are widely available, they are little used in routine clinical practice due to the resources and training required (e.g. Wykes & Sturt, 1986). Hence, in addition to being a useful quick measure in population-based surveys (Phelan *et al*, 1994), the GAF has a role as a routine measure in clinical practice, which then facilitates monitoring of individual patients over time and of levels of morbidity in particular clinical services. The current data indicate that it is worthwhile obtaining all three GAF scores when assessing patients, as their relationships with medication, and to a lesser degree with prior support levels, differ.

References

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

Dunn, G. (1989) *The Design and Analysis of Reliability Studies*. London: Edward Arnold.

—— (1992) The design and analysis of reliability studies. *Statistical Methods in Medical Research*, **1**, 123–157.

Dworkin, R. J., Friedman, L. C., Telschow, R. L., *et al* (1990) The longitudinal use of the Global Assessment scale in multiple rater situations. *Community Mental Health Journal*, **26**, 335–344.

Endicott, J., Spitzer, R. L., Fleiss, J. L., *et al* (1976) The global assessment scale. *Archives of General Psychiatry*, **33**, 766–771.

Goldman, H. H., Skodol, A. E. & Lave, T. R. (1992) Revising axis V for DSM-IV: A review of measures of social functioning. *American Journal of Psychiatry*, **149**, 1148–1156.

Lyness, J. M., Caine, E. D., Conwell, Y., *et al* (1993) Depressive symptoms, medical illness and functional status in depressed psychiatric patients. *American Journal of Psychiatry*, **150**, 910–915.

Phelan, M., Wykes, T. & Goldman, H. (1994) Global functioning scales: A review. *Social Psychiatry and Psychiatric Epidemiology*, **29**, 205–211.

Shanks, J. (1994) How are things? Developing outcome measures for mental health services. *Health Trends* (in press).

Wykes, T. & Sturt, E. (1986) The measurement of social behaviour in psychiatric patients: An assessment of the reliability and validity of SBS. *British Journal of Psychiatry*, **148**, 1–11.

Steven H. Jones, PhD, **Graham Thornicroft**, FRCPsych, **Graham Dunn**, PhD, Institute of Psychiatry, London; **Michael Coffey**, RMN, Maudsley Hospital, London

Correspondence: Dr Jones, Department of Psychology, Institute of Psychiatry, DeCrespigny Park, London SE5 8AF

(First received 14 June 1994, accepted 12 September 1994)