Reliability of Global Assessment of Functioning Ratings Made by Clinical Psychiatric Staff

Per Söderberg, M.Sc.
Stefan Tungström, M.Sc.
Bengt Åke Armelius, Ph.D.

Objective: **In the Swedish psychiatric care system, systematic follow-up of clinical work with patients is becoming a part of regular service, and a number of care providers are using the Global Assessment of Functioning (GAF) to measure outcomes. This study investigated the reliability of the GAF and analyzed certain factors that affect measurement errors when the scale is used by regular psychiatric staff.** ***Methods:*** **Eighty-one raters from various psychiatric outpatient clinics rated eight case vignettes. Interrater reliability was assessed by using intraclass correlation coefficients (ICCs), and factors associated with reliability were analyzed by using raters' unique residual values.** ***Results:*** **The results showed that staff who are responsible for assessing first-time patients at outpatient psychiatric clinics and making diagnoses are using the GAF with satisfactory reliability ($ICC_{1,1}$=.81). The factors associated with reliability were raters' subjective attitude toward the GAF and motivation to use the scale and other measurement instruments in psychiatry.** ***Conclusions:*** **GAF ratings made by an individual rater can be used to measure changes and outcomes at the group level. However, the measurement error is too large for assessment of change for an individual patient, in which case it might be necessary to use several raters. If raters are positively inclined to use rating instruments, measurement errors are minimized and reliability is maximized. (*Psychiatric Services* 56:434–438, 2005)**

In recent years, the Swedish psychiatric care system has become more interested in quality improvement and systematic follow-up of treatment results. Today, most patients who are referred to the psychiatric care system are being routinely assessed with the *DSM* system (1). The Global Assessment of Functioning (GAF), which is listed as axis V in *DSM IV-TR,* is used to measure the patient's overall level of functioning at a particular point in time and thus constitutes a tool for developing procedures for measuring outcomes.

In our clinic, as well as in many others, GAF ratings are made as part of the initial assessment and the discharge assessment. The assessments are saved in databases together with other clinical data, which makes it possible to use the ratings for evaluating treatment effects. However, to be certain that relatively small differences in GAF scores can be validly interpreted as treatment effects, the reliability of the GAF ratings must be high. If not, there is an obvious risk that any change in GAF scores from before treatment to after treatment is just an effect of measurement error or chance. It is therefore important to know whether GAF ratings that are made by regular clinical staff meet satisfactory standards of reliability. It is also important to know which rater factors are associated with the reliability of the GAF assessments, because some of them can be used to improve reliability—for example, by means of training or selection of raters.

Considering how extensively and routinely the GAF is used in clinical practice and how often it is cited in international studies, surprisingly few studies have investigated the scale's reliability. In the original reliability study to which all later versions of *DSM* refer (1–3), results from five reliability studies of the Global Assessment Scale (GAS) (a forerunner to the GAF) are presented (4). Interrater reliability measured with intraclass correlation coefficients (ICCs) for the GAS ranged from .61 to .91 in the substudies, and the value for the instrument's standard error ranged from 5 to 8 points. The standard error can be used to compute the amount of uncertainty around a single rating. Usually, GAF score±2SE is used as a

*Mr. Söderberg and Mr. Tungström are affiliated with the psychiatric research and development department of Säter, Sweden, and the department of psychology of the University of Umeå in Sweden, with which Dr. Armelius is affiliated. Send correspondence to Mr. Söderberg at Psykiatrins Utvechlingsenhet, Box 350, S-783 27 Säter, Sweden (e-mail, per.soderberg@ltdalarna.se).**This article is part of a special section on the Global Assessment of Functioning scale.*

confidence interval. Thus a GAF rating of 60 might just as well have been somewhere between 50 and 70 with a standard error of 5. Other studies of the reliability of the GAF have shown divergent results. Reliability is clearly lower in studies performed in clinical settings (5–8), with reliability estimates between .54 and .65, compared with studies performed in research settings—an ICC of .9 in a study by Tracy and colleagues (9) and an ICC of .86 in a study by Hilsenroth and colleagues (10).

Luborsky and Bachrach (11) studied factors associated with the reliability of the closely related Health Sickness Rating Scale (HSRS). They found that reliability was associated primarily with the raters' experience, knowledge, and training in using the scale. Dworkin and colleagues (12) also showed that the raters' education and training were positively related to the reliability of the GAS. The very good results for reliability reported by Tracy and colleagues (9) and Hilsenroth and colleagues (10) are linked to the structured training programs carried out in conjunction with the research projects.

Loevdahl and Friis (7) examined whether there were differences in reliability between occupational groups and found no differences between psychiatrists and psychiatric nurses. Harel and colleagues (13) also studied the reliability of the GAF in relation to occupational groups. They found that psychiatrists and social workers rated patients with good reliability and that occupation was not related to reliability.

In summary, the rater factors that seem to be associated with interrater reliability of the GAF are the raters' level of education, amount of training in using the GAF, and general clinical experience. Occupation does not seem to be related to the reliability of the scale. The fact that reliability studies based on research data show better reliability than studies based on clinical assessments may be explained by the higher levels of training and education among research raters but may also be due to raters' motivation and incentive. The results of a research project are dependent on the reliability of the assessments, which constitutes a clear incentive to rate accurately. In clinical settings, there is usually not such a strong link to motivation.

Whether the GAF can be used to measure outcome and treatment effect is directly dependent on the reliability of the scale in clinical use. Can the changes the patient has achieved from initiation to termination of treatment as demonstrated by GAF assessments be said to constitute a reliable change? The main objectives of this study were to estimate the reliability of GAF assessments made by clinical staff and to clarify what rater factors might be associated with reliability. In addition, we wanted to investigate whether raters' occupation, frequency, experience, and attitude toward the GAF were associated with reliability.

Methods

Setting

The study was conducted in the psychiatric care system in the Swedish province of Dalarna in March 2002. The system's management gave permission for clinicians to participate on a voluntary basis. All outpatient facilities were invited to participate in the study, and 11 of 17 facilities chose to participate. Lack of participation among facilities may have been due to any of several factors—for example, facility management may not have given priority to the study.

Case vignettes

We used eight case vignettes to provide a representative range of GAF assessments for a psychiatric setting. Four vignettes were videotaped interviews, and the other four vignettes were written. The videos, which were based on real cases and in which professional actors played the roles of patients, were specially produced for training of psychiatric assessments. Each video showed an initial diagnostic assessment, and an experienced clinician responded as though this were an actual patient. The interview was edited to be of about 15 minutes' duration. The written case vignettes were translations from the *DSM-IV* casebook, a study by Spitzer and colleagues (14), and other clinical cases that had been presented in a similar way.

The representativeness of the sample was controlled by comparing the means and standard deviations in the GAF assessments made in this study with data from another large-scale national project (15). The mean scores were 55.7±14 in our study and 55.5±13 in the national study. Thus the case vignettes in the study reported here may be regarded as a representative sample of an outpatient population within the psychiatric care system.

The raters

In the 11 facilities that participated in the study, 81 of 88 available raters participated. During the study period, the total number of raters in the 17 facilities of the outpatient psychiatric organization was 144.

In terms of individual levels of experience in the study group, the raters ranged from neophytes to very experienced raters, with an average of 4.8 years of experience using the GAF and a range of 0 to 15 years. As a comparison, the average number of years of experience in the group of all 144 raters was 4.7. Thirty-seven of the raters in the study group (45 percent) performed GAF assessments at least once a week, whereas the other raters made assessments less frequently (39 percent made assessments at least once a week). A comparison of the occupational groups in the study group with the occupational groups in the larger population showed no significant differences. Thus the study group satisfactorily represented the total rater population in terms of occupation, experience, and frequency of use of the GAF.

Several raters did not turn in their answers on the questionnaire about attitudes toward rating scales, which resulted in an internal dropout rate of 31 percent. Most of those who dropped out came from two facilities, and the reason given for dropping out was that the instructions were unclear when data were collected. A comparison of responders and nonresponders indicated no significant differences between the groups in terms of occupation or sex, and there did not seem to be any systematic reason for dropping out among those who dropped out.

Table 1

Descriptive data for 81 clinicians who made Global Assessment of Functioning (GAF) ratings in a study of the reliability of the scale

Occupation	N	Years' experience		Ratings per week		Attitude to the GAF		
		>5	≤5	≥1	<1	Not positive	Positive	Dropped out
Psychologists	23	13	10	10	13	9	9	5
Social workers	8	4	4	4	4	2	4	2
Psychiatric nurses	31	15	16	19	12	11	11	9
Psychiatric technicians	15	7	8	10	5	2	5	8
Other[a]	4	1	3	1	3	2	1	1

[a] Consulting psychiatrists and occupational therapists

Descriptive data on raters' occupation, experience, frequency, and attitudes toward rating scales are shown in Table 1.

Attitudes toward the GAF and follow-up

The raters used the instructions from *DSM-IV-TR* (1). We developed a questionnaire containing five statements to measure raters' attitudes toward the GAF and follow-up, which was based on a questionnaire developed in the national study (15): "I think GAF assessments are good treatment planning tools"; "We should use more instruments to follow up treatment outcomes"; "I do my utmost to perform good GAF assessments"; "GAF ratings provide important knowledge to me as a treatment provider"; and "A GAF rating is a good indicator of psychiatric illness." Each question was answered on a scale of 1 to 5, with higher scores indicating a more positive attitude. The reliability of the instrument was tested with data from a large group of ordinary psychiatric staff from the national study (N=348) and was found to be .74 (Cronbach's alpha).

Procedure

All participants assessed the eight case vignettes individually, immediately after watching or reading the case vignette. The raters were given about five minutes to perform the rating, which is the normal time used to perform a GAF assessment in clinical settings. The raters had access to the GAF in *DSM-IV.* The participants answered the questionnaire along with some questions about professional skills and personal experience making GAF assessments.

Data analysis

First, we analyzed interrater reliability. Two ICCs were computed according to the model of Shrout and Fleiss (16). The first was $ICC_{1,1}$, which makes it possible to generalize the results to the population of clinical raters. The second was $ICC_{3,1}$, which gives an estimate of the consistency of the ratings for an individual rater.

Our main interest in this study was of course how staff generally perform when doing their assessments in routine clinical practice. Thus $ICC_{1,1}$ was of primary interest, because it can tell us the reliability of any rating done within the population of raters in the outpatient psychiatric care system.

The other formula, $ICC_{3,1}$, measures the consistency of a particular individual's ratings, because the formula allows raters to have a personal systematic deviation compared with the true value (mean value per case).

To analyze what factors were associated with reliability, we needed an index showing how much a single rater's assessments deviated from the "true" GAF scores on each of the case vignettes. Because we defined the true value as the mean rating of all raters on each vignette, the index we sought was one that could be computed as the sum of the individual rater's deviations from the mean rating for the eight vignettes. Computationally this sum is equal to the residual as described in the article by Shrout and Fleiss (16). The rater's unique "error term" for a vignette is calculated from the two-way analysis of variance (ANOVA) on which $ICC_{3,1}$ is based: error term=X – (grand mean + systematic deviation of the case vignette from all case vignettes + systematic deviation of rater from all raters), where X is the rated GAF value and grand mean is the mean of all ratings for all case vignettes (equation 1).

To estimate the residuals we first squared and then took the square root of the individual error terms and then added the error terms for the eight case vignettes:

$$\text{Rater's residual}=\sum_{\text{case } 1\text{–}8}\sqrt{\text{Error term}^2} \quad \text{(equation 2)}$$

The analysis of what factors may be associated with reliability was made by means of a three-way ANOVA, with the residual as the dependent variable and the factors that might be associated with reliability as independent variables (attitude, experience, and frequency).

The variables were dichotomized. Experience (more or less than five years), frequency (more or less than once a week), and attitude (average value above the mean score for all five statements [>3.3 points]) were grouped into the positive group (Table 1).

The association of occupation with reliability was analyzed by means of a one-way ANOVA with the residual as the dependent variable and occupational groups as the independent variable. All statistics were calculated with use of SPSS 11.5.

Results

Reliability

$ICC_{1,1}$ for the eight case vignettes and 81 raters was .81. All paired combinations of raters were within a 95 percent confidence interval of .65 to .95. This result can be generalized to the population of raters within the outpatient psychiatric care system. The result for $ICC_{3,1}$ was .83, which shows that there were certain systematic differences between raters, so that some systematically rated patients slightly higher or lower than the "true" value.

To compute the standard error for the GAF ratings, we needed to know the size of the standard deviation of GAF ratings in an outpatient setting. A good estimate of the standard deviation for the population of psychiatric outpatients in Sweden has been computed in the national study mentioned above (15). In that project, the standard deviation was computed to be 13 points; using this estimate gave us a standard error of 5.7 points in our study. With a 95 percent confidence interval, the "true" individual GAF rating falls within 11.4 of the observed rating. The result shows that a GAF assessment made by a single rater is surrounded by relatively large uncertainty.

Rater factors associated with reliability

The residual was the dependent variable, and the three factors expected to be associated with reliability were the independent variables in a three-way ANOVA. The mean and standard deviation for the residuals for the three variables are shown in Table 2, and the results of the variance analysis are shown in Table 3. There was a significant difference between the groups on attitude ($F=7.62$, $df=1, 48$, $p<.008$), whereby a positive attitude was associated with significantly lower value on the residual. Greater experience tended to be associated with greater reliability, but frequency and the interactions between the variables showed no relation to reliability.

The mean scores for the occupational groups are presented in Table 4. The results of the ANOVA showed a significant difference between occupational groups ($F=2.76$, $df=4, 76$, $p<.034$), whereby social workers had lower mean scores than psychiatric technicians (post hoc test, Bonferroni, $p<.048$). The occupational groups did not differ significantly in experience or attitude toward the GAF.

Table 2

Mean and standard deviation of the residual for experience, attitude to the rating scale, and frequency of ratings in a study of the reliability of the Global Assessment of Functioning (N=56)

Variable	N	Mean	SD
Experience (years)			
More than five	35	32.3	11.2
Five or less	21	39.3	16.3
Attitude to the scale			
Positive	30	30.4	9.6
Not positive	26	40.2	15.9
Frequency of ratings			
At least once a week	28	34.9	13.7
Less than once a week	28	35.0	13.9

Discussion

This study had some limitations that need to be kept in mind in discussing the results. The primary limitation was that we did not ask the raters to rate actual patients but, rather, eight case vignettes. One consequence of this approach is that the information base for the ratings was not as rich as it would be in a real-life interview. However, it is difficult to tell whether this limitation would have increased or decreased the consistency among the raters in this study.

Another limitation is that, in a research situation, the raters are probably more concentrated and attentive to making accurate ratings. In addition, all the raters used *DSM-IV,* which is probably not the case in a clinical situation. This aspect of the study design might have produced greater reliability than in routine clinical work. In summary, we think these factors might have contributed to a somewhat inflated estimate of reliability in our study compared with a regular clinical assessment situation. On the other hand, the large number of raters and the variations in occupation, experience, and attitude toward the GAF seem clearly representative of a regular outpatient staff.

Nevertheless, the main result of this study indicates that staff members who are responsible for assessing patients in the outpatient psychiatric care system can use the GAF with satisfactory reliability. The results may be generalized to other raters in the outpatient psychiatric care system who routinely and con-

Table 3

Three-way analysis of variance for residuals with main effects and interaction effects for rating experience, attitude to the rating scale, and frequency of ratings in a study of reliability of the Global Assessment of Functioning

Source of variance	Sum of square	Mean square	F	df	p
Experience	517.72	517.72	3.19	1	.08
Attitude to the scale	1,237.95	1,237.95	7.64	1	.008
Frequency of ratings	.04	.04	0	1	.987
Experience × attitude	140.19	140.19	.86	1	.357
Experience × frequency	412.69	412.69	2.55	1	.117
Frequency × attitude	104.57	104.57	.64	1	.426
Experience × frequency × attitude	14.20	14.20	.09	1	.769
Error	7,781.53	162.11		48	
Total	78,593.76			56	

Table 4

Mean scores and standard deviations of the residual for 81 clinicians who made Global Assessment of Functioning Ratings in a study of the reliability of the scale

Occupational group	N	Mean	SD
Psychologists	23	34	14
Social workers	8	27	7
Psychiatric nurses	31	37	11
Psychiatric technicians	15	44	17
Other[a]	4	44	23

[a] Consulting psychiatrists and occupational therapists

tinually use the GAF. The reliability of the GAF ratings was .81, which is within an acceptable range of clinical ratings (17) and somewhat higher than estimates from other clinical studies.

If GAF ratings are to be used for comparing aggregated data, whereby a number of ratings are averaged and compared with another set of ratings, the precision will be high enough to enable relatively small differences to be detected, because precision depends on both the reliability of the scale and the number of patients in each group. Such applications of GAF ratings might be to compare aggregated before-and-after treatment data for a group of patients or to compare levels of GAF ratings for different diagnostic groups.

If GAF ratings are to be used to assess an individual patient, precision will usually not be high enough. However, there are some ways to remedy the situation. One is to use more than one clinical rater for each occasion and take the average of all ratings. This approach will improve reliability in the same way that increasing the number of items in a psychological test does. The increase in reliability can be computed by means of the Spearman-Brown formulae.

The only factor that was associated with reliability was attitude toward the GAF and the use of measurement instruments in psychiatry in general. If the rater has a positive attitude in this regard, the reliability of his or her rating is improved. An understanding of this relationship is vital in the clinical setting, because responsibility for performing GAF assessments is not always linked to any incentive or feedback. Clinical facilities should perhaps increase motivation to perform good assessments through more systematic use of the results of GAF assessments—for example, in annual reports. Motivation might also be increased by giving rewards to skillful raters. Another possibility is that raters who cannot learn to rate well should not be allowed to make clinical ratings.

The result that experience using the GAF was associated with only marginally improved reliability should be interpreted with some caution, because most raters in the study group had a relatively long duration of experience making assessments (an average of 4.8 years), and the group with little or no experience was small.

Conclusions

This study showed that clinical raters can use the GAF in a sufficiently reliable way to enable use of the ratings at an aggregated level. To use the scale for an individual patient, special procedures are required. However, a higher level of reliability of clinical ratings is desirable, because there is room for improvement. Raters' motivation levels seem to be the primary factor in improving the reliability of ratings. ♦

Acknowledgment

This study was supported by grant 420011 from the Center for Clinical Research in Dalarna, Sweden.

References

1. Diagnostic and Statistical Manual of Mental Disorders, 4th ed, Text Revision. Washington, DC, American Psychiatric Association, 2000
2. Diagnostic and Statistical Manual of Mental Disorders, 3rd ed. Washington, DC, American Psychiatric Association, 1987
3. Diagnostic and Statistical Manual of Mental Disorders, 4th ed. Washington, DC, American Psychiatric Association, 1994
4. Endicott J, Spitzer R, Fleiss J, et al: The Global Assessment Scale: a procedure for measuring overall severity of psychiatric disturbance. Archives of General Psychiatry 33:766–771, 1976
5. Hall R: Global Assessment of Functioning: a modified scale. Psychosomatics 36:267–275, 1995
6. Jones SH, Thornicroft G, Coffey M, et al: A brief mental health outcome scale: reliability and validity of the Global Assessment of Functioning (GAF). British Journal of Psychiatry 166:654–659, 1995
7. Loevdahl H, Friis S: Routine Evaluation of Mental Health: reliable information or worthless "guesstimates"? Acta Psychiatrica Scandinavica 3:125–128, 1996
8. Michels R, Siebel U, Freyberger HJ, et al: The multiaxial system of ICD-10: evaluation of a preliminary draft in a multicentric field trial. Psychopathology 29:347–356, 1996
9. Tracy K, Adler LA, Rotrosen J, et al: Interrater reliability issues in multicenter trials, part 1. Psychopharmacology Bulletin 33: 53–57, 1997
10. Hilsenroth MJ, Ackerman SJ, Blagys MD, et al: Reliability and validity of DSM-IV axis V. American Journal of Psychiatry 157: 1858–1863, 2000
11. Luborsky L, Bachrach H: Factors influencing clinicians' judgments of mental health. Archives of General Psychiatry 31:292–299, 1974
12. Dworkin RJ, Friedman LC, Telschow RL, et al: The longitudinal use of the Global Assessment Scale in multiple-rater situations. Community Mental Health Journal 26: 335–344, 1990
13. Harel TZ, Smith DW, Rowles JM, et al: A comparison of psychiatrists' clinical-impression-based and social workers' computer-generated GAF scores. Psychiatric Services 53:340–342, 2002
14. Spitzer RL, Gibbon M, Skodol AE, et al: DSM-IV Casebook: A Learning Companion to the Diagnostic and Statistical Manual of Mental Disorders, 4th edition. Washington, DC, American Psychiatric Press, 1994
15. Söderberg P: GAF projektet, Psykiatrins patienter, ett jämförelsematerial mellan olika psykiatriska öppenvårdsmottagningar i Sverige [The patients in psychiatry, comparison of data between different psychiatric outpatient clinics in Sweden.] NYSAM-Rapport. Säter, Sweden, Psykiatrins Utvecklingsenhet, 2001
16. Shrout PE, Fleiss JL: Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86:420–428, 1979
17. Nunally JC, Bernstein ICH: Psychometric Theory, 3rd ed. New York, McGraw-Hill, 1994