Understanding Psychiatric Impairment

COURSE OUTLINE

INTRODUCTION AND COURSE OVERVIEW

Although the DSM-5 has been released, the DSM-IV (and the GAF) will likely continue to be required for use in California psychiatric work injury cases. The DSM-5 no longer uses the multi-axial system or the GAF. Even so, anyone completing medical-legal evaluations of industrial psychiatric injuries must be completely familiar with the concepts presented in this course.

The Global Assessment of Functioning Scale (GAF) is a standard method for a clinician to judge a patient's overall level of psychosocial functioning. The GAF requires a clinician to develop an overall judgment about the patient's current psychological, social, and occupational functioning. This global rating is made on a scale from 1-100, with 1 being the lowest level of functioning and 100 being the highest level of functioning. The primary purpose of the GAF is a quick and efficient method of assessing and summarizing a patient's current psychiatric status (symptoms and functioning) and to assess change. It is most commonly used to assess patients at the beginning of psychiatric treatment, to monitor their progress throughout the intervention, and to provide a status at discharge.

Currently, the GAF is the most widely used method for assessing impairment among patients with psychiatric disorders (Moos et.al. 2002). The GAF was introduced as a new rating scale of overall psychiatric disturbance as Axis V in the Diagnostic and Statistical Manual of Mental Disorders (DSM III-R, American Psychiatric Association, 1987). The GAF Scale was retained in the DSM-IV (1994). Interestingly, Goldman (1992) stated that the GAF was "not widely used" as of the early 1990's. As will be discussed subsequently, mandates for its use by such organizations as the VA Medical Center system and managed care companies have no doubt contributed to its current popularity. The GAF was retained in the DSM-IV-TR (Text Revision, 2000) with a slight modification in the instructions. The modification was necessary due to confusion regarding the time frame for the GAF rating (how was "current" defined) and how the clinician is to appropriately integrate the contributions of a patient's psychiatric symptoms and functioning to the final

GAF score (First and Pincus, 2002). In 2005, the GAF was adopted by the State of California as the primary method for determining permanent psychiatric disabilities in the workers' compensation population. As discussed in the Schedule For Rating Permanent Disabilities (SRPD, 2005), psychiatric impairment is to be evaluated using the GAF, which is then converted to a whole person impairment (WPI). DSM-5 does not contain a multi-axial system or the GAF. All of these issues will be discussed in great detail in this course.

The course will begin with an overview of the DSM-IV Multiaxial Diagnostic System. The GAF comprises Axis V of this system. For those in the mental health field, this will be a review. The history of the development of the GAF Scale in its current form will then be discussed. This review begins with a predecessor of the GAF, the original 100 point Health Sickness Rating Scale (HSRS) developed by Luborsky (1962). Gaining an understanding of the various revisions of the scale can help give the clinician insight into some of its current strengths and weaknesses. The scale began as the Health Sickness Rating Scale in 1962 which was revised to form the Global Assessment Scale (GAS; Endicott et al., 1976). Subsequently, the GAS was revised and developed as the GAF for inclusion in DSM III-R (1987). The GAF remained essentially the same for inclusion in DSM-IV and DSM IV-TR with the exception of a slight modification in the instructions. The GAF was abandoned in DSM-5 but that is not relevant to the QME assessment since the use DSM-IV and use of the GAF will likely continue.

As part of reviewing the diagnostic system of the DSM, the course will discuss the scoring method for the GAF. Although most clinicians who use the GAF frequently believe that they are scoring it in a valid fashion, a review of the research reveals otherwise. The course will discuss specifics of how to properly score the assessment along with common pitfalls. Later in the course, I will discuss alternative scoring conceptualizations that will help the clinician arrive at a more reliable and valid GAF result.

The course will review some of the most common uses of the GAF especially related to those that have dramatically increased its popularity. These include its use in treatment decisions by many managed care organizations, its use being mandated by the VA Medical Center system, as well as its inclusion in the Workers' Compensation Reform legislation in California in 2005.

The course will discuss, in detail, the problems associated with the GAF in its current form. Being aware of the problems and weaknesses of any assessment can help a clinician arrive at a more valid result and avoid any pitfalls. Areas of problems that have been pointed out with the GAF in the research literature include: (1) The scale collapses three dimensions of function into one score; (2) The reliability of the scale is not good except for clinicians who have undergone extensive training as part of research projects; and, (3) the GAF shows essentially no predictive validity. Being aware of these important issues is important to help the clinician use the GAF properly.

The course will then provide a detailed discussion about how to improve one's use of the GAF including such things as carefully following the GAF instructions, utilizing the "split method" of scoring, reviewing a modified GAF that scores all three dimensions independently, being sure to obtain quality information for GAF determination, presenting training vignettes that have been used by the VA Medical Center system, and reviewing research utilizing clinicians trained in the use of the GAF to elucidate how their scores might correlate with those in the injured worker population.

The course will conclude with an excellent research article published by BioMed Central as an open source publication in the Annals of General Psychiatry entitled, Guidelines for Rating Global Assessment of Functioning (GAF) by I.H. Monrad Aas.

LEARNING OBJECTIVES

Explain the four step GAF scoring method
Describe the history of the GAF beginning with the HSRS (1962)
List the three major problem categories with the GAF
List four methods of improving the quality of a GAF rating
Discuss suggestions for improving the GAF

DSM-IV-TR MULTIAXIAL DIAGNOSTIC SYSTEM

In the DSM-IV (1994, TR: 2000), a Multiaxial System is utilized which involves assessing various domains of information across several axes. This was done to assist the clinician in planning treatment and to predict outcome. The five Axes included in the DSM-IV Multiaxial Classification can be seen in Table 1.

Table 1. DSM-IV Multiaxial System

Axis I:	Clinical Disorders
Axis II:	Personality Disorder
Axis III:	General Medical Condition
Axis IV:	Psychosocial and Environmental Problems
Axis V:	Global Assessment of Functioning

Axis I: Clinical Disorders. The various disorders and conditions in the diagnostic classification system are recorded on Axis I except for personality disorders and mental retardation (which are reported on Axis II). The major groups of disorders that are reported on Axis I are listed in Table 2. According to the Manual, when an individual has more than one Axis I disorder, they should all be reported. The primary Axis I diagnosis is listed first.

Table 2. Axis I: Clinical Disorders

Disorders Usually First Diagnosed in Infancy, Childhood, or Adolescence (excluding Mental Retardation which is diagnosed on Axis II)
Delirium, Dementia, and Amnestic and Other Cognitive Disorders
Mental Disorder Not Due to a General Medical Condition
Substance-Related Disorders
Schizophrenia and Other Psychotic Disorders
Mood Disorders
Anxiety Disorders
Somatoform Disorders
Factitious Disorders
Dissociative Disorders

Sexual and Gender Identity Disorders

Eating Disorders

Sleep Disorders

Impulse-Control Disorders Not Elsewhere Classified

Adjustment Disorders

Other Conditions That May be the Focus of Clinical Attention

Axis II: Personality Disorders and Mental Retardation. Personality disorders and mental retardation are reported on Axis II. The disorders to be reported on Axis II are listed in Table 3.

Table 3. Axis II: Personality Disorders and Mental Retardation

Mental Retardation

Cluster A Personality Disorders (odd, eccentric)

Paranoid
Schizoid
Schizotypal

Cluster B Personality Disorders (dramatic, erratic)

Antisocial
Borderline
Histrionic
Narcissistic

Cluster C Personality Disorders (anxious, fearful)

Avoidant
Dependent
Obsessive-Compulsive

It is not uncommon for the patient to have more than one Axis II diagnoses, and all of these should be reported. Aside from making an actual personality diagnosis, Axis II may also be used to indicate maladaptive personality features that do not meet the threshold for a personality disorder.

Axis III: General Medical Conditions. General Medical Conditions that are possibly relevant to the management of the patient's psychiatric disorder are recorded on Axis III. The broad categories of Axis III diagnoses can be seen in Table 4. General Medical conditions are recorded in an effort to encourage a thorough evaluation and to promote communication among healthcare providers.

Table 4. Axis III: General Medical Condition
Infectious and Parasitic Diseases
Neoplasms
Endocrine, Nutritional, and Metabolic Diseases and Immunity Disorders
Diseases of the Blood and Blood-Forming Organs
Diseases of the Nervous System and Sense Organs
Diseases of the Circulatory System
Diseases of the Respiratory System
Diseases of the Digestive System
Diseases of the Genitourinary System
Complications of Pregnancy, Childbirth, and the Puerperium
Diseases of the Skin and Subcutaneous Tissue
Diseases of the Musculoskeletal System and Connective Tissue
Congenital Abnormalities
Certain Conditions Originating in the Perinatal Period
Symptoms, Signs, and Ill-Defined Conditions
Injury and Poisoning

Axis IV: Psychosocial and Environmental Problems. Axis IV is for recording psychosocial and environmental problems that may impact the diagnosis, treatment, and prognosis of the psychiatric diagnosis or mental disorders (Axis I and II). Psychosocial or environmental problems are generally a negative life event, although they might include a "positive stressor" if it leads to a problem. Psychosocial problems may play an etiologic role in the initiation or exacerbation of a mental disorder, or they may occur as a consequence of the patient's psychopathology. Psychosocial and environmental problems are grouped into various categories and these are listed in Table 5.

Table 5. Axis IV: Psychosocial and Environmental Problems

Problems with the primary support group
Problems related to the social environment
Educational problems
Occupational problems
Housing problems
Economic problems
Problems with access to health care services
Problems related to interaction with the legal system
Other psychosocial and environmental problems

Axis V: Global Assessment of Functioning (GAF).

Axis V is made up of the GAF Scale which is the clinician's judgment of the patient's overall level of functioning. The GAF Scale is a global rating of the patient's psychological, social, and occupational functioning. The GAF instructions guide the clinician to rate the patient with respect only to psychological, social and occupational functioning. The instructions specify, "Do not include impairment in functioning due to physical (or environmental) limitation". The GAF is scored from 1 (most severe) to 100 (highest level of function). A score of 0 is assigned if there is not enough information to make an assessment. A summary of the GAF can be seen in Table 6.

Table 6: Summary of the GAF

91-100	no impairment
71-90	absent or minimal symptoms/impairment
61-70	mild symptoms or impairment
51-60	moderate symptoms or impairment
41-50	severe symptoms or impairment
1-40	pervasive symptoms or impairment
0	inadequate information for rating

The GAF is divided into 10 ranges of functioning and each decile has two components: general descriptions of severity of psychological symptoms and behavioral descriptors of social-occupational functioning. The clinician first rates the patient relative to the decile within which he or she falls. The GAF rating is within a particular decile if either the symptom severity OR the level of function falls within that range. The clinician then decides the exact GAF score from within the decile (tending more toward the adjacent decile above or below). The final score is the most severe condition of psychological symptoms or the social-occupational level of function.

The GAF has psychological and behavioral descriptors for each decile to help the clinician with decision-making and to increase reliability. In these deciles the terms, "Mild", "Moderate", and "Severe" are used. The DSM-IV-TR (2000) provides definitions of these terms as can be seen in Table 7, but these have not been operationally defined.

Table 7: DSM-IV-TR Definitions

Mild- few, if any, symptoms excess of those required to make the diagnosis are present, and symptoms result in no more than minor impairment in social or occupational functioning

Moderate-symptoms or functional impairment between "mild" and "severe" are present

Severe- many symptoms in excess of those required to make the diagnosis, or several symptoms that are particularly severe, are present, or the symptoms result in marked impairment in social and occupational functioning

General categories of function for GAF ratings have also been established both clinically and in the research. These general categorizations can be found in Table 8.

Table 8: General Categories of GAF Ratings
Superior Functioning (91-100) Superior functioning in a wide range of activities, life’s problems never seem to get out of hand, is sought out by others because of his or her many positive qualities. No symptoms.
Minimal Impairment (71-90) 71-80   If symptoms are present, they are transient and expectable reactions to psychosocial stressors with no more than slight functional impairment 81-90   Absent or minimal symptoms with good functioning in all areas
Mild Impairment (61-70) Some mild symptoms (e.g., depressed mood and mild insomnia) OR some difficulty in social, occupational, or school functioning (e.g., occasional truancy, or theft within the household), but generally functioning pretty well, has some meaningful interpersonal relationships
Moderate Impairment (51-60) Moderate symptoms (e.g. flat affect and circumstantial speech, occasional panic attacks) OR moderate difficulty in social, occupational, or school functioning (e.g. few friends, conflicts with peers or co-workers)
Severe Impairment (41-50) Serious symptoms (e.g. suicidal ideation, severe obsessional rituals, frequent shoplifting) OR any serious impairment in social, occupational or school functioning (e.g. no friends, unable to keep a job)

Pervasive Impairment (1-40)

1-30: inability to function is almost all areas (e.g. Danger to self/other; Poor reality testing, unable to care for self/family)

31-40: impairment in reality testing or major impairment in several areas of functioning

To assist with scoring, and to ensure that no elements of the GAF rating are overlooked, a four step scoring method has been included (See Table 9).

Table 9. Four Step GAF Scoring Method

Step 1: Starting at the top level, evaluate each range by asking, "is either the individual's symptom severity OR level of functioning worse than what is indicated in the range description?

Step 2: Keep moving down the scale until the range that best matches the individual's symptom severity OR the level of functioning is reached, whichever is worse.

Step 3: Look at the next lower range as a double-check against having stopped prematurely. This range should be too severe on both symptom severity and level of functioning. If it is, the appropriate range has been reached (continue with step 4). If not, go back to step 2 and continue moving down the scale.

Step 4: To determine the specific GAF rating within the selected 10-point range, consider whether the individual is functioning at the higher or lower end of the 10-point range. For example, consider an individual who hears voices that do not influence his behavior (e.g. someone with long-standing Schizophrenia who accepts his hallucinations as part of his illness). If the voices occur relatively infrequently (once a week or less), a rating of 39 or 40 might be most appropriate. In contrast, if the individual hears voices almost continuously, a rating of 31 or 32 would be more appropriate.

As can be seen, the GAF is based upon taking into account information about severity of psychological symptoms, social-interpersonal functioning, and occupational functioning. These three domains encompass a wide range of assessment data, examples of which can be seen in Table 10.

Table 10: Examples of Psychological, Social, and Occupational Functioning Data

Psychological Symptoms Depression Anxiety Panic Attacks Suicidality Obsessive Hallucinations Delusions
Social and Interpersonal Hobbies Activities of Daily Living Family Relationships Ability to Develop and Maintain Friendships Awareness of Social Norms Communication Skills Follows Rules of Society Personal Hygiene and Self-Care
Occupational Work or School Attendance Work Decisions Organizational Skills Relationships with Co-Workers Ability to Work Independently Ability to Follow Directions Ability to Work Full or Part-Time

According to the DSM-IV, the GAF Scale ratings should, in most instances, be for the current period (e.g. one week). However, there is also the option of providing a GAF for the "current period" and highest level over the previous year.

THE GAF, AMA IMPAIRMENT GUIDE, AND THE SRPD

The California Worker Compensation Reform of 2005 mandated that the AMA Impairment Guides be used to evaluate injured workers. In the Guides, Chapter 14 ("Mental and Behavioral Disorders") addresses psychiatric impairments. Unlike other chapters, "Numerical impairment rating are not included; however, instructions are given for how to assess an individual's abilities to perform activities of daily living" (AMA, 2000, p. 357). However, the Guides provide a method for rating mental impairment (no impairment to extreme impairment) on each of four dimensions (Activities of Daily Living, Social functioning, Concentration, and Adaptation; See page 363). This can be seen in Table 11.

Table 11: AMA Classes of Psychiatric Impairment

Area of Function	Class 1 No Impairment	Class 2 Mild Impairment	Class 3 Moderate Impairment	Class 4 Marked Impairment	Class 5 Extreme Impairment
Activities of Daily Living Social Concentration Adaptation	No Impairment Noted	Impairment levels are compatible with most useful functioning	Impairment levels are compatible with some but not all useful functioning	Impairment levels significantly impede useful functioning	Impairment levels preclude useful functioning

Due to the fact that the AMA Guides do not provide a numerical impairment rating, the GAF is used. This is included in the "Rating Psychiatric Impairment" section of the Schedule for Rating Permanent Disabilities (2005). This specifically includes the GAF Scale as well as the instructions from DSM-IV-TR. The instructions and guidelines for the GAF can be found in Tables 6 and 7 and these are also included in the SRPD. These additional instructions relative to the methodology of using the GAF were included in an attempt to increase its reliability and validity since these psychometric issues have plagued all versions of this instrument.

It is important to note that the terms such as "Mild and Moderate" as used in the AMA Guides are not, in any way, operationally equivalent to the same terms as used in the GAF. Also, I am not aware of any research demonstrating that the five classes of impairment as listed in the AMA Guides (No Impairment to Extreme Impairment) have any relationship to the five general categories of GAF rating (Superior Function or No Impairment to Pervasive Impairment). I have reviewed many QME and AME evaluation reports that attempt to equate these measures which, as far as I know, is not valid. Completing the AMA ratings for each of the four categories may help the practitioner conceptualize impairment for the three dimensions of the GAF, but that is all.

HISTORY OF THE GAF

Having an understanding of the development of the GAF in its current form is important to illustrate some of the problems that have plagued this assessment instrument. Many of the revisions done to the Scale over the years have attempted to address these various problems.

The Health-Sickness Rating Scale (HSRS) was first published by Dr. Luborsky in 1962 (Luborsky, 1962). The HSRS Scale was based on clinical research at the Menninger Foundation beginning in 1949 which sought to develop a standardized method for determining a patient's overall mental health, including level of function. The HSRS is a scale which ranges from 0-100 with scores in the upper range representing minimal psychological symptoms and a high level of function, and scores in the lower range representing a more severe level of psychological symptoms and a diminished level of function. The HSRS produced one global score assessing these two dimensions.

Subsequently, Endicott et al. (1976) developed the Global Assessment Scale (GAS) which is a revision of the HSRS. The GAS has values that range from 1 (the sickest patient) to 100 (a person with no symptoms). The GAS is divided into ten equal intervals with ten values in each interval. Criteria that define each score in each interval are listed. Endicott et al. (1976) performed the first series of reliability studies on the GAS. For the non-statistician reader, reliability concepts are presented in Table 12. The researchers found intra-class correlation coefficients (ICC's) ranging from .61 to .91 with one study showing ICC's of .60, which is moderate at best. The higher values are certainly reasonable, but these were obtained by clinicians trained in the GAS.

Table 12: Reliability Concepts for the Non-Statistician

For those who are not statisticians, reliability is the extent to which a test is repeatable and yields consistent scores. There are many types of reliability including test-retest, alternate form, split half, inter-rater, and internal consistency. Inter-rater reliability or intra-class correlation (ICC) is defined as the agreement between two independent “raters” (or administration of a test) assessing the same variable (e.g. impairment, depression). This is most often expressed as a correlation coefficient such as the Pearson r which can range from 0 (no relationship or agreement) to 1 (perfect agreement).

Take for example a bathroom scale as a measurement tool. You would expect that if you weighed yourself once, and then again very shortly thereafter, the results should be the same (r = 1.0). This is analogous to the same patient being given the GAF by two independent raters. However, what if on the first weighing you were 150 pounds and the on the second done shortly thereafter (no meal in between) you weighed 165 pounds. You would be very concerned about the accuracy of this measure. It is not likely that your weight had changed, rather the measurement tool (bathroom scale or GAF) is flawed. In addition, with this scale you cannot know your “true” weight. Therefore, the scale has no validity. The same holds true for the GAF or any other psychological measure. The poorer the reliability, the less meaningful the results. In addition, the validity of a measure is always limited by its reliability.

Inter-class correlation coefficients (r) for rating scales are considered excellent if greater than 0.74, good if ranging from .60 to .74, and fair if ranging from .40 to .59.

Most of these ratings in the GAS research were completed by a small number of very well trained interviewers. As such, critics have raised questions about the degree to which these results were generalizable to typical clinical settings in which users are not specifically trained in the GAS. Subsequent studies found a high degree of variability in the reliability of the GAS. Dworkin et al. (1990) concluded that specific training of clinicians was necessary in order to maximize inter-rater reliability in situations in which an individual patient may be rated at various times by multiple interviewers. These researchers concluded that the training would help ensure that the differences in the GAS scores were actually due to a change in the patient’s overall level of global functioning (e.g. improvement in response to treatment) rather than simply problems with the reliability of the scale (measurement error).

The DSM-III (1980) did not include any type of GAS or GAF Scale on Axis V. Rather, a patient's overall level of "adaptive functioning" was rated on a 1 (superior) to 7 (grossly impaired) scale. The rating was to include "highest level of adaptive functioning past year."

The DSM-III-R (1987) used a modified version of the GAS, renamed the GAF, as Axis V. The GAF in DSM-III-R was scored on a scale from 1 to 90 and included behavioral descriptors. Data on the basic reliability and validity of the GAF were not provided in the DSM III-R.

In the DSM-IV (1994), the GAF was retained, but modified to a scale from 1 to 100 (a score of 0 indicates inadequate information for assessment). The GAF included behavioral descriptors for the decile ranges. In addition, the instructions for use are essentially the same as what is used currently.

The DSM-IV-TR (Text Revision) was released in 2000. The text revision was published to address some of the problems that were identified in the DSM-IV (First and Pincus, 2002; DSM-IV-TR ,2000). Based on the progress of research in the psychiatric literature, it was not time for a complete revision (e.g. DSM-V) which is reported to be due sometime after 2013. The DSM-IV-TR addressed some of the reported problems with the GAF Scale. According to First and Pincus (2002), there were two problems with the GAF that were identified. One source of confusion was how to operationalize the "current" time frame for the GAF. From the instructions in DSM-IV, it was unclear to the clinician as to the definition of "current" (e.g. specifically during the clinical interview, over the past week, over the past month, etc.). Therefore, in the DSM IV-TR, a sentence was added specifying that the "current period" is sometimes operationalized as the lowest level of functioning for the past week. Clinicians are given the option of recording the time period for which the GAF is completed. None of the DSM's include reliability or validity data for the GAF.

Another source of confusion with the GAF instructions in the DSM IV was the problem of integrating "disparate contributions of a patient's psychiatric symptoms and functioning to the final GAF score" (First and Pincus, 2002, page 291). The following example is given by the authors:

> For example, what should the final GAF score be for a patient who is a significant danger to himself, which would justify a GAF rating below 20, but is otherwise functioning while at work and with his family, reflecting a GAF rating above 60? Some clinicians mistakenly use an

> average of the two, which in this case would result in a GAF score of around 40. In fact, the final GAF score should always reflect the lower of the two ratings. In this case, the GAF score should be below 20, despite the patient's higher social and occupational functioning.

A paragraph relative to this issue was added in DSM IV-TR to clearly outline this convention.

CURRENT USES OF THE GAF

As discussed previously, some authors (Goldman, 1992) specifically stated that the GAF Scale was "not widely used" at that time. In contrast, more recent researchers have concluded that it is currently one of the most widely used psychiatric assessments scales. Clearly, some of its meteoric rise to being one of the most popular psychiatric rating scales has included various systems that have mandated its use. For instance, many managed care provider networks and insurance carriers will utilize the GAF Scale to make treatment determinations including such things as response to psychological treatment interventions, the need for inpatient psychiatric hospitalization, and the need for continued psychiatric hospitalization. In addition, the VA Medical Center system mandated that clinicians use the GAF as part of the diagnostic assessment of all mental health patients (see Moos et al., 2002 for a discussion). Among other things, VA clinicians are required to obtain a GAF rating every 90 days for all of their mental health patients (Niv et al., 2007).

Lastly, with the workers' compensation reform in California (2005), the Schedule for Rating Permanent Disabilities (2005) mandated that "psychiatric impairment shall be evaluated by the physician using the Global Assessment of Function (GAF) Scale shown below. The resultant GAF score shall then be converted to a whole person impairment rating using the GAF conversion Table below" (page 1-12). Given all of these factors, it is no wonder that the GAF has become so popular. After the GAF for the injured worker is determined, it is then converted to a Whole Person Impairment (WPI) and some of these conversion values can be seen in Table 13. After the WPI is determined, the value is modified by diminished Future Earning Capacity (FEC), the Occupational Adjustment, and the Age Adjustment. It is outside the focus of this course to discuss these issues and, typically, calculations beyond the GAF and the WPI are not included in the evaluation report. For detailed information about these other adjustments, please see the SRPD.

Table 13: Conversion of GAF to Whole Person Impairment

GAF	WPI	GAF	WPI	GAF	WPI
70	0	59	17	49	32
69	2	58	18	48	34
68	3	57	20	47	36
67	5	56	21	46	38
66	6	55	23	45	40
65	8	54	24	44	42
64	9	53	26	43	44
63	11	52	27	42	46
62	12	51	29	41	48
61	14	50	30	40	51
60	15			39	53

The GAF, and its predecessors, were actually designed to meet the needs of not only the treating clinician, but also these types of administrative systems. In its original form, Luborsky (1962) sought to develop a standardized assessment that would yield a single global score reflecting a patient's overall level of mental health or psychiatric function. The score was to be a combination of psychological symptoms, as well as social and occupational functioning. The idea was that this would allow for rapid communication relative to a patient's status. Aside from summarizing the various dimensions of function, the scale necessarily needed to be easily administered and not require any special training (ease of use). This would allow for its use across a number of evaluation and treatment environments (e.g. inpatient, outpatient), patient groups (e.g. ranging from adjustment disorders to schizophrenia), and types of clinicians (e.g. psychologists, psychiatrists, social worker, psychiatric nurse, physician, etc.). The idea was to develop a scale that did not require special training, but would have adequate reliability and validity. The GAF has many strengths and these are listed in Table 14.

Table 14: Strengths of the GAF
Easy and quick to administer
Summarizes a great amount of information into one global score
Was designed to require no specialized training
Is widely used
Strong correlation with severity of psychiatric symptoms
Can be used in a variety of settings
Can track changes in a patient’s status

PROBLEMS WITH THE GAF

According to the research, the GAF has three broad problem areas which can be summarized as follows:

(1) Three different dimensions of functioning are collapsed into one composite score
(2) The inter-rater reliability of the GAF
(3) Its validity (specifically predictive validity).

These broad areas will be discussed in detail in the following.

The Global Score Problem

A major problem with the GAF scale is that it integrates three different dimensions of functioning into one composite or total score. These dimensions include psychological symptoms, social and interpersonal functioning, and occupational functioning. A wide range of research suggests that these three dimensions do not necessarily co-vary with each other. In fact, this is reflected in the example in First and Pincus (2002). This is the example of an individual who is a significant danger to himself (GAF rating=20) but is otherwise functioning well at work and with his family (GAF=60). Other similar examples are provided in the GAF scoring instructions in the DSM-IV-TR (2000). This scoring methodology of the GAF does not allow for differentiation amongst these dimensions and, instead, forces the clinician to focus either on psychological symptoms or functioning, whichever is worse. Although many studies have addressed this problem, I will outline the following as examples.

Hilsenroth et.al. (2000) investigated the reliability and convergent and discriminate validity of the GAF Scale. For those non-statisticians who are taking the course, these terms are defined in Table 15.

Table 15: Validity Concepts for the Non-Statistician

Validity is the extent to which a test (or rating scale) is measuring what it is supposed to measure. There are many types of validity including face, construct, criterion (concurrent and predictive), convergent, and discriminant. Those relevant to this course are:

Face validity is the least important aspect of validity and is the extent to which a test appears (on its face) to measure what is intended. In many case, test authors specifically design a test with low face validity so it is less vulnerable to manipulation by the patient. One example is the MMPI-2 in which it is virtually impossible to tell what the questions are measuring. On the other hand, a simple depression questionnaire such as the Beck Depression Inventory (BDI-2) has high face validity. If desired, on the BDI-2 or similar instrument, the patient can present any symptom pattern, whether it is accurate or not

Concurrent validity is the degree to which the test correlates with other expected manifestations of the construct under investigation.

Predictive validity is the degree to which the test predicts an expected outcome or other variable (e.g. an individual's performance, symptoms, response to treatment, etc.).

Convergent validity is the degree to which the test under investigation delivers results that are consistent with other tests of the same or related constructs. This validity also tests whether the measure correlates to some external "criterion" which is often a known "gold standard." One example might be the certain results of a neuropsychological test battery and an MRI.

Discriminant validity is being able to demonstrate that a test doesn't measure what it is not supposed to measure. For instance, a test of depression that correlates highly with an anxiety test will not have good discriminant validity (it cannot discriminate).

In addition to the GAF, the Hilsenroth study also utilized two other important measures, the Global Assessment of Relational Functioning Scale (GARF) and the Social and Occupational Functioning Assessment Scale (SOFA). The GARF is used to indicate an overall judgment of the functioning of a family or other ongoing relationship on a hypothetical continuum ranging from competent, optimal relational functioning, to a disrupted, dysfunctional relationship. The scale relates the degree of relational functioning from optimal to disruptive by using the three major content areas of problem solving, organization, and emotional climate. This assessment is the GAF-equivalent for evaluating the dimension of social functioning. A summary version of the GARF can be seen in Table 16.

Table 16: Summary of the GARF

81-100. Overall the relational unit is functioning satisfactorily from self-report of participants and from perspective of observers.
61-80. Overall the functioning of the relational unit is somewhat unsatisfactory. Over a period of time, many but not all difficulties are resolved without complaints.
41-60. Overall the relational unit has occasional time of satisfying and competent functioning together, but clearly dysfunctional, unsatisfying relationships tend to predominate.
21-40. Overall the relational unit is obviously and seriously dysfunctional; forms and time periods of satisfactory relating are rare.
1-20. Overall the relational unit has become too dysfunctional to retain continuity of contact and attachment.
0 -Inadequate information.

The SOFA is designed to asses an individual's level of social and occupational functioning not directly influenced by the overall severity of psychiatric symptoms. The SOFA also considers the effects of the individual's general medical condition in the evaluation of social and occupational functioning. Analogous to the GAF, this scale is thought to more independently assess a patient's level of occupational function and attempts to partial out the impact of psychiatric symptoms (See Table 17). As such, Hilsenroth et al. (2000) were using measurements that would ostensibly assess the three dimensions that are collapsed in the single GAF global score. They were also choosing the instruments based on the fact that the GAF rating has been consistently found to correlate most highly with a patient's severity of psychiatric symptoms rather than the other two dimensions. The GARF and the SOFA are actually included in the DSM-IV as experimental measures.

Table 17: Summary of the SOFA
91-100. Superior functioning in a wide range of activities.
81-90. Good function in all areas, occupationally and socially effective.
71-80. No more than slight impairment in social, occupational, or school functioning.
61-70. Some difficulty in social, occupational, or school functioning, but generally functioning well, has some meaningful interpersonal relationships.
51-60. Moderate difficulty in social, occupational, or school functioning.
41-50. Serious impairment in social, occupational, or school functioning.
31-40. Major impairment in several areas, such as work or school, family relations.
21-30. Inability to function in almost all areas.

11-20. Occasionally fails to maintain minimal personal hygiene; unable to function independently.

1-10. Persistent inability to maintain minimal personal hygiene. Unable to function without harming self or others or without considerable external support.

Hilsenroth et al. (2000) evaluated 44 patients admitted to a university based outpatient community clinic utilizing the GAF, GARF, and SOFA. The study participants carried a variety of diagnoses in categories of mood, anxiety, substance-related, and adjustment. In the study, each patient completed a videotaped semi-structured clinical interview. After the interview was completed, the clinician rated the patient on the three scales. Prior to participating in the study, the clinicians participated in both individual and group training sessions on scoring the three scales (GAF, GARF, and SOFA). After the first clinician completed the videotaped interview and the scale ratings, a second clinician viewed the videotape and independently rated the patient on the three scales. The second, external rater was unaware of the patient's diagnosis, self-report data, and first clinician's ratings for the three scales. The study design allowed for assessing such things as inter-rater reliability for each scale as well as the correlation between one scale and another.

The mean scores for the three scales can be found in Table 18. As can be seen, the average GAF score is about one would expect for this patient population. Again, for those non-statisticians, an example of two standard deviations around the mean for the GAF would be as follows: a mean of 64.5 plus or minus 14 (two SD's) includes 98% of all the patients' scores.

Table 18: Mean GAF, GARF, and SOFA Scores

GAF	64.5 (SD 7.1)
GARF	62.6 (SD 15.1)
SOFA	62.6 (SD 10.7)

Table 19 shows the correlation between the first clinician's ratings and the ratings obtained by the "external rater" who subsequently viewed the videotape (second rating). These values constitute inter-rater reliabilities and are considered very reasonable for a psychometric instrument. Again, these results are consistent with other findings relative to the GAF that suggest that one can achieve high inter-rater reliabilities when clinicians are specifically trained to the instrument. This was also found for the GARF and the SOFA.

Table 19: Correlations Between Clinical and Videotape Ratings

GAF	r = .86
GARF	r = .85
SOFA	r = .89

Convergent and discriminate validity (See Table 15 for definitions) were assessed using multiple statistical methods including factor analysis, correlations amongst the scales, and correlations with other measures including the SCL-90-R Global Severity Index (a measure of psychological distress), the Social Adjustment Scale (SAS) Global Score (a measure of social impairment), and the Inventory of Interpersonal Problems (IIP) Total Score (a measure of interpersonal functioning). The relationship of the GAF to the GARF was significant (r = 0.60, P<0.0001) as was its relationship to the SOFA (r = .60, P<0.0001). However, the SOFA and the GARF demonstrated a low correlation (r = 0.34, P=0.02.) As concluded by the authors, the results of the study suggest that the GARF and the SOFA are each more related to the GAF individually than they are to each other. This suggests that these two scales are evaluating something different from one another and to a lesser extent from the GAF. Conceptually, this makes sense given the fact that the GAF is a composite of dimensions specifically measured by the GARF, SOFA, and severity of psychological symptoms.

To further assess convergent and discriminate validity, the three scales were also correlated with other measures as discussed previously. The results of some of these analyses are presented in Table 20. Significant correlations (p<.01) are marked (**). As can be seen, consistent with other studies, the GAF correlated with severity of psychological symptoms or distress. The SOFA was significantly correlated to the Social Adjustment Scale and

Inventory of Interpersonal Problems (the negative correlation result is due to how the results are expressed numerically).

Table 20: Hilsenroth Study Results (correlations)

	Global Severity Index (psychological)	Social Adjustment Scale (social impairment)	Interpersonal Problems (relationship functioning)
GAF	**r = .46 ****	r = -.31	r = -.16
GARF	r = -.16	r = -.24	r = -.06
SOFA	r = -.37	**r = -.47 ****	**r = -.46 ****

The data include the finding that the GAF Scale showed the largest significant relationship to a patient's report of psychiatric symptoms (SCL-90-R Global Severity Index), but did not show a specific significant association with social impairment (SAS Global Score) or interpersonal impairment (IIP Total Score). This is consistent with previous research that has demonstrated that the GAF Global Score is most often highly correlated with the patient's severity of psychiatric symptoms and not generally correlated with social and/or occupational functioning.

In a second study, Hay et al. (2003) evaluated the predictive validity of the GAF, the GARF, and the SOFA. In this study, a total of 97 psychiatric patients were followed for up to two years to evaluate outcome and contrast the validity of the GAF, SOFA, and GARF. Results demonstrated that the SOFA and the GAF scores on psychiatric admission were significantly negatively correlated with duration of hospital admission. These results make sense given the fact that a patient who is functioning at a higher level (SOFA) and showing less psychiatric severity (GAF) would show a shorter hospital stay (note: the reason for the negative correlation is that numerically higher SOFAS and GAF scores represent better function and mental health). The researchers also found that the SOFA ratings at psychiatric hospital discharge were significantly and negatively correlated with overall psychiatric outcome at two year follow-up. They conclude that the SOFA (a measure of adaptive functioning) had better predictive and concurrent validity than the GAF or the GARF.

A large-scale study of the GAF was conducted by Moos et al. (2002). This study was done using the VA Medical Center network data set since routine GAF evaluations are required. The researchers obtained GAF results used to assess global functioning for 9854 patients with psychiatric or substance abuse disorders, or both. These assessments were done by clinicians across 148 VA facilities. The clinicians were experienced mental health professionals who routinely utilize the GAF within the context of routine clinical diagnostic interviews. In the study, patients were classified according to five categories of GAF scores, and these can be seen in Table 21. Scores from 91-100 are indicative of no symptoms at all. Statistical analysis was completed across the five groups of patients based on these GAF categorizations. In addition to the GAF, the researchers also collected demographic data as well as data related to receipt of services such as inpatient or residential care (number of days), outpatient care (number of days), as well as data on the patient's symptoms and social and occupational functioning. Data analysis also included information about alcohol and drug use.

Table 21: GAF Combined Scores (Moos, 2002)

71-90	Minimal Impairment
61-70	Mild Impairment
51-60	Moderate Impairment
41-50	Serious Impairment
1-40	Pervasive Impairment

Multiple regression analysis was utilized to identify the best independent predictors of GAF ratings. As discussed by the researchers (page 733), "when entered first in the regressions, the social or occupational functioning indexes accounted for only 1% of the variance in GAF ratings. Patient's psychiatric diagnoses, previous inpatient care, psychiatric symptoms, substance use, and substance related problems were each significantly associated with higher levels of global impairment." The researchers go on to state that "after these variables were entered, employment status was the only social or occupational index that independently predicted global functioning and it accounted for less than 1% of the variance in clinician's GAF ratings." The researchers found the same results for the continuous GAF as opposed to the categorized GAF ratings. The authors conclude that, "however, in this study, clinician's rating of global impairment were more closely associated with patient's diagnoses, previous treatment, and severity of symptoms than with their social or occupational functioning" (page 735). They also conclude that "our findings and the results of these studies indicate that GAF rating provide little or no information about social or occupational functioning that is independent of clinician's judgment about diagnoses and the severity of symptoms" (page 735). They also found that the GAF ratings were not predictive of treatment outcomes.

These studies underscore the conclusion that the global score generated by the GAF Scale is problematic in that it attempts to incorporate three different dimensions of function including severity of psychological symptoms, social and interpersonal functioning, and occupational functioning. The research demonstrates that the GAF is primarily tapping into the severity of psychological symptoms. In many ways, this is not surprising since the majority of the data available to the clinician relative to the GAF assessment is likely to be related to the severity of psychological symptoms. Severity of psychological symptoms is most often assessed through the clinical interview, mental status examination, and objective psychological testing. Gathering detailed and objective data on the other dimensions (social and occupational) is much more difficult. As such, most clinicians likely develop their GAF ratings based on the clinical data related to severity of psychological symptoms. Researchers are clearly aware of this problem and the DSM-IV has included the two experimental measures which attempt to assess both social and occupational function. As will be discussed in more detail subsequently, it is important for the clinician to be aware of these potential problems with the GAF and the over reliance and focus on psychological symptoms.

Reliability of the GAF

Interviewer rater scales such as the GAF and its predecessors are notoriously vulnerable to problems of low inter-rater reliability because they tend not to operationally define terms and are used by examiners with different levels of training and experience. This is a significant issue since scores for the scales need to be comparable and meaningful across situations (e.g. treatment programs, research projects, etc.) for the scale to have any value. Variability can occur for many reasons including that some raters may have a propensity to make high GAF ratings, whereas others may have a tendency to make low ratings. As we discussed previously reliability is critical for a scale or test to have any use or meaning (See Table 12). In fact, the validity of the measure is limited by its reliability both from a mathematical and conceptual standpoint. All of the various versions of the GAF (beginning with the HSRS, and including the GAS, GAF in DSM-IV and the current GAF in DSM-IV-TR with modified instructions) have been plagued by problems with inter-rater reliability. In fact, the GAF instructions have become more and more structured in an effort to address this problem.

Adequate inter-rater reliability can be achieved for the GAF if the clinicians undergo structured training. This generally involves the presentation of clinical vignettes with a subsequent review of the rationale behind why a specific GAF was assigned. However, the GAF was designed to be quick, easy to use, and without specific training. Therefore, in clinical practice, very few practitioners have actually undergone the type of training that is common in the published research articles. This is nicely exemplified in a study completed by Bates et al. (2002). The researchers examined the impact of a brief training program on clinicians using the GAF. In the study, 31 staff members within one VA Medical Center were presented two vignettes without any training in the GAF scale. The clinicians were asked to provide a GAF rating for each of the vignettes. Subsequently, they participated in a brief training session aimed at increasing the reliability of GAF assessments. After the training was completed, the clinicians again rated the two vignettes using the GAF. This allowed the researchers to compare the results for pre and post training. The investigators were really interested in two issues: (1) What were the inter-rater correlations pre-training versus post-training; and (2) What "strategies" were clinicians using to arrive at the GAF rating prior to their training.

Table 22: GAF Strategies Used Before Training

Highest area of functioning
Lowest area of functioning
Average of areas of functioning
Least severe of symptoms
Worst severe of symptoms
Average of symptoms
Least severe of either symptom or function
Worst of either symptom or function **
Average of symptoms and function

The nine strategies used by the clinicians prior to GAF training can be found in Table 22. As can be seen, the clinicians utilized a number of different strategies, only one of which is consistent with the scoring instructions for the GAF (**). As can be seen in Table 23, incorrect strategies were used 90% of the time and the most common strategy was to average the severity of symptoms and functioning level. The correct strategy was utilized less than 10% of the time by untrained clinicians. Inter-rater reliability was consistent with previous research in showing poor results prior to training and improved results after training.

Table 23: Effect of Training on GAF Strategies
Before Training Incorrect strategies were used 90% of the time The most common strategy was to average the severity of symptoms and level of function The correct strategy was only used 9.7% of the time
After Training The correct strategy was used 64% of the time Incorrect strategies were still used 36% of the time.

Table 23 also demonstrates the effect of the training on clinicians' GAF scores. Even after training, only 64% of clinicians used the correct strategy (of course, this means 36% continued to use incorrect strategies). After training, the obtained GAF scores were significantly different from what was obtained prior to the training. The post-training GAF scores were much closer to the "criteria" scores. The authors conclude that "the study highlights common errors and points to the need for formal training in the use of the scale".

In a more recent study, Vatnaland et al. (2006) asked the question, "are GAF scores reliable in routine clinical use?" In their introduction and review of the literature, the authors make the point that the GAF scale has been considered as a reliable tool, but most of the studies of GAF reliability have been based on special conditions including prior training, test awareness, and under strictly controlled research conditions. The study sought to assess the reliability of the GAF as it might be commonly used in a clinical situation.

In the review of existing research on GAF inter-rater reliability, the authors determined that all but two of the studies concluded that the reliability is "excellent" as measured by intra-class correlations (ICC greater than 0.74). However, they point out two main problems with this body of research:

Raters often undergo prior calibration and training, and are generally selected from dedicated students or researchers. The authors conclude that the positive results may not be generalizable to other settings and may reflect what has been termed "within-center inter-rater reliability."

The authors also point out that most studies report results from clinician-raters who are highly aware that their GAF scores are being monitored. Such test awareness interacts with the rating process.

The study sought to determine the inter-rater reliability of the GAF as used in routine clinical practice where none of the above conditions exist. In the study, 100 consecutive psychiatric admissions were assessed and assigned a GAF score by three different raters, both at admission and discharge. The same individual staff member did not necessarily obtain GAF scores at admission and discharge. The clinicians did not utilize any type of structured interview guide or other tools in the process of assessing the GAF scores. In addition, formal training in the use of the GAF Scale had not taken place. For each patient, a "criteria" GAF was also established by two psychiatrists trained in the use of the GAF. These criteria GAF scores were determined for both admission and discharge status.

Table 24: GAF Clinical Correlations

	Admission	Discharge
Clinicians and Expert #1	r = .39	r = .56
Clinicians and Expert #2	r = .39	r = .59
Expert #1 and Expert #2	r = .81	r = .85

Table 24 shows the inter-rater reliability between the non-trained clinicians and psychiatrist number one (admission and discharge), non-trained clinicians and psychiatrist number two (admission and discharge), and the relationship between the GAF scores determined by the two experienced psychiatrists. As can be seen, the correlations amongst the clinicians and the criteria scores were poor. The inter-rater reliabilities between the two experienced psychiatrists were quite high and consistent with previous research investigating clinicians who have been trained in the use of the GAF. The authors conclude:

> "the results reported above suggested there are critical issues concerning the reliability of the GAF when applied in a realistic clinical context. ICC coefficients between scores obtained by standard department procedures and those by the two research raters at admission were 0.39 and clearly inadequate. In terms of standard practice, this level is less than what would be accepted as a fair agreement beyond chance. This indicates that only about 40% of rater differences reflect real differences in subject conditions. This study again underscores that the likelihood that GAF ratings in real world settings tend to have a fairly low level of reliability."

These issues raise concern about the use of the GAF in the workers' compensation system. As discussed previously, the GAF rating is converted to a whole person impairment and is used to determine permanent psychiatric disability. Therefore, inter-rater reliability as well as agreement with the criteria rating ("true" GAF value) is critical. It is important for clinicians completing GAF ratings within the workers' compensation system to be aware of these reliability problems with the GAF especially when used by practitioners outside of a research setting and who have not undergone specialized training. Clearly, this represents the vast majority of those who are providing GAF assessments of injured workers. It also underscores the complexity of assigning a GAF score. It certainly goes beyond choosing a number from the scale based on clinical instincts that almost always reflect severity of psychological symptoms to the exclusion of the other dimensions. The last section of this course will review methods for improving the clinician's use of the GAF both in terms of reliability and validity.

Validity of the GAF

As can be seen in Table 15, there are various types of validity when discussing psychometric testing. In general, validity refers to the extent to which a test for scale measures what it is supposed to measure. Conceptually, this is somewhat difficult to determine relative to the GAF since it is a composite of three dimensions including psychological symptom severity, social and interpersonal functioning, and occupational functioning. The research reviewed did not demonstrate that a construct of what the GAF is "suppose to measure" has been established. Therefore, most studies investigating the validity of the GAF look at such things such as,

> Does it correlate with what we would expect it to correlate with (convergent validity).
>
> Does it not correlate with measures we would expect it not to correlate with (divergent or discriminate validity).
>
> Does it predict what we would expect it to predict (predictive validity).

Probably the largest study of predictive validity of the GAF has been carried out by Moos et al. (2000, 2002) using a VA Medical Center database. As discussed previously, all VA mental health clinicians are required to utilize the GAF when treating patients in this system. This provides a valuable set of data that is amenable to investigating the usefulness of the GAF. In one study (2002), the researchers analyzed the GAF and other data for 9854 patients with psychiatric or substance abuse disorders or both. Since the use of the GAF was mandated by the VA Medical System, scores were available at multiple time points including entry into treatment and follow-up assessments six to 12 months later. In addition, other measures of psychiatric symptoms, as well as social and occupational functioning were contained within the data set. In analyzing the GAF scores, Moos utilized a method common in previous research in which the ratings are combined into fewer categories. The researchers divided the GAF into five categories and these were presented previously in Table 21.

Consistent with previous research, the authors concluded "however, in this study, clinicians ratings of global impairment were more closely associated with patient's diagnoses, previous treatment, and severity of symptoms then with their social or occupational functioning" (p. 735). Again, the research consistently demonstrates that the GAF is primarily a measure of severity of psychological symptoms to the exclusion of social and occupational functioning. They go on to state that "once these clinical and symptom-related factors are considered, indexes of social and occupational functioning made only negligible contributions to the GAF ratings" (p. 735). They go on to state that "our findings and the results of these studies indicate that GAF ratings provide little or no information about social or occupational functioning that is independent of clinician's judgment about diagnosis and the severity of symptoms" (p. 735).

Relative to the GAF predicting treatment outcome, the authors conclude that, "moreover, we found little or no relationship between GAF ratings and either symptom outcomes or social or occupational outcomes. This result was the same when we used the continuous GAF scores for the five categories of GAF scores" (p. 735). The authors go on to state that this finding was a replication of a previous study completed which found minimal associations between clinicians' ratings of patient's current level of function and patient's self-rated symptoms and functioning at follow-up (Moos et al., 2000). They conclude that "in conjunction with the lack of previous positive findings that link GAF ratings to outcomes, these findings cast doubt on the value of including GAF ratings as predictors of treatment outcome in an outcomes monitoring system" (p. 735). They state that "although intuitively appealing, a brief uni-dimensional rating of global functioning cannot capture changes in psychological, social, and occupational functioning that are only moderately inter-related at best."

These results are important to keep in mind when utilizing the GAF in the workers' compensation system. The Moos et al. studies (2000, 2002) suggest that a global rating such as the GAF is not predictive of such things as social or occupational functioning. In fact, research has demonstrated that, in all except for the most severe of psychiatric cases, psychological measures are not predictive of future occupational functioning. Certainly, for someone who is rated at the very high level of psychiatric impairment based on psychological symptoms (e.g. a very low GAF in the 1-20 range), prediction of occupational and social functioning in the future is relatively straight forward. However, when one is faced with a heterogeneous patient population showing a variety of symptoms, the predictive power of the GAF and psychological tests seem to fall apart. This problem is echoed by MacDonald-Wilson et al. (2001) who state that "adding to the difficulties in assessing work function amongst individuals with a psychiatric disability is a question of whether and how well psychiatric diagnoses and symptoms (i.e. psychiatric impairment) can predict work capacity or functioning" (p. 221). The authors go on to state that "a variety of early studies reported little relationship between future work performance and various assessments of psychiatric symptoms. From these studies, there appear to be no symptoms or symptom patterns that were consistently related to work performance" (p. 222). In reviewing the research, they did state that several long term follow-up studies suggest that psychotic-like features and symptoms were associated with poor role functioning and less likelihood of being employed. This is consistent with what we discussed earlier in that the GAF is likely predictive of future psychiatric disability if a patient is scoring in the very low ranges (high psychiatric impairment). MacDonald-Wilson et al. (2001) concludes that "taken together, these studies do not support the conclusion that psychiatric diagnosis alone is a good predictor of vocational capacity. While there is sufficient evidence to suggest that a diagnosis of a psychotic disorder is associated with somewhat poor vocational outcomes, these relationships appear modest. Similar conclusions can be drawn about symptoms. Psychiatric symptoms, unless severe, bear a small relationship to vocational functioning.

IMPROVING THE USE OF THE GAF

The use of the GAF in determining permanent psychiatric impairment is mandated in the California Workers' Compensation System. As discussed previously, the GAF is converted to a whole person impairment, according to the Schedule for Rating Permanent Disabilities (2005). Therefore, reliable and valid use of the GAF is certainly important. As we have seen, there are many problems with the GAF Scale.

Use of the GAF in a Forensic Setting

If you function as a QME, you are performing evaluations and rendering opinions within the context of a forensic setting. By nature, this is often an adversarial environment (applicant and defense). Therefore, your conclusions must be defensible and this includes the GAF determination. Having an understanding of the problems inherent in the GAF can help the practitioner produce reliable, valid, and defensible GAF determinations.

As we have reviewed previously, there has been a myriad of research using the GAF in psychiatric inpatient settings, university based clinics, and the VA Medical Center system. I could not locate any articles that address its use within a workers' compensation system. Whenever one is functioning within a workers' compensation system, factors such as patient credibility and the validity of self-reported information must be taken into account and objectively assessed. The GAF contains no validity scales and is largely based on self-report data from the patient. Its use within a forensic setting is therefore problematic due to its very nature. This also makes it more difficult to defend one's conclusion relative to this instrument. However, there are ways in which the use of the GAF can be improved by the individual practitioner as well as making the GAF conclusions more defensible. These will be reviewed in the following section.

Use the GAF According to the Instructions

This issue seems almost ridiculous to mention but it is important to remember to use the GAF according to the instructions. As we discovered in the review of research literature relative to the GAF, one of the most common problems is that clinicians simply do not follow the instructions. This was poignantly highlighted in the study of Bates et al. (2002). As you will recall, nine strategies were used by untrained clinicians to achieve their GAF ratings and of these, only one was consistent with following the directions for the GAF Scale. Even after training, only 64% of clinicians followed the directions. As the researchers pointed out, the most common strategy was to average the severity of symptoms and functioning level. Although I am speculating, this may be due to a clinician's inclination to incorporate all of the available data into one global measure rather than choosing one dimension over another. Regardless of the reasons, this strategy is incorrect, according to the instructions. Interestingly, these researchers also found that 36% of the clinicians continued to use incorrect strategies even after structured training on the GAF.

If this study demonstrated that untrained clinicians were using incorrect strategies 90% of the time, the base rate is likely equal to that or higher in routine use (since the clinicians in the study knew they were being monitored). The frequency of incorrect use in a forensic setting, such as completing a QME evaluation, is simply unknown. However, I believe we can assume that it is likely quite high. Therefore, the first strategy in improving the quality of a GAF rating is to simply follow the directions as outlined in the DSM-IV-TR and the SRPD. The four step process as outlined in these publications (See Table 9), as well as the enhanced instructions in DSM-IV-TR, will certainly help with this process. As can be seen in the four step process, the clinician is to choose a GAF level that reflects either the individual's symptom severity OR level of functioning, whichever is worse.

Another component of the GAF instructions is that the clinician is not to include impairment due to physical (or environmental) limitations. Many researchers have questioned this aspect of the GAF since they do not believe that physical impairment can accurately be separated from psychiatric impairment (it does appear to represent mind-body dualism). Even so, according to the GAF instructions, one must attempt to do so as diligently as possible. I could find no research or other literature that provides the clinician with any empirical guidance as to how this is to be accomplished aside from a completely subjective assessment (a clinical "guesstamate").

The "Split Method"

To improve the quality of your GAF score you may want to consider using the "split" method as initially suggested by First (1995). Dr. First has recommended that the clinician should "treat the GAF as it were two scales: one for symptom severity and another for level of function. Then… make one rating for severity and a second for level of functioning. The worst of the two can be used as the GAF" (p. 259). It has been suggested that this is an especially useful approach when there is some discrepancy between a patient's symptomatology and level of functioning (e.g. a patient with psychotic symptoms who nevertheless functions fairly independently). The research has also suggested that this rating rule can help to counteract an apparent tendency to try and somehow either combine symptomatology and functioning or take an average of the two. Certainly, this approach is useful in being a structured reminder to the clinician to assess both symptomatology and function, then choose the worst of the two. However, it does continue to combine both social and occupational functioning into one variable, and these certainly may co-vary differently. However, this is consistent with the directions of the GAF and may be a problem inherent in this scale.

The MIRECC GAF Scale

The VA Mental Illness Research, Education, and Clinical Centers (MIRECC) has developed a modified GAF scale that independently assesses all three components of the traditional GAF (psychological, social, and occupational). As discussed by the authors, the DSM-IV-TR directs raters to base the GAF score on the worst functioning of these three domains (I believe it's actually two domains, symptom severity versus function). As such, the GAF score typically represents one dimension, and clinicians do not know which dimension is represented or how the patient fairs on other dimensions, rendering the GAF limited in its utility (Niv et al. 2007, p. 529). Niv (2007) developed the MIRECC GAF which has behavioral descriptors for each of the three domains of function. An overview of the rating system can be found in Table 26 and a copy of the actual scale can be found here.

Table 26: MIRECC GAF Overview

Group and GAF Range	Occupational Function	Social Function	Symptomatic Function
Fully Functional (71-100)	Works consistently Cares for children consistently Attends school consistently	Superior functioning Socially effective Slight impairment	None Very Minimal Symptoms in reaction to stressors (1 to 2 days maximum).

Borderline Functional (51-70)	Misses work fairly frequently Inconsistently able to attend to child care Misses school frequently	Frequent interpersonal conflicts or withdrawal, but still able to maintain some meaningful interpersonal relationships	Mild (e.g. persistent and mild depressed mood) Moderate (moderate depression, occasional panic attacks, flat affect)
Dsyfunctional (21-50)	Works consistently in sheltered work Intermittent work in sheltered work No work activities	Able to have coherent conversation Some difficulty with conversation Serious difficulty sustaining a coherent conversation	Serious (e.g. suicidal thoughts, severe obsessions, frequent intoxication) Impairment in reality testing or communication Behavior is influenced by delusion or hallucinations; serious impairment in communication or judgment
Dangerousness (1-20)	Not able to provide for his/her own food or clothing	Only able to interact with other people for a brief period of time	Some dangerousness to self or others; gross communication impairment Persistent and

			imminent danger of hurting self or others

The MIRECC GAF has been found to have excellent inter-rater reliability for all three of the subscales (r = .98 to .99) when used by specifically trained clinicians. Analysis of convergent and discriminate validity were also completed. The authors concluded that "results demonstrated good convergent and discriminate validity, with 40% of the variance accounted for by work and school status". In other words, the three dimensions correlated with external measures in the expected pattern (e.g. occupational GAF rating with occupational variables, but not with other psychological variables, etc.).

In addition, the researchers investigated predictive validity of the instrument. Outcome measures included such things as work status, presence of family support, presence of a close friend, and psychological symptoms. The authors stated that "in terms of predicting work status at follow-up, the MIRECC GAF occupational ratings were significantly predictive, whereas MIRECC GAF social and GAF symptoms ratings were not" (Niv et al, 2007, p. 533). Similarly, the MIRECC GAF social scores were significant predictors of follow-up family and social support. The occupational and psychological symptom scales were not predictive of family support or a close friend at follow-up.

In even more detail, this study underscores the tri-dimensional nature of the GAF global score. The conceptualization of the MIRECC GAF goes beyond the instructions for the traditional GAF by splitting apart social and occupational functioning. Even so, being aware of these data can help the clinician think in terms of the various dimensions represented in the GAF global score and taking care to assess these carefully and independently to arrive at a valid conclusion.

Quality of the Information upon which the GAF Rating is Based

One of the critical factors in completing psychological evaluations (including impairment) within a forensic setting is firmly establishing the credibility of the patient as well as relying as much as possible on objective data. As the old saying goes, "Garbage In-Garbage Out". If the quality of the data upon which a GAF rating is based is unknown, there is the risk that it is unreliable, biased and invalid; as such, the quality of the GAF rating cannot be established. The GAF rating is only as good as the data upon which it is based.

It is beyond the scope of this course to review all of the research literature and approaches to establishing patient credibility relative to psychological assessment and the reader is referred elsewhere (Rogers, 2008). The importance of establishing the credibility of the patient is due to the fact that the GAF rating is often based primarily on self-report data or questionnaires that are highly face valid. If one can establish the credibility of the patient, then all of the other data can certainly be determined to be of higher value. Often, credibility is established through the use of standardized psychological testing that includes sophisticated validity scales. Of course, the gold standard in this arena is the MMPI-2. Often, clinicians will add other tests of credibility or symptom embellishment especially in the realm of cognitive functioning. Whatever method is chosen, the clinician should utilize a standardized approach in an attempt to establish the patient's credibility of self-report and response to face valid instruments. If the patient shows "symptom embellishment" or amplification, then this must be taken into account when determining the GAF. This issue of patient motivation is also discussed in the AMA Impairment Guides (p. 358).

Once credibility is established, the clinician can improve his or her use of the GAF by purposely assessing all three of the dimensions that make up the global score. In routine QME clinical practice, it appears that this is rarely done. Certainly, data related to severity of psychological symptoms is readily available. Therefore, the issue is to increase the focus on assessment of social and occupational functioning. This might be done through the use of an instrument such as the MIRECC GAF or occupational and social functioning might be completed through the use of brief questionnaires that are readily available. Some examples of these questionnaires are outlined in Chapter 1 of the AMA Impairment Guides and other examples can be found throughout the research we have reviewed previously.

Review GAF Examples and Vignettes Presented in the Research Literature

Another method for improving one's quality of GAF rating is to review the literature for example GAF ratings based on various clinical samples and vignettes. The vast majority of these articles have used clinicians that were specially trained in the use of the GAF. By having an understanding of what type of clinical presentation represents a criterion GAF score, the user can develop higher quality scores even without the benefit of structured training.

One method for accomplishing this is to look at some of the training vignettes that have been used by the VAMC along with the criteria GAF scores assigned to these examples. Some of these training vignettes are presented in Table 27 along with the GAF score assigned to each case. I could not locate example vignettes in the higher levels of functioning (e.g. greater than the decile of 50-60).

Table 27: VAMC Example GAF Training Vignettes

GAF = 20

Mr. A is a single veteran in his mid-60's who was admitted to the IPCC program about 2 years ago following lengthy hospital stays and repeated failures to adjust to residential care placements. One residential care sponsor described him as being very dependent and requiring constant supervision and attention. Mr. A was re-hospitalized in 1992 when he assaulted a residential care sponsor after she told him to take a shower because he had been incontinent of feces. During that admission he developed somatic delusions about having cancer, his ADLs continued to be very poor, and he began to smear feces. With IPCC support, he was discharged in Nov. 1995 and placed in a rest home because he required a high level of care and supervision of his ADLs. Mr. A attends activities at the community-based IPCC day program three times a week. His speech is tangential and irrelevant at times, but he will usually cooperate if given explicit directions. Due to the fact that Mr. A has been able to live outside the hospital for the past 2 years, he remains very dependent on the home and IPCC staff for all his needs. His smearing of feces continues to be a problem.

GAF = 22

Mr. B is a veteran in his early 40's with a diagnosis of schizoaffective disorder who has had multiple psychiatric admissions, including 11 in the past two years for severe delusions (e.g., he believes he is a character from the Little Rascals, that he has a girlfriend who is a famous TV actress, and that he owns seven businesses). He has extreme money management skills problems (e.g., he is unable to purchase food and has been evicted for failure to pay his rent). He has a history of giving large sums of money away (e.g. $500-$1,000 at a time) to people (often drug users) he just met off streets and considers his "friends" for religious reasons. He refuses to

participate in the community-based IPCC day programs and prefers to stay in bed much of the day. He is noncompliant with taking psychotropic medications and attending outpatient appointments at the hospital. Recently, a conservator was appointed to handle his funds.
GAF = 25 Mr. C is a single veteran in his late 50’s who was admitted to the IPCC program on July 1, 1996 after thirty years of continuous hospitalization at the VAMC. He now lives in a residential care home and attends the IPCC community based day program five days a week. He needs lots of prompting to attend to ADLs and uses an assisted transport system to attend the day program. He hoards large amounts of money on his person and refuses to open a bank account because he believes "the banks are controlled by the Mafia." He continues to express bizarre delusions of a religious grandiose nature, believing that he is the Son of God and that he has the ability to communicate with animals and extraterrestrial beings.
GAF= 28 Mr. D is a veteran in his mid 40’s who was discharged two years ago after 18 years of hospitalization. He resides in a residential care home and attends IPCC community-based day programming 5 days per week. He is in a structured money management program and receives concrete rewards for completion of daily ADLs. He is able to negotiate the city transportation system and is very capable of accessing community resources. He is extremely delusional, believing his body is made out of glass or wood and that many people are his mother (including the queen mother in England). He is also very thought-disordered and tangential; suspected brain damage contributes to speech oddities.
GAF = 32 Mr. E is a single veteran in his mid 50’s who has a long history of schizoaffective illness characterized by frequent mood swings. When in a manic phase, he exercises very poor judgment regarding financial matters and his behavior is very inappropriate. When depressed he becomes sullen and withdrawn. He also experiences auditory hallucinations. He has had numerous and lengthy hospitalizations for his illness, although far fewer and much shorter lengths of stay since entry into the IPCC program in 1989. He has not been able to hold a job in many years. His personal hygiene is quite

poor. He resides in a boarding home where he receives assistance with his medications.

GAF = 35

Mr. F is a veteran in his early 50’s who has had multiple admissions to the hospital. Two admissions during the past were for an increase in auditory hallucinations that were telling him to harm himself and others. He continues to hear voices but is able to separate out that they are not real and he will not act on them. In some areas he functions independently; for example, he showers and changes clothes daily, attends mass daily, and has a group of church friends with whom he has coffee. He is compliant with appointments and medications (even though he insists that religion is better than the meds), and gets along well with the other day program members and with the residents at his residential care home. At the community-based IPCC day program, he follows through with his work assignment and attends group where he raises pertinent issues about community living and relates his personal experiences. However, he is very religiously preoccupied and has delusions about electricity, telephones and vehicles that are impairing (e.g., because of his delusions he will not ride in cars). As a result, he remains dependent on the program staff for some essential services.

GAF = 45

Mr. G is a veteran in his early 50’s who has had multiple hospitalizations due to non-compliance with medications. When he was living in his own condo he would stop taking meds and become very paranoid, with hallucinations. He is now in an apartment within a residential care home where meals are provided and meds are supervised. He continues to experience some symptoms of paranoia and will have an occasional hallucination in which someone is calling his name. However, he drives his own car and attends the IPCC community-based day program, where he is in charge of collecting lunch monies and keeping data for the program's point-based reward system. He usually keeps to his own at the day program and what conversations he has tend to be short. He visits his family twice a month.

GAF = 52

Mr. H is a divorced in-country Vietnam Veteran in his late 40’s with a dual diagnosis of schizoaffective disorder and alcoholism; although, he has been abstinent from alcohol for several years and has not been overtly psychotic since entry into the IPCC program in 1990. He becomes very anxious in social situations or when confronted with a stressful situation. He had not worked for several years until two years ago when he obtained part-time employment bagging groceries at a supermarket. He had some significant difficulties coping with the job and at one point became depressed and required admission to inpatient psychiatry. He currently is unemployed because he quit his part-time job and attempted to work full time, but then quit working altogether when he felt unable to cope with the demands of a job. He resides in a rooming home and manages his own medications and funds.

Another technique in this category includes looking at the average GAF ratings assigned by trained clinicians to various psychiatric diagnostic groups within the context of research studies. This can be helpful since, as we have discussed, psychiatric diagnosis is highly correlated with GAF. Some examples of these studies are presented in Table 28. This type of information can give the practitioner a general idea of GAF ranges for various diagnostic groups that are entering different types of treatments. Of course, this type of data is largely focused on severity of psychiatric symptoms since social and occupational functioning is rarely reported. However, that is generally the case with the GAF anyway.

Table 28: Example Research Studies Using the GAF Determined by Trained Clinicians

Zohar et al. (2002) studied the GAF ratings of patients being discharged from an inpatient psychiatric facility after stays ranging from 1-30 days. Patients carried a variety of diagnoses and mean GAF scores were: substance abuse (59), schizophrenia (58), and mood disorders (60). Example GAF scores at discharge related to length of stay are: one to three days (67), four to seven days (58), and 15 to 30 days (55).

Psychiatric outpatients (N=44) presenting initially to an outpatient clinic for treatment carrying various DSM-IV Axis I diagnoses including depressive disorders, anxiety disorders, adjustment disorders, substance related disorders, and Axis II personality disorders, achieved an average GAF score of 64.5 (Hilsenroth et al., 2000).

In a study by Piersma and Boes (1997) GAF ratings for adult psychiatric inpatients were investigated at admission and at discharge. The mean GAF rating for the adult psychiatric inpatients at the time of admission was 45, which increased to 60 at the time of discharge. For those adults requiring only partial hospitalization, the GAF average was 55 at admission and 65 at discharge.

Harel et al. (2002) studied the GAF scores of patients being discharged from an inpatient psychiatric facility after undergoing appropriate treatment. A variety of diagnoses are represented including affective disorders, substance abuse disorders, and schizophrenia. The average discharge GAF across all diagnoses was 57.7 and for mood disorders it was 60.4. This can give an idea of GAF scores of those patients that required inpatient psychiatric treatment.

In a study investigating new patients (N=44) presenting to a psychiatric clinic and carrying a diagnosis of a Major Depressive episode, they were found to have an average GAF of 55.9 prior to treatment (Uehara et al, 1997).

Narud et al. (2005) investigated 91 patients presenting to psychiatric outpatient clinics. Multiple Axis I diagnoses were represented with depressive disorders, anxiety disorders, and alcohol dependence being the most common. The initial evaluation GAF mean was 55.4, before treatment.

Thienhaus et al. (1990) studied the effects of ECT and found an average GAF pre-treatment of 53 and post-treatment of 68. This suggests GAF scores for depression that is refractory to the usual treatment approaches.

Garcia-Cabeza et al. (2001) studied responses to various antipsychotic medications in a group of patients with schizophrenia (including subtypes such as paranoid, disorganized, catatonic, etc) with an average disease duration of 10 years. The sample size was 2657 and each patient received a GAF rating at the beginning of treatment. The average GAF was 44.1.

PART I - SUMMARY AND CONCLUSIONS

The Global Assessment of Functioning Scale (GAF) is a standard method for a clinician to judge a patient’s overall level of psychosocial functioning. The GAF requires a clinician to develop an overall judgment about the patient’s current psychological, social, and occupational functioning. These dimensions are collapsed into a single global score. In 2005, the GAF was adopted by the State of California as the primary method for determining permanent psychiatric disability in the workers’ compensation population. As discussed in the SRPD (2005), psychiatric impairment is to be evaluated using the GAF, which is then converted to a whole person impairment (WPI). In this course, several problems with the GAF were discussed including attempting to include three areas of function in one score, the inter-rater reliability, and its validity. Being aware of these limitations can help the clinician use the GAF in a more accurate fashion. Suggestions for improving one's GAF skills include carefully following the instructions, using the "split method" to help with scoring, relying on objective assessment of the three dimensions, and being familiar with research results for the GAF when used by trained clinicians.

PART II - INTRODUCTION TO THE RESEARCH ARTICLE

This section is mostly based on an open source article published in the Annals of General Psychiatry (Guidelines for rating Global Assessment of Functioning: GAF; by I.H.M. Aas; 2011). This article provides one of the most comprehensive and up-to-date reviews of the GAF we have found. Since the GAF is an integral part of the workers compensation system, it is important to be familiar with all of the issues and problems related to its use. As discussed in the article, "The present study aimed to identify the current status of guidelines for rating GAF, and relevant factors and gaps in knowledge for the development of improved guidelines." The article is an excellent review and includes over 100 references. It discusses important issues such as the following. It should be noted that there are no answers to most of these problems, but being aware of their existence can help the rater take them into account.

Problems with the GAF

Unknown reliability and validity across settings, patient populations, and purposes

The finding that different professions tend to assign different scores

Subjectivity of GAF ratings

Common violations of GAF rating instructions such as combining severity of symptoms and level of function rather than choosing which is worse

The effect on GAF scoring of starting at the top of the scale rather than the bottom or middle

The confusion surrounding scoring for different time periods (current, past year, etc.)

The problem of fine discriminations in scoring within a decile. Research suggests a tendency to score at the decile or a migration to the mid-decile.

Cultural influences on GAF rating

REVIEW OF GAF RESEARCH

The following is an article published by BioMed Central as an open source publication in the Annals of General Psychiatry. It is reproduced here in its entirely, without modification. The article can be found here and/or here.

Article: Guidelines for Rating Global Assessment of Functioning (GAF)

Author: I.H. Monrad Aas

ABSTRACT

Background

Global Assessment of Functioning (GAF) is a scoring system for the severity of illness in psychiatry. It is used clinically in many countries, as well as in research, but studies have shown several problems with GAF, for example concerning its validity and reliability. Guidelines for rating are important. The present study aimed to identify the current status of guidelines for rating GAF, and relevant factors and gaps in knowledge for the development of improved guidelines.

Methods

A thorough literature search was conducted.

Results

Few studies of existing guidelines have been conducted; existing guidelines are short; and rating has a subjective element. Seven main categories were identified as being important in relation to further development of guidelines:

Important Issues Related to the GAF

general points about guidelines for rating GAF
introduction to guidelines, with ground rules
starting scoring at the top, middle or bottom level of the scale
scoring for different time periods and of different values (highest, lowest or average)
the finer grading of the scale
different guidelines for different conditions
different languages and cultures

Little information is available about how rules for rating are understood by different raters: the final score may be affected by whether the rater starts at the top, middle or bottom of the scale; there is little data on which value/combination of GAF values to record; guidelines for scoring within 10-point intervals are limited; there is little empirical information concerning the suitability of existing guidelines for different conditions and patient characteristics; and little is known about the effects of translation into different languages or of different cultural understanding.

Conclusions

Few studies have dealt specifically with guidelines for rating GAF. Current guidelines for rating GAF are not comprehensive, and relevant points for new guidelines are presented. Theoretical and empirical studies, and international expert panels would be valuable, as well as production of a manual with more information about scoring. Computerized assessment may well be the future.

BACKGROUND

Reliable assessment of the problems patients face is important. With regard to the assessment instruments, guidelines for their use are also important [1-5]. Work has been carried out internationally to develop guidelines for psychological tests [6-8], but it is considered that a gap exists between existing standards and the need for regulation of the assessment process. Standardized scoring procedures are important, as they can reduce unintended bias [9-11]. There are many assessment procedures available in psychiatry, but little work has been done with guidelines for these methods [8].

In psychiatry, the severity of illness can be scored by Global Assessment of Functioning (GAF).GAF is known worldwide and it is Axis V of the internationally accepted Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition Text Revision (DSM-IV-TR) [12]. The GAF instrument was analyzed in a previous study [13], but questions have been raised as to whether clinician's rate GAF appropriately [14]. GAF is intended to be a generic rather than a diagnosis-specific scoring system. It is constructed as an overall (global) measure of how patients are doing and rates psychological, social, and occupational functioning, covering the range from positive mental health to severe psychopathology. Internationally, GAF recorded values can be either a single score (only the most severe of the symptom and functioning values is recorded) or separate scores for symptoms (GAF-S) and functioning (GAF-F).For both the GAF-S and GAF-F scales, there are 100 scoring possibilities (1-100).

An advantage of GAF is its simplicity [13], but problems have been found with its reliability and validity. Reliability studies show the extreme 20% of raters account for more than 50% of the spread of scores, and deviations can be 20 points or more [15,16]. Overall reliability can be good, but is not sufficient in the routine clinical setting [16-21] and is too low for assessment of change for the individual patient [20]. Concurrent validity [17,18,22-34] and predictive validity [19,23,25,27,35-37] are problematic. There are few empirical results for GAF sensitivity [13]. In general, psychiatric evaluation is too dependent on subjectivity, as assessors may rate psychiatric impairments according to their own experience and attitudes [3]. Rating GAF is no exception to this element of subjective judgment [13]; there is evidence that different professions assign different scores [38,39] and that the scores can be influenced by disagreement on criteria for rating [16], lack of training [22], or problems related to the intrinsic properties of GAF itself [13]. It has also been reported that site of investigation can explain some of the variability [34].

In the present study, guidelines are defined as written instructions that give guidance or recommendations for scoring and consist of some steps that are accepted by clinicians and the scientific community.

Guidelines are important for quality assurance of the assessment [40], and research has demonstrated that variation in guidelines influences the responses given by patients [41]. It should, therefore, be possible to develop better instructions for scoring of GAF [42].

The aims of the present study were to identify the current status of guidelines for rating GAF, points that are relevant for new guidelines, and gaps in knowledge that are of interest for the development of improved guidelines. Gaps in knowledge are defined as points concerning guidelines for scoring GAF where no, or little, research has been done and where it is likely that further development would play a role for improved scoring.

Methods

A literature review [43-47] was carried out. This was conducted by both hand searching and a search of bibliographic databases in several steps, where steps (a) and (b) represent the necessary 'end of the thread' to start the literature search:

(a) from previous work [13], the author had access to literature about relevant issues, namely literature about GAF and other scoring systems, which also includes information about methodology;

(b) browsing through journals, which has been recommended as a useful first step before computer searching [44], where each issue of a set of journals for the period January 2000 to December 2009 was searched (Acta Psychiatrica Scandinavica, American Journal of Psychiatry, Applied Psychological Measurement, Archives of General Psychiatry, BMC Psychiatry, British Journal of Psychiatry, Comprehensive Psychiatry, European Journal of Psychological Assessment, European Psychiatry, Evidence-Based Mental Health, International Journal of Testing, Journal of Psychiatric Research, Psychiatric Bulletin, Psychiatric Services, Social Psychiatry and Psychiatric Epidemiology, and Journal of Clinical Psychiatry);

(c) thorough hand searching: after identification of publications by steps (a) and (b), their reference lists were hand searched for more literature and, by reading total publications, a search for citations to other studies was also conducted.

Each time a relevant publication was identified, the same search for new literature was performed. After several rounds of such hand searching, new relevant references became difficult to find and the search proceeded to steps (d) to (i): (d) search in PubMed, which used experiences from research on search strategies [48,49]. A search was carried out for English language articles from the period January 1990 to December 2009. Search terms were: 'Global Assessment of Functioning OR GAF' AND combined with nine search terms (guidelines, standard, reliability, validity, sensitivity, literature review, systematic review, psychometrics, methodology) in nine separate searches. A total of 1,694 studies were identified by this method; (e) Possible missing publications remaining after steps (a) to (d) were controlled for by an Advanced Search in Google Scholar (for both books and articles) for the period from January 1990 to the day the search was performed (22 April 2010). The search terms 'Global Assessment of Functioning psychiatry' (used in 1 common search) identified 17,300 items (mostly publications), and the first 1,000 were screened for relevance. Google Scholar gives information about the number of links to each publication (this is effectively a citation tracking with the most frequently cited publications listed first). The Google Scholar search did not identify any studies that had not been already identified by steps (a) to (d); (f) A search in PsycINFO: this used experiences from research on search strategies [48,49]. A search was carried out for English language articles from the period January 1990 to 28 April 2010. Search terms were: ' Global Assessment of Functioning OR GAF AND' combined with seven search terms ('guidelines', 'instructions', 'standard', 'norm', 'process AND rating', 'process AND scoring', 'methodology') in seven separate searches. A total of 69 studies were identified by this search; (g) A search in The Campbell Collaboration Library of Systematic Reviews was carried out on 22 April 2010. The all-text searches were not limited to a specific time period. Five separate searches were performed (search terms: 'GAF', 'Global Assessment of Functioning', 'psychiatry systematic review', 'psychiatry literature review', 'psychiatry review'). However, this search identified no relevant studies; (h) The abstracts from steps (d) to (f) were screened, with the purpose of identifying literature concerning guidelines for GAF. When this screening started, the researcher was experienced from reading literature from steps (a) to (c). Abstracts were evaluated for inclusion by looking for information on the following issues in relation to GAF: guidelines, instructions, process of rating, methodology, psychometrics (studies with information on validity and reliability), history of GAF, and modifications/changes made. When the screening of abstracts was finished, selected publications were read in their entirety, but it became clear that most of the relevant literature had already been identified by steps (a) to (c); (i) For the selected publications from step (h), the reference lists were hand searched for more literature. New publications that were relevant for inclusion were difficult to find, and the literature search was complete.

The final two steps were as follows: (j) the contribution of each selected publication to the knowledge base for the present study was summarized [44]. Emphasis was placed on points that were relevant for new guidelines and analysis was performed to identify gaps in knowledge; (k) The final set of selected publications is the reference list of the present study. Included publications are original research papers, books, articles and book reviews.

Results

The literature review identified seven main categories, with a number of points (covered individually below) considered important in relation to further development of guidelines:

Important Issues Related to GAF and Development of Future Guidelines

(1) general points about guidelines for rating GAF
(2) introduction to guidelines, with ground rules
(3) starting at the top, middle or bottom level of the scale
(4) scoring for different time periods and of different values (highest, lowest or average)
(5) the finer grading of the scale
(6) different guidelines for different conditions
(7) different languages and cultures

Where the presentation of problems concerning guidelines does not require any distinction between the single-scale and dual-scale GAF, no remarks are made about this. Guidelines for scoring single-scale and dual- scale GAF can be quite similar. When the single scale is used, 'whichever is the worse' of the symptom and functioning values is the single value recorded (according to the manual for DSM-IV-TR) [12].

General Points about Guidelines for Rating GAF

Brief guidelines for rating GAF exist, but their lack of depth is likely to result in subjectivity in rating [5].They are also different in several respects. An early version of GAF (the Global Assessment Scale(GAS)had scoring instructions [50],but the publication of DSM-IV-TR updated GAF, with significant changes in these rating instructions [12,27].The Veterans Administration in the US [5,22] and Norwegian psychiatry services [51] have guidelines. Other systems based on GAF also have guidelines, for example the Modified GAF [24] and Kennedy Axis V [52].

In practice, experienced clinicians operate by forming initial hypotheses and testing them through assessment [53], but they can be faced with dilemmas about which GAF value to choose. If guidelines are going to be of value for rating, they need to be clear, specific and complete. The process of scoring must take account all of the specific properties of GAF [13]. Working with guidelines for psychological tests could form the learning base for further work with guidelines for GAF; for example, the International Test Commission has developed guidelines for using psychological tests [6,7,54,55] and several of the points in these guidelines apply to assessments used in psychiatry.

When assessment instruments are developed, study of the assessment process should be a standard procedure [9], but there has been little interest in guidelines for GAF scoring. International panels of experts have played a limited role in guideline development, and few have compared the content of existing guidelines or investigated what the correct norm for the scoring process should be [3,14,39]. There is limited empirical research on the actual process of scoring, and one study has shown that the actual process agrees well with the concept of GAF [14]; however, the actual process is not necessarily the same as the prescribed process [14]. Before training, practitioners will often choose an incorrect strategy for scoring GAF [22]; for example, they may use the average of the functioning and symptom scores (for the single-scale GAF, only one value is recorded), the least severe of symptoms, or the highest area of functioning [22].

Gap in knowledge. In the historical development of GAF, there has been little research on existing guidelines. Few studies have compared the effect of using different existing guidelines for rating and the effect of systematically varying guidelines. We do not know which norms for the guidelines are best or whether changed and extended guidelines would improve rating.

Introduction to Guidelines, with Ground Rules

The introduction to guidelines should give raters a basic understanding of the guidelines, other specifications and what to look for when scoring GAF. However, existing guidelines for rating GAF have different introductions [5,12,50,51]. When different introductions lead raters thinking in different directions, an effect on GAF scores is likely. Developing a good concise introduction should not be considered an irrelevant detail; if it is weak and poorly defined there is a risk that raters will use their individual perspectives to make judgments and use norms from other sources; for example, a clinician working mainly with severely ill patients may unintentionally use this experience as a norm for the less severely ill [5]. However, this has been given little attention in international publications.

The introductory paragraph in a guideline for rating GAF could start by explaining the purpose of rating GAF, for example to score the overall level of functioning or severity of illness [50] and why GAF values are important.

Then, a key purpose for the guideline should be given, for example to enhance assessment by describing competent instrument use, to help in standardizing rating so that influence of change in the assessor is minimized, and to help in assigning more accurate scores [6,7,56].

In the second paragraph, a definition of what GAF is can be given [13] and an image of the scale(s) provided (with anchor points, key words and examples). The next point could be ground rules for the rating itself. As GAF means rating functioning and symptoms, these terms should be defined, with examples of symptoms and functioning that should and should not be taken into consideration. When rating, all the available information that is important for GAF-S and GAF-F should be considered [14,29], but this information should then be sufficient for good overall judgment of both symptoms and functioning. In both the DSM-IV-TR and the Norwegian instructions, there is a ground rule: 'consider psychological, social, and occupational functioning on a hypothetical continuum of mental health-illness' [12,51,57], but there is little published analysis of how this ground rule is understood by different assessors and how well it works in practice. According to the Norwegian guidelines, this ground rule means that symptoms (and functioning) should be viewed in their broader context, for example the need for treatment [51]. According to the DSM-IV-TR [12], the GAF value is useful in planning treatment, measuring the impact of treatment, and predicting outcome, but there is limited information available on the adequacy of GAF in prediction of outcome [19]. Information concerning the choice of level of care for different ratings could be given, for example a patient with a score of 1-30 is a potential candidate for inpatient care, a patient with a score of 31-69 a potential candidate for outpatient care, and a patient with a score of 70 and higher may be functioning too well to be a candidate for any treatment.

Gap in knowledge. Introductions to guidelines have been given little attention in international literature. Ground rules for rating have been little analyzed and there is little information about how they are understood by different raters. It is not known what the result would be if international consensus panels of experts worked with ground rules.

Starting Scoring at the Top, Middle or Bottom Level of the Scale

It is known from methodology studies of questionnaire design that the ordering of response categories is a problem. Studies show a tendency to choose the both first listed response category ('primacy' effect) and the last listed response option ('recency' effect). Primacy effects are more likely in self-completion surveys [58]. A similarity in methodology problems exists for GAF and questionnaires [13]. Clinicians perform the rating by asking questions, and the GAF's deciles (with anchor points) are used as response categories. There is no common international norm for where to start; existing guidelines for GAF:

(a) recommend starting at the top level of the scale with evaluation of whether the patient is worse than indicated by each of the decile's anchor points [12]; or

(b) recommend starting at the bottom level [51]; or

(c) give no instructions for where to start [5].

It may be hypothesized that starting from the top results in higher values than starting from the bottom and it is known that with questionnaires even seemingly minor changes can have a major impact [59]. An alternative approach would be to start in the middle of the scale (GAF = 50) and ask if the severity is worse or the patient is more healthy and then keep moving down or up the scale until the range that best matches the individual's symptom severity or level of functioning is reached. To double check, a look at the next upper or lower range would be taken.

Gap in knowledge. Information concerning the effects of starting the rating process at top, middle or bottom level is difficult to find.

Scoring for Different Time Periods and of Different Values - Which Time Period?

In psychiatry, symptoms can change over time, for example over 24 h [16]. According to the DSM-IV-TR manual [12], the GAF score (in most instances) should be the level at the time of evaluation. The current level of functioning can be operationalized to the lowest level of functioning for the last week [12,38,50,51], which may be used to represent a baseline before onset of treatment [60]. It has also been suggested that symptom scales for the degree of severity of current illness should cover the past 3 days [61], but in acute care departments, even shorter time periods can be relevant [51].

The score for the last week may conflict with the patient's previous mental health, and fluctuations in the patient's condition may need to be scored several times over a longer period of time [62]. If this is not done, clinically useful information might be lost [63]. Scoring can also be done for time periods, for example for the last week and the past year [23]; this may cause considerable differences in scores [61] and so, when relevant, scoring can be done for more than one time period [23]. Examples of proposed time periods are: last year, last 6 months, at least a few months during the past year, and the preceding month [12,21,29,42,51].

Knowledge of the course of different conditions over time is essential [64]; for some patients and studies, scoring for longer periods may be appropriate. Longitudinal descriptions of the psychopathology can add information. The importance of premorbid level of functioning has been little explored and is rarely documented [3], but for chronic conditions, it is logical to consider adding scores for longer periods [65]. Depression can be scored by, for example: depression in the past year for 2 weeks or more, for much of the time in the past year, or for most of the days over a 2-year period [65]. For bipolar disorder, scoring of current symptoms is not enough and it is necessary to check for a past history of mania [66]. If psychosis has lasted for a longer period, the GAF score should be lower than the score given at admission for a first-time psychosis. For personality disorders, the stability of personality is a defining feature and a longitudinal perspective is essential in diagnosing [67]: scoring can be done for the past several years, the past 5 years, the 2 years before the interview, or the 'usual self' [67].

When the effect of treatment is being studied, GAF should be scored both before and after treatment [12]; scoring periods of between 3 and 12 months after discharge are suggested [65]. For patients under treatment for a longer period, scoring can be done every 2 or 3 months [63]. For example, outpatients who have not been given a GAF score in the last 90 days should be given a new score [42,68].

Gap in knowledge. The longitudinal dimension of using different GAF scores for different disorders has been little explored and existing guidelines give little instruction. There is little research data available about the time period that should be used for GAF rating or the criteria for choosing a specific time period. It is not known whether scoring should be done for the same time period for the GAF-S and GAF-F scales, whether scoring should be done for different time periods for the higher and lower ends of each GAF scale, or whether scoring should be done for different time periods for different anchor points.

Which Value (lowest, highest or average)?

The aim of scoring should be to give a true image of the patient's mental health that will be useful for clinicians and research. As the severity of illness can vary over time, the question of which GAF value to record becomes relevant. Simple alternatives are the lowest, highest or average GAF for a time period. According to scoring instructions for GAF, when the current level of functioning is scored, the lowest score for the last week should be used; the lowest level of functioning is chosen because of its clinical relevance [51]. Rating GAF may mean choosing the lowest score for other specified time periods, for example the lowest level in the past month or for the worst week during the month prior to interview [3,37,39,63,69].

However, assigning the lowest GAF score is not without problems. It may give a wrong impression of both the overall mental situation and the present status [42]; the highest level of functioning should not be disregarded [12,31,39,57,70] as it may predict outcome [71]. For example, the highest level of functioning for at least a few months during the last year may be very predictive of outcome [19,52] and indicate the potential level of functioning [60]. Also, it has been reported that the highest level of functioning during the past year can be highly correlated with current level [19]. If the patient is not well described by either the highest or the lowest GAF for the last week, a solution may be to use more scores; for example, scores such as highest and lowest for the last year, the highest and lowest the patient has ever had, or scores for when the patient is symptomatic and asymptomatic. Rating of average functioning has also been proposed [29,50], for example, the average level of functioning during the previous 3 weeks [5,57]. If such scores describe the patient well, they can be added.

Internationally, both the single-scale and dual-scale GAF are in use. For the single-scale GAF, according to the manual for DSM-IV-TR [12] only one value should be recorded, namely, 'whichever is the worse’ of the symptom and functioning values [5,12,21,22]. It is assumed that the GAF-S and GAF-F are comparable scales [16,27], so recording only the most severe of the GAF-S and GAF-F scores is in accordance with the general principle of using the most severe condition as the overall score [16]; however, the difference between the two scales is disregarded so it is not clear which factor of symptoms and functioning is being measured [52]. An alternative could be to record the average of symptoms and functioning levels [72], but this raises the question of whether or not symptoms and functioning have equal weight, and the importance of any weighting effect [73]. Although the values on each scale may be close [29], symptoms and functioning are different aspects of patient condition and they do not necessarily vary together [23], so in some countries a dual scale GAF is used where both GAF-S and GAF-F are recorded [13].

In the clinical setting, comments can be added to a GAF score on why a particular score was chosen, which may be important when others take over treatment. It may also have an educational effect, add meaning to the scores, and improve inter-rater reliability [42]. However, it would be helpful if guidelines included a norm for the choice of score with more detailed information about which score to record; this is not an easy task, as mental illness is a multifaceted and complex problem. Deciding the criteria for such a norm is problematic.

Gap in knowledge. It is difficult to find empirical research aimed at finding the right GAF value (lowest, highest, or average), or combination of GAF values, to record for different applications. The potential applications for GAF scoring are wide ranging and include different diagnostic categories, the chronic and acutely ill, treatment decisions, prediction or measurement of outcome, choice of level of care, and measurement of case mix. Little is known about which score gives the best inter-rater reliability and validity, and it is not known whether separate GAF-S and GAF-F, or the lower of the two scores is best for treatment decisions and measurement of outcome, or how much weight should be given to GAF-S versus GAF-F for such applications.

The Finer Grading of the Scale

The DSM-IV-TR, Veterans Administration and Norwegian guidelines have instructions for scoring within 10- point intervals, but instructions are limited [5,12,13,51]. Scoring within the 10-point intervals is open to subjective judgment and finer distinctions readily become somewhat random. In practice, clinicians tend to score around the decile or mid-decile divisions of the scale [42]. Patients who are scored in the same 10-point interval should be relatively homogenous in functioning, but functioning is a construct with many facets and when information for a more accurate score is lacking, intermediate scores in the deciles are chosen [63,74].

It is possible that more detailed verbal instructions would result in more accurate scores. An alternative to having more anchor points is to use categorical scales for scoring within the 10-point intervals, in which case the anchor points (with key words and examples of symptoms and functioning items) should be graded [13,75]. Both symptoms and functioning can be graded in different ways [76]. A categorical scale requires a decision about the number of categories; such scales often have five categories, for example: very marked, marked, neither marked nor weak, weak, or very weak. Numbers of categories other than five can also be considered [61,77]. More experienced raters may be able to make finer distinctions and score correctly with more categories, but scoring in the clinic is often carried out by people with different educational backgrounds [15,16,19-21,29]. An alternative procedure for scoring within 10-point intervals is found in the 'modified GAF' [24], which uses the number of criteria met: for example, for the interval 41-50, when one criterion is met the score should be 48-50 and when two criteria are met it should be 44-47.

Gap in knowledge. In the history of GAF, systematic work to improve scoring within 10-point intervals is limited and it is not known how to best score within 10-point intervals. This also applies to the use of categorical scales for scoring, which requires considerations concerning the nature and number of categories.

Different Guidelines for Different Conditions

There can be a vast difference between the mental states of different patients. However, a dual-scale GAF scoring uses two straight lines (that is, a multidimensional phenomenon is scored in a two-dimensional way), which may not reflect this complexity. The answer to the problem is not necessarily to have more scales covering different aspects of, for example functioning, as this would require a more complex scoring process [13]. However, if guidelines for rating are not good enough, the value of an assessment instrument is reduced. It does seem appropriate to consider development of guidelines for different conditions.

Panels of experts aided by empirical data could develop norms with ranges of relevant GAF values. The comprehensibility of anchor points (with key words and examples) for different diagnostic groups should be considered and it would be helpful to include examples of patients scored and not scored in each decile [13,77]. The reliability of scores is not necessarily the same for all diagnostic groups. To ensure assignment of the correct GAF value, advice could be given on how to obtain good information for each patient (for example which psychiatric interview to use). For some diagnostic groups, this can mean collecting more information than for others. Guidelines should have information on how to take different comorbid conditions into consideration.

If different GAF values are expected for different ages and sexes, this should be noted in the guidelines, but there is little information available about this. Different norms of functioning can represent different baselines against which the patient is evaluated, so, for example, instruments should be adapted to assessing older patients, to include scoring of dementia and happiness at the end of life [9]. Guidelines could also be different for different situations, for example for admission to inpatient departments and for community studies [13].

GAF should score impairment due to mental condition, but the effect of somatic and mental impairment can be interrelated and it can be difficult to distinguish between them [14]. The GAF rating should not be influenced by considerations on prognosis, previous diagnosis, presumed nature of the underlying disorder, or whether or not the patient is receiving medication or some other form of help [5,12,50,51].

Gap in knowledge. There is limited empirical information concerning the suitability of existing guidelines for different conditions, different groups of patients and patients with several other characteristics. The effect of adapting guidelines to these variations is not known. Having different guidelines for symptoms and functioning has been little explored.

Different Languages and Cultures

GAF has been translated into many languages, but languages encode meaning in different ways. Instruments should be adapted to different cultures and languages [6,7,40,73,78].

People from different cultures can answer in different ways when questions are asked, for a number of reasons [73,79], and this can have consequences for GAF values. It is important to understand illness explanations and help-seeking behaviors [80] within the patients' cultural framework and patients should be evaluated against what is 'normal' in their own culture. Cultural factors can be important for attitudes to disorder [81-83], and the use of GAF in multiethnic societies presents challenges to assessment [9].

Language differences may also present problems; a patient may be clearly psychotic when interviewed in their own language, but not when interviewed in a foreign language [83]. When translated into other languages, the guidelines for rating GAF, interviews for rating GAF, and GAF itself (for example anchor points with key words and examples) can be influenced. Translation of assessment instruments can involve translation, back translation, review and modification and guidelines are available for translating tests and assessment instruments [9,84].

Gap in knowledge. Little is known about the importance of translation and culture for GAF guidelines. The safety of international comparisons should be questioned. Meta-analyses based on data from countries with different languages and cultures may be influenced by these differences.

Further Development for GAF

We are a long way from having a comprehensive set of heuristic guidelines that could support the assessor in executing the scoring process [85], but progress in the study of the assessment process is anticipated [9]. Guidelines should be based on both theory, and empirical knowledge [85] about how each guideline works in practice. Development of new guidelines for GAF would be facilitated by first reviewing the literature about guidelines for psychological assessment, and extracting relevant points [6,7]. New empirical research could then be performed, for example by performing qualitative studies of the actual process of scoring, to search for items that are relevant for guidelines, while bearing in mind that if the scoring process is made too complex, errors are more likely to be introduced [76]. The existence of international guidelines would provide support to the implementation and use of the guidelines in different countries. Guidelines should reflect consensus on practice [7] and a draft of new guidelines for GAF should therefore be circulated widely to provide ample opportunity for comments [56]. A GAF scale with new guidelines should also be tested out for reliability and validity for different diagnoses, with different scorers, across different sites and with different patient populations. To study the effects of varying guidelines, knowledge of 'true' values would be useful and mean scores from expert panels can work as reference norms [29].

When designing a norm for the scoring process, it is important to consider which process can best achieve the aims. It is essential to first define the purpose of a scoring system. For example, a system that is mainly intended for clinical use should be viewed by clinicians as sensible and easy to use. However, having a short version of the guidelines for the clinic and more detailed guidelines for research could result in scores that are not directly comparable; evidence-based treatment is, by definition, based on research and this could pose a problem for its implementation.

A manual with more information about GAF and scoring of GAF could also be developed alongside the guidelines [86]. The requirement for guidelines to be short and concise makes it necessary to decide which information should be given in the guidelines and which in the manual. The manual can serve as principal source of information and might contain information about issues relating to GAF, such as history of its development; the theoretical basis; the comprehensiveness of GAF for different conditions; the reliability and validity of GAF with explanations for problems; statistical information for different diagnostic groups (mean value, standard deviation, range and statistical distribution, whether normal or skewed, and in which direction); information about which methods to use together with GAF (multimethod assessment is common); GAF values compared to values from other methods; implications of different GAF scores for treatment, with examples and thresholds of severity values defining when treatment is desirable; management use of GAF (for example in planning and comparison of case mix) [87]; rating by teams and individuals; use of GAF for patients with different cultural and linguistic backgrounds; and training material with descriptions of several cases with assigned GAF values.

Computerization of assessment may well be the future. Assigning scores could begin with a visible GAF scale on the screen, where placing the cursor at different places along the scale reveals different windows with information about the criteria for scoring; clicking the mouse in one of these windows could make even more detailed information available in another window. The use of electronic patient records represents a possibility for new quality assurance methods. Some diagnoses are not combinable with high GAF scores; if such a diagnosis has been given, a warning could pop up on the screen if a GAF score that is too high is given. If a low GAF-S is given, a warning could pop up if a high GAF-F is given. A reminder may come up if the psychiatric record is completed for a new patient without having entered a GAF score. When a GAF score has not been given for an outpatient for the last 3 months, a reminder could pop up on the screen. Computer-based scoring of GAF can give high correlation with scoring based on clinical impression [88], but difficulties with computer-assisted assessment suggest a number of guidelines for users [41]. The International Test Commission has developed guidelines on computer-based and internet-delivered testing [89-94], but these guidelines were not developed with GAF in mind.

Work with a scoring instrument is not complete without testing or pilot study [82, 95]. Alterations to the scoring process are not necessarily always improvements, and a pilot study is needed to reveal any additional changes that are necessary.

DISCUSSION

Methods

Literature reviews can play a role in development of guidelines [96].The present study can be defined as a systematic review [48,49].Several important criteria for review articles are satisfied, such as defining the problem, informing the reader of the status of current research, identifying gaps and suggesting the next step [97].

An encompassing hand search of literature was done because it was considered that some relevant publications were likely not to be included in computerized databases. A combination of searching reference lists and reading publications has been considered the most thorough way of hand searching [98]. PubMed includes more than 500 psychology-related journals [99], but as the search showed, few publications deal specifically with guidelines for rating GAF, the search was continued in other databases. The citation tracking in Google Scholar is not completely reliable when it comes to listing the most frequently cited first, but screening of the first 1,000 results represents a thorough Google Scholar search. The search in PsycINFO added little new knowledge. The search in The Campbell Collaboration Library of Systematic Reviews added no new studies. The searches in PubMed, Google Scholar, The Campbell Collaboration Library of Systematic Reviews, and PsycINFO are reproducible. The search in PubMed, Google Scholar, and PsycINFO revealed that most of the publications were already identified by the thorough hand search (step(c) in Methods). In step (i), a stage was reached where new perspectives could not be identified by reading more publications; the situation is described by the term 'saturation' from qualitative research. It is not considered likely that publications that could have changed the results were missed as a result of the search process. The design and conduct of the present study protected against bias [47,48].

Better Guidelines for GAF

The literature review identified the state of knowledge for GAF guidelines and a review of this type can be valuable in work to develop better guidelines. In the history of GAF, limited focus has been given to development of guidelines and currently available guidelines are short. In the clinic, the primary goal of the assessment process is to contribute to the solution of a person's problems [100]. A generic and global scoring system, such as GAF, that covers the range from positive mental health to severe psychopathology has advantages for clinical practice (for example, routine quality assessment of treatment, supplementing scales that give more detail) [75], research (for example, comparison of treatment outcome across diagnoses), and policy and management planning (for example, allocation of resources, measurement of case mix in psychiatric organizations). For GAF to have such a broad range of applications, it must be good enough for the purpose. It is important not to simply dismiss GAF because of problems concerning either the instrument itself [13] or guidelines; existing scales can be dismissed too lightly [72].

A scoring system must be robust enough to allow for scorer bias and more random errors of measurement. If GAF is not good enough, a given change in GAF value would not necessarily reflect a corresponding change in severity. Subjectivity in scoring should be kept to a minimum; some scorers can be unwilling to give a low score because of the negative labeling of clients [22] and clinicians who do most of their work with one patient category may use their experience as a norm. Improved consistency of scoring can be achieved locally by delivering courses in rating GAF [22], but the risk of variation between different local standards will remain. Improved guidelines have the potential to reduce such bias.

The aim of better guidelines is to make scores more reliable, to improve comparability of scores (for example across organizations and from different studies), to make combination of scores in meta-analysis safer, help in assigning more accurate scores (choosing better between individual points in the 10-point ranges), to provide more accurate information for the choice of intervention and evaluation of treatment results, and to be of help in the education and training of assessors. However, it is not a matter of course that new guidelines will give much better GAF scores.

The clinical situation is not just about having a perfect scoring system; it is equally important to earn the respect and trust of the patient [70]. New guidelines should not be destructive for the clinician-patient relationship. They should also be adaptable and tolerate changes in clinical practices; information for scoring should be easy to obtain; and the scoring process should not be too time consuming. Evidence-based medicine has shown that examples of successful implementation of guidelines exist, but also that implementation is not always successful [101]. It is important that once new guidelines for GAF have been developed, they are implemented effectively.

Factors Other than the Process of Scoring

The present review has focused on guidelines for rating GAF, but other factors can also play a part in the choice of GAF value. Factors that have not been treated include:

(1) characteristics of the patient interview and the importance of collecting information from different sources;

(2) characteristics of the rater, i.e. professional background, training and motivation, groups, or individuals score; and

(3) properties of GAF(discussed in a previous study) [7,13,19,20,23,34,36,39,57,58,61,77,102-105].

CONCLUSIONS

The guidelines that are currently available for rating GAF are not the result of a sophisticated development, but guidelines are important for reliable assessments. There are few published studies dealing specifically with guidelines for rating GAF. This study presents a number of points that are relevant for new guidelines and show a significant potential for development.

International panels of experts have a role to play, and a manual for GAF can be developed. Computerization of the scoring process can offer advantages for rating. In light of the current situation, care should be exercised when comparing outcomes across facilities and also with international comparison, and meta-analyses. More work is needed to develop improved guidelines for rating GAF.

The Fine Print

The following is related specifically to the article by IH Monrad Aas.

Correspondence: monrad.aas@piv.no

Affiliation: Department of Research, Vestfold Mental Health Care Trust, Tönsberg, Norway

Acknowledgements: I thank Dr Penny Howes (Medical and Scientific Editing Service, UK) who provided assistance with the language.

Funding: Vestfold Mental Health Care Trust funded the study.

Competing interests: The author declares that he has no competing interests.

Received: 9 November 2010 Accepted: 20 January 2011 Published: 20 January 2011

WebURL:www.annals-general-psychiatry.com/content/10/1/2

doi:10.1186/1744-859X-10-2
Cite this article as: Aas: Guidelines for rating Global Assessment of Functioning (GAF). Annals of General Psychiatry 2011 10 :2.

REFERENCES - Part I

American Medical Association (AMA, 2000). The Guides to the Evaluation of Permanent Impairment, Fifth Edition. American Medical Association.

American Psychiatric Association. (1980, 1987, 1994, 2000). Diagnostic and Statistical Manual of Mental Disorders (III, III-R, IV, IV-TR Editions). Washington, DC: Authors.

Bates et al. (2002). Effects of brief training on application of the Global Assessment of Functioning Scale. Psychological Reports, 91, 999-1006.

Dworkin et al. (1990). The longitudinal use of the Global Assessment Scale in multiple-rater situations. Community Mental Health Journal, 26, 335-341.

Endicott et al. (1976). The Global Assessment Scale: A procedure for measuring overall severity of psychiatric disturbance. Archives of General Psychiatry, 33, 766-771.

First, M.B. and Pincus, H.A. (2002). The DSM-IV Text Revision: Rationale and potential impact on clinical practice. Psychiatric Services, 53, 288-292.

Garcia-Cabeza, I., et al. (2001). Subjective response to antipsychotic treatment and compliance in schizophrenia. A naturalistic study comparing olanapine, risperidone and haloperdol. BMC Psychiatry, 1:7. (www.biomedcentral.com/1471-244X/1/7, accessed 02-20-2010).

Goldman, Skodol and Lave (1992). Revising axis V for DSM-IV: a review of measures of social functioning. American Journal of Psychiatry, 149, 1148-1156.

Hall, RCW. (1995). Global Assessment of Functioning: A modified to scale. Psychosomatics, 36, 267-275.

Harel, TZ et al (2002). A comparison of psychiatrists' clinical-impression-based and social workers' computer-generated GAF scores. Psychiatric Services, 53, 340-342.

Hay et al. (2003). A two-year follow-up study and prospective evaluation of the DSM-IV Axis V. Psychiatric Services, 54, 1028-1030.

Howes, JL et al (1997). Outcome evaluation of a short term mental health day treatment program. Canadian Journal of Psychiatry, 42, 502-508.

Hilsenroth, MJ et al (2000). Reliability and validity of DSM-IV Axis V. Am. Journal of Psychiatry, 157, 1858-1863.

Kessler et al (2003). Screening for serious mental illness in the general population. Archives of General Psychiatry, 60, 184-189.

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

MacDonald-Wilson et al. (2001). Unique issues in assessing work function among individuals with psychiatric disabilities. Journal of Occupational Rehabilitation, 11, 217-232.

Moos, RH, McCoy, L, Moos, BS (2000). Global assessment of functioning (GAF) ratings: Determinants and role as predictors of one-year treatment outcomes. Journal of Clinical Psychology, 56, 449-461.

Moos, RH, Nichol, AC, Moos, BS (2002). Global assessment of functioning ratings and the allocation and outcomes of mental health services. Psychiatric Services, 53, 730-727.

Narud, K et al. (2005). Quality of life in patients with personality disorders seen at an ordinary psychiatric outpatient clinic. BMC Psychiatry, 5:10, www.biomedcentral.com/1471-244X/5/10. (accessed 2-20-2010)

Niv et al. (2007). The MIRECC Version of the Global Assessment of Functioning Scale: Reliability and Validity. Psychiatric Services, 58, 529-535.

Piersma and Boes (1997). The GAF and psychiatric outcome: A descriptive report. Community Mental Health, 46, 117-121.

Rogers, R. (2008). Clinical Assessment of Malingering and Deception, Third Edition. New York: Guilford.

Roy-Byrne et al (1996). Evidence for limited validity of the revised global assessment of functioning scale. Psychiatric Services, 47, 864-866.

Soderberg et al. (2005). Reliability of Global Assessment of Functioning Ratings made by clinical psychiatric staff. Psychiatric Services, 56, 434-438.

Thienhaus, OJ et al. (1990). A study of the clinical efficacy of maintenance ECT. Journal of Clinical Psychiatry, 51, 141-144.

Uehara, T (1997). Correlations among depression rating scales and a self-rating anxiety scale in depressive outpatients. The International Forum for Psychiatry. (accessed 2-20-2010).

Vatnaland et al. (2007). Are GAF scores reliable in routine clinical use? Acta Psychiatric Scandivania, 115, 326-330.

REFERENCES - PART II

1. Hagmeister C, Westhoff K: Teaching and learning psychological assessment: aspects of the client's question. Eur J Psychol Assess 2002, 18:252-258.

2. Kici G, Westhoff K: Evaluation of requirements for the assessment and construction of interview guides in psychological assessment. Eur J Psychol Assess 2004, 20:83-98.

3. Ryu SG, Hong N, Jung HY, Hwang S-C, Jung H-Y, Jeong D, Rah UW, Suh D- S: Developing Korean Academy of Medical Sciences guideline for rating the impairment in mental and behavioural disorders: a comparative study of KNPA's new guidelines and AMA's 6th guides. J Korean Med Sci 2009, 24 (Suppl 2):S338-342.

4. Sawyer J: Measurement and prediction, clinical and statistical. Psychol Bull 1966, 66:178-200.

5. Watson P, McFall M, McBrine C, Schnurr PP, Friedman MJ, Keane T, Hamblen JL: Best practice manual for posttraumatic stress disorder (PTSD) compensation and pension examinations. 2002 [http://www.avapl. org/pub/PTSD%20Manual%20final%206.pdf].

6. Bartram D: The development of international guidelines on test use: the International Test Commission project. Int J Testing 2001, 1:33-53.

7. Bartram D: Guidelines for test users: a review of national and international initiatives. Eur J Psychol Assess 2001, 17:173-186.

8. Watson P, McFall M, McBrine C, Schnurr PP, Friedman MJ, Keane T, Hamblen JL: Guidelines for the assessment process (GAP): a proposal for discussion. Eur J Psychol Assess 2001, 17:187-200.

9. Fernández-Ballesteros R: Psychological assessment: future challenges and progresses. Eur Psychol 1999, 4:248-262.

10. Meyer GJ, Finn SE, Eyde LD, Kay GG, Moreland KL, Dies RR, Eisman EJ, Kubiszyn TW, Reed GM: Psychological testing and psychological assessment. A review of evidence and issues. Am Psychol 2001, 56:128-165.

11. Shermis MD: Book review. Int J Testing 2007, 7:409-411.

12. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR) Washington, DC, USA: American Psychiatric Association; 2000.

13. Aas IHM: Global Assessment of Functioning (GAF): properties and frontier of current knowledge. Ann Gen Psychiatry 2010, 9:20.

14. Yamauchi K, Ono Y, Ikegami N: The actual process of rating the Global Assessment of Functioning scale. Compr Psychiatry 2001, 42:403-409.

15. Loevdahl H, Friis S: Routine evaluation of mental health:reliable information or worthless ‘ guesstimates ’ ? Acta Psychiatr Scand 1996, 93:125-128.

16. Vatnaland T, Vatnaland J, Friis S, Opjordsmoen S: Are GAF scores reliable in routine clinical use? Acta Psychiatr Scand 2007, 115:326-330.

17. Burlingame GM, Dunn TW, Chen S, Lehman A, Axman R, Earnshaw D, Rees FM: Selection of outcome assessment instruments for inpatients with severe and persistent mental illness. Psychiatr Serv 2005, 56:444-451.

18. Hilsenroth MJ, Ackerman SJ, Blagys MD, Baumann BD, Baity MR, Smith SR, Price JL, Smith CL, Heindselman TL, Mount MK, Holdwick DJ Jr: Reliability and validity of DSM-IV axis V. Am J Psychiatry 2000, 157 :1858-1863.
Aas AnnalsofGeneralPsychiatry 2011, 10:2 http://www.annals-general-psychiatry.com/content/10/1/2 Page 9 of 11

19. Moos R, McCoy L, Moos BS: Global Assessment of Functioning (GAF) ratings: determinants and role as predictors of one-year treatment outcomes. J Clin Psychol 2000, 56:449-461.

20. Söderberg P, Tungström S, Armelius BÅ: Reliability of Global Assessment of Functioning ratings made by clinical psychiatric staff. Psychiatr Serv 2005, 56:434-438.

21. Startup M, Jackson MC, Bendix S: The concurrent validity of the Global Assessment of Functioning (GAF). Br J Clin Psychol 2002, 41:417-422.

22. Bates LW, Lyons JA, Shaw JB: Effects of brief training on application of the global assessment of functioning scale. Psychol Rep 2002, 91:999-1006.

23. Goldman HH, Skodol AE, Lave TR: Revising axis V for DSM-IV: a review of measures of social functioning. Am J Psychiatry 1992, 149:1148-1156.

24. Hall RCW: Global Assessment of Functioning. A modified scale. Psychosomatics 1995, 36:267-275.

25. Hay P, Katsikitis M, Begg J, Da Costa J, Blumenfeld N: A two-year follow-up study and prospective evaluation of the DSM-IV Axis V. Psychiatr Serv 2003, 54:1028-1030.

26. Jones SH, Thorncroft G, Coffey M, Dung G: A brief mental health outcome scale reliability and validity of the Global Assessment of Functioning (GAF). Br J Psychiatry 1995, 166:654-659.

27. Niv N, Cohen AN, Sullivan G, Young A: The MIRECC Version of the Global Assessment of Functioning scale: Reliability and validity. Psychiatr Serv 2007, 58:529-535.

28. Patterson DA, Lee M-S: Field trial of the Global Assessment of Functioning Scale-Modified. Am J Psychiatry 1995, 152:1386-1388.

29. Pedersen G, Hagtvedt KA, Karterud S: Generalizability studies of the Global Assessment of Functioning- split version. Compr Psychiatry 2007, 48:88-94.

30. Piersma HL, Boes JL: Agreement between patient self-report and clinician rating: concurrence between the BSI and the GAF among psychiatric inpatients. J Clin Psychol 1995, 51:153-157.

31. Robert P, Aubin V, Dumarcet M, Braccini T, Souetre E, Darcourt G: Effect of symptoms on the assessment of social functioning: comparison between Axis V of DSM III-R and the psychosocial aptitude rating scale. Eur Psychiatry 1991, 6:67-71.

32. Roy-Byrne P, Dagadakis C, Unutzer J, Ries R: Evidence for limited validity of the revised Global Assessment of Functioning Scale. Psychiatr Serv 1996, 47:864-866.

33. Salvi G, Leese M, Slade M: Routine use of mental health outcome assessments: choosing the measure. Br J Psychiatry 2005, 186:144-152.

34. Tungström S, Söderberg P, Armelius B-Å: Relationship between the Global Assessment of Functioning and other DSM Axes in routine clinical work. Psychiatr Serv 2005, 56 :439-443.

35. Bacon SF, Collins MJ, Plake EV: Does the Global Assessment of Functioning assess functioning? J Ment Health Counsel 2002, 24 :202-212.

36. Fallmyr Ø, Repål A: Evaluering av GAF-skåring som del av Minste Basis Datasett. Tidsskrift for Norsk Psykologforening 2002, 39:1118-1119.

37. Parker G, O ’ Donell M, Hadzi-Pavlovic D, Proberts M: Assessing outcome in community mental health patients: a comparative analysis of measures. Int J Soc Psychiatry 2002, 48:11-19.

38. Laderman ER, Stein SM, Papanastassiou M: Flattened hierarchies and equality in clinical judgement. Therapeut Commun 1999, 2081-92.

39. Schorre BEH, Vandvik IH: Global assessment of psychosocial functioning in child and adolescent psychiatry. A review of three unidimensional scales (CGAS, GAF, GAPD). Eur Child Adolesc Psychiatry 2004, 13:273-286.

40. Kersting M, Hornke LF: Improving the quality of proficiency assessment: the German standardization approach. Psychol Sci 2006, 48:85-98.

41. Groth-Marnat G: Handbook of Psychological Assessment Hoboken, NJ, USA: John Wiley & Sons Inc; 2009.

42. Rosse RB, Deutsch SI: Use of the Global Assessment of Functioning scale in the VHA: moving toward improved precision. Veterans Health Syst J 2000, 5:50-58.

43. Breslow RA, Ross SA, Weed DL: Quality of reviews in epidemiology. Am J Public Health 1998, 88:475-477.

44. Cooper H: Synthesizing Research. A guide for literature reviews Thousand Oaks, CA, USA: Sage Publications; 1998.

45. Garrard J: Health Sciences Literature Review Made Easy. The Matrix Method Sudbury, MA, USA: Jones and Bartlett Publishers; 2007.

46. Hart C: Doing a Literature Review. Releasing the Social Science Research Imagination London, UK: Sage Publications Ltd; 1998.

47. Oxman AD: Systematic reviews: checklists for review articles. BMJ 1994, 309: 648-651.

48. Egger M, Jüni P, Bartlett C, Holenstein F, Sterne J: How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Health Technol Assess 2003, 7:1-76.

49. Shojania KG, Bero LA: Taking advantage of the explosion of systematic reviews:an efficient MEDLINE search strategy. Eff Clin Pract 2001, 4:157-162.

50. Endicott J, Spitzer RL, Fleiss JL, Cohen J: The Global Assessment Scale, a procedure for measuring overall severity of psychiatric disturbance. Arch Gen Psychiatry 1976, 33:766-771.

51. Karterud S, Pedersen G, Løvdal H, Friis S S-GAF: Global Funksjonsskåring - Splittet Versjon [Global Assessment of Functioning - Split version]. Bakgrunn og skåringsveiledning Oslo, Norway: Klinikk for Psykiatri, Ullevål sykehus; 1998.

52. Kennedy JA: Mastering the Kennedy Axis V. A new psychiatric assessment of patient functioning Washington DC, USA: American Psychiatric Publishing, Inc; 2003.

53. Poole R, Higgo R: Psychiatric Interviewing and Assessment Cambridge, UK: Cambridge University Press; 2006.

54. Foxcroft CD: Reflections on implementing the ITC ' s international guidelines for test use. Int J Testing 2001, 1:235-244.

55. International Test Commission: International guidelines for test use. Int J Testing 2001, 1 :93-113. 56. Bartram D: The need for international guidelines on standards for test use:a review of European and international initiatives. Eur Psychol 1998, 3:155-163.

57. Rey JM, Starling J, Weaver C, Dossetor DR, Plapp JM: Inter-rater reliability of global assessment of functioning in a clinical setting. J Child Psychol Psychiatry 1995, 36:787-792.

58. McColl E, Jacoby A, Thomas L, Soutter J, Bamford C, Steen N, Thomas R, Harvey E, Garratt A, Bond J: Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. Health Technol Assess 2001, 5:1-256.

59. Goodman R, Iervolino AC, Collishaw S, Pickles A, Maughan B: Seemingly minor changes to a questionnaire can make a big difference to mean scores: a cautionary tale. Soc Psychiatr Psychiatr Epidemiol 2007, 42:322-327.

60. First MB: Mastering DSM-IV Axis V. J Pract Psychiatry Behav Health 1995, 1:258-259.

61. Bech P, Malt UF, Dencker SJ, Ahlfors UG, Elgen K, Lewander T, Lundell A, Simpson GM, Lingjærde O: Scales for assessment of diagnosis and severity of mental disorders. Acta Psychiatr Scand 1993, 87(Suppl 372):3-86.

62. Hesse M, Rasmussen J, Pedersen MK: Standardised assessment of personality - a study of validity and reliability in substance abusers. BMC Psychiatry 2008, 8 :7.

63. Dworkin RJ, Friedman LC, Telschow RL, Grant KD, Moffic HS, Sloan VJ: The longitudinal use of the Global Assessment scale in multiple-rater situations. Community Ment Health J 1990, 26:335-444.

64. American Medical Association: Guides to the Evaluation ofPermanent Impairment. 2 edition. Chicago, IL, USA: American Medical Association; 1993.

65. Bowling A: Measuring Disease. A Review of Disease-Specific Quality of Life Measurement Scales Buckingham, UK: Open University Press; 1997.

66. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, Hergueta T, Baker R, Dunbar GC: The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic interview for DSM-IV and ICD-10. J Clin Psychiatry 1998, 59 (Suppl 20):22-33.

67. Zimmerman M: Diagnosing personality disorders. Arch Gen Psychiatry 1994, 51:225-245.

68. Greenberg GA, Rosenheck RA: Using the GAF as a national mental health outcome measure in the Department of Veterans Affairs. Psychiatr Serv 2005, 56:420-426.

69. Williams JBW, Gibbon M, First MB, Spitzer RL, Davis M, Borus J, Howes MJ, Kane J, Pope HG, Rounsaville B, Wittchen H-U: The structured clinical interview for DSM-III-R (SCID), II: multisite test-retest reliability. Arch Gen Psychiatry 1992, 49:630-636.

70. Mackinnon RA, Michels R, Buckley PJ: The Psychiatric Interview in Clinical Practice. 2 edition. Washington, DC, USA: American Psychiatric Publishing Inc; 2006.

71. Dixon S: Book review. Psychiatr Serv 2004, 55 :196-197. Aas Annals of General Psychiatry 2011, 10:2 http://www.annals-general-psychiatry.com/content/10/1/2 Page 10 of 11

72. Piersma HL, Boes JL: The GAF and psychiatric outcome: a descriptive report. Community Ment Health J 1997, 33:35-41.
73. Bowling A: Measuring Health. A Review of Quality of Life Measurement Scales Buckingham, UK: Open University Press; 1993.

74. Streiner DL, Norman GR: Health Measurement Scales. A Practical Guide to Their Development and Use Oxford, UK: Oxford University Press; 1994.

75. Andersson B-E: Som man frågar får man svar - en introduktion i intervju - och enkätteknik Kristianstad, Sween: Rabén Prisma; 1994.

76. Rogers R: Handbook of Diagnostic and Structured Interviewing New York, USA: The Guilford Press; 2001.

77. Lingjærde O, Bech P, Malt U, Dencker SJ, Elgen K, Ahlfors UG: Skalaer for diagnostikk og sykdomsgradering ved psykiatriske tilstander. Del 1: Metodologiske aspekter. Nord J Psychiatry 1989, 43(Suppl 19) :1-39.

78. Gregoire J, Hambleton RK: Advances in test adaptation research: a special issue. Int J Testing 2009, 9:75-77.

79. Van De Vijver F, Leung K: Methods and Data Analysis for Cross-cultural Research London, UK: Sage; 1997.

80. Lingjærde O, Bech P, Malt U, Dencker SJ, Elgen K, Ahlfors UG: Essentials of the World Psychiatric Association ʼ s International Guidelines

for Diagnostic Assessment (IGDA). Br J Psychiatry 2003, 182 (Suppl 45):s37-s57.

81. Hansagi H, Allebeck P: Enkät och intervju inom hälso - och sjukvård. Handbok för forskning och utvecklingsarbete Lund, Sweden: Studentlitteratur; 1994.

82. Del Castillo JC: The influence of language upon symptomatology in foreign-born patients. Am J Psychiatry 1970, 127:242-234.

83. Payer L: Medicine and culture. Notions of Health and Sickness in Britain, the US, France and West Germany London, UK: Victor Gollancz Ltd; 1989.

84. Solano-Flores G, Backhoff E, Contrea-Niño LA: Theory of test translation error. Int J Testing 2009, 9:78-91.

85. Bruyn EEJ: A normative-prescriptive view on clinical psychodiagnostic decision making. Eur J Psychol Assess 1992, 3:163-171.

86. Harel TZ, Smith DW, Rowles JM: A comparison of psychiatrists' clinical-impression-based and social workers' computer-generated GAF scores. Psychiatr Serv 2002, 53:340-342.

87. Kuhlman TL, Sincaban VA, Bernstein MJ: Team use of the Global Assessment scale for inpatient planning and evaluation. Hosp Community Psychiatry 1990, 41:416-19.

88. Naglieri JA: Psychometric issues in the assessment of impairment. In Assessing Impairment. Edited by: Goldstein S, Naglieri JA. New York, USA: Springer; 2009:49-57.

89. Coyne I, Bartram D: Design and development of the ITC guidelines on computer-based and Internet-delivered testing. Int J Testing 2006, 6:133-142.

90. Foxcroft CD, Davies C: Taking ownership of the ITC ' s guidelines on computer-based and Internet- delivered testing: a South African application. Int J Testing 2006, 6:173-80.

91. International Test Commission: International guidelines on computer-based and Internet-delivered testing. Int J Testing 2006, 6:143-171.

92. Lievens F: The ITC guidelines on computer-based and Internet-delivered testing: where do we go from here? Int J Testing 2006, 6:189-194.

93. Sale R: International guidelines on computer-based and Internet-delivered testing: a practitioner's perspective. Int J Testing 2006, 6:181-188.

94. Scheuerman F, Pereira AG: Towards a Research Agenda on Computer-based Assessment. Challenges and Needs for European Educational Measurement Luxembourg: European Commission, Joint Research Centre, Institute for the Protection and Security of the Citizen, European Communities; 2008.

95. Del Greco L, Eastridge L, Marchand B, Szentveri K: Questionnaire development: 4. Preparation for analysis. Can Med Assoc J 1987, 136:927-928.

96. Reed GM, McLaughlin CJ, Newman R: The development and evaluation of guidelines for professional practice. Am Psychol 2002, 57:1041-1047.

97. Bern DJ: Writing a review article for PsychologicalBulletin . Psychol Bull 1995, 118:172-177.

98. Conn VC, Isaramalai S, Rath S, Jantarakupt P, Wadhawan R, Dash Y: Beyond MEDLINE for literature searches. J Nurs Scholarsh 2003, 35:177-182.

99. Arnold SJ, Bender VF, Brown SA: A review and comparison of psychology- related electronic resources. J Elect Res Med Lib 2006, 3:61-79.

100. Bruyn EEJ: Assessment process. In Encyclopedia of Psychological Assessment. Edited by: Fernández-Ballesteros R. Thousand Oaks, CA, USA: Sage; 2003:93-97.

101. Forsner T, Wisted AÅ, Brommels M, Forsell Y: An approach to measure compliance to clinical guidelines in psychiatric care. BMC Psychiatry 2008, 8:64.

102. Hilsenroth MJ, Ackerman SJ, Blagys MD, Price JL: Dr Hilsenroth and colleagues reply. Am J Psychiatry 2001, 158 :1936-1937. 103. Pedersen G, (Ed): Personlighetsfortsyrrelser. Forståelse, evaluering, kombinert gruppebehandling Oslo, Norway: Pax Forlag; 2000, 237-239.

104. Spitzer RL, Forman JB: DSM-III field trials, II: initial experience with the multiaxial system. Am J Psychiatry 1979, 136:818-820.

105. Widiger TA, Clark LE: Toward DSM-V and the classification of psychopathology. Psychol Bull 2000, 126:946-963.