

03 - 5.3 Psychiatric Rating Scales

5.3 Psychiatric Rating Scales

government is a complaint-driven system in which the OCR will respond to complaints made by patients concerning confidentiality violations or denied access to records, all of which are covered under HIPAA. In such cases, OCR may follow up and audit compliance. The APA's Committee on Confidentiality, along with legal experts, has developed a set of sample forms. They are part of the APA's HIPAA educational packet, which can be obtained on the APA web site (www.psych.org/). On the web site, there are also recommendations for enabling physicians to comply with HIPAA.

REFERENCES Dougall N, Lambert P, Maxwell M, Dawson A, Sinnott R, McCafferty S, Springbett A. Deaths by suicide and their relationship with general and psychiatric hospital discharge: 30-year record linkage study. *Br J Psychiatry* . 2014;204(4). Simon RI. *Clinical Psychiatry and the Law*. American Psychiatric Pub; 2003.

5.3 Psychiatric Rating Scales The term psychiatric rating scales encompasses a variety of questionnaires, interviews, checklists, outcome assessments, and other instruments that are available to inform psychiatric practice, research, and administration.

Psychiatrists must keep up with major developments in rating scales for several reasons. Most critically, many such scales are useful in psychiatric practice for monitoring patients over time or for providing information that is more comprehensive than what is generally obtained in a routine clinical interview. In addition, health care administrators and payors are increasingly requiring standardized assessments to justify the need for services or to assess quality of care. Lastly, but equally important, rating scales are used in research that informs the practice of psychiatry, so familiarity with them provides a deeper understanding of the results of that research and the degree to which it applies to psychiatric practice. **POTENTIAL BENEFITS AND LIMITATIONS OF RATING SCALES IN PSYCHIATRY** The key role of rating scales in psychiatry and elsewhere is to standardize the information collected across time and by various observers. This standardization ensures a consistent, comprehensive evaluation that may aid treatment planning by establishing a diagnosis, ensuring a thorough description of symptoms, identifying comorbid conditions, and characterizing other factors affecting treatment response. In addition, the use of a rating scale can establish a baseline for follow up of the progression of an illness over time or in response to specific interventions. This is particularly useful when more than one clinician is involved—for instance, in a group practice or in the conduct of psychiatric research. In addition to standardization, most rating scales also offer the user the advantages of a formal evaluation of the measure's performance characteristics. This allows the

clinician to know to what extent a given scale produces reproducible results (reliability) and how it compares to more definitive or established ways of measuring the same thing (validity).

TYPES OF SCALES AND WHAT THEY MEASURE

Scales are used in psychiatric research and practice to achieve a variety of goals. They also cover a broad range of areas and use a broad range of procedures and formats.

Measurement Goals

Most psychiatric rating scales in common use fall into one or more of the following categories: making a diagnosis; measuring severity and tracking change in specific symptoms, in general functioning, or in overall outcome; and screening for conditions that may or may not be present.

Constructs Assessed

Psychiatric practitioners and investigators assess a broad range of areas, referred to as constructs, to underscore the fact that they are not simple, direct observations of nature. These include diagnoses, signs and symptoms, severity, functional impairment, quality of life, and many others. Some of these constructs are fairly complex and are divided into two or more domains (e.g., positive and negative symptoms in schizophrenia or mood and neurovegetative symptoms in major depression).

Categorical versus Continuous Classification

Some constructs are viewed as categorical or classifying, whereas others are seen as continuous or measuring. Categorical constructs describe the presence or absence of a given attribute (e.g., competency to stand trial) or the category best suited to a given individual among a finite set of options (e.g., assigning a diagnosis). Continuous measures provide a quantitative assessment along a continuum of intensity, frequency, or severity. In addition to symptom severity and functional status, multidimensional personality traits, cognitive status, social support, and many other attributes are generally measured continuously. The distinction between categorical and continuous measures is by no means absolute. Ordinal classification, which uses a finite, ordered set of categories (e.g., unaffected, mild, moderate, or severe) stands between the two.

Measurement Procedures

Rating scales differ in measurement methods. Issues to be considered include format, raters, and sources of information.

Format.

Rating scales are available in a variety of formats. Some are simply checklists or guides to observation that help the clinician achieve a standardized rating.

Others are self-administered questionnaires or tests. Still others are formal interviews that may be fully structured (i.e., specifying the exact wording of questions to be asked) or partly structured (i.e., providing only some specific wording, along with suggestions for additional questions or probes).

Raters.

Some instruments are designed to be administered by doctoral-level clinicians only, whereas others may be administered by psychiatric nurses or social workers with more limited clinical experience. Still other instruments are designed primarily for use by lay raters with little or no experience with psychopathology.

Source of Information.

Instruments also vary in the source of information used to make the ratings. Information may be obtained solely from the patient, who generally knows the most about his or her condition. In some instruments, some or all of the information may be obtained from a knowledgeable informant. When the construct involves limited insight (e.g., cognitive disorders or mania) or significant social undesirability (e.g., antisocial personality or substance abuse), other informants may be preferable. Informants may also be helpful when the subject has limited ability to recall or report symptoms (e.g., delirium, dementia, or any disorder in young children). Some rating scales also allow or require information to be included from medical records or from patient observation.

ASSESSMENT OF RATING SCALES

In clinical research, rating scales are mandatory to ensure interpretable and potentially generalizable results and are selected based on coverage of the relevant constructs, expense (based on the nature of the raters, purchase price if any, and necessary training), length and administration time, comprehensibility to the intended audience, and quality of the ratings provided. In clinical practice,

one considers these factors and, also, whether a scale would provide more or better information than what would be obtained in ordinary clinical practice or would contribute to the efficiency of obtaining that information. In either case, the assessment of quality is based on psychometric, or mind-measuring, properties. **Psychometric Properties** The two principal psychometric properties of a measure are reliability and validity. Although these words are used almost interchangeably in everyday speech, they are distinct in the context of evaluating rating scales. To be useful, scales should be reliable, or consistent and repeatable even if performed by different raters at different times or under different conditions, and they should be valid, or accurate in representing the true state of nature. **Reliability.** Reliability refers to the consistency or repeatability of ratings and is largely empirical. An instrument is more likely to be reliable if the instructions and

questions are clearly and simply worded and the format is easy to understand and score. There are three standard ways to assess reliability: internal consistency, interrater, and test-retest. **Internal Consistency.** Internal consistency assesses agreement among the individual items in a measure. This provides information about reliability, because each item is viewed as a single measurement of the underlying construct. Thus, the coherence of the items suggests that each is measuring the same thing. **Interrater and Test-Retest Reliability.** Interrater (also called interjudge or joint) reliability is a measure of agreement between two or more observers evaluating the same subjects using the same information. Estimates may vary with assessment conditions—for instance, estimates of interrater reliability based on videotaped interviews tend to be higher than those based on interviews conducted by one of the raters. Test-retest evaluations measure reliability only to the extent that the subject's true condition remains stable in the time interval. **Issues in Interpreting Reliability Data.** When interpreting reliability data, it is important to bear in mind that reliability estimates published in the literature may not generalize to other settings. Factors to consider are the nature of the sample, the training and experience of the raters, and the test conditions. Issues regarding the sample are especially critical. In particular, reliability tends to be higher in samples with high variability in which it is easier to discriminate among individuals. **Validity.** Validity refers to conformity with truth, or a gold standard that can stand for truth. In the categorical context, it refers to whether an instrument can make correct classifications. In the continuous context, it refers to accuracy, or whether the score assigned can be said to represent the true state of nature. Although reliability is an empirical question, validity is partly theoretical—for many constructs measured in psychiatry, there is no underlying absolute truth. Even so, some measures yield more useful and meaningful data than others do. Validity assessment is generally divided into face and content validity, criterion validity, and construct validity. **FACE AND CONTENT VALIDITY.** Face validity refers to whether the items appear to assess the construct in question. Although a rating scale may purport to measure a construct of interest, a review of the items may reveal that it embodies a very different conceptualization of the construct. For instance, an insight scale may define insight in either psychoanalytic or neurological terms. However, items with a transparent relationship to the construct may be a disadvantage when measuring socially undesirable traits, such as substance abuse or malingering. Content validity is similar to face validity but describes whether the measure provides good balanced coverage of the construct and is less focused on whether the items give the appearance of validity. Content validity is often assessed with formal procedures such as expert consensus or

factor analysis. **CRITERION VALIDITY.** Criterion validity (sometimes called predictive or concurrent validity) refers to whether or not the measure agrees with a gold standard or criterion of accuracy.

Suitable gold standards include the long form of an established instrument for a new, shorter version, a clinician-rated measure for a self-report form, and blood or urine tests for measures of drug use. For diagnostic interviews, the generally accepted gold standard is the Longitudinal, Expert, All Data (LEAD) standard, which incorporates expert clinical evaluation, longitudinal data, medical records, family history, and any other sources of information. **CONSTRUCT VALIDITY.** When an adequate gold standard is not available—a frequent state of affairs in psychiatry—or when additional validity data are desired, construct validity must be assessed. To accomplish this, one can compare the measure to external validators, attributes that bear a well-characterized relationship to the construct under study but are not measured directly by the instrument. External validators used to validate psychiatric diagnostic criteria and the diagnostic instruments that aim to operationalize them include course of illness, family history, and treatment response. For example, when compared with schizophrenia measures, mania measures are expected to identify more individuals with a remitting course, a family history of major mood disorders, and a good response to lithium. **SELECTION OF PSYCHIATRIC RATING SCALES** The scales discussed below cover various areas such as diagnosis, functioning, and symptom severity, among others. Selections were made based on coverage of major areas and common use in clinical research or current (or potential) use in clinical practice. Only a few of the many scales available in each category are discussed here. **Disability Assessment** One of the most widely used scales to measure disability was developed by the World Health Association (WHO), known as the WHO Disability Assessment Schedule, now in its second iteration (WHODAS 2.0). It is self-administered and measures disability along a number of parameters such as cognition, interpersonal relations, work and social impairment, among many others. It can be taken at intervals along the course of a person's illness and is reliable in tracking changes that indicate a positive or negative response to therapeutic interventions or course of illness (Table 5.3-1). Table 5.3-1 WHODAS 2.0

A number of assessment scales were developed for inclusion in the 5th edition of the Diagnostic and Statistical Manual of Mental Disorders of the American Psychiatric Association, (DSM-5); however, they were developed by and intended for use by research psychiatrists and are not as well tested as the WHO scales. It is expected that, in time, they will eventually be better adapted for clinical use. Some clinicians may wish to use the scales known as Cross-Cutting Symptom Measure Scales, but at this time the WHO scale is recommended for general use. **Psychiatric Diagnosis Instruments** assessing psychiatric diagnosis are central to psychiatric research and may be useful in clinical practice as well. However, they tend to be rather long, especially with individuals reporting many symptoms, potentially requiring many follow-up questions. When such instruments are evaluated, it is important to ensure they implement the current diagnostic criteria and cover the diagnostic areas of interest. **Structured Clinical Interview for DSM (SCID).** The SCID begins with a section on demographic information and clinical background. Then there are seven diagnostic modules focused on different diagnostic groups: mood, psychotic, substance abuse, anxiety, somatic, eating, and adjustment disorders; the modules can be administered separately. Both required and optional probes are provided, and skip outs are suggested where no further questioning is warranted. All available information, including that from hospital records, informants, and patient observation, should be used to rate the SCID. The SCID is designed to be administered by experienced clinicians and is generally not recommended for use by lay interviewers. In addition, formal training in the SCID is required, and training books and videos are available to facilitate this. Although the primary focus is research with psychiatric patients, a nonpatient version (with no reference to a chief complaint) and a more clinical version (without as much

detailed subtyping) are also available. Reliability data on the SCID suggest that it performs better on more severe disorders (e.g., bipolar disorder or alcohol dependence) than on milder ones (e.g., dysthymia). Validity data are limited, as the SCID is more often used as the gold standard to evaluate other instruments. It is considered the standard interview to verify diagnosis in clinical trials and is extensively used in other forms of psychiatric research. Although its length precludes its use in routine clinical practice, the SCID can sometimes be useful to ensure a systematic evaluation in psychiatric patients—for instance, on admission to an inpatient unit or at intake into an outpatient clinic. It is also used in forensic practice to ensure a formal and reproducible examination. Psychotic Disorders A variety of instruments are used for patients with psychotic disorders. Those discussed here are symptom severity measures. A developing consensus suggests that the

distinction between positive and negative symptoms in schizophrenia is worthwhile, and more recently developed instruments implement this distinction. Brief Psychiatric Rating Scale (BPRS). The BPRS (Table 5.3-2) was developed in the late 1960s as a short scale for measuring the severity of psychiatric symptomatology. It was developed primarily to assess change in psychotic inpatients and covers a broad range of areas, including thought disturbance, emotional withdrawal and retardation, anxiety and depression, and hostility and suspiciousness. Reliability of the BPRS is good to excellent when raters are experienced, but this is difficult to achieve without substantial training; a semistructured interview has been developed to increase reliability. Validity is also good as measured by correlations with other measures of symptom severity, especially those assessing schizophrenia symptomatology. The BPRS has been used extensively for decades as an outcome measure in treatment studies of schizophrenia; it functions well as a measure of change in this context and offers the advantage of comparability with earlier trials. However, it has been largely supplanted in more recent clinical trials by the newer measures described below. In addition, given its focus on psychosis and associated symptoms, it is only suitable for patients with fairly significant impairment. Its use in clinical practice is less well supported, in part because considerable training is required to achieve the necessary reliability. Table 5.3-2 Brief Psychiatric Rating Scale

Positive and Negative Syndrome Scale (PANSS). The PANSS was developed in the late 1980s to remedy perceived deficits in the BPRS in the assessment of positive and negative symptoms of schizophrenia and other psychotic disorders by adding additional items and providing careful anchors for each. The PANSS requires a clinician rater because considerable probing and clinical judgment are required. A semistructured interview guide is available. Reliability for each scale has been shown to be fairly high, with excellent internal consistency and interrater reliability. Validity also appears good based on correlation with other symptom severity measures and factor analytic validation of the subscales. The PANSS has become the standard tool for assessing clinical outcome in treatment studies of schizophrenia and other psychotic disorders and has been shown to be easy to administer reliably and sensitive to change with treatment. Its high reliability and good coverage of both positive and negative symptoms make it excellent for this purpose. It may also be useful for tracking severity in clinical practice, and its clear anchors make it easy to use in this setting. Scale for the Assessment of Positive Symptoms (SAPS) and Scale for the Assessment of Negative Symptoms (SANS). The SAPS and SANS (Tables 5.3-3 and 5.3-4) were designed to provide a detailed assessment of positive and negative

symptoms of schizophrenia and may be used separately or in tandem. SAPS assesses hallucinations, delusions, bizarre behavior, and thought disorder, and SANS assesses affective flattening, poverty of speech, apathy, anhedonia, and inattentiveness. The SAPS and SANS are mainly used to monitor treatment effects in clinical research. Table 5.3-3 Scale for the Assessment of Positive Symptoms (SAPS)

Table 5.3-4 Scale for the Assessment of Negative Symptoms (SANS)

Mood Disorders The domain of mood disorders includes both unipolar and bipolar disorder, and the instruments described here assess depression and mania. For mania, the issues are similar to those for psychotic disorders in that limited insight and agitation may hinder accurate symptom reporting, so clinician ratings including observational data are

generally required. Rating depression, on the other hand, depends, to a substantial extent, on subjective assessment of mood states, so interviews and self-report instruments are both common. Because depression is common in the general population and involves significant morbidity and even mortality, screening instruments— especially those using a self-report format—are potentially quite useful in primary care and community settings. Hamilton Rating Scale for Depression (HAM-D). The HAM-D was developed in the early 1960s to monitor the severity of major depression, with a focus on somatic symptomatology. The 17-item version is the most commonly used version, although versions with different numbers of items, including the 24-item version in Table 5.3-5, have been used in many studies as well. The 17-item version does not include some of the symptoms for depression in DSM-III and its successors, most notably the so-called reverse neurovegetative signs (increased sleep, increased appetite, and psychomotor retardation). The HAM-D was designed for clinician raters but has been used by trained lay administrators as well. Ratings are completed by the examiner based on the patient interview and observations. A structured interview guide has been developed to improve reliability. The ratings can be completed in 15 to 20 minutes. Reliability is good to excellent, particularly when the structured interview version is used. Validity appears good based on correlation with other depression symptom measures. The HAM-D has been used extensively to evaluate change in response to pharmacological and other interventions and, thus, offers the advantage of comparability across a broad range of treatment trials. It is more problematic in the elderly and the medically ill, in whom the presence of somatic symptoms may not be indicative of major depression. Table 5.3-5 Hamilton Rating Scale for Depression

Beck Depression Inventory (BDI). The BDI was developed in the early 1960s to rate depression severity, with a focus on behavioral and cognitive dimensions of depression. The current version, the Beck-II, has added more coverage of somatic symptoms and covers the most recent 2 weeks. Earlier versions are focused on the past week or even shorter intervals, which may be preferable for monitoring treatment response. The scale can be completed in 5 to 10 minutes. Internal consistency has been high in numerous studies. Test-retest reliability is not consistently high, but this may reflect changes in underlying symptoms. Validity is supported by correlation with other depression measures. The principal use of the BDI is as an outcome measure in clinical trials of interventions for major depression, including psychotherapeutic interventions. Because it is a self-report instrument, it is sometimes used to screen for major depression. **Anxiety Disorders** The anxiety disorders addressed by the measures below include panic disorder, generalized anxiety

disorder, posttraumatic stress disorder (PTSD), and obsessivecompulsive disorder (OCD). When anxiety measures are examined, it is important to be aware that there have been significant changes over time in how anxiety disorders are defined. Both panic and OCD are relatively recently recognized, and the conceptualization of generalized anxiety disorder has shifted over time. Thus, older measures have somewhat less relevance for diagnostic purposes, although they may identify symptoms causing considerable distress. Whether reported during an interview or on a self-report rating scale, virtually all measures in this domain, like the measures of depression discussed above, depend on subjective descriptions of inner states.

Hamilton Anxiety Rating Scale (HAM-A). The HAM-A (Table 5.3-6) was developed in the late 1950s to assess anxiety symptoms, both somatic and cognitive. Because the conceptualization of anxiety has changed considerably, the HAM-A provides limited coverage of the “worry” required for a diagnosis of generalized anxiety disorder and does not include the episodic anxiety found in panic disorder. A score of 14 has been suggested as the threshold for clinically significant anxiety, but scores of 5 or less are typical in individuals in the community. The scale is designed to be administered by a clinician, and formal training or the use of a structured interview guide is required to achieve high reliability. A computer-administered version is also available. Reliability is fairly good based on internal consistency, interrater, and test-retest studies. However, given the lack of specific anchors, reliability should not be assumed to be high across different users in the absence of formal training. Validity appears good based on correlation with other anxiety scales but is limited by the relative lack of coverage of domains critical to the modern understanding of anxiety disorders. Even so, the HAM-A has been used extensively to monitor treatment response in clinical trials of generalized anxiety disorder and may also be useful for this purpose in clinical settings. Table 5.3-6 Hamilton Anxiety Rating Scale

Panic Disorder Severity Scale (PDSS). The PDSS was developed in the 1990s as a brief rating scale for the severity of panic disorder. It was based on the Yale-Brown Obsessive-Compulsive Scale and has seven items, each of which is rated on an itemspecific, 5-point Likert scale. The seven items address frequency of attacks, distress associated with attacks, anticipatory anxiety, phobic avoidance, and impairment. Reliability is excellent based on interrater studies, but, in keeping with the small number of items and multiple dimensions, internal consistency is limited. Validity is supported by correlations with other anxiety measures, both at the total and item levels; lack of correlation with the HAM-D; and, more recently, by brain imaging studies. Growing experience with the PDSS suggests that it is sensitive to change with treatment and is useful as a change measure in clinical trials or other outcome studies for panic disorder, as well as for monitoring panic disorder in clinical practice. Clinician-Administered PTSD Scale (CAPS). The CAPS includes 17 items required to make the diagnosis, covering all four criteria: (1) the event itself, (2) reexperiencing of the event, (3) avoidance, and (4) increased arousal. The diagnosis requires evidence of a traumatic event, one symptom of re-experiencing, three of avoidance, and two of arousal (typically, an item is counted if frequency is rated at least 1 and intensity is at least 2). The items can also be used to generate a total PTSD severity score obtained by summing the frequency and intensity scales for each item. The CAPS also includes several global rating scales for the impact of PTSD symptomatology on social and occupational functioning, for general severity, for recent changes, and for the validity of the patient’s report. The CAPS must be administered by a trained clinician and requires 45 to 60 minutes to complete, with follow-up examinations somewhat briefer. It has demonstrated reliability and validity in multiple settings and multiple languages, although it has

had more limited testing in the setting of sexual and criminal assault. It performs well in the research setting for diagnosis and severity assessment but is generally too lengthy for use in clinical practice. Yale-Brown Obsessive-Compulsive Scale (YBOCS). The YBOCS was developed in the late 1980s to measure the severity of symptoms in OCD. It has ten items rated based on a semistructured interview. The first five items concern obsessions: the amount of time that they consume, the degree to which they interfere with normal functioning, the distress that they cause, the patient's attempts to resist them, and the patient's ability to control them. The remaining five items ask parallel questions about compulsions. The semistructured interview and ratings can be completed in 15 minutes or less. A self-administered version has recently been developed and can be completed in 10 to 15 minutes. Computerized and telephone use have also been found to provide acceptable ratings. Reliability studies of the YBOCS show good internal consistency, interrater reliability, and test-retest reliability over a 1-week interval. Validity appears good, although data are fairly limited in this developing field. The YBOCS has become the standard instrument for assessing OCD severity and is used in virtually every drug trial. It may also be used clinically to monitor treatment response.

Substance Use Disorders Substance use disorders include abuse and dependence on both alcohol and drugs. These disorders, particularly those involving alcohol, are common and debilitating in the general population, so screening instruments are particularly helpful. Because these behaviors are socially undesirable, underreporting of symptoms is a significant problem; thus, the validity of all substance use measures is limited by the honesty of the patient. Validation against drug tests or other measures is of great value, particularly when working with patients who have known substance abuse. **CAGE**. The CAGE was developed in the mid-1970s to serve as a very brief screen for significant alcohol problems in a variety of settings, which could then be followed up by clinical inquiry. CAGE is an acronym for the four questions that comprise the instrument: (1) Have you ever felt you should Cut down on your drinking? (2) Have people Annoyed you by criticizing your drinking? (3) Have you ever felt bad or Guilty about your drinking? (4) Have you ever had a drink first thing in the morning to steady your nerves or to get rid of a hangover (Eye-opener)? Each "yes" answer is scored as 1, and these are summed to generate a total score. Scores of 1 or more warrant follow up, and scores of 2 or more strongly suggest significant alcohol problems. The instrument can be administered in a minute or less, either orally or on paper. Reliability has not been formally assessed. Validity has been assessed against a clinical diagnosis of alcohol abuse or dependence, and these four questions perform surprisingly well. Using a threshold score of 1, the CAGE achieves excellent sensitivity and fair to good specificity. A threshold of 2 provides still greater specificity but at the cost of a drop in sensitivity. The CAGE performs well as an extremely brief screening instrument for use in primary care or in psychiatric practice focused on problems unrelated to alcohol. However, it has limited ability to pick up early indicators of problem drinking that might be the focus of preventive efforts. **Addiction Severity Index (ASI)**. The ASI was developed in the early 1980s to serve as a quantitative measure of symptoms and functional impairment due to alcohol or drug disorders. It covers demographics, alcohol use, drug use, psychiatric status, medical status, employment, legal status, and family and social issues. Frequency, duration, and severity are assessed. It includes both subjective and objective items reported by the patient and observations made by the interviewer. **Eating Disorders** Eating disorders include anorexia nervosa, bulimia, and binge-eating disorder. A wide variety of instruments, particularly self-report scales, are available. Because of the secrecy that may surround dieting, bingeing, purging, and other symptoms, validation against other indicators (e.g., body weight for anorexia or dental examination

for bulimia) may be very helpful. Such validation is particularly critical for patients with anorexia, who may lack insight into their difficulties.

Eating Disorders Examination (EDE). The EDE was developed in 1987 as the first interviewer-based comprehensive assessment of eating disorders, including diagnosis, severity, and an assessment of subthreshold symptoms. A self-report version (the EDE-Q) as well as an interview for children have since been developed. The EDE focuses on symptoms during the preceding 4 weeks, although longer-term questions are included to assess diagnostic criteria for eating disorders. Each item on the EDE has a required probe with suggested follow-up questions to judge severity, frequency, or both, which are then rated on a 7-point Likert scale. For the self-report version, subjects are asked to make similar ratings of frequency or severity. The instrument provides both global severity ratings and ratings on four subscales: restraint, eating concern, weight concern, and shape concern. The interview, which must be administered by a trained clinician, requires 30 to 60 minutes to complete, whereas the self-report version can be completed more quickly. Reliability and validity data for both the EDE and EDE-Q are excellent, although the EDE-Q may have greater sensitivity for binge-eating disorder. The EDE performs well in both the diagnosis and the detailed assessment of eating disorders in the research context. It also has the sensitivity to change as is required for use in clinical trials or monitoring of individual therapy. Even in the research setting, however, the EDE is fairly lengthy for repeated use, and the EDE-Q may be preferable for some purposes. Although the EDE is too lengthy for routine clinical practice, the EDE or EDE-Q might be helpful in providing a comprehensive assessment of a patient with a suspected eating disorder, particularly during an evaluation visit or on entry into an inpatient facility.

Bulimia Test-Revised (BULIT-R). The BULIT-R was developed in the mid-1980s to provide both a categorical and a continuous assessment of bulimia. Patients with bulimia typically score above 110, whereas patients without disordered eating typically score below 60. The instrument can be completed in approximately 10 minutes. The BULIT-R shows high reliability based on studies of internal consistency and test-retest reliability in multiple studies. Validity is supported by high correlations with other bulimia assessments. The recommended cutoff of 104 suggested to identify probable cases of bulimia shows high sensitivity and specificity for a clinical diagnosis of bulimia nervosa. With cutoffs between 98 and 104, the BULIT-R has been used successfully to screen for cases of bulimia nervosa. As with any screening procedure, follow-up by clinical examination is indicated for individuals scoring positive; clinical follow-up is particularly critical because the BULIT-R does not distinguish clearly between different types of eating disorders. The BULIT-R may also be useful in clinical and research practice to track symptoms over time or in response to treatment, although more detailed measures of the frequency and severity of bingeing and purging may be preferable in research settings.

Cognitive Disorders A wide variety of measures of dementia are available. Most involve cognitive testing

and provide objective, quantifiable data. However, scores vary by educational level in subjects without dementia, so these instruments tend to be most useful when the patient's own baseline scores are known. Other measures focus on functional status, which can be assessed based on a comparison with a description of the subject's baseline function; these types of measures generally require a knowledgeable informant and, thus, may be more cumbersome to administer but tend to be less subject to educational biases. A third type of measure focuses on the associated behavioral symptoms that are frequently seen in demented patients. Mini-Mental State Examination (MMSE). The MMSE is a 30-point cognitive test developed in the mid-1970s to provide a bedside assessment

of a broad array of cognitive function, including orientation, attention, memory, construction, and language. It can be administered in less than 10 minutes by a busy doctor or a technician and scored rapidly by hand. The MMSE has been extensively studied and shows excellent reliability when raters refer to consistent scoring rules. Validity appears good based on correlations with a wide variety of more comprehensive measures of mental functioning and clinicopathological correlations. Since its development in 1975, the MMSE was widely distributed in textbooks, pocket guides, and on web sites and has been used at the bedside. In 2001 the authors granted a worldwide exclusive license to Psychological Assessment Resources (PAR) to publish, distribute, and manage all intellectual property rights to the test. A licensed version of the MMSE must now be purchased from PAR per test. The MMSE form is gradually disappearing from textbooks, web sites, and clinical tool kits. In an article in the *New England Journal of Medicine* (2011;365:2447-2449) John C. Newman and Robin Feldman concluded: "The restrictions on the MMSE's use present clinicians with difficult choices: increase practice costs and complexity, risk copyright infringement, or sacrifice 30 years of practical experience and validation to adopt new cognitive assessment tools." Neuropsychiatric Inventory (NPI). The NPI was developed in the mid-1990s to assess a wide range of behavioral symptoms that are often seen in Alzheimer's disease and other dementing disorders. The current version rates 12 areas: delusions, hallucinations, dysphoria, anxiety, agitation/aggression, euphoria, disinhibition, irritability/lability, apathy, aberrant motor behavior, nocturnal disturbances, appetite and eating. The standard NPI is an interview with a caregiver or other informant that can be performed by a clinician or trained lay interviewer and requires 15 to 20 minutes to complete. There is also a nursing home interview version, the NPI-NH, and a self-report questionnaire, the NPI-Q. For each area, the NPI asks whether a symptom is present and, if so, assesses frequency, severity, and associated caregiver distress. The instrument has demonstrated reliability and validity and is useful to screen for problem behaviors in both clinical and research settings. Because of the detailed frequency and severity ratings, it is also useful to monitor change with treatment. Scored General Intelligence Test (SGIT). This test was developed and

validated by N. D. C. Lewis at the New York State Psychiatric Institute in the 1930s. It is one of the few tests that attempts to measure general intelligence that can be administered by the clinician during the psychiatric interview. A decline in general intelligence will be seen in cognitive disorders, and the SGIT can alert the clinician to begin a workup for disease states that interfere with cognition. This test deserves more widespread use (Table 5.3-7). Table 5.3-7 Scored General Intelligence Test (SGIT) Personality Disorders and Personality Traits Personality may be conceptualized categorically as personality disorders or dimensionally as personality traits, which may be viewed as normal or pathological.

The focus here is on personality disorders and the maladaptive traits generally viewed as their milder forms. There are ten personality disorders that are divided into three clusters. Patients tend not to fall neatly into DSM personality categories; instead, most patients who meet the criteria for one personality disorder also meet the criteria for at least one other, particularly within the same cluster. This and other limitations in the validity of the constructs themselves make it difficult to achieve validity in personality measures. Personality measures include both interviews and self-report instruments. Self-report measures are appealing in that they require less time and may appear less threatening to the patient. However, they tend to overdiagnose personality disorders. Because many of the symptoms suggesting personality problems are socially undesirable and

because patients' insight tends to be limited, clinician-administered instruments, which allow for probing and patient observation, may provide more accurate data. Personality Disorder Questionnaire (PDQ). The PDQ was developed in the late 1980s as a simple self-report questionnaire designed to provide categorical and dimensional assessment of personality disorders. The PDQ includes 85 yes-no items designed primarily to assess the diagnostic criteria for personality disorders. Within the 85 items, two validity scales are embedded to identify underreporting, lying, and inattention. There is also a brief clinician-administered Clinical Significance Scale to address the impact of any personality disorder identified by the self-report PDQ. The PDQ can provide categorical diagnoses, a scaled score for each, or an overall index of personality disturbance based on the sum of all of the diagnostic criteria. Overall scores range from 0 to 79; normal controls tend to score below 20, personality disordered patients generally score above 30, and psychotherapy outpatients without such disorders tend to score in the 20 to 30 range. Childhood Disorders A wide variety of instruments are available to assess mental disorders in children. Despite this rich array of instruments, however, the evaluation of children remains difficult for several reasons. First, the child psychiatric nosology is at an earlier stage of development, and construct validity is often problematic. Second, because children change markedly with age, it is virtually impossible to design a measure that covers children of all ages. Lastly, because children, particularly young children, have limited ability to report their symptoms, other informants are necessary. This often creates problems because there are frequent disagreements among child, parent, and teacher reports of symptoms, and the optimal way to combine information is unclear. Child Behavior Checklist (CBCL). The CBCL is a family of self-rated instruments that survey a broad range of difficulties encountered in children from preschool through adolescence. One version of the CBCL is designed for completion by parents of children between 4 and 18 years of age. Another version is available for parents of children between 2 and 3 years of age. The Youth Self-Report is completed by

children between 11 and 18 years of age, and the Teacher Report Form is completed by teachers of school-age children. The scale includes not only problem behaviors, but also academic and social strengths. Each version includes approximately 100 items scored on a 3-point Likert scale. Scoring can be done by hand or computer, and normative data are available for each of the three subscales: problem behaviors, academic functioning, and adaptive behaviors. A computerized version is also available. The CBCL does not generate diagnoses but, instead, suggests cutoff scores for problems in the "clinical range." Parent, teacher, and child versions all show high reliability on the problem subscale, but the three informants frequently do not agree with one another. The CBCL may be useful in clinical settings as an adjunct to clinical evaluation, as it provides a good overall view of symptomatology and may also be used to track change over time. It is used frequently for similar purposes in research involving children and, thus, can be compared with clinical experience. The instrument does not, however, provide diagnostic information, and its length limits its efficiency for tracking purposes. Diagnostic Interview Schedule for Children (DISC). The current DISC, the DISC-IV, covers a broad range of DSM diagnoses, both current and lifetime. It has nearly 3,000 questions but is structured with a series of stem questions that serve as gateways to each diagnostic area, with the remainder of each section skipped if the subject answers no. Subjects who enter each section have very few skips, so complete diagnostic and symptom scale information can be obtained. Child, parent, and teacher versions are available. Computer programs are available to implement diagnostic criteria and generate severity scales based on each version or to combine parent and child information. A typical DISC interview may take more than 1 hour for

a child, plus an additional hour for a parent. However, because of the stem question structure, the actual time varies widely with the number of symptoms endorsed. The DISC was designed for lay interviewers. It is fairly complicated to administer, and formal training programs are highly recommended. Reliability of the DISC is only fair to good and generally better for the combined child and parent interview. Validity judged against a clinical interview by a child psychiatrist is also fair to good—better for some diagnoses and better for the combined interview. The DISC is well tolerated by parents and children and can be used to supplement a clinical interview to ensure comprehensive diagnostic coverage. Because of its inflexibility, some clinicians find it uncomfortable to use, and its length makes it less than optimal for use in clinical practice.

However, it is used frequently in a variety of research settings. **Conners Rating Scales.** The Conners Rating Scales are a family of instruments designed to measure a range of childhood and adolescent psychopathology but are most commonly used in the assessment of attention-deficit/hyperactivity disorder (ADHD). The main uses of the Conners Rating Scales are in screening for ADHD in school or clinic populations and following changes in symptom severity over time; sensitivity to change in response to specific therapies has been demonstrated for most versions of the Conners Rating Scales. There are teacher, parent, and self-report (for adolescents) versions and

both short (as few as ten items) and long (as many as 80 items, with multiple subscales) forms. Reliability data are excellent for the Conners Rating Scales. However, the teacher and parent versions tend to show poor agreement. Validity data suggest that the Conners Rating Scales are excellent at discriminating between ADHD patients and normal controls. **Autism Diagnostic Interview-Revised (ADI-R).** The Autism Diagnostic Interview (ADI) was developed in 1989 as a clinical assessment of autism and related disorders. The ADI-R was developed in 2003 with an aim to provide a shorter instrument with better ability to discriminate autism from other developmental disorders. The instrument has 93 items, is designed for individuals with a mental age greater than 18 months, and covers three broad areas, consistent with the diagnostic criteria for autism: language and communication; reciprocal social interactions; and restricted, repetitive, and stereotyped behaviors and interests. There are three versions: one for lifetime diagnosis, one for current diagnosis, and one for patients under age 4 focused on an initial diagnosis. It must be administered by a clinician trained in its use and takes about 90 minutes to complete. When clinicians are properly trained, it has good to excellent reliability and validity but performs poorly in the setting of severe developmental disabilities. It is generally intended for the research setting when a thorough assessment of autism is required but may have use in clinical practice as well.

REFERENCES Aggarwal NK, Zhang XY, Stefanovics E, Chen da C, Xiu MH, Xu K, Rosenheck RA. Rater evaluations for psychiatric instruments and cultural differences: The positive and negative syndrome scale in China and the United States. *J Nerv Ment Dis.* 2012;200(9):814. Blacker D. Psychiatric rating scales. In: Sadock BJ, Sadock VA, Ruiz P, eds. *Kaplan & Sadock's Comprehensive Textbook of Psychiatry.* 9th ed. Philadelphia: Lippincott Williams & Wilkins; 2009:1032. Gearing RE, Townsend L, Elkins J, El-Bassel N, Osterberg L. Strategies to Predict, Measure, and Improve Psychosocial Treatment Adherence. *Harv Rev Psychiatry.* 2014;22:31–45. Gibbons RD, Weiss DJ, Pilkonis PA, Frank E, Moore T, Kim JB, Kupfer DJ. Development of a computerized adaptive test for depression. *Arch Gen Psychiatry.* 2012;69(11):1104. Leentjens AFG, Dujardin K, Marsh L, Richard IH, Starkstein SE, Martinez-Martin P. Anxiety rating scales in Parkinson's disease: A validation study of the Hamilton anxiety rating scale, the Beck anxiety inventory, and the hospital anxiety and depression scale. *Mov Disord.* 2011;26:407. McDowell I, Newell C. *Measuring Health: A Guide to Rating Scales and Questionnaires.* New York: Oxford University Press; 2006. Posner K, Brown GK,

Stanley B, Brent DA, Yershova KV, Oquendo MA, Currier GW, Melvin GA, Greenhill L, Shen S, Mann JJ. The Columbia–Suicide Severity Rating Scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. *Am J Psychiatry*. 2011;168:1266. Purgato M, Barbui C. Dichotomizing rating scale scores in psychiatry: A bad idea? *Epidemiol Psychiatric Sci*. 2013;22(1):17– 19. Rush J, First MB, Blacker D, eds. *Handbook of Psychiatric Measures*. 2nd ed. Washington, DC: American Psychiatric Press;

107.

Tolin DF, Frost RO, Steketee G. A brief interview for assessing compulsive hoarding: The Hoarding Rating Scale-Interview. *Psychiatry Rev*. 2010;178:147. Wilson KCM, Green B, Mottram P. Overview of rating scales in old age psychiatry. In: Abou-Saleh MT, Katona C, Kumar A, eds. *Principles and Practice of Geriatric Psychiatry*. 3rd ed. Hoboken, NJ: Wiley; 2011.

Revision #1

Created 2026-01-04 19:50:39 UTC by Omar Ayman

Updated 2026-01-04 19:50:39 UTC by Omar Ayman