European Psychiatry

www.cambridge.org/epa

Review/Meta-analyses

Cite this article: Demyttenaere K, Jaspers L (2020). Trends in (not) using scales in major depression: A categorization and clinical orientation. *European Psychiatry*, **63**(1), e91, 1–6 https://doi.org/10.1192/j.eurpsy.2020.87

Received: 26 May 2020
Revised: 12 August 2020
Accepted: 31 August 2020

Key words:
Assessment; depression; experience sampling; rating; scale

Author for correspondence:
Koen Demyttenaere,
E-mail: koen.demyttenaere@uzleuven.be

© The Author(s), 2020. Published by Cambridge University Press on behalf of European Psychiatry. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (http://creativecommons.org/licenses/by-nc-sa/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is included and the original work is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use.

Trends in (not) using scales in major depression: A categorization and clinical orientation

Koen Demyttenaere[1*] and Liesbeth Jaspers[2]

[1]Faculty of Medicine, Department of Neurosciences, Research Group Psychiatry and University Psychiatric Center, KU Leuven, Leuven, Belgium and [2]Medical Centre Sint Jozef, Munsterbilzen, Belgium

Abstract

Background. Standard depression rating scales like the Hamilton Depression Rating Scale and the Montgomery–Åsberg Depression Rating Scale were developed more than 40 years ago. They are mandatory in clinical trials but are for a variety of reasons seldom used in clinical practice. Moreover, most clinicians are less familiar with more recent trends or with some dilemmas in assessment tools for major depression.
Methods. Narrative review.
Results. Asssessment tools can be observer-rating or self-rating scales, disease-specific or non–disease-specific scales, subjective scales or objective lab assessments, standard questionnaires or experience sampling methods. An overarching question is to what degree current assessment methods really address the individual patient's needs and treatment expectations.
Conclusions. The present paper aims to offer a framework for understanding the current trends in assessment tools that can orientate and guide the clinician.

Introduction

Depression rating scales have acquired an indispensable role in clinical trials [1], in which they are used to select eligible patients and to assess changes in symptoms and in symptom intensity during treatment [2]. Depression treatment guidelines strongly recommend the use of measurement tools to monitor the course of treatment [3,4], while in some countries, health care providers even link the use of validated questionnaires to funding [5,6].

On the contrary, most clinicians do not use scales in everyday practice. In the United Kingdom, as much as 88.7% of psychiatrists never or occasionally use standardized measures in patients with depression or an anxiety disorder [7]. In the United States, 82% of psychiatrists never, rarely or only sometimes use scales to monitor outcome in depressed patients [8]. Some clinicians report doubts on the validity of available tools or fear that using scales is too time-consuming [8,9]. Others worry about potential (mis)use in the current management-benchmarking-ranking culture [10]. Developments as pay for quality could moreover guide clinicians to prioritize what can be measured, to consider unimportant what cannot be measured, and to direct organizational efforts toward what is easily quantified. Others consider themselves as insufficiently trained to apply scales correctly [8,9]. Many caregivers do trust more on their own clinical judgment while blaming the reductionist nature of scales, insufficiently able to display the complex state of their patients [7–9,11]. Max Hamilton already warned that rating a patient risks to fit him "into a Procrustean bed" [12] meaning that, as Procrustes amputated the limbs of his guests to adjust them to his bed, clinicians can ignore vital patient information because it does not correspond with the content of a scale.

Since the Hamilton Depression Rating Scale [12] and the Montgomery–Åsberg Depression Rating Scale [13], many other depression rating scales have been proposed: from observer-rating to self-rating scales, from disease-specific to non–disease-specific scales, from "subjective" questionnaires to "objective" lab assessments, from questionnaires to experience sampling. One overarching concern is that information delivered by scales is not always relevant to patients, families, and even to clinicians.

The present paper aims to summarize the trends in assessment tools for unipolar major depression in order to provide an orientating framework to the practicing clinician.

Methods

This paper is neither a compendium nor a systematic review of assessment scales for unipolar major depression. It is a selective review aiming to help the clinician/researcher in choosing a scale by providing an orientational framework wherein the existing scales can be positioned and categorized: observer-rating versus self-rating scales, disease-specific versus non–disease-specific scales, site rating versus centralized rating, "subjective" questionnaire rating versus "objective" (lab) assessment, and questionnaires versus experience sampling method (Table 1). This

Table 1. Categorization of assessment scales available for patients with unipolar depression.

Observer-rating scales * More time consuming * Open to observer bias * Larger effect sizes	**Self-rating scales** * Less time consuming * Open to patient bias * Smaller effect sizes
Disease-specific scales * Higher specificity for defined patient populations	**Non–disease-specific scales** * Allowing comparisons between different patient populations
Site rating * Open to investigator bias * Assessment within therapeutic relation	**Centralized rating** * Not open to investigator bias * Assessment outside therapeutic relation
Questionnaires * Work with what is introspected and remembered * More global impression	**Experience sampling methods (ESM)** * Work with what is captured in the moment * Repetitive measurement with more precision

framework is illustrated by papers based on Pubmed searches and is followed by an overarching comment on the relevance of these scales from a depressed patient perspective.

Observer-Rating versus Self-Rating Scales

A first positioning has to be made between observer-rating and self-rating scales. Observer-rating scales benefit from the experience of the rater, supposed to be free from patient bias [12,14], while self-rating scales are less time-consuming and supposed to be free from clinician bias [8,14].

The Hamilton Rating Scale for Depression (HAMD [12]) and the Montgomery–Asberg Depression Rating Scale (MADRS [13]) are the first and second most commonly used clinician rating scales in depression treatment studies [15]. Hamilton designed a tool to quantify results of clinical interviews in patients with established depression [16]. But since the HAMD has many anxiety and sleep items, the scale reflects the efficacy (and sedative side effects) of the tricyclics [17,18]. Sensitivity to change was at the origin of the development of the MADRS [13]. For the construction of the scale, the authors selected the 10 items of the much larger Comprehensive Psychiatric Rating Scale that changed most during treatment with various antidepressants. One can hence conclude that both the HAMD and the MADRS are "antidepressant friendly" scales. And since the HAMD merges depressive and anxious symptoms as well as neurovegetative symptoms, it seems to put all depressions into one basket: a more anxious depression or a depression with neurovegetative symptoms will both be more severe depressions [19]. Since the MADRS reflects the effects of a variety of antidepressants (with different mode of action), it seems to put all antidepressants in one basket and therefore cannot answer the question whether antidepressants with a different mode of action target different symptoms or different symptom clusters [19]).

Similar comment can be made on self-rating questionnaires. The Beck Depression Inventory (BDI) is a widely used self-rating instrument [20–22] focusing on cognitive symptoms and is therefore "cognitive behavioral psychotherapy friendly" [23].

Moreover, discrepancies can be found between how observer-rating and self-rating instruments detect change. Cuijpers et al. [24] compared the effect sizes generated by self-report scales and clinician-rated scales and found that clinician-rated instruments consistently result in significantly higher effect sizes than self-report instruments from the same studies ($\Delta g = 0.20$; 95%, CI 0.10–0.30). On the contrary, Zimmerman found overall comparable effect sizes and percentage of responders (≥50% reduction in baseline scores) in routine clinical practice (away from a "sponsored" study context) [25].

Interestingly, discrepancies are also found in observer-rated and self-rated versions of the same scale (MADRS vs. MADRS-S). In a randomized controlled trial comparing escitalopram and citalopram, responses were lower using the self-rating version than on the clinician-rating version (response rate on MADRS-S: 66.4 and 53.9% for escitalopram and citalopram, respectively [$p = 0.043$], vs. 76.1 and 61.5% on the MADRS [$p = 0.009$]) [26].

Disease-Specific versus Non–Disease-Specific Scales

A second positioning has to be made between disease-specific scales that focus on disease-specific symptoms, and non–disease-specific scales that assess the "overall" impression of clinical status or "overall" impression of clinical change.

Within the so-called disease-specific scales (for major depression), some authors advocate the use of scales with an even higher specificity for specific subpopulations: more age specific (e.g., the Geriatric Depression Scale [27]), more psychiatric comorbidity specific (e.g., the Calgary Depression Scale for Schizophrenia [28]), more somatic comorbidity specific (e.g., the Post-Stroke Depression Rating Scale [29]), and more life phase specific (e.g., the Meno-D for perimenopausal depression [30]).

The Clinical Global Impression (CGI) scale was originally developed to provide a brief, stand-alone assessment of the clinician's view of the patient's global functioning prior to and after initiating a study medication [31]. The CGI is concise and simple: it is a non–disease-specific tool that measures global illness severity (CGI-S) and global improvement (CGI-I). The CGI-S is rated with scores from 1 (normal) through to 7 (among the most severely ill patients). The CGI-I is also rated with scores from 1 (very much improved) through 7 (very much worse) [31]. In the past years, the need for instruments with similar user friendliness but with improved inter-rater reliability has led to a partial return to more disease-specific and transdiagnostic versions of the CGI [32–36].

There is ongoing controversy about what is a clinically meaningful change in score on a rating scale: response (a 50% reduction of the baseline score or "much improved" or "very much improved") or remission (a score below a cut-off value or "very much improved") [37]. The question remains whether non–disease-specific scales differ in their ability to detect meaningful change in the condition of patients and to what degree they depend upon baseline severity of depression.

Investigators were asked to rank-order elements that determined their CGI scores: symptom severity and functional status were the two most important drivers, and strikingly less importance was given to self-report symptoms scores [38,39] indicating low attention to the patient perspective.

In 2016, Bobo et al. equated HAMD-17 response percentages with CGI-I scores in antidepressant trials and confirmed the consensus definition of response on standard scales (50% improvement): "much improved" ratings (CGI-I responders) corresponded with 50–57% improvement. Differentiating one step further, absolute changes in HAMD-17 and CGI-I scores have been compared in patients with higher or lower depression severity at baseline. Patients with higher depression severity needed a decrease of 13–14 points to be considered "much improved," while the lower

severity group only needed a nine-point decrease [40,41]. This effect disappeared when the relative change on HAMD scores was considered. The more severe the depression severity, the larger should be the improvement before the clinician decides on a "much improved" status [41].

Site Rating versus Centralized Rating

At least in clinical research, a third positioning has to be made between site rating and centralized rating. The development of centralized rating tried to overcome the problem of many failed or negative pharmacological trials. One of the contributing factors of trial failure is measurement methodology: poor interrater reliability leading to smaller between-groups effect sizes, baseline score inflation, and rater expectancy effects leading to decreased signal detection [42].

Centralized rating deploys highly skilled, site-independent raters, who asses patients through video- or teleconferencing [43–45], and they are blinded for inclusion criteria, study visit, and study site location. The comparison of these two assessment modalities (centralized vs. site rating) learned that 35% of the study subjects (included by the site raters) would not have entered the study (by the centralized raters). Moreover, site raters found significantly more placebo responders than central raters did (respectively, 28% vs. 14%, $p < 0.001$). Finally, this difference in placebo response between site raters and central raters disappeared when the analysis was conducted in the 65% of patients that would have been included by both site and central raters [44].

Targum and colleagues added the modality of self-rating to the comparison of site and central rating in three arms with placebo, 15 mg buspirone, or a combination of buspirone 15 mg and melatonine 3 mg. The difference in response rates between the combination treatment (buspirone and melatonine) and placebo was 15.9% when done by site raters and 7.1% when done by central raters. However, these differences between the two treatment arms increased (19.4% instead of 15.9% when done by site raters and 15.2% instead of 7.1% when done by centralized raters) when a "dual scoring" method was used: that is, excluding patients who at baseline had remarkably discordance (more than 1 standard deviation from baseline means) between site raters and central raters. The "dual scoring" method resulted in higher treatment response rates and lower placebo response rates (resp. 48.6% vs. 29.2% in site ratings, and resp. 48.57% vs. 33.33% in central ratings) suggesting that more advanced rating methodology could be useful in future clinical trials [45].

Subjective Questionnaire Rating versus Objective (Lab) Assessment

A fourth positioning has to be made between more subjective questionnaire rating and more objective lab assessment. Some more biological-oriented psychiatrists blame the field for the lack of objective parameters while expressing their suspicion toward the subjectivity of rating scales and hope for biological measures (blood tests, imaging, genetics, etc.). More psychotherapeutically oriented psychiatrists on the contrary are convinced that the essence of psychotherapy is in working with subjectivity. A somewhat intermediate trend is to complement questionnaires with more objective lab testing.

One example of the differentiation between subjective and objective rating has been investigated in the assessment of cognitive symptoms in depression. One assessment method is the Perceived Deficits Questionnaire (PDQ), a brief screening instrument designed to measure perceived cognitive impairment (originally in patients with multiple sclerosis. This questionnaire comprises four subscales: attention/concentration, prospective memory, planning/organization, and retrospective memory [46]. Another assessment method is more objective testing like the Digit Symbol Substitution Test supposed to assess executive functioning, psychomotor speed, attention, and memory [47], or like the Rey Auditory Verbal Learning Test supposed to assess acquisition and delayed recall [48]. We use the wording "supposed to assess" since basic motivation or giving up at failure always interfere with these so-called objective cognitive tests. A marked correlation was found between subjectively perceived cognitive deficits on the PDQ and both depression and self-efficacy scores but no relationship with objective cognitive performance [49]. A similar effect was seen in remitted unipolar and bipolar patients, where subjective cognitive dysfunction was correlated with depression severity but was not differentiating between unipolar and bipolar patients; this contradicts objective cognitive assessments generally showing a greater dysfunction in bipolar disorder [50]. These findings suggest that subjective ratings of cognitive functioning are more strongly influenced by mood symptoms than objective ratings of cognitive functioning. Attempts have been made to disentangle the cognitive and the other depressive symptoms in a vortioxetine trial where path analysis showed that part of the subjective/objective cognitive improvement was independent from the improvement in depressive symptom severity [51]. This suggests that for both subjective and objective measures of cognitive functioning, cognitive improvement can be disentangled from the improvement in the other depressive symptoms like lack of motivation or lack of energy.

Another example of the differentiation between subjective and objective rating has been investigated in the assessment of anhedonia. Anhedonia is a core symptom of depression, maybe even the most specific depressive symptom, but receives remarkably poor attention in standard observer scales as HAMD-17 or MADRS. In both scales, only one item is (partially) dedicated to anhedonia. To address this deficiency, scales that focus on the assessment of hedonic tone in depression such as the Snaith–Hamilton Pleasure scale (SHAPS [52]), the Temporal Experience of Pleasure Scale (TEPS [53]), and Leuven Affect and Pleasure Scale (LAPS [54]) have been developed. These self-report scales try to cover the multidimensional concept of anhedonia. The SHAPS assesses both sensory and social anhedonia but offers no differentiation between anticipatory and consummatory elements. The TEPS does address these aspects but solely for sensory anhedonia while the LAPS covers all dimensions.

Some researchers in the cognitive field moved away from assessing anhedonia with subjective questionnaires to develop more objective, laboratory-based anhedonia measures [55–58]. They operationalize hedonic capacity as responsiveness to reinforcing stimuli, assessed by a signal detection task. Pizzagalli, for instance, uses a signal detection task generating a differential monetary reward after correct identification of one of two possible stimuli. Normally, subjects develop a preference (bias) to the stimulus that is associated with more frequent awards. Absence of a response bias was found in participants with elevated depressive symptoms [58] and in patients with major depressive disorder [57]. Only moderate differences were found on the BDI melancholic subscore of the BDI anhedonia subscore for subjects showing a positive or negative response bias showing that the "objective" test results only partially overlap with the "subjective" test results.

Questionnaires versus Experience Sampling Method

A fifth positioning has to be made between questionnaires assessing mood states during a certain time interval and experience sampling assessing and aggregating mood states based upon multiple time points per day. Standard depression rating scales have the problem of a time frame: how could depressed patients who tend to (over) generalize be able to correctly report how they felt during the past week or during the past 2 weeks? This resulted in the development of the experience sampling method (ESM), aiming to assemble information of subjective experience of patients via collection of self-reports on activities, emotions, or other elements of daily life at various points throughout the day. ESM is considered as a more sophisticated version of the diary approach, subjects being invited to repeatedly answer short questionnaires, preferably timed randomly with restricted intervals to avoid behavioral adaptation to fixed intervals [59,60]. It has been suggested that ESM "allows us to capture the film rather than a snapshot of daily life reality of patients" [61].

Because of the repeated measures over time in the continuously changing context of daily life, ESM is supposed to have multiple benefits such as a higher ecological validity and a higher sensitivity to (subtle) change(s). It is seen as a method less dependent of participants memory, less vulnerable to assessment error, suitable to assess dynamic processes (e.g., how long does it last to be able to experience positive mood after a negative mood inducing event), and able to provide a view on variability in mental states. It also allows some "contextual" analysis by giving the possibility of linking emotions and affect to situational aspects (e.g., being at home or being at work while experiencing emotions). When used in clinical practice, ESM could increase the engagement of patients in the treatment process although the latter still has to be confirmed [60]. It is certainly more precise, but the question can again be raised whether more precise is more "meaningful" to patients and to physicians. One can easily assume that ESM will be more easily integrated in cognitive behavior approaches than in family therapy or psychodynamic therapy.

But some doubts and some possible disadvantages of ESM have also been described [59]. One practical concern is the participant burden: being invited multiple times per day to fill out (even brief) assessments on your mobile can be intrusive and disruptive (e.g., on inopportune moments or in inopportune settings) and hence become a burden; several studies indeed showed rather high drop-out rates. A more fundamental comment is that measuring "in the moment" does not enable to capture the patient's reflection on the measured phenomenon, while the latter is the basis for psychotherapeutical work [59,60,62]. Moreover, the aggregation and time courses of the patient's self-assessments can be poorly correlated with the memories of introspected experiences which again is the basis for psychotherapeutical work. The issue of "reactivity-induction" by bringing a certain content under the subject's attention and possibly moving it from a preconscious/unconscious to a conscious level is less clear-cut and subject of an interesting debate. Another issue is that the so-called "contextual" assessment is extremely limited and therefore not very meaningful (assessed while "being at work" does not differentiate between probably important contextual aspects of that moment on the workplace).

Until today, ESM research in depression has mainly focused on the role and interaction of positive and negative affect and on the effect of (physical) activity to affect [63]. It is commonplace to state that patients with major depressive disorder suffer from reduced positive and increased negative affect [64]. A refinement illustrated by ESM research found that stress generates stronger negative affect in MDD patients compared with controls, while the stress reactive decrease in positive affect was comparable in depressed patients and controls [65]. ESM has been used to document time courses of positive and negative affect in depressed patients, in remitted patients, and in controls but also to look at patterns predicting response. However, some of these studies get so methodologically refined that it becomes difficult to draw clinical relevant conclusions: one example is a study where it was shown that in recurrent-episode future responders, the daily maximum positive affect increase resulted in significantly lower levels of subsequent negative affect over the next few hours compared to future nonresponders or compared to first-episode responders [66].

Whether ESM will be a real assessment breakthrough and a real therapeutic breakthrough or whether it is mainly an academic sophistication and mainly a computer science–driven approach still has to be elucidated.

Is What is Commonly Assessed What Matters to Patients?

An overarching question is to whose reification each assessment tool contributes: to their author(s), to a specific theoretical framework, to a specific therapeutic effect, to the Diagnostic and Statistical Manual (DSM), or to the patient's expectations?

Max Hamilton, who developed observer-rating scales, stated in 1977: "I have some antipathy to self-rating scales….self-rating scales provide an excellent excuse for the investigator to avoid interviewing his patient…" which could be considered a conflict of interest. On the contrary, Mark Zimmerman who developed several self-rating scales stated: "clinician-rated scales are time consuming, require training to ensure the ratings are reliable and valid, and may be prone to clinician bias. Self-report questionnaires are inexpensive in terms of professional time needed for incorporation into the clinical encounter, they do not require special training for administration, and they correlate highly with clinician ratings. Moreover, self-report scales are free of clinician bias and are therefore free from the potential risk of clinician overestimation of patient improvement (which might occur when there is incentive to document treatment success)"[25].

The 21 items of the BDI-I [21] were originally biased toward cognitive behavior theory and therapy and comprise many cognitive items, but the BDI-II changed the time frame (during the last 2 weeks instead of during the last week in BDI-I) [20] and changed some items in order to reflect more closely DSM-IV symptomatic diagnostic criteria for major depressive disorder. One step further in the reification of DSM was the development of the nine-item Patient Health Questionnaire mirroring the nine DSM criteria [67]. The HAMD items closely reflect the effects (efficacy as well as sedative side effects) of tricyclics, while the MADRS closely reflects the improvements obtained with a variety of antidepressants.

Important discrepancies do exist between the content of most depression scales and what matters to patients [68]. Patients rather want to know what are the chances they can get back to work, whether they will be able to fully resume their role as a partner or parent, and whether they will be able again to engage in pleasant activities [69,70]. When patients and caregivers were asked what they consider important in being cured from depression, caregivers emphasize the reduction of depressive symptoms, while patients take a greater interest in restoration of a meaningful life and in return of positive affect [71]. However, the concept of positive affect (and associated concepts: hedonic tone, pleasure, motivation, and reward) is at the risk of simplification: it has been suggested that a

better disentangling of these concepts is helpful in understanding their neurobiological underpinnings [72].

Several attempts were made to develop scales based on patient's expectations. The Remission from Depression Questionnaire [73] also assesses positive mental health, functioning, life satisfaction, and general sense of well-being and the LAPS [54] assessing positive and negative affect, hedonic tone, (cognitive) functioning, meaningfulness of life, and happiness.

Conclusions

Assessment of severity of depressive symptomatology and of changes in severity during treatment is still suboptimal. It is remarkable that many clinicians do not routinely use scales in their daily practices: they should use at least one quantitative measure to assess clinical changes during treatment while accepting the reductionistic nature of it. Which scale should be used is maybe of only secondary importance compared to using at least one, despite being aware of the limitations. The present paper aims to give a framework facilitating the clinician's or researcher's orientation among scales commonly used in depression research: the choice is between observer-rating and self-rating scales, between disease-specific and non–disease-specific scales, between site rating and centralized rating, between subjective and objective (lab) rating, and between questionnaires versus experience sampling methods. The use of depression rating scales is highly recommended in clinical practice, as long as one realizes and accepts that "a rating scale is only a particular device for recording information about a patient…for clinical purposes, the best way of describing a patient is by a free and full psychiatric case history"[12].

Conflicts of interest. The authors declare no conflicts of interest.

References

[1] Hughes JR, O'Hara MW, Rehm LP. Measurement of depression in clinical trials: an overview. J Clin Psychiatry. 1982;43:85–8.
[2] Snaith RP. Present use of the Hamilton Depression Rating Scale: observation on method of assessment in research of depressive disorders. Br J Psychiatry. 1996;168:594–7.
[3] American Psychiatric Association. Practice guideline for the treatment of patients with major depressive disorder. Washington, DC: American Psychiatric Association; 2010. Available from: https://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/mdd.pdf.
[4] National Institute for Health and Care Excellence. Depression in adults: recognition and management. London, UK: National Institute of Health and Care Excellence; 2009. Available from: https://www.nice.org.uk/guidance/CG90.
[5] Kilbourne AM, Beck K, Spaeth-Rublee B, Ramanuj P, O'Brien RW, Tomoyasu N, et al. Measuring and improving the quality of mental health care: a global perspective. World Psychiatry. 2018;17:30–8.
[6] Macdonald AJ, Elphick M. Combining routine outcomes measurement and 'Payment by Results': will it work and is it worth it? Br J Psychiatry. 2011;199:178–9.
[7] Gilbody SM, House AO, Sheldon TA. Psychiatrists in the UK do not use outcomes measures. National survey. Br J Psychiatry. 2002a;180:101–3.
[8] Zimmerman M, McGlinchey JB. Why don't psychiatrists use scales to measure outcome when treating depressed patients? J Clin Psychiatry. 2008;69:1916–9.
[9] Hatfield DR, Ogles BM. Why some clinicians use outcome measures and others do not. Adm Policy Ment Health. 2007;34:283–91.
[10] Dowrick C, Leydon GM, McBride A, Howe A, Burgess H, Clarke P, et al. Patients' and doctors' views on depression severity questionnaires incentivised in UK quality and outcomes framework: qualitative study. BMJ. 2009;338:b663.
[11] Frank AW. The standpoint of storyteller. Qual Health Res. 2000;10:354–65.
[12] Hamilton M. The role of rating scales in psychiatry. Psychol Med. 1976;6:347–9.
[13] Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry. 1979;134:382–9.
[14] Möller HJ. Rating depressed patients: observer- vs self-assessment. Eur Psychiatry. 2000;15:160–72.
[15] Zimmerman M, Clark HL, Multach MD, Walsh E, Rosenstein LK, Gazarian D. Have treatment studies of depression become even less generalizable? A review of the inclusion and exclusion criteria used in placebo-controlled antidepressant efficacy trials published during the past 20 years. Mayo Clin Proc. 2015;90:1180–6.
[16] Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960;23:56–62.
[17] Tansey T. Review of D. Healy, 'The antidepressant era'. Hist Psychiatry. 1998;9:536.
[18] Worboys M. The Hamilton rating scale for depression: the making of a "gold standard" and the unmaking of a chronic illness, 1960–1980. Chronic Illn. 2013;9:202–19.
[19] Demyttenaere K, De Fruyt J. Getting what you ask for: on the selectivity of depression rating scales. Psychother Psychosom. 2003;72:61–70.
[20] Beck AT, Steer RA, Brown GK. BDI-II, Beck Depression Inventory: manual. San Antonio, TX: Psychological Corporation, 1996.
[21] Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry. 1961;4:561–71.
[22] Dozois DJA, Covin R. The Beck Depression Inventory-II (BDI-II), Beck Hopelessness Scale (BHS), and Beck Scale for Suicide Ideation (BSS). In: Comprehensive handbook of psychological assessment, Vol. 2: Personality assessment. Hoboken, NJ: John Wiley & Sons Inc, 2004; p. 50–69.
[23] Hagen B. Measuring melancholy: a critique of the Beck Depression Inventory and its use in mental health nursing. Int J Ment Health Nurs. 2007;16:108–15.
[24] Cuijpers P, Li J, Hofmann SG, Andersson G. Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: a meta-analysis. Clin Psychol Rev. 2011;30:768–78.
[25] Zimmerman M, Walsh E, Friedman M, Boerescu DA, Attiullah N. Are self-report scales as effective as clinician rating scales in measuring treatment response in routine clinical practice? J Affect Disord. 2018;225:449–52.
[26] Fantino B, Moore N. The self-reported Montgomery-Asberg depression rating scale is a useful evaluative tool in major depressive disorder. BMC Psychiatry. 2009;9:26.
[27] Mitchell AJ, Bird B, Rizzo M, Meader N. Which version of the geriatric depression scale is most useful in medical settings and nursing homes? Diagnostic validity meta-analysis. Am J Geriatr Psychiatry. 2010;18:1066–77.
[28] Addington D, Addington J, Schissel B. A depression rating scale for schizophrenia. Schizophr Res. 1990;3:247–51.
[29] Gainotti G, Azzoni A, Zazzano C, Lannillotta M, Marra C, Gasparini F. The post-stroke depression rating scale: a test specifically devised to investigate affective disorders of stroke patients. J Clin Exp Neuropsychol. 1997;19(3):340–56.
[30] Kulkarni J, Gavrilidis E, Hub-daib AR, Bleeker C, Worsley R, Gurvich C. Development and validation of a new rating scale for perimenopausal depression – the Meno-D. Transl Psychiatry. 2018;8:123. doi: 10.1038/s41398-018-0172-0.
[31] Guy W. ECDEU assessment manual for psychopharmacology. Rockville, MD: U.S. Department of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, National Institute of Mental Health, Psychopharmacology Research Branch, Division of Extramural Research Programs, 1976.
[32] Dunlop BW, Gray J, Rapaport MH. Transdiagnostic clinical global impression scoring for routine clinical settings. Behav Sci (Basel). 2017;7:40.
[33] Haro JM, Kamath SA, Ochoa S, Novick D, Rele K, Fargas A, et al. The clinical global impression-schizophrenia scale: a simple instrument to measure the diversity of symptoms present in schizophrenia. Acta Psychiatr Scand Suppl. 2003;416:16–23.

[34] Malow BA, Connolly HV, Weiss SK, Halbower A, Goldman S, Hyman SL, et al. The pediatric sleep clinical global impressions scale-a new tool to measure pediatric insomnia in autism spectrum disorders. J Dev Behav Pediatr. 2016;37:370–6.
[35] Spearing MK, Post RM, Leverich GS, Brandt D, Nolen W. Modification of the Clinical Global Impressions (CGI) Scale for use in bipolar illness (BP): the CGI-BP. Psychiatry Res. 1997;73:159–71.
[36] Targum SD, Hassman H, Pinho M, Fava M. Development of a clinical global impression scale for fatigue. J Psychiatr Res. 2012;46:370–4.
[37] Hawley CJ, Gale TM, Sivakumaran T, Hertfordshire Neuroscience Research Group. Defining remission by cut off score on the MADRS: selecting the optimal value. J Affect Disord. 2002;72:177–84.
[38] Forkmann T, Scherer A, Boecker M, Pawelzik M, Jostes R, Gauggel S. The Clinical Global Impression Scale and the influence of patient or staff perspective on outcome. BMC Psychiatry. 2011;11:83.
[39] Leon AC, Shear MK, Klerman GL, Portera L, Rosenbaum JF, Goldenberg I. A comparison of symptom determinants of patient and clinician global ratings in patients with panic disorder and depression. J Clin Psychopharmacol. 1993;13:327–31.
[40] Bobo WV, Anglero GC, Jenkins G, Hall-Flavin DK, Weinshilboum R, Biernacka JM. Validation of the 17-item Hamilton Depression Rating Scale definition of response for adults with major depressive disorder using equipercentile linking to Clinical Global Impression scale ratings: analysis of Pharmacogenomic Research Network Antidepressant Medication Pharmacogenomic Study (PGRN-AMPS) data. Hum Psychopharmacol. 2016;31:185–92.
[41] Leucht S, Fennema H, Engel R, Kaspers–Janssen M, Lepping P, Szegedi A. What does the HAMD mean? J Affect Disord. 2013;148:243–8.
[42] Papakostas GI, Ostergaard SD, Iovieno N. The nature of placebo response in clinical studies of major depressive disorder. J Clin Psychiatry. 2015;76: 456–66.
[43] Freeman MP, Pooley J, Flynn MJ, Baer L, Mischoulon D, Mou D, et al. Guarding the gate: remote structured assessments to enhance enrollment precision in depression trials. J Clin Psychopharmacol. 2017;37:176–81.
[44] Kobak KA, Leuchter A, DeBrota D, Engelhardt N, Williams JBW, Cook IA, et al. Site versus centralized raters in a clinical depression trial: impact on patient selection and placebo response. J Clin Psychopharmacol. 2010; 30:193–7.
[45] Targum SD, Wedel PC, Robinson J, Daniel DG, Busner J, Bleicher LS, et al. A comparative analysis between site-based and centralized ratings and patient self-ratings in a clinical trial of major depressive disorder. J Psychiatr Res. 2013;47:944–54.
[46] Sullivan MJ, Edgley K, Dehoux E. A survey of multiple sclerosis: I. Perceived cognitive problems and compensatory strategy use. Canada: Canadian Assn. for Research in Rehabilitation, 1990;p. 99–105.
[47] Jaeger J. Digit symbol substitution test: the case for sensitivity over specificity in neuropsychological testing. J Clin Psychopharmacol. 2018;38:513–9.
[48] Vakil E, Blachstein H. Rey auditory-verbal learning test: structure analysis. J Clin Psychol. 1993;49:883–90.
[49] Strober LB, Binder A, Nikelshpur OM, Chiaravalloti N, DeLuca J. The perceived deficits questionnaire: perception, deficit, or distress? Int J MS Care. 2016;18:183–90.
[50] Miskowiak K, Vinberg M, Christensen EM, Kessing LV. Is there a difference in subjective experience of cognitive function in patients with unipolar disorder versus bipolar disorder? Nord J Psychiatry. 2012;66:389–95.
[51] McIntyre RS, Lophaven S, Olsen CK. A randomized, double-blind, placebo-controlled study of vortioxetine on cognitive function in depressed adults. Int J Neuropsychopharmacol. 2014;17:1557–67.
[52] Snaith RP, Hamilton M, Morley S, Humayan A, Hargreaves D, Trigwell P. A scale for the assessment of hedonic tone the Snaith-Hamilton pleasure scale. Br J Psychiatry. 1995;167:99–103.
[53] Gard DE, Gard MG, Kring AM, John OP. Anticipatory and consummatory components of the experience of pleasure: a scale development study. J Res Pers. 2006;40:16.
[54] Demyttenaere K, Mortier P, Kiekens G, Bruffaerts R. Is there enough "interest in and pleasure in" the concept of depression? The development of the Leuven Affect and Pleasure Scale (LAPS). CNS Spectr. 2019;24: 265–74.
[55] Henriques JB, Davidson RJ. Decreased responsiveness to reward in depression. Cogn Emot. 2000;14:711–24.
[56] Henriques JB, Glowacki JM, Davidson RJ. Reward fails to alter response bias in depression. J Abnorm Psychol. 1994;103:460–6.
[57] Pizzagalli DA, Iosifescu D, Hallett LA, Ratner KG, Fava M. Reduced hedonic capacity in major depressive disorder: evidence from a probabilistic reward task. J Psychiatr Res. 2008;43:76–87.
[58] Pizzagalli DA, Jahn AL, O'Shea JP. Toward an objective characterization of an anhedonic phenotype: a signal-detection approach. Biol Psychiatry. 2005;57:319–27.
[59] van Berkel N, Ferreira DS, Kostakos V. The experience sampling method on mobile devices. ACM Comput Surv. 2017;50:93.
[60] Verhagen SJ, Hasmi L, Drukker M, van Os J, Delespaul PAEG. Use of the experience sampling method in the context of clinical trials. Evid Based Ment Health. 2016;19:86–9.
[61] Myin-Germeys I, Oorschot M, Collip D, Lataster J, Delespaul P, van Os J. Experience sampling research in psychopathology: opening the black box of daily life. Psychol Med. 2009;39:1533–47.
[62] Engelbert M, Carruthers P. Descriptive experience sampling: what is it good for? J Concious Stud. 2011;18:130–49.
[63] Armey MF, Schatten HT, Haradhvala N, Miller IW. Ecological Momentary Assessment (EMA) of depression-related phenomena. Curr Opin Psychol. 2015;4:21–5.
[64] Watson D, Clark LA, Weber K, Assenheimer JS, Strauss ME, McCormick RA. Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. J Abnorm Psychol. 1995;104:15–25.
[65] Myin-Germeys I, Peeters F, Havermans R, Nicholson NA, DeVries MW, Delespaul P, et al. Emotional reactivity to daily life stress in psychosis and affective disorder: an experience sampling study. Acta Psychiatr Scand. 2003;107:124–31.
[66] Wichers M, Peeters F, Rutten BP, Jacobs N, Derom C, Theiry E, et al. A time-lagged momentary assessment study on daily life physical activity and affect. Health Psychol. 2012;31:135–44.
[67] Kroenke K, Spitzer RL, Williams JBW. The PhQ-9. Validity of a brief depression severity measure. J Gen Intern Med. 2001;16:606–13.
[68] Fava GA, Tomba E, Tossani E. Innovative trends in the design of therapeutic trials in psychopharmacology and psychotherapy. Prog Neuropsychopharmacol Biol Psychiatry. 2013;40:306–11.
[69] Gilbody S, Wahlbeck K, Adams C. Randomized controlled trials in schizophrenia: a critical perspective on the literature. Acta Psychiatr Scand. 2002b;105:243–51.
[70] Zimmerman M, Martinez JH, Young D, Chelminski I, Dalrymple K. Severity classification on the Hamilton Depression Rating Scale. J Affect Disord. 2013;150:384–8.
[71] Demyttenaere K, Donneau AF, Albert A, Ansseau M, Constant E, van Heeringen K. What is important in being cured from depression? Discordance between physicians and patients (1). J Affect Disord. 2015;174: 390–6.
[72] Moccia L, Mazza M, Di Nicola M, Janiri L. The experience of pleasure: a perspective between neuroscience and psychoanalysis. Front Hum Neurosci. 2018;12:259. doi: 10.3389/fnhum.2018.00359.
[73] Zimmerman M, Galione J, Attiullah N, Friedman M, Toba C, Boerescu D, et al. Depressed patients perspectives of two measures of outcome: the Quick Inventory of Depressive Symptomatology (QIDS) and the Remission from Depression Questionnaire (RDQ). Ann Clin Psychiatry. 2011; 23:208e12.