Assessing treatment effects in clinical trials with the Discan metric of the Sheehan Disability Scale

Kathy Harnett Sheehan and David V. Sheehan

The Sheehan Disability Scale (SDS) is a patient-rated, discretized analog measure of functional disability in work, social, and family life. Its increasing use in clinical trials in psychiatry suggests a need to assess its responsiveness and interpretability. In this paper we identify and review studies in which the SDS was used as a treatment outcome measure. Our objectives are (i) to evaluate the sensitivity of the SDS to treatment effects and (ii) to examine potential thresholds or cutoff scores for remission and response. Studies for the review were retrieved from the National Library of Medicine's PubMed database (1966 to 21 March 2007) and other sources. All studies had to use the SDS, be double-blind, controlled or large open-label trials in English. Studies assessing nonpharmacological treatments, long-term trials (>12 weeks), small *n* trials (less than 20 patients per treatment arm) and trials for conditions other than one of the anxiety disorders, depression, or premenstrual dysphoric disorder were excluded. Extracted data included the diagnostic target of treatment, *n*, study design, and method of analysis. Initial, endpoint and/or mean change scores were extracted from tables, text, or extrapolated from figures. In all, 37 studies meeting the inclusion criteria were retrieved and reviewed. All of the studies treated the SDS as a numeric scale and analyzed mean change or endpoint differences with parametric statistics. Three provided additional outcome data using nonparametric response or remission criteria. Overall, the SDS performed well in discriminating between active and inactive treatments. The results indicate that the SDS is sensitive to treatment effects. To establish reliable and valid cutoff scores for remission and response, there is a need to supplement parametric analyses using mean change and endpoint differences with nonparametric analyses showing the percentage meeting specified response and remission criteria. In addition, the percentages with endpoint scores of zero should be reported. *Int Clin Psychopharmacol* 23:70–83 © 2008 Wolters Kluwer Health | Lippincott Williams & Wilkins.

International Clinical Psychopharmacology 2008, **23**:70–83

Keywords: anxiety disorders, depression, disability, Discan, impairment, mood disorders, outcomes, scales

University of South Florida College of Medicine, Tampa, Florida, USA

Correspondence to David V. Sheehan, MD, MBA, USF Institute for Research in Psychiatry, 3515 East Fletcher Avenue, Tampa, FL 33613, USA
Tel: +1 813 974 4544; fax: +1 813 974 4575;
e-mail: dsheehan@health.usf.edu

Received 12 June 2007 **Accepted** 27 September 2007

Introduction

The disabling effects of mental illness are high. Using data developed by the World Health Organization (WHO), the World Bank, and Harvard University, Murray and Lopez (1996) reported that mental illness, including suicide, ranked second in the burden of disease, behind all cardiac conditions, in established market economies such as the United States in 1996. More recently, mental disorders have been found to be the leading cause of disability in the United States and Canada for ages 15–44 years (WHO, 2004). Patients with major depression, for example, are known to have increased rates of impaired occupational functioning, marital problems, and social isolation (Wells *et al.*, 1989; Broadhead *et al.*, 1990; Olfson *et al.*, 1997). Significant impairment in functioning has also been observed in patients with anxiety disorders including generalized anxiety (Leon *et al.*, 1995, Olfson *et al.*, 1997; Kessler *et al.*, 1999), panic disorder (PD) (Markowitz *et al.*, 1989; Hollifield *et al.*, 1997; Marcus *et al.*, 1997), social anxiety disorder (SAD) (Lecrubier *et al.*, 2000; Fehm *et al.*, 2005), and posttraumatic stress disorder (PTSD) (Davidson *et al.*, 1991). Moreover, there is evidence that recovery is associated with improvement in functioning (Mintz *et al.*, 1992; Schulberg *et al.*, 1996; Katzelnick *et al.*, 2000). These considerations have led to increasing recognition of the importance of measuring the unique individual ramifications of mental illness in terms of disability and handicap (Rosenberg, 2000).

The Sheehan Disability Scale (SDS) was originally developed to measure impairment in treatment outcome studies in psychiatry. Despite its growing use in psychopharmacology and other treatment trials, its sensitivity to treatment effects across a range of disorders has not been well documented. In addition, little has been written about its interpretability. The latter is important since, as Juniper *et al.* (1994) point out, clinicians, patients, and others (pharmaceutical companies, insurance payers, regulators, and government bodies) not only want to know if a scale can separate active and inactive treatments but also if the magnitude of the effect is meaningful for patients and providers.

This paper provides a review of the responsiveness and interpretability of the SDS in treatment outcome studies in psychiatry. As the SDS was designed in a unique format, and the format has implications for how the scale is analyzed to assess treatment effects, the paper also describes the development and design of the SDS.

Background: design and development of the Sheehan Disability Scale

Design

As shown in Fig. 1, the SDS is an unweighted composite of three self-rated items of family, work, and social impairment in the previous week. Each item is preceded by a lead-in question: 'The symptoms have disrupted your (work/studies, social life, family life/home responsibilities)'. For each item, 11 potential responses ranging from 0 (not at all), 1–3 (mildly), 4–6 (moderately), 7–9 (markedly), to 10 (extremely) are presented along a continuum graphically represented by a horizontal line. Work is specified as paid, unpaid volunteer work or training and respondents have the option to skip this item if they have not worked/or studied at all in the last week for reasons unrelated to their disorder, for example, normal retirement. Subscale scores for work, social, and family disability are calculated separately. Total scores, ranging from 0 to 30, are calculated only for patients who rate all three items (older retired people may choose to skip the work item; alternatively, if they are engaged in unpaid volunteer work, they may complete all three subscales, including work).

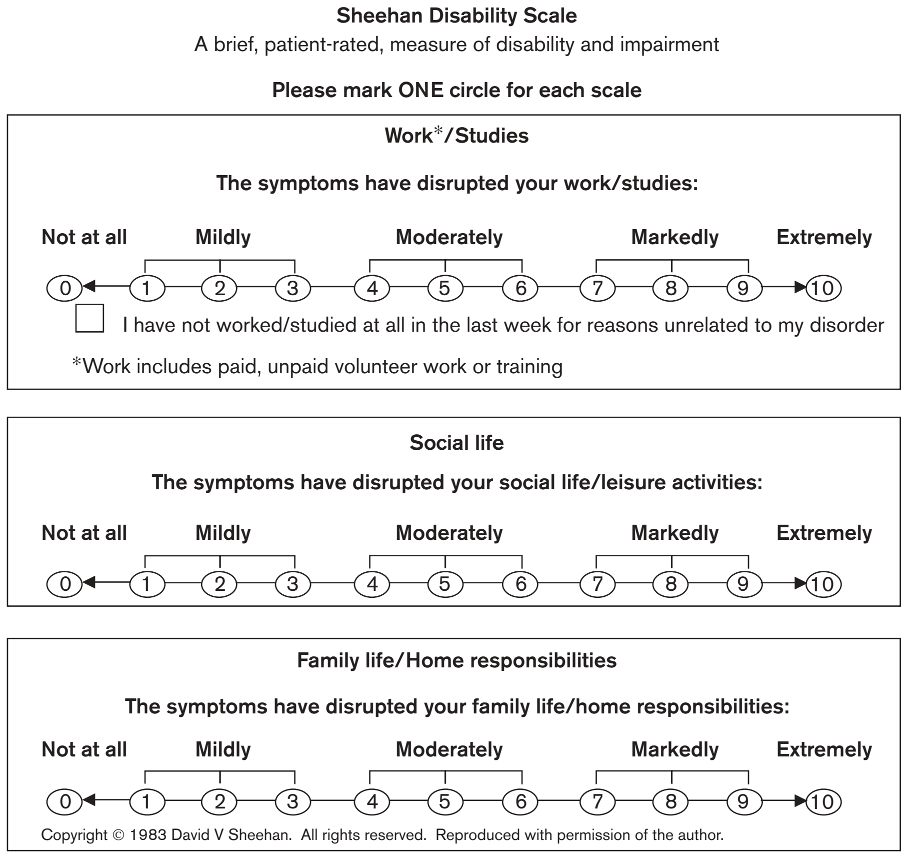
Development and rationale for scale

The SDS was originally developed in 1981 to provide a brief but reliable and valid self-rated measure of impairment in functioning for patients in treatment outcome studies in psychiatry (Sheehan, 1983; Sheehan *et al.*, 1996). At that time there was increasing recognition of the need to document and track mental health-related impairment. Most of the existing scales, however, were long and difficult to administer. For example, the Social Adjustment Scale (SAS) consisted of over 50 items (Weissman and Bothwell, 1976) whereas the Medical Outcome Study (MOS) SF-36 had 36 items (Ware and Sherbourne, 1992) and the Global Assessment Scale required differentiating among 21 items to produce a score from 0 to 100 (Endicott *et al.*, 1976).

Format selection

The first task was to choose a scale format. This choice was guided by the following principles. First, in keeping with the increasing advocacy of 'clinimetric' scales in

Fig. 1



Sheehan Disability Scale

A brief, patient-rated, measure of disability and impairment

Please mark ONE circle for each scale

Work*/Studies

The symptoms have disrupted your work/studies:

Not at all	Mildly			Moderately			Markedly			Extremely
0	1	2	3	4	5	6	7	8	9	10

□ I have not worked/studied at all in the last week for reasons unrelated to my disorder

*Work includes paid, unpaid volunteer work or training

Social life

The symptoms have disrupted your social life/leisure activities:

Not at all	Mildly			Moderately			Markedly			Extremely
0	1	2	3	4	5	6	7	8	9	10

Family life/Home responsibilities

The symptoms have disrupted your family life/home responsibilities:

Not at all	Mildly			Moderately			Markedly			Extremely
0	1	2	3	4	5	6	7	8	9	10

Copyright © 1983 David V Sheehan. All rights reserved. Reproduced with permission of the author.

The Sheehan Disability Scale (SDS).

psychiatry (Feinstein, 1983), we felt that the scale should be patient centered, that is, it should address patient perceptions, within their own frame of reference, rather than abstract constructs of disability. Second, we believed that it should be generic, that is, it should be able to measure impairment over a range of psychiatric diagnoses and mental health problems. Third, it should be easy to use. It should not be long and it should be formatted in a way that patients with different cognitive orientations (verbal, numeric, and visual spatial) can follow. Some people rate numerically ('he was 6′3″ tall'), others use verbal descriptive anchors ('he was very tall'), and still others communicate by rating space visually (using their hands to point while commenting 'he was this tall'). Some choose a combination of two or all three methods. Finally, it should allow enough points of discrimination to make fine distinctions in degrees and types of impairment.

At the time of the development of the SDS, the sensitivity of psychiatric rating scales to treatment effects (responsiveness) in clinical trials was receiving increasing attention. Montgomery and Asberg (1979) were among the first to point out that whereas many scales used in psychiatry were valuable in assessing the severity of a disorder, they were not necessarily designed to discriminate between responders and nonresponders (Montgomery and Asberg, 1979). This may be because the scales have too many constructs, the constructs are not meaningful to individual patients, or because some symptoms of a disorder are oversampled, for example, anxiety and insomnia items on the Hamilton Depression Scale (HAM-D). Scales developed for groups using conventional psychometric approaches are also less likely to be patient-centered (use the patient's own frame of reference) and may underreport unique but important individual-specific changes that occur with treatment (Bech, 2004; Faravelli, 2004).

A large number of the rating scales used in psychiatry follow a Likert or Likert-type format. In such scales, the patient is presented with a series of questions or items and for each item is asked to choose one of a usually small series of discrete numbered verbal responses, for example, 1 'not at all', 2 'slightly' 3 'moderately' 4 'severely'. In general the item responses are summed and the summed results are compared. Although Likert and Likert-type scales (such as the HAM-A and HAM-D) tend to be reliable, the limited number of available responses for each item makes it difficult for patients to make fine distinctions, and these distinctions may be critical to detecting treatment change and treatment differences. To improve on precision, some researchers have turned to visual analog scales (VAS), such as mood thermometers. In this approach, the patient is presented with a graphic continuum in the form of a line, usually horizontal and 100 mm, anchored by descriptors such as 'Not at all' and 'Extremely' at the ends. The score on such a scale is the distance in millimeters between the left end and the point recorded. Although VAS scales are more precise than Likert scales, they have been found to be less reliable (Singh and Bilsbury, 1989a).

To achieve the objectives we defined, we opted for a different format, that of the discretized analog scale (Discan) described by Singh and Bilsbury (1982, 1989b). Discan scales were originally developed to overcome some of the limitations of the numeric Likert scales on the one hand and graphic VAS on the other and provide a simpler, more clinimetric or patient-centered approach to rating symptoms in clinical trials (Bilsbury and Richmond, 2002). The scale format generalizes the Personal Questionnaire methodology of Shapiro (1961) and the Mixed Standard Sequential Pair Comparisons technique proposed by Singh and Bilsbury (1982). The Discan scale is similar to the Likert scale in that it uses verbal descriptors accompanied by numbers to elicit a finite number of responses. It, however, incorporates the graphical properties of the VAS by providing a visual line (usually horizontal) to represent a continuum of responses and, as a rule it offers the potential for a larger number of distinctive graduated responses compared with the Likert scale. In the strictest form of Discan, the patient is presented with cards and asked to choose between pairs of fine-tuned responses within categories, for example, of mild, moderate, or severe. In its modified form, as used in the SDS, the Discan scale uses a metric that integrates and presents to the patient at the same time the opportunity to rate verbally, numerically, and visually spatially.

Item selection

Our next task was to select items to identify and quantify impairment. After reviewing other impairment instruments and consulting with patients and colleagues we decided to build the scale as a composite of three items regarding work, social, and family impairment. To capture the effects of psychiatric symptoms on impairment, the items were initially worded: 'To what extent have emotional symptoms disrupted your (work, family, or social) life in the last month?'

Item scaling

Next, we had to decide on the number of points to use to rate each of the three items. In a literature review, Cox (1980) concluded that there is no single number of points for a rating scale that is appropriate for all situations. Although, in general, he recommended the use of 5–9 numbered points, Friedman and Friedman (1986) found that in many situations 11-point scaling produces more valid results. We elected to use an 11-point (0–10) scale with the number '0' signifying no impairment and the number '10' signifying the highest ('extremely') impairment because such a scale is easy for patients to read and understand. The points on such a scale are also easy to convert into percentages.

Graphic line

Having decided on the number of points for each item, we inserted a horizontal visual analog line as a visual spatial guide and placed the numbers, left to right (from 0 to 10) along the line.

Verbal descriptors

Finally, we chose to set off points 1–3, 4–6, and 7–9 with ordered verbal descriptors ('mildy', 'moderately', and 'markedly') to signify increasing degrees of perceived impairment and make it easier for patients who rate verbally to orient themselves to the numbered points on the scale and visual spatial layout.

Generally scales should be balanced with equal numbers of positive and negative response choices (Friedman *et al.*, 1981). Our use of a verbal anchor at either end of the scale and three additional verbal descriptors, each referring to three numbered points, was guided by considerations of scale balance. Using fewer (e.g. 2) or more (e.g. 4 or 5) descriptors for these points would have led to unequal numbers of points for each descriptor. Schwartz *et al.* (1991) observe that scale respondents use numeric values to 'disambiguate' the meaning of scale labels. Although the verbal descriptors we used might appear to be open to different interpretations, we felt that the numeric values for each of the descriptors provided sufficient discrimination for most patients.

Position of descriptors and numbers

Researchers have found that scale respondents are affected by both the verbal descriptor and position effects (Wildt and Mazis, 1978). We placed the verbal descriptors at the midpoint of each series (1–3, 4–6, 7–9) with brackets to indicate that the descriptors referred to all three numbers in the designated series. We, however, placed the numeric values (0–10) at equidistant intervals on the graphic line to indicate that the distance from one value to the next should be viewed as equal. That is, we used the positions of the numeric values to indicate that the amount of difference between the selection of a 3 over a 2 was the same as the difference between the selection of a 4 over a 3 – even though, in the latter case, choosing the higher number would shift the response into a different category according to the verbal descriptor.

Whether respondents are more affected by verbal descriptors (labels) than by position effects has been a matter of debate (Friedman *et al.*, 2003). To correct for the possibility that respondents might view the verbal descriptors ('mild', 'moderate', etc.) as more relevant than the numbers, we made the verbal descriptors smaller than the numbers. In addition, we provided graphic circles for each number and we indicated in the instructions that the respondent was to mark a circle (i.e. numeric value) for each scaled item.

Implications of format for analysis

The unique format of the Discan scale has occasionally raised questions about how it should be analyzed. Some researchers believe that this type of scale lies in between the ordinal and interval scale and recommend that it be analyzed using nonparametric statistics, for example, statistics based on ranks (Bilsbury and Richmond, 2002). Our own experience with the SDS is that most researchers treat it as an interval scale and use parametric statistics [*t*-tests and analyses of variance (ANOVAs)] to measure change and treatment differences. In view of the increasing emphasis in clinical trials on finding 'meaningful' or 'relevant' or 'clinically important' change' as opposed to just statistical change (Middel and Von Sonderen, 2002), however, there may be a need to establish thresholds or benchmarks for identifying meaningful or important change when using modified Discan scales, such as the SDS, and to consider the addition of nonparametric statistics to determine meaningful response or remission.

Pilot testing

In a series of informal pilot tests, we modified and refined the introductory wording and verbal descriptors in response to patient and clinician feedback. For example, we shortened the wording for each item and we added a clarification of 'work' and a check box for patients who had not worked for reasons unrelated to their disorder in the previous week.

Psychometric testing

The scale was used in the early 1980s in several PD treatment studies including one of the first trials of alprazolam in the treatment of PD (Sheehan, 1985), the subsequent Cross National Collaborative Panic Study (Ballenger *et al.*, 1988), and the Panic Depression Study (Keller *et al.*, 1993). The latter two trials, with *n*'s of 476 and 122, respectively, were subsequently used to investigate the scale's internal consistency, construct validity (factor structure and sensitivity to change), criterion-related validity, and discriminant validity. The results of these tests indicated that the reliability (internal consistency among items and measurement error in each item) of the scale was acceptable and the factor structure of the items and the sensitivity to change of their composite had satisfactory construct validity. The criterion-related validity was further substantiated by the significant relationship between patient symptomatology and impairment. Despite what might appear to be ambiguous anchor points (e.g. 'not at all', 'moderately', 'extremely'), the scale was found to adequately measure impairment and it was recommended for inclusion as an outcome in treatment outcome studies for panic and other psychiatric disorders (Leon *et al.*, 1992).

The psychometric properties of the SDS were further evaluated in a primary care sample of 1001 patients. The internal consistency reliability was found to be high in this sample and empirical support was provided for its construct validity and sensitivity for identifying patient with mental health-related functional impairment. Overall, more than 80% of the patients with one or more of six psychiatric disorders had an elevated SDS score, and nearly 50% of those with elevated scale scores met diagnostic criteria for at least one psychiatric disorder. Furthermore, patients with any of six psychiatric disorders (alcohol dependence, drug dependence, generalized anxiety disorder (GAD), major depressive disorder (MDD), obsessive-compulsive disorder (OCD), and PD) had significantly higher impairment scores than those who did not have the disorders. Validity was further indicated by the fact that the scores of participants who reported 'problems getting along with partner' were higher on both the family impairment item and the total score than the scores of participants who denied having this type of problem. In addition, those who reported 'missing work in the past month due to emotional problems' had significantly higher scores on both the work impairment item and the total score. Overall, a total score of ≥ 5 was found to strongly predict impairment (Leon *et al.*, 1997).

Population norms

In their report on data from the National Comorbidity Survey Replication Study, Kessler *et al.* (2003, 2006) found that 59.3% of adults 18 and over with major depression and 84% of those with panic disorder with agoraphobia had total SDS scores of 21 or higher, indicating marked or extreme impairment. In contrast, only 11% of individuals with panic attacks alone had marked or extreme impairment on the SDS (Kessler *et al.*, 2006).

In an analysis of 1001 adults 18 years and over in primary care, Leon *et al.* (1997) found mean scores of 2.01, 2.02, and 2.08 for the 0–10 work, social, and family subscales, respectively, and 6.08 for the 0–30 total SDS scale. For primary care patients not meeting criteria for any one of six psychiatric disorders they studied, the mean total SDS score was 3.7. Mean total scores for patients meeting criteria for specific disorders were 11.6 for alcohol dependence, 12.3 for drug dependence, 14.0 for GAD, 16.2 for major depression, 14.6 for OCD, and 16.5 for PD. In a separate study of 1266 primary care patients aged 18–70 years, Olfson *et al.* (2000) reported results using a 2-item (0–20) SDS that excluded the work subscale. He found that for patients without a psychiatric disorder, the mean for the 2-item SDS was 1.7 ± 3.8. Comparable means on the 2-item scale for patients were 5.4 for major depression only, 10.8 for major depression with other disorders, 7.8 for PD, only 10.4 for PD with other disorders, 4.9 for GAD only, 10.1 for GAD with other disorders, 2.0 for substance use only and 11.7 for substance use with other disorders. SDS scores for elderly primary care patients with depression have also been reported. In a study of 1801 patients aged 60 years or older who met diagnostic criteria for major depression or dysthymia, Noel *et al.* (2004) reported an average SDS score of 4.6 across the three SDS subscales. This would equate approximately to a total SDS score in this population of 13.8.

Translations and use

The SDS is now available in 48 languages and has become one of the most frequently used measures of impairment in psychopharmacology trials. It is also being used with increasing frequency in studies of nonpsychopharmacological treatments in psychiatry, including cognitive behavioral treatment (CBT) for schizophrenia (McQuaid *et al.*, 2000), social anxiety (Herbert *et al.*, 2005), and PD (Galassi *et al.*, 2007) and psychodynamic psychotherapy for PD (Milrod *et al.*, 2000, 2001).

The SDS has been incorporated as a severity measure into the diagnostic modules of the most recent version of the Composite International Diagnostic Interview (CIDI 2.1) developed by the WHO and the World Mental Health Organization (World Mental Health, 2004) and it was used in the WHO 2001–2003 survey of the prevalence, severity and unmet need for treatment of mental disorders in over 60 463 adults in 14 countries in the Americas, Europe, the Middle East, Africa, and Asia (WHO World Mental Health Survey Consortium, 2004) and in the National Comorbidity Survey Replication Study in the United States (Kessler *et al.*, 2003, 2006).

Methods

We used PubMed (1966 to 21 March 2007) to identify clinical trials in which the SDS was used to measure disability in treatment studies. We identified other trials from posters presented at international and national meetings.

Inclusion criteria

Studies were included if they were double-blind, controlled, or very large open-label trials for the following disorders: MDD, any of the five principal anxiety disorders, including PD, SAD, GAD, OCD, and PTSD, or premenstrual dysphoric disorder (PMDD), and were published in English. Studies assessing the effects of nonpharmacological treatments, long-term trials (> 12 weeks), small *n* trials (< 20 per treatment arm) were excluded.

Data extraction and assessment

We extracted the diagnostic target of treatment, *n*, study design and method of analyzing outcomes on the SDS. We developed a checklist to evaluate and compare how clinical change and drug placebo differences (responsiveness) on the SDS were assessed and how the results were reported in the trials. In particular, we examined the method of analysis (parametric or nonparametric), whether the analysis was of mean change and/or endpoint differences across treatments, whether total scores and/or the three subscale scores (work, social, and family disability) were analyzed, and whether remission or percentage response rates were reported.

As calculations of effect size (ES) and standard mean response contribute to the interpretability of results and such calculations depend on knowledge of standard deviations (SDs), we also examined whether the SD was documented.

In addition, to help establish benchmarks for responsiveness, we extracted initial and endpoint as well as mean change total and subscale scores from tables and text or extrapolated these data from figures. For studies with more than two treatment groups, for example, fixed dose studies with three or more arms, we extracted data for the most relevant contrast (e.g. highest dose active treatment vs. placebo).

Criteria for responsiveness

Liang (1995) has defined responsiveness as the ability of a scale to detect meaningful clinical changes over time when evaluating the benefits of a medical intervention. Responsiveness may be tested by assessing change on the outcome of interest in a single group or by comparing change across two or more groups. Tests of responsiveness with a Discan scale such as the SDS can be conducted in a variety of ways.

Remission and response

Some researchers distinguish between categories of remission, response, and nonresponse. Remission, which is viewed as the optimal outcome (Keller, 2003), is usually defined as a state of minimal to no symptoms and a return to normal functioning (Frank *et al.*, 1991). Remission criteria are usually based on categorical thresholds or cutoff scores. One convention is the 70% rule. Using this criterion, a '1' or 'very much improved' on the 7-point Clinical Global Impression Scale may be used to denote remission. Two other conventions for remission in depression treatment include scores of ≤ 7 on the 17-item Hamilton Depression Rating Scale (HAM-D-17) and scores of ≤ 10 on the Montgomery Asberg Depression Rating Scale (MADRS) (Lecrubier, 2002). These thresholds approximate a 70% improvement. Response is typically defined as at least a 50% reduction in a symptom score from baseline. As Lecrubier (2002) points out, response is not the same as remission. A person could have a 50% reduction in symptoms (from a high initial score on the HAM-D) and still have substantial symptomatology.

As noted by Kelsey (2001) and Thase (2003), normalization of functioning is a critical aspect of remission. To date, however, there are no agreed upon criteria for remission in regard to disability. One convention, on the basis of the data from Leon *et al.* (1997) in a primary care sample, is that a total SDS score of < 5 signifies remission. Another convention, on the basis of the guidelines published by Ballenger (1999), is that remission is established only when the SDS subscale scores for work, social, and family disability are each ≤ 1. Our own experience (see Discussion) is that these thresholds may be overly ambitious. As for response, whereas a 50% reduction on the SDS total score or on the subscales might appear to be a good outcome, it needs to be emphasized that such an outcome may still signify considerable disability if the initial SDS scores are high.

In tests of remission and response, such as those described here, the McNemar's χ^2 or the Cochran Q test are appropriate for dependent samples (to assess change in one group over time) and the χ^2 test or the Fisher exact test may be used to test differences across groups.

Mean change and endpoint differences

A mean change (from baseline to endpoint) score is often used to measure responsiveness (Deyo *et al.*, 1991). Typically, paired *t*-tests or repeated measures ANOVAs are used to assess the statistical significance of change in one group over time whereas independent *t*-tests or ANOVAs are used to evaluate differences in mean change across two or more treatments or to assess endpoint differences. Alternatively, as Discan scales have been viewed as being in between the ordinal and the interval scale, some observers recommend that they be analyzed using tests based on medians or ranks, for example, the Sign test or the Wilcoxan's-matched pairs test to assess change in one group over time or the Median test or Kruskal–Wallis analysis of ranks to evaluate differences across two or more groups (Bilsbury and Richmond, 2002). Unfortunately, strict reliance on *P* values for tests such as these can be problematic for two reasons. The mean change (or endpoint difference) may be small in cases where negative changes in some patients cancel out the positive changes in others. In contrast, a mean change score may be large but patients may still be symptomatic.

Effect size and standardized response mean

Mean differences in outcomes can also be standardized to quantify a treatment's effect in terms of the SD. Standardizing mean change over time with an SD permits comparison of different outcomes, independent of the measuring units. One example is Cohen's d estimate of effect. For independent groups (e.g. treatment and control), this statistic is calculated as the posttest mean minus the pretest mean divided by the pretest SD (Kazis *et al.*, 1989). For dependent groups (e.g. to assess change in one group over time), Cohen proposed that the difference or change within participants needed to be adjusted by first dividing the difference (change over treatment) by the square root of $1-r$ with r denoting the correlation between the pretest and the posttest means. The resulting statistical measure, known as ES, communicates something very different from the *P* value, which indicates the obtained probability of a type I error in a test of statistical significance. As Middel and Von Sonderern (2002) point out, if a *P* value is significant, rejecting the null hypothesis does not necessarily mean that the effect was important. Nor does a nonsignificant *P* value necessarily indicate an unimportant result. For his ES calculations, Cohen (1977) produced conventions for those values that constitute a 'trivial' (ES < 0.20), 'small' (ES $\geq 0.20 < 0.50$), 'medium' (ES $\geq 0.50 < 0.80$), and a 'large' effect (ES ≥ 0.80). No consensus, however, exists on how to interpret the meaningfulness of ESs. Although some researchers believe that an ES greater than 0.80 denotes clinically important change, others argue that this is not always the case and to report ESs without the appropriate statistical tests and associated *P* values can be misleading.

Other statistical tests that can be used to measure responsiveness are the standardized response mean (SRM) calculated as the mean change from pretest to posttest divided by the SD of change over the same time period (Liang *et al.*, 1990; Liang, 1995), relative efficiency (Liang *et al.*, 1985), and Norman's responsiveness statistic calculated as the variance due to change divided by the variance due to change + error variance (Streiner and Norman, 1995).

Criteria for interpretability

In clinical trials, it is not difficult to determine the significance of a change in symptoms or impairment, but placing the magnitude of these changes in a context that is meaningful to clinicians, patients, and others (pharmaceutical companies, payers, regulators) is often more difficult (Juniper *et al.*, 1994). These considerations have led to an increasing emphasis on interpretability, a concept is usually defined as the degree to which one can assign qualitative meaning to quantitative scores. One approach to assessing the interpretability of a scale is to examine the magnitude of change that corresponds to a minimal important difference. In other fields of medicine, there is an increasing emphasis on knowing whether a treatment produces a minimal clinically important difference (MCID). Guyatt *et al.* (1987) and Jaeschke *et al.* (1989), for example, have stressed the usefulness of knowing what is a minimally important difference which they define as the smallest difference in score, which patients perceive as beneficial, and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management. We rated if the authors presented a MCID or other information was documented that could aid in interpreting the outcome scores, for example, presentation of means and SDs of scores before and after treatment. To decide whether a change in disability, for example, is important, it is helpful to have basic information, for example, means and SDs of patient scores before and after treatment. When mean changes are reported, it is also valuable to have the SD of the change score to be able to interpret the magnitude of change.

Results

Database

The PubMed search of the term 'Sheehan Disability Scale' identified 109 publications. Of these, 69 were identified as clinical trials and 48 were identified as randomized clinical trials. We excluded 27 publications (six could not be obtained; one was not in English; one was a small open-label study; in five the *n* was too small, i.e. < 20 per treatment arm; five were retrospective or exploratory analyses; two were extension studies; four focused on consult liaison or collaborative care; and three were for other diagnoses, e.g. gambling, insomnia, fibromyalgia). Further searches provided three additional publications and 12 reports of SDS data from two posters presented at national or international meetings using data on file in pharmaceutical companies for a total of 37 studies.

The final list included three major depression trials, five GAD trials, eight PD trials, 10 trials of treatments for SAD, one for OCD, four for PTSD, and six for PMDD. Two trials, one for depression and one for PTSD, were large open-label trials. One of the trials for depression compared two active drugs without a placebo control and one of the trials for PMDD compared the effects of two different menstrual phase onsets of paroxetine administration without a placebo. The remainder were all double-blind and placebo controlled. In all but one of the trials, the active medication was an SSRI or SNRI. The total *n* for the combined studies was 1609 (Table 1).

Baseline functioning

Mean total SDS scores at baseline were available for 12 of the 37 studies including three panic studies (Hoehn-Saric *et al.*, 1993; Sheehan *et al.*, 1993; Asnis *et al.*, 2001), three SAD studies (Stein *et al.*, 2002; Davidson *et al.*, 2004; Westenberg *et al.*, 2004), three GAD studies (Pollack *et al.*, 2001; Rickels *et al.*, 2003; Allgulander *et al.*, 2006), two PTSD studies (Tucker *et al.*, 2001; Davidson *et al.*, 2005), and one MDD study (Smeraldi, 1998). As shown in Fig. 2, the MDD trial had the highest total SDS score (23.1) at baseline. Mean total SDS scores at baseline ranged from 15.5 to 16.8 for panic, from 17.1 to 19.2 for SAD, from 13.9 to 15.1 for GAD, and from 16.7 to 17.3 for PTSD. Average total scores across the reviewed studies by disorder are shown in Fig. 2.

Table 1 Database used for review of SDS

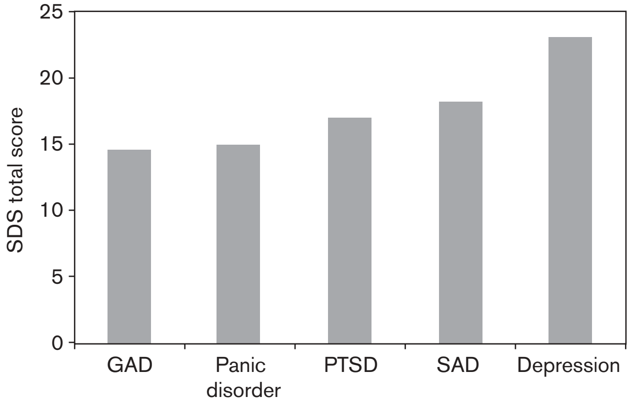
Author	Active medication (s)	Duration (weeks)	Design	*n*
PD				
Hoehn-Saric *et al.* (1993)	Fluvoxamine flexible dose	Eight weeks	DBPC	36
Sheehan *et al.* (1993)	Alprazolam flexible dose 1.5–10 mg/day; buspirone 15–100 mg/day	Eight weeks	DBPC	92
Michelson *et al.* (1998)	Fluoxetine fixed dose 10 or 20 mg/day	Ten weeks	DBPC	243
Asnis *et al.* (2001)	Fluvoxamine flexible dose 100–300 mg/day	Eight weeks	DBPC	188
Versiani *et al.* (2002)	Reboxetine	Eight weeks	DBPC	82
Sheehan *et al.* 2006:				
Lecrubier *et al.* (1997)	Paroxetine flexible dose 10–60 mg/day	Twelve weeks	DBPC	241
Data on file GSK	Paroxetine flexible dose 10–60 mg/day	Twelve weeks	DBPC	143
Ballenger *et al.* (1998)	Paroxetine fixed dose 10, 20, or 40 mg/day	Ten weeks	DBPC	268
SAD				
Stein MB, *et al.* (1999)	Fluvoxamine flexible dose 50–300 mg/day	Twelve weeks	DBPC	92
Liebowitz *et al.* (2002)	Paroxetine fixed dose 20, 40, or 60 mg/day	Twelve weeks	DBPC	384
Stein *et al.* (2002)	Paroxetine flexible dose 20–50 mg/day	Twelve weeks	DBPC	437
Davidson *et al.* (2004)	Fluvoxamine	Twelve weeks	DBPC	279
Lader *et al.* (2004)	Escitalopram fixed dose 5, 10, or 20 mg/day	Twelve weeks	DBPC	839
Westenberg *et al.* (2004)	Fluvoxamine CR flexible dose 100–300 mg/day	Twelve weeks	DBPC	300
Kasper *et al.* (2005)	Escitalopram, flexible dose 10–20 mg/day	Twelve weeks	DBPC	358
Sheehan *et al.* (2006):				
Baldwin *et al.* (1999)	Paroxetine fixed dose 10, 20, or 40 mg/day	Twelve weeks	DBPC	364
Baldwin *et al.* (1999)	Paroxetine flexible dose 20–50 mg/day	Twelve weeks	DBPC	274
Stein MB, *et al.* (1998)	Paroxetine flexible dose 20–50 mg/day	Twelve weeks	DBPC	187
GAD				
Pollack *et al.* (2001)	Paroxetine flexible dose 20–50 mg/day	Eight weeks	DBPC	324
Rickels *et al.* (2003)	Paroxetine fixed dose 20 or 40 mg/day	Eight weeks	DBPC	566
Rickels *et al.* (2004)	Venlafaxine ER flexible dose 75–225 mg/day	Twelve weeks	DBPC	261
Allgulander *et al.* (2006)	Duloxetine fixed dose 60 or 120 mg/day	Nine weeks	DBPC	507
Sheehan *et al.* (2006):		Eight weeks	DBPC	364
Hewett *et al.* (2001)	Paroxetine flexible dose 20–50 mg/day			
PTSD				
Tucker *et al.* (2001)	Paroxetine flexible dose 20–50 mg/day	Twelve weeks	DBPC	307
Marshall *et al.* (2001)	Paroxetine fixed dose 20 or 40 mg/day	Twelve weeks	DBPC	551
Davidson *et al.* (2005)	Tiagabine, Fluoxetine, Sertraline	Twelve weeks	OL	93
Stein MB, *et al.* (2003)	Paroxetine flexible dose 20–50 mg/day	Twelve weeks	DBPC	322
OCD				
Montgomery *et al.* (2001)	Citalopram fixed dose 20, 40, or 60 mg/day	Twelve weeks	DBPC	401
PMDD				
Cohen *et al.* (2004)	Paroxetine CR fixed dose 12.5 or 25 mg/day	Three cycles	DBPC	327
Freeman *et al.* (2005)	Escitalopram flexible dose 10–20 mg/day	Three cycles	DB	27
Landen *et al.* (2007)	Paroxetine CR intermittent or continuous flexible dose 10–40 mg/day	Three cycles	DBPC	167
Sheehan *et al.* (2006):				
Yonkers *et al.* (2003)	Paroxetine CR fixed dose 12.5 or 25 mg/day (all three studies)	Three cycles	DBPC	358
Yonkers *et al.* (2003)			DBPC	359
Bellew *et al.* (2003)			DBPC	366
MDD				
Smeraldi (1998)	Amisulpride fixed dose 50 mg/day or fluoxetine fixed dose 20 mg/day	Twelve weeks	DB	281
Detke (2004)	Duloxetine fixed dose 80 or 120 mg/day or paroxetine fixed dose 20 mg/day	Eight weeks	DBPC	367
Rush and Bose (2005)	Escitalopram flexible dose 10–20 mg/day	Eight weeks	OL	5453

DB, double-blind; DBPC, double-blind, placebo-controlled; GAD, generalized anxiety disorder; MDD, major depressive disorder; OCD, obsessive–compulsive disorder; OL, open label; PD, panic disorder; PMDD, premenstrual dysphoric disorder; PTSD, posttraumatic stress disorder; SAD, social anxiety disorder; SDS, Sheehan Disability Scale.

Mean subscale scores at baseline were available for nine of the 37 studies including one panic study (Sheehan *et al.*, 1993), four SAD studies (Stein *et al.*, 1999; Liebowitz *et al.*, 2002; Lader *et al.*, 2004; Kasper *et al.*, 2005), one GAD study (Rickels *et al.*, 2004), one PMDD study (Landen *et al.*, 2007), one OCD study (Montgomery *et al.*, 2001), and one MDD study (Rush and Bose, 2005). As shown in Fig. 3, mean baseline subscores for work and social disability were almost uniformly in the high moderate to marked range for all of the disorders. Mean family disability scores showed more variation with GAD and SAD characterized by relatively low scores and PMDD and MDD showing much higher ones (Table 2).

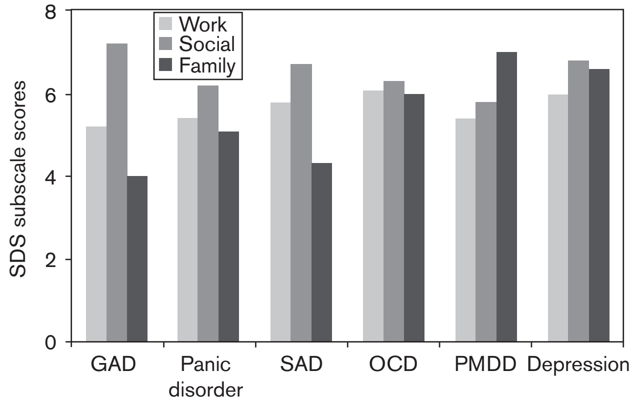
Three studies reported functional impairment at baseline in categorical as well as numeric terms. In a depression study, Rush and Bose (2005) provided the percent distributions of marked to extreme (score of 7–10) disability on the work, social, and family life subscales as 47, 61, and 58%, respectively. In a panic study, Sheehan *et al.* (1993) reported the percent distributions of moderate to extreme disability (score of 4 or higher) on these subscales as 67, 85, 67, and 85%. In another panic study, Asnis *et al.* (2001) provided the percentage with zero (or no disability) at baseline on the total SDS as 3.7% for the active treatment and 0 for placebo.

Fig. 2



Mean total Sheehan Disability Scale (SDS) scores by disorder at baseline. GAD, generalized anxiety disorder; PTSD, posttraumatic stress disorder; SAD, social anxiety disorder.

Fig. 3



Mean work, social, and family disability. Subscale scores at baseline. GAD, generalized anxiety disorder; OCD, obsessive-compulsive disorder; PMDD, premenstrual dysphoric disorder; SAD, social anxiety disorder; SDS, Sheehan Disability Scale.

Responsiveness

Remission and response

Only one of the 37 studies provided remission data. In that study, Asnis *et al.* (2001) reported the endpoint distributions of patients on active medication and placebo with no disability (score of 0) on the total SDS. Using the nonparametric Cochran–Mantel–Haenszel test, they found that the difference (29 vs. 14%) was statistically significant. Two studies supplemented mean endpoint data with categorical response data. In the open trial on escitalopram for depression Rush and Bose (2005) provided the percent distributions with mild or no disability (score of 0–3) at endpoint on the work, social, and family subscales as 58, 51, and 52% and indicated that only 5–6% continued to report extreme impairment. In our own study of alprazolam, buspirone and placebo for PD, we reported the percentages of patients in each group who continued to have moderate to extreme disability (score of 4 or higher) at endpoint on each of the subscales. We analyzed these differences, which favored alprazolam, with χ^2 (Sheehan *et al.*, 1993).

Table 2 **Baseline total SDS and subscale scores by type of disorder**

Type of disorder	Mean total disability	Mean work disability	Mean social disability	Mean family disability
GAD	14.6	5.2	7.2	4.0
Panic disorder	15.0	5.4	6.2	5.1
PTSD	17.0	–	–	–
SAD	18.2	5.8	6.7	4.3
OCD	–	6.1	6.3	6.0
PMDD	–	5.4	5.8	7
Depression	23.1	6	6.8	6.6

GAD, generalized anxiety disorder; OCD, obsessive–compulsive disorder; PMDD, premenstrual dysphoric disorder; PTSD, posttraumatic stress disorder; SAD, social anxiety disorder; SDS, Sheehan Disability Scale.

Endpoint differences

Six studies including two panic (Hoehn-Saric *et al.*, 1993; Asnis *et al.*, 2001), two SAD (Stein *et al.*, 2002; Davidson *et al.*, 2004), one PMDD (Freeman *et al.*, 2005), and one MDD (Smeraldi, 1998) study provided endpoint scores for the total SDS. Of these, four were placebo controlled. In two, significant ($P < 0.05$) drug placebo differences in the total SDS score were detected at endpoint (Asnis *et al.*, 2001; Davidson *et al.*, 2004). The magnitude of difference between drug and placebo was relatively small, that is, 2–3 points.

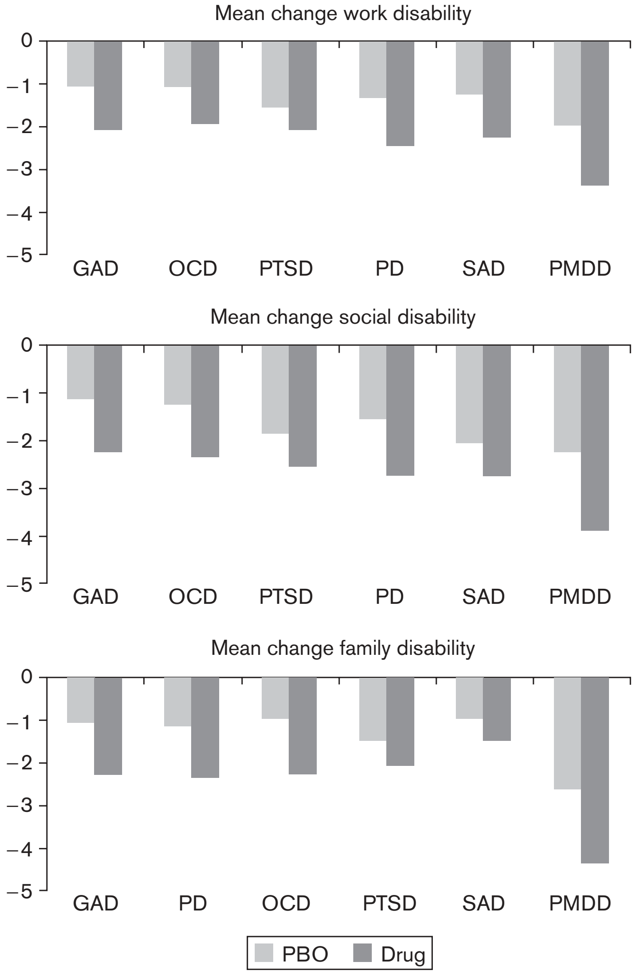
Four studies provided endpoint scores for the work, social, and family disability subscales (Sheehan *et al.*, 1993; Stein *et al.*, 1999; Liebowitz *et al.*, 2002; Rickels *et al.*, 2004). Of these, only three reported statistical comparisons between active drug and placebo. Significant ($P < 0.05$) drug placebo differences in favor of drug at endpoint were found on the work disability subscale in two of the studies (Sheehan *et al.*, 1993; Rickels *et al.*, 2004) on the social disability subscale in all three studies, and on the family subscale in one study (Sheehan *et al.*, 1993). The magnitude of drug placebo difference, when significant, ranged from 1 to 2 points, on these subscales.

Mean change

Six trials including one panic (Michelson *et al.*, 1998), one SAD (Westenberg *et al.*, 2004), two GAD (Pollack *et al.*, 2001; Rickels *et al.*, 2003), and two PTSD (Marshall *et al.*, 2001; Tucker *et al.*, 2001) trials reported mean change scores on the total SDS for drug and placebo. In all six (100%), significant drug placebo differences in favor of the drug were detected for at least one dose level of the active drug. Mean change in the SDS total score across studies for active treatments was 6.2 (range: 5.3–7.8) and for placebo was 4.2 (range: 2.8–5.8).

Mean change scores for the work, social, and family subscales were provided for 25 (67%) of the 37 studies (four panic, seven SAD, five GAD, two PTSD, five PMDD, one MDD, one OCD). Of these, all except one (the MDD study) were placebo controlled. Overall, the effects of the active treatment were significantly superior to those of placebo at the $P < 0.05$ level or less in 18 of the 25 studies (75%) for work disability, in 19 (79%) for social disability and in 18 (75%) for family disability. As shown in Fig. 4, drug placebo differences in mean change on the subscales typically ranged from 1 to 2 points. Differences in mean change, however, were clearly greater on some of the subscales for some disorders compared with others.

Fig. 4



Mean change on work, social, and family. Disability subscales by disorder. GAD, generalized anxiety disorder; OCD, obsessive-compulsive disorder; PBO, placebo; PD, panic disorder; PMDD, premenstrual dysphoric disorder; PTSD, posttraumatic stress disorder; SAD, social anxiety disorder.

Effect size and standardized response mean

Effect sizes (ESs) could be calculated for only four studies for the total SDS (Hoehn-Saric *et al.*, 1993; Smeraldi, 1998; Asnis *et al.*, 2001; Pollack *et al.*, 2001). In these studies, the ESs for the SDS for active treatments were all 0.9 or higher, indicating large effects. These large effects were similar to those for the HAM-A (1.0 vs. 1.1 in the Pollack study and 2.0 vs. 2.1 in the Smeraldi study) and for the CAS (0.9 vs. 1.1 in the Asnis study) but higher than those for the MADRS (0.9 vs. 0.64 in the Asnis study and higher than the ERD (2.0 vs. 1.15) in the Smeraldi study.

Standardized response means (SRMs) could also be calculated for only four studies for the total SDS (Michelson *et al.*, 1998; Marshall *et al.*, 2001; Rickels *et al.*, 2003; Westenberg *et al.*, 2004). The standard mean response was 0.75 for the Michelson study, 0.8 for the Rickels study and over 0.9 for the Westenberg and Marshall studies.

Interpretability

Interpreting the size of an effect or the magnitude or response in a way that can be standardized across studies requires information about the SD or standard error of the mean at baseline or that of the mean change. Although 12 (32%) of the 37 studies provided mean baseline SDS total scores, only six provided the SDs of these scores. Similarly, although most of the studies reported endpoint or mean change scores either for the total SDS or for the subscales or both, only a few reported the SD of that change. As a result, the ES or SRM could only be calculated for a small minority of the studies. None of the studies reported what they considered to be a MCID.

Discussion

The SDS was developed in the early 1980s to overcome some of the limitations of existing scales for impairment. Since then, other scales such as the Work and Social Adjustment Scale (WSAS), developed in 1986 by Marks (Mundt *et al.*, 2002), and the Quality of Life Enjoyment and Satisfaction Questionnaire (Q-LES-Q), developed by Endicott *et al.* (1993), have expanded the repertoire of brief but valid and reliable scales that are sensitive to change in functional impairment. Reliable and valid measurements of impairment continue to be necessary to help researchers assess new treatments for psychiatric disorders. No consensus as yet, however, has been reached on a standard to measure responsiveness. One objective of this study was to examine the ways in which responsiveness is analyzed using the Discan metric of the SDS.

Our results indicate that very few analyses of the SDS address remission or even response as a percentage in short-term clinical trials in the anxiety and mood disorders. This may be in part because whereas normalization of functioning is a critical aspect of remission, there are no agreed upon criteria for remission or response in relation to disability. In addition, there is good evidence that remission and response in disability and impairment lag well behind remission and response in the symptoms of psychiatric disorders (Hirschfeld *et al.*, 2002).

According to some reports, reductions in work impairment may not peak until 4–6 months after treatment has started, and in some cases the lag may be as long as 8 months (Weissman *et al.*, 1974; Giller *et al.*, 1988; Mintz *et al.*, 1992). Nonetheless, documentation of the percentage of patients who reach a predetermined threshold of remission or response in short-term treatment trials may be helpful in establishing short-term and longer-term benchmarks for disability.

Such data would also be helpful in calculating the 'number needed to treat' (NNT), an important but underused measure of relative treatment effectiveness (Pinson and Gray, 2003) that is increasingly being asked for in reporting the results of clinical trials (Altman *et al.*, 2001). Unfortunately, only two of the studies we reviewed (Sheehan *et al.*, 1993; Asnis *et al.*, 2001) provided SDS total or subscale percentage responses for active medication and placebo and therefore permitted calculation of the NNT to achieve remission or response on the SDS. The results of these calculations, however, were instructive. For the former study, the NNT was 7, indicating that for every seven PD patients treated with fluvoxamine for 8 weeks, there would have been one with a zero or no disability at endpoint on the total SDS. For the latter study, the NNTs were 4 for alprazolam for the work disability subscale and 3 for the social and family subscales, indicating that for every four PD patients treated with alprazolam for 8 weeks, there would have been four patients with an endpoint score of ≤ 3 (mild disability or less) on the work subscale and three patients with endpoint scores of ≤ 3 on the social and family subscales. By contrast, the NNTs in that study for buspirone for the work, social, and family subscales were 14, 50, and 9, indicating that many more patients would have needed to be treated to achieve endpoint score of ≤ 3 on these SDS subscales.

Although absolute cutoff scores such as ≤ 1 on each of the SDS subscales and < 5 on the total SDS have been proposed as potential remission criteria, our experience is that less restrictive criteria are more appropriate. In a recent reanalysis of 915 observations from a PD study (Sheehan *et al.*, 1993), we examined the mean SDS total and subscale scores for remission and response, designated as ≤ 7 and 8–14 on the HAM-A. For remission, categorized as ≤ 7 on the HAM-A (104 observations), the mean SDS scores were 1.9, 2.1, 1.7 on the work, social, and family subscales and the mean total SDS score was 5.9. For response, categorized as 8–14 on the HAM-A (252 observations), the mean SDS scores for the work, social, and family subscales were 3.7, 4.0, and 3.3 and the mean total SDS score was 11.0. On the basis of these preliminary observations, which need to be validated using data from other studies and data for other disorders, we recommend that scores of ≤ 2 on the subscales and ≤ 6 on the total SDS be used for remission and that scores of ≤ 4 on the subscales and ≤ 12 on the total SDS be used for response.

Overall, the SDS proved sensitive to treatment effects (responsive) when mean change or endpoint differences were analyzed. In addition, the evidence from three of the studies we reviewed indicated that group comparisons at endpoint of patients with no disability (score of 0), mild disability (score of 3 or less), or alternatively those with continued moderate to extreme disability (score of 4 or higher) also reliably discriminate between responders from nonresponders.

For studies using mean change or endpoint differences, active treatments appear to separate out statistically from inactive ones when there is a group difference of ~ 4 points in the total SDS or 1–2 points in a subscale. These numbers may be helpful in establishing benchmarks for MCIDs for the SDS in clinical treatment studies in psychiatry.

Most of the studies we reviewed failed to provide SDs for baseline, endpoint or mean change scores. These statistics would be helpful in generating ESs and standardized mean responses for treatments so that different treatments could be compared within and across populations and disorders. At this time, there is no consensus on standards for evaluating or interpreting the ESs of treatment-related research. Still, there is an increasing demand for information about the magnitude of change related to different treatments and how relevant or meaningful such change is. Although calculation of ESs and SRMs can never substitute for statistical testing, they may have value in assessing the relative importance of treatment effects and making results more interpretable. Impairment in work, social, or family functioning may be a function of continuing symptoms or of other characteristics of patients, for example, age, sex, socioeconomic status, physical health, or a combination of these. Better understanding of patient characteristics contributing to continued impairment in the presence of reduced symptomatology would be helpful.

Our review was limited by the inclusion criteria we used. We limited our analysis to short-term clinical psychopharmacological trials for anxiety disorders, depression, and PMDD. Publication bias may also have distorted our results as published studies are more likely than nonpublished ones to report treatment effects. Nonetheless, the evidence suggests that the SDS, with its Discan metric, is sensitive to treatment effects. Future research should examine whether the SDS is equally sensitive in long-term outcome studies, in studies for other disorders, and in studies using nonpsychopharmacological treatments, for example, CBT and cognitive behavioral group treatment (CBGT).

Conclusion

The modified Discan design used in the SDS has emerged as a concise efficient metric that is very sensitive to placebo drug differences in treatment outcome studies. Future reports with this metric would be enriched by more precise and standardized reporting of results along the lines recommended above and may permit comparison of results across studies with different treatments (including nonpsychopharmacological treatments such as CBT) and with different disorders. In addition, there is a need to supplement parametric analyses of mean change and endpoint differences with nonparametric data showing the percentage meeting response and remission criteria. In addition, the percentages with endpoint scores of zero should be reported.

Acknowledgement

This study is supported by a research grant from Merck & Co. Inc., Pennsylvania, USA.

References

Allgulander C, Koponen H, Erickson J, Pritchett Y, Detke M, Ball S, *et al.* (2006). Duloxetine as an effective treatment for improving painful physical symptoms and functioning associated with generalized anxiety disorder. Poster presented at 26th Annual Conference of ADAA, March 2006, Miami, Florida.

Altman DG, Schulz KE, Moher D, Egger M, Davidoff F, Elbourne D, *et al.* (2001). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* **134**:663–694.

Asnis GM, Hameedi FA, Goddard AW, Potkin SG, Black D, Jameel M, *et al.* (2001). Fluvoxamine in the treatment of panic disorder: a multi-center, double-blind, placebo-controlled study in outpatients. *Psychiatry Res* **103**: 1–14.

Baldwin D, Bobes J, Stein DJ, Scharwachter I, Faure M (1999). Paroxetine in social phobia/social anxiety disorder. Randomised, double-blind, placebo-controlled study. Paroxetine Study Group. *Br Psychiatry* **175**:120–126.

Ballenger JC (1999). Clinical guidelines for establishing remission in patients with depression and anxiety. *J Clin Psychiatry* **60** (**Suppl 20**):29–34.

Ballenger JC, Burrows GD, DuPont RL, Lesser IM, Noyes R, Pecknold JC (1988). Alprazolam in panic disorder and agoraphobia: results from a multicenter trial: I. efficacy in short-term treatment. *Arch Gen Psychiatry* **45**:413–422.

Ballenger JC, Wheadon DE, Steiner M, Bushnell W, Gergel IP (1998). Double-blind, fixed-dose, placebo-controlled study of paroxetine in the treatment of panic disorder. *Am J Psychiatry* **155**:36–42.

Bech P (2004). Modern psychometrics in clinimetrics: impact on clinical trials of antidepressants. *Psychother Psychosom* **73**:134–138.

Bellew KM, Cohen LS, Lambert J, Bridges IM, McCafferty JP. Longterm treatment of PMDD. American Psychiatric Association Annual Meeting. San Francisco, California, USA, 2003.

Bilsbury CD, Richmond AC (2002). A staging approach to measuring patient-centered subjective outcomes. *Acta Psychiatr Scand* **106** (**s414**):5–40.

Broadhead WE, Blazer DG, George LK, Tse CKJ (1990). Depression, disability days and days lost from work in a prospective epidemiologic survey. *JAMA* **264**:2524–2528.

Cohen J (1977). *Statistical power analysis for the behavioral sciences*, rev. ed. New York: Academic Press.

Cohen LS, Soares CN, Yonkers KA, Bellew KM, Bridges IM, Steiner M (2004). Paroxetine controlled release for premenstrual dysphoric disorder: a double-blind, placebo controlled trial. *Psychosom Med* **66**:707–713.

Cox EP (1980). The optimal number of response alternatives for a scale: a review. *J Mark Res* **17**:407–422.

Davidson J, Yaryura-Tobias J, DuPont R, Stallings L, Barbato LM, van der Hoop RG, *et al.* (2004). Fluvoxamine-controlled release formulation for the treatment of generalized social anxiety disorder. *J Clin Psychopharmacol* **24**:118–125.

Davidson JR, Hughes D, Blazer D, George LK (1991). Posttraumatic stress disorder in the community: an epidemiological study. *Psych Med* **21**:1–19.

Davidson JR, Payne VM, Connor KM, Foa EB, Rothbaum BO, Hertzberg MA, *et al.* (2005). Trauma, resilience and saliostasis: effects of treatment in post-traumatic stress disorder. *Int Clin Psychopharm* **20**:43–48.

Detke MJ, Wiltse CG, Mallinckrodt CH, McNamara RK, Demitrack MA, Bitter I (2004). Duloxetine in the acute and long-term treatment of major depressive disorder: a placebo- and paroxetine-controlled trial. *Eur Neuropsychopharm* **14**:457–470.

Deyo R, Diehr P, Patrick DL (1991). Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. *Control Clin Trials* **12 (4 Suppl)**:142S–158S.

Endicott J, Spitzer RL, Fleiss JF, Cohen J (1976). The Global Assessment Scale: a procedure for measuring overall severity of psychiatric disturbance. *Arch Gen Psychiatry* **33**:766–771.

Endicott J, Nee J, Harrison W, Blumenthal R (1993). Quality of life enjoyment and satisfaction questionnaire: a new measure. *Psychopharmacol Bull* **29**: 321–326.a.

Faravelli C (2004). Assessment of psychopathology. *Psychother Psychosom* **7**:139–141.

Fehm L, Pelissolo A, Furmark T, Wittchen HU (2005). Size and burden of social phobia in Europe. *Eur Neuropsychopharmacol* **15**:453–462.

Feinstein AR (1983). An additional science for clinical medicine. IV. The development of clinimetrics. *Ann Intern Med* **99**:843–848.

Frank E, Prien RF, Jarrett RB, Keller MB, Kupfer DJ, Lavori PW, *et al.* (1991). Conceptualization and rationale for consensus definitions of terms in major depressive disorder. Remission, recovery, relapse, and recurrence. *Arch Gen Psychiatry* **48**:851–855.

Freeman EW, Sondheimer SJ, Sammel MD, Ferdousi T, Lin H (2005). A preliminary study of luteal phase vs. symptom-onset dosing with escitalopram for premenstrual dysphoric disorder. *J Clin Psychiatry* **66**:769–773.

Friedman HH, Friedman LW (1986). On the danger of using too few points in a rating scale: a test of validity. *J Data Collection* **26**:60–63.

Friedman HH, Wilamowsky Y, Friedman LW (1981). A comparison of balanced and unbalanced rating scales. *Mid-Atlantic J Bus* **19** (**Summer**):1–7.

Friedman HH, Cohen D, Amoo T (2003). Label or position: which has the greater impact on subjects' responses to a rating scale? *J Int Mark Mark Res* **28**: 77–81.

Galassi F, Quercioli S, Charismas D, Niccolai V, Barciulli E (2007). Cognitive-behavioral group treatment for panic disorder with agoraphobia. *J Clin Psychol* **63**:409–416.

Giller E Jr, Bialos D, Riddle MA, Waldo MC (1988). MAOI treatment response: multiaxial assessment. *J Affect Disord* **14**:171–175.

Guyatt G, Walter S, Norman G (1987). Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis* **2**:171–178.

Herbert JD, Gaudiano BA, Rheingold AA, Myers VH, Dalrymple K, Nolan E (2005). Social skills training augments the effectiveness of Cognitive Behavioral Group Therapy for social anxiety disorder. *Beh Ther* **36**: 125–138.

Hewett K, Adams A, Bryson H, et al. Generalized anxiety disorder-efficacy of Paroxetine. Presented at the 7th World Congress of Biological Psychiatry, Berlin, Germany, 1–6 July 2001.

Hirschfeld RM, Dunner DL, Keitner G, Klein DN, Koran LM, Kornstein SG (2002). Does psychosocial functioning improve independent of depressive symptoms? A comparison of nefazodone, psychotherapy, and their combination. *Biol Psychiatry* **51**:123–133.

Hoehn-Saric R, McLeod DR, Hipsley PA (1993). Effect of fluvoxamine on panic disorder. *J Clin Psychopharmacol* **13**:321–326.

Hollifield M, Katon W, Skipper B, Chapman T, Ballenger JC, Mannuzza S, *et al.* (1997). Panic disorder and quality of life: variables predictive of functional impairment. *Am J Psychiatry* **154**:766–772.

Jaeschke R, Singer J, Guyatt G (1989). Measurement of health status: ascertaining the minimal clinically important difference. *Control Clin Trials* **10**:407–415.

Juniper EF, Guyatt GH, Willan A, Griffith LE (1994). Determining a minimal important change in a disease-specific quality of life questionnaire. *J Clin Epidemiol* **47**:81–87.

Kasper S, Stein DJ, Loft H, Nil R (2005). Escitalopram in the treatment of social anxiety disorder: randomised, placebo-controlled, flexible-dosage study. *Br J Psychiatry* **186**:222–226.

Katzelnick DJ, Simon GE, Pearson SO, Manning WG, Helstad CP, Henk HJ. (2000). Randomized trial of a depression management program in high utilizers of medical care. *Arch Fam Med* **9**:345–351.

Kazis LE, Anderson JJ, Meenan RF (1989). Effect sizes for interpreting changes in health status. *Medical Care* **27**:S178–S189.

Keller M, Lavori PW, Goldenberg IM, Baker LA, Pollack MH, Sachs GS, *et al.* (1993). Influence of depression on the treatment of panic disorder with impramine, alprazolam and placebo. *J Affect Disord* **28**:27–38.

Keller MB (2003). Past, present, and future directions for defining optimal treatment outcome in depression: remission and beyond. *JAMA* **289**: 3152–3160.

Kelsey JE (2001). Clinician perspective on achieving and maintaining remission in depression. *J Clin Psychiatry* **62** (**Suppl 26**):16–22.

Kessler RC, DuPont RL, Berglund P, Wittchen HU (1999). Impairment in pure and comorbid generalized anxiety disorder and major depression at 12 months in two national surveys. *Am J Psychiatry* **156**:1915–1923.

Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Rush AJ, *et al.* (2003). The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *JAMA* **289**:3095–3105.

Kessler R, Chiu WT, Jin R, Ruscio AM, Shear K, Walters EE (2006). The epidemiology of panic attacks, panic disorder, and agoraphobia in the National Comorbidity Survey Replication. *Arch Gen Psychiatry* **63**:415–424.

Lader M, Stender K, Burger V, Nil R (2004). Efficacy and tolerability of escitalopram in 12- and 24-week treatment of social anxiety disorder: randomised, double-blind, placebo-controlled, fixed-dose study. *Depress Anxiety* **19**:241–248.

Landen M, Nissbrandt H, Allgulander C, Sorvik K, Ysander C, Eriksson E (2007). Placebo-controlled trial comparing intermittent and continuous paroxetine in premenstrual dysphoric disorder. *Neuropsychopharmacology* **32**:153–161.

Lecrubier Y (2002). How do you define remission? *Acta Psychiatr Scand* **106** (**S415**):7–11.

Lecrubier Y, Judge R (1997). Long-term evaluation of paroxetine, clomipramine and placebo in panic disorder. Collaborative Paroxetine Panic Study Investigators. *Acta Psychiatr Scand* **95**:153–160.

Lecrubier Y, Wittchen H, Faravelli C, Bobes J, Patel A, Knapp M (2000). A European perspective on social anxiety disorder. *Eur Psychiatry* **15**:5–15.

Leon AC, Shear K, Portera L, Klerman GL (1992). Assessing impairment in patients with panic disorder: the Sheehan Disability Scale. *Soc Psychiatry Psychiatr Epidemiol* **27**:78–82.

Leon AC, Portera L, Weissman MM (1995). The social costs of anxiety disorders. *Br J Psychiatry* **166** (**Suppl 27**):19–22.

Leon AC, Olfson M, Portera L, Farber L, Sheehan DV (1997). Assessing psychiatric impairment in primary care with the Sheehan Disability Scale. *Int J Psychiatry Med* **27**:93–105.

Liang MH (1995). Evaluating measurement responsiveness. *J Rheumatol* **22**:1191–1192.

Liang MH, Larson MG, Cullen KE, Schwartz JA (1985). Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. *Arthritis Rheum* **28**:542–547.

Liang MH, Fossel AH, Larson MG (1990). Comparisons of five health status instruments for orthopedic evaluation. *Med Care* **28**:632–642.

Liebowitz MR, Stein MB, Tancer M, Carpenter D, Oakes R, Pitts CD (2002). A randomized, double-blind, fixed-dose comparison of paroxetine and placebo in the treatment of generalized social anxiety disorder. *J Clin Psychiatry* **63**:66–74.

Marcus SC, Olfson M, Pincus HA, Shear MK, Zarin DA (1997). Self-reported anxiety, general medical conditions, and disability bed days. *Am J Psychiatry* **154**:1766–1768.

Markowitz JS, Weissman MM, Oulette R, Lish JD, Klerman GL (1989). Quality of life in panic disorder. *Arch Gen Psychiatry* **46**:984–992.

Marshall RD, Beebe KL, Oldham M, Zaninelli R (2001). Efficacy and safety of paroxetine treatment for chronic PTSD: a fixed dose, placebo controlled study. *Am J Psychiatry* **158**:1982–1988.

McQuaid JR, Granholm E, McClure FS, Roepke S, Pedrelli P, Patterson TL, Jeste DV (2000). Development of an integrated cognitive behavioral and social skills training intervention for older patients with schizophrenia. *J Psychother Pract Res* **9**:149–156.

Michelson D, Lydiard RB, Pollack MH, Tamura RN, Hoog SL, Tepner R, *et al.* (1998). Outcome assessment and clinical improvement in panic disorder: evidence from a randomized controlled trial of fluoxetine and placebo. The Fluoxetine Panic Disorder Study Group. *Am J Psychiatry* **155**:1570–1577.

Middel B, Von Sonderen E. (2002). Statistical significant change vs. relevant or important change in (quasi) experimental design: some conceptual and methodological problems in estimating magnitude of intervention-related change in health services research. *Int J Integr Care* **2**:1–18.

Milrod B, Busch F, Leon AC, Shapiro T, Aronson A, Roiphe J, *et al.* (2000). Open trial of psychodynamic psychotherapy for panic disorder: a pilot study. *Am J Psychiatry* **157**:1878–1880.

Milrod B, Busch F, Leon AC, Aronson A, Roiphe J, Rudden M *et al.* (2001). A pilot open trial of brief psychodynamic psychotherapy for panic disorder. *J Psychother Prac Res* **10**:239–245.

Mintz J, Mintz LI, Arruda MJ, Hwang SS (1992). Treatments of depression and the functional capacity to work. *Arch Gen Psychiatry* **49**:761–768.

Montgomery SA, Asberg M (1979). A new depression scale designed to be sensitive to change. *Br J Psychiatry* **134**:382–389.

Montgomery SA, Kasper S, Stein DJ, Bang Hedegaard K, Lemming OM (2001). Citalopram 20, 40 and 60 mg are all effective and well tolerated compared with placebo in obsessive-compulsive disorder. *Int Clin Psychopharmacol* **16**:75–86.

Mundt JC, Marks IM, Shear MK, Greist JH (2002). The Work and Social Adjustment Scale: a simple measure of impairment in functioning. *Br J Psychiatry* **180**:461–464.

Murray CJL, Lopez AD. (1996). *The global burden of disease*. Cambridge, Massachusetts: Harvard University Press.

Noel PH, Williams JW, Unutzer J, Worchel J, Lee S, Cornell J, *et al.* (2004). Depression and comorbid illness in elderly primary care patients: impact on multiple domains of health status and well-being. *Ann Fam Med* **2**:555–562.

Olfson M, Fireman B, Weissman M, Leon AC, Sheehan DV, Kathol RG, *et al.* (1997). Mental disorders and disability among patients in primary care group practice. *Am J Psychiatry* **154**:1734–1740.

Olfson M, Shea S, Feder A, Fuentes M, Normura Y, Gameroff MA, Weissman M (2000). Prevalence of anxiety, depression and substance use disorders in an urban general medical practice. *Arch Fam Med* **9**:876–883.

Pinson L, Gray GE (2003). Number needed to treat: an underused measure of treatment effect. *Psychiatr Serv* **54**:145–146, 154.

Pollack MH, Zaninelli R, Goddard A, McCafferty JP, Bellew KM, Burnham DB, *et al.* (2001). Paroxetine in the treatment of generalized anxiety disorder: results of a placebo-controlled, flexible-dosage trial. *J Clin Psychiatry* **62**:350–357.

Rickels K, Zaninelli R, McCafferty J, Bellew K, Iyengar M, Sheehan D (2003). Paroxetine treatment of generalized anxiety disorder: a double-blind, placebo-controlled study. *Am J Psychiatry* **160**:749–756.

Rickels K, Mangano R, Khan A (2004). A double-blind, placebo-controlled study of a flexible dose of venlafaxine ER in adult outpatients with generalized social anxiety disorder. *J Clin Psychopharmacol* **24**:488–496.

Rosenberg R (2000). Outcome measures of antidepressive therapy. *Acta Psychiatr Scand* **101** (**s402**):41–44.

Rush AJ, Bose A (2005). Escitalopram in clinical practice: results of an open-label trial in a naturalistic setting. *Depress Anxiety* **21**:26–32.

Schulberg HC, Block MR, Madonia MJ, Scott CP, Rodriguez E, Imber SD, *et al.* (1996). Treating major depression in primary care practice; eight month clinical outcome. *Arch Gen Psychiatry* **53**:913–919.

Schwartz N, Knauper B, Hippler HJ, Noelle-Neumann E, Clark L (1991). Rating scales: numeric values may change the meaning of scale labels. *Public Opin Q* **55**:570–582.

Shapiro MB (1961). A method of measuring psychological change specific to the individual psychiatric patient. *Br J Med Psychol* **34**:151–155.

Sheehan D, Christie J, Dube E (2006). Paroxetine improves the functional disability associated with mood and anxiety disorders. Poster presented at 26th Annual Conference of ADAA, March 2006, Miami, Florida.

Sheehan DV (1983). *The anxiety disease*. New York: Charles Scribner and Sons.

Sheehan DV (1985). Monoamine oxidase inhibitors and alprazolam in the treatment of panic disorder and agoraphobia. *Psychiatr Clin North Am* **8**:49–62.

Sheehan DV, Raj AB, Harnett-Sheehan K, Soto S, Knapp E (1993). The relative efficacy of high-dose buspirone and alprazolam in the treatment of panic disorder: a double-blind placebo controlled study. *Acta Psychiatr Scand* **88**:1–11.

Sheehan DV, Harnett-Sheehan K, Raj BA (1996). The measurement of disability. *Int Clin Psychopharmacol* **11** (**Suppl 3**):89–95.

Singh AC, Bilsbury CD (1982). Scaling subjective variables by SPC (sequential pair comparisons). *Behav Psychother* **10**:128–145.

Singh AC, Bilsbury CD (1989a). Measurement of subjective variables: the Discan method. *Acta Psychiatr Scand* **79** (**Suppl 347**):1–38.

Singh AC, Bilsbury CD (1989b). Measuring levels of experiential states in clinical applications by Discan: a discretized analog method. *Behav Psychother* **17**:27–41.

Smeraldi E (1998). Amisulpride vs. fluoxetine in patients with dysthymia or major depression in partial remission: a double-blind, comparative study. *J Affect Disord* **48**:47–56.

Stein DJ, Davidson J, Seedat S, Beebe K (2003). Paroxetine in the treatment of post-traumatic stress disorder: pooled analysis of placebo-controlled studies. *Expert Opin Pharmacother* **4**:1829–1838.

Stein DJ, Versiani M, Hair T, Kumar R (2002). Efficacy of paroxetine for relapse prevention in social anxiety disorder: a 24-week study. *Arch Gen Psychiatry* **59**:1111–1118.

Stein MB, Fyer AJ, Davidson JR, Pollack MH, Wiita B (1999). Fluvoxamine treatment of social phobia (social anxiety disorder): a double-blind, placebo-controlled study. *Am J Psychiatry* **156**:756–760.

Stein MB, Liebowitz MR, Lydiard RB, Pitts CD, Bushnell W, Gergel I (1998). Paroxetine treatment of generalized social phobia (social anxiety disorder): a randomized controlled trial. *JAMA* **280**:708–713.

Streiner DL, Norman GR (1995). *Health measurement scales: a practical guide to their development and use*. New York: Oxford University Press; pp. 163–180.

Thase ME (2003). Effectiveness of antidepressants: comparative remission rates. *J Clin Psychiatry* **64** (**Suppl 2**):3–7.

Tucker P, Zaninelli R, Yehuda R, Ruggiero L, Dillingham K, Pitts CD (2001). Paroxetine in the treatment of chronic posttraumatic stress disorder: results of a placebo-controlled, flexible-dosage trial. *J Clin Psychiatry* **62**:860–868.

Versiani M, Cassano G, Perugi G, Benedetti A, Mastalli L, Nardi A, *et al*. (2002). Reboxetine, a selective norepinephrine reuptake inhibitor, is an effective and well-tolerated treatment for panic disorder. *J Clin Psychiatry* **63**:31–37.

Ware JE, Sherbourne CD (1992). The MOS 36-item short form health survey (SF-36): I conceptual framework and item selection. *Med Care* **30**:473–483.

Weissman MM, Bothwell S (1976). Assessment of social adjustment by patient self report. *Arch Gen Psychiatry* **33**:111–1115.

Weissman MM, Klerman GL, Paykel ES, Prusoff B, Hanson B (1974). Treatment effects on the social adjustment of depressed patients. *Arch Gen Psychiatry* **30**:771–778.

Wells KB, Stewart A, Hays RD (1989). The functioning and well-being of depressed patients: results of the Medical Outcome Study. *JAMA* **262**: 914–919.

Westenberg HG, Stein DJ, Yang H, Li D, Barbato LM (2004). A double-blind placebo-controlled study of controlled release fluvoxamine for the treatment of generalized social anxiety disorder. *J Clin Psychopharmacol* **24**:49–55.

Wildt AR, Mazis MB (1978). Determinants of scale response: label vs. position. *J Mark Res* **15**:261–267.

World Health Organization. The World Health Report 2004: Changing History, Annex Table 3: Burden of disease in DALYs by cause, sex, and mortality stratum in WHO regions, estimates for 2002. Geneva: WHO.

The WHO World Mental Health Survey Consortium (2004). Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. *JAMA* **291**:2581–2590.

World Mental Health (2004). About the WMH-CIDI. http://www.hcp.med.harvard.edu/wmhcidi/about.php.

Yonkers KA, Bellew KM, Rolfe TE, Perera P. Pooled analysis of three large clinical trials in the treatment of PMDD. American Psychiatric Association (APA) Annual Meeting, May 17–22, 2003, San Francisco, California, USA.