How should the severity of depression be rated on self-report depression scales?

Mark Zimmerman[a,b,*], Sophie Kerr[a,b]

[a] Department of Psychiatry and Human Behavior, Brown Medical School, Rhode Island Hospital, Providence, United States
[b] Department of Psychiatry, Rhode Island Hospital, Providence, United States

ARTICLE INFO

ABSTRACT

Almost all depression measures have been developed without discussing how to best conceptualize and assess the severity of depression. The most valid rating format of depression severity scales is unsettled and has been little studied. In the present study from the Rhode Island Methods to Improve Diagnostic Assessment and Services (MIDAS) project, we compared the validity of alternative approaches towards rating the severity of depressive symptoms. Data was collected using TurkPrime. One hundred eighty-five participants currently in treatment for a mental health problem completed a self-report measure of depression and rated the symptoms on two 4-point ordinal scales assessing symptom frequency and symptom intensity. The respondents also rated their global level of depression severity and completed a measure of psychosocial functioning and quality of life. The symptom ratings based on intensity and frequency were highly correlated with each other, and equally highly correlated with subjects' global rating of overall severity of depression, as well as ratings of psychosocial functioning and quality of life. A composite index of severity based on the sum of frequency and severity ratings was no more highly correlated with the external validators. The results of the present study suggest that ratings of depressive symptoms based on either symptom intensity or symptom frequency are equally valid.

1. Introduction

The severity of depression has been most frequently quantified on paper-and-pencil and clinician-administered rating scales. There are numerous instruments assessing depressive symptoms, with variability amongst them in the time frame covered (the two most common time frames being the past one or two weeks), rating guidelines (most scales use Likert-type ratings based on symptom frequency, persistence or intensity), and item content.

Descriptions of scale construction typically focus on the content of the measure. Some measures are linked to the symptom criteria that are used to diagnose depression (Kroenke et al., 2001; Olsen et al., 2003; Rush et al., 2003; Zimmerman et al., 2008), whereas others assess a broad range of features that patients indicate are most important in measuring outcome (Zimmerman et al., 2011) or assess a range of diagnostic and associated symptoms of depression (Rush et al., 1996). In developing the Multidimensional Depression Assessment Scale, Cheung and Power (2012) reviewed the content of fifteen depression scales and how the scale they had developed would address a content gap. There was no discussion, however, of rating formats and why a symptom frequency format was chosen for their measure rather than a rating

format based on symptom intensity. Rarely do scale developers discuss the reason for choosing a particular rating format.

Little research has examined which rating parameters provide the most valid indicator of depression severity. Is the severity of depression best conceptualized as the number of symptoms (i.e., present or absent), frequency of symptoms (e.g., every day vs. half the days vs. few days), persistence of symptoms (e.g., always present vs. often present vs. sometimes present), or intensity of symptoms (e.g., severe vs. moderate vs. mild)? Williams et al. (2008), in standardizing the scoring of the Hamilton Depression Rating Scale (HAMD), created a grid scoring format to incorporate information regarding symptom frequency/persistence and intensity in the ratings. In developing the Patient-Reported Outcomes Measurement Information System (PROMIS) depression scale, Pilkonis et al. (2011) reviewed studies comparing alternative response options and concluded that frequency scaling outperformed intensity ratings, though the studies reviewed were not studies of depression ratings.

The most valid rating format of depression severity scales is unsettled and has been little studied. We are aware of only one study comparing alternative rating formats in assessing depression. Parker and colleagues examined whether it is important to consider both

---

intensity and frequency constructs and found that symptom intensity was a better indicator of severity than symptom frequency (Parker et al., 1994).

In the present study from the Rhode Island Methods to Improve Diagnostic Assessment and Services (MIDAS) project, we compared the validity of alternative approaches towards rating the severity of depressive symptoms. We hypothesized that ratings based on both symptom intensity and symptom frequency would be more highly correlated with ratings of psychosocial impairment and quality of life than ratings based on symptom intensity alone or symptom frequency alone.

2. Methods

Data was collected using TurkPrime, an integration of research platforms that allows for efficient collection of large samples of data. TurkPrime manages a panel of Mechanical Turk respondents and partners with multiple other online market research platforms to achieve a global reach of over 15 million respondents. The respondents are offered the opportunity to complete surveys and other brief tasks online from the location of their choosing for compensation. The amount and type of compensation (e.g. cash, gift cards) depends on which online platform the participants were drawn from. Researchers were provided with a financial quote for data collection and TurkPrime managed the compensation of participants. The total cost was less than $8.00 per participant.

The respondent base of TurkPrime is referred to as the Prime Panel and participants must be 18 years old. TurkPrime integrates with the widely used Amazon Mechanical Turk (MTurk) as well as other online platforms and provides services to improve the quality and functionality of data collection. Unlike MTurk, TurkPrime offers participants the opportunity to provide demographic and behavioral information on an ongoing basis for no compensation and then uses this information to later recruit certain types of samples (e.g., Davidai, 2018; Skrzynski et al., 2018). This approach is preferable to studies that explicitly state inclusion or exclusion criteria in which participants are more likely to respond dishonestly in order to be able to participate in the study (Chandler and Shapiro, 2016).

For the current study, TurkPrime provided researchers with a financial quote to recruit and compensate a sample of 180 participants currently in treatment for a mental health problem. TurkPrime assessed current psychiatric treatment as one of their ongoing pre-screening questions using the statement, "Are you currently in treatment for a mental health problem?" Those who endorsed current psychiatric treatment were presented with a link to complete our survey on Qualtrics Survey Software as one of the available tasks on their online dashboard. After following the link, participants were given the opportunity to consent. Those who refused consent were navigated away from the study. Participants then completed the survey which included demographic questions, questions regarding their psychiatric treatment history, a depression scale with both intensity and frequency ratings of symptoms, and scales measuring functioning and quality of life. The survey concluded with a debriefing message including the research coordinator's email to contact with questions and concerns as well as resources to locate mental health professionals in their area.

The survey included 3 'attention check' questions as a measure of validity to ensure that participants were reading and responding to each item. The attention check questions followed the same format as the other survey question but directed the respondent to select a specific answer. For example, the attention check item embedded in the depression scale frequency items stated "Please select 0 days." There was no symptom listed for this item, rather just the instruction to select 0 days. Analogously, on the intensity subscale the respondent was instructed to select "not present" for an item where a symptom was not listed. Following the depression scale an item was included which again simply directed the respondent to select a particular response option.

Participants were excluded from analyses if they failed these attention checks or did not complete the majority of the survey. This allows researchers to check whether participants are reading each question and exclude those who do not follow the instructions in the "attention check" items. TurkPrime recruited additional participants to replace the excluded subjects. Given a slight overflow in testing, this process yielded our sample of 185 participants.

2.1. Measures

The depression scale was based on a modification of the Clinically Useful Depression Outcome Scale (CUDOS) (Zimmerman et al., 2008). The CUDOS includes 16 items assessing the DSM-IV/DSM-5 symptom criteria for MDD. Compound diagnostic criteria referring to more than one construct (e.g. problems concentrating or making decisions; insomnia or hypersomnia) were subdivided into their respective components and a CUDOS item assesses each component. The individual symptoms assessed by the CUDOS are: depressed mood, loss of interest in usual activities, low energy, psychomotor agitation, psychomotor retardation, guilt, worthlessness, thoughts of death, suicidal ideation, impaired concentration, indecisiveness, decreased appetite, increased appetite, insomnia, hypersomnia, and hopelessness. To the original version of the CUDOS we added 2 items: loss of pleasure in usual activities and low motivation.

The respondent was instructed to rate the symptom items on two 4-point ordinal scales indicating "how well the item describes you during the past week, including today." Symptom frequency was rated as follows: 1 = 0 days, 2 = 1–2 days, 3 = 3–5 days, 4 = 6–7 days). Symptom severity was rated as 1 = not present, 2 = mild, 3 = moderate, 4 = severe. The 2 sets of symptom ratings were displayed side-by-side.

After completion of the symptom ratings the subjects rated their global level of depression severity. The subjects were asked to rate the "overall level of severity of your symptoms of depression during the past week including today" on a 5-point ordinal scale (none, minimal, mild, moderate, severe).

Psychosocial impairment due to depression and quality of life were assessed with the subscales from the Diagnostic Inventory for Depression (DID) (Zimmerman et al., 2004). The six-item psychosocial functioning subscale assesses the amount of difficulty symptoms of depression have caused in usual daily responsibilities, relationships with significant others such as spouse, relationships with close family members, relationships with friends, participation in leisure activities, and overall level of function. Items are rated on a 5-point ordinal scale (0 = no difficulty; 4 = extreme difficulty). The quality of life subscale assesses satisfaction with the same areas covered by the psychosocial functioning scale as well as global satisfaction with mental health and physical health. Items are rated on a 5-point Likert scale (0 = very satisfied; 4 = very dissatisfied).

2.2. Data analysis

We computed a total score for the intensity ratings, the frequency ratings, and the sum of intensity and frequency ratings. For the 3 subscales (intensity, frequency, sum) we conducted 3 analyses. First, at both the item level and total subscale level, we examined the correlation between the frequency and intensity ratings, and the correlation between each of these subscales with the sum total of frequency and intensity scores. Second, we examined the internal consistency and item-scale correlations for all 3 subscales. And third, we examined the correlations between each subscale and global rating of depression severity, psychosocial functioning, and quality of life. We compared the correlations between these variables with the frequency, intensity, and sum ratings by calculating the difference between the Fisher z transformations of the correlation coefficients and dividing the difference by the standard error (Steiger, 1980).