Psychotherapy Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tpsr20

Patients' judgments of the importance of treatment-induced reductions in symptoms of depression: The role of specific symptoms, magnitudes of change, and post-treatment levels

Thomas T. Kim, Colin Xu & Robert J. Derubeis

To cite this article: Thomas T. Kim, Colin Xu & Robert J. Derubeis (2021): Patients' judgments of the importance of treatment-induced reductions in symptoms of depression: The role of specific symptoms, magnitudes of change, and post-treatment levels, Psychotherapy Research, DOI: 10.1080/10503307.2021.1938731

To link to this article: https://doi.org/10.1080/10503307.2021.1938731

View supplementary material

Published online: 14 Jun 2021.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tpsr20

Psychotherapy Research, 2021
https://doi.org/10.1080/10503307.2021.1938731

Routledge
Taylor & Francis Group
Check for updates

EMPIRICAL PAPER

Patients' judgments of the importance of treatment-induced reductions in symptoms of depression: The role of specific symptoms, magnitudes of change, and post-treatment levels

THOMAS T. KIM, COLIN XU, & ROBERT J. DERUBEIS

Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA

(Received 17 February 2021; revised 29 May 2021; accepted 31 May 2021)

Abstract
Objective: An implicit assumption in the use of depressive severity measures to assess change during treatment, such as the Hamilton Rating Scale for Depression (HRSD), is that reductions from pre- to post-treatment that are equal to each other are of equal value. However, stakeholders' valuations of changes might depart substantially from this assumption. **Method:** Vignettes were constructed that reflected the six possible 1, 2, and 3-point reductions on five cognitive and four somatic symptoms derived from the HRSD. Former or currently depressed patients provided judgments of the importance of the symptom reductions. Mean importance ratings were modeled using symptom category and the pre/post-treatment combination. Differences were explored using the Tukey method. **Results:** Results indicated that mean ratings, from most to least important, were: Anxiety, Suicide, Depressed Mood, Work, and Guilt (the cognitive symptoms) followed by Somatic, Sleep, Appetite & Weight, and Retardation (the somatic symptoms). Participants valued reductions that resulted in posttreatment scores of zero more than expected, given the magnitude of the reductions. **Conclusions:** The value of reductions in symptoms captured by the HRSD, as judged by patients, appears to differ as a function of symptom category and the post-treatment score. Similar patterns might characterize other measures of depression severity.

Keywords: Hamilton rating scale for depression; measurement of depressive severity

Clinical or methodological significance of this article: An implicit assumption in the use of depressive severity measures to assess change during treatment, such as the Hamilton Rating Scale for Depression (HRSD), is that reductions from pre- to post-treatment that are equal to each other are of equal value. However, we were not aware of any empirical work that has examined this assumption extensively of the HRSD, or of any widely used measure of depression severity. Results from this study suggest that former or currently depressed patients may not judge changes of the same magnitude on the HRSD to be equal in value; furthermore, judgments from these patients appeared to differ as a function of symptom category and post-treatment score.

Introduction

One of the earliest and most popular measures of depressive symptom severity is the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960). Studies have shown the HRSD to be reliable, and it exhibits a high degree of concurrent and differential validity (Williams, 1988). The HRSD was originally designed to yield a total depression severity score by adding the scores from 17 items. The symptoms evaluated by the HRSD are characterized by "anchor-point descriptions that increase in intensity; clinicians are to consider both the intensity and frequency of a symptom when assigning it a rating value" (Williams, 1988, p. 742). Each item is scored either on a 0 to 4 scale or a 0 to 2 scale. Scores on the five-point scale correspond with "absent," "mild or trivial," "moderate" (which reflects scores of both 2 and 3)

Correspondence concerning this article should be addressed to Thomas T. Kim, University of Pennsylvania, Levin 456, 425 S. University Ave, Philadelphia, PA, USA. Email: thomastk@sas.upenn.edu

© 2021 Society for Psychotherapy Research

and "severe." Hamilton (1960) used the three-point scale – whose scores correspond with "absent," "slight or doubtful," and "clearly present" – for symptoms difficult or impossible to quantify. Modern versions follow the original HRSD's scheme. An interview guide is typically used to obtain the information required to make the distinctions (Williams, 1988). Note that investigators have added items to the scale, resulting in several other versions of the HRSD (e.g., 24-, 25-, 28-, and 31-item scales; see Carmody et al., 2006; Zimmerman et al., 2005).

A common use of the HRSD is in the evaluation of outcomes of treatments for depression (Fried & Nesse, 2015; Williams, 1988). Many randomized clinical trials comparing treatments for depression have used the total HRSD, at baseline and post-treatment (as well as intermediate points), to characterize each individual's response to treatment. However, despite its widespread use and its adoption as the standard in the field in evaluating outcome of antidepressant treatment, there have been many critiques of the HRSD, even prompting consideration of its replacement as the primary outcome measure in treatment studies of depression (Zimmerman et al., 2005).

One set of criticisms of the HRSD concerns its psychometric properties. An implicit assumption about the scaling properties of constituent items on symptom severity measures such as the HRSD is that each numerical point (either at baseline, at post-treatment, or in the change from pre- to post-treatment) has equal value. A reduction in one symptom (e.g., a two-point improvement on suicide) is valued equivalently as the same reduction in another (e.g., a two-point improvement on depressed mood). Gibbons et al. (1993) used item response theory to assess the dimensionality of the HRSD for inpatients and outpatients with depression. Their analyses yielded a five-dimensional solution in which eight symptoms (depressed mood, guilt, suicide, work and interests, agitation, anxiety – psychic, anxiety – somatic, and genital symptoms) appeared to define a unidimensional index of global depression severity. The other symptoms assessed – primarily somatic or vegetative symptoms – did not contribute to this index. A shorter version of the HRSD (Bech et al., 1975) comprises six items (depressed mood, guilt, work and interests, anxiety – psychic, retardation, and somatic symptoms – general), in which four of the six overlapped with symptoms identified in Gibbons, Clark, and Kupfer's unidimensional index. This scale has been found to be more unidimensional (Bagby et al., 2004; Bech et al., 1984) and even more sensitive in detecting drug and placebo differences (Faries et al., 2000) than the original HRSD.

These results suggest that equivalent reductions in specific symptoms could be expected to be valued quite differently. Although neither Gibbons et al. (1993) nor Bech et al. (1984) examined the relative importance of changes in severity score from pre- to post-treatment, one might expect a similar pattern, specifically that reductions in those 10 symptoms might be judged as more important than others.

Another assumption in analyses of outcome data is that patients whose pre- and post-treatment total scores are identical (e.g., pre-treatment scores of 27 and post-treatment scores of 9) will have experienced improvement that is equivalent in value, irrespective of whether the reduction comprises many small changes across all symptoms, or fewer, larger changes in some symptoms along with little or no changes in others. This assumption is problematic, unless reductions in the constituent items behave like interval scales such that, for example, a reduction from 1 to 0 on an item would be experienced as roughly equal in importance to reductions from 2 to 1 or from 3 to 2 (Harwell & Gatti, 2001). Moreover, the value of a given level of symptom reduction should conform to the magnitude of that reduction such that, for example, a one-point reduction should be judged as less important than a two-point reduction. We are not aware of any empirical work that has examined these assumptions of the HRSD or of any widely used measure of depression severity.

The Present Study

We asked participants with a history of treatment for depression to judge the importance of reductions in nine different symptoms of depression across six different pre/post treatment combinations. The stimuli presented to participants were derived from the HRSD (Hamilton, 1960; Williams, 1988), allowing us to: (a) evaluate the relative importance of change in the nine symptoms, as judged by participants; and (b) examine how well participants' judgments of pre- to post-treatment reductions within symptoms approximated the patterns expected from an interval scale. Specifically, we examined whether changes reflecting higher numeric reductions were judged consistently to be more important than those reflecting lower numeric reductions, and whether stimuli representing equivalent point reductions were judged as equally important, irrespective of the pre-treatment or post-treatment values represented.

Methods

Materials and Procedure

Participants completed an online survey that presented descriptions of symptom reductions in depression experienced by a hypothetical patient. The descriptions represented various levels of symptom severity experienced by the patient prior to the initiation of treatment and at the end of a four-month course of treatment. Participants were asked to render a judgement of the importance of the improvement represented in each vignette. From the original 17-item HRSD (1960), five items (depressed mood, guilt, suicide, work and interests, and retardation) were presented as symptoms on their own. We did not represent agitation, hypochondriasis, or insight in the materials we presented to participants because in two datasets available to us (from DeRubeis et al., 2005; Hollon et al., 2014), ratings on these items were rarely above zero, even among patients with high HRSD total scores.

The remaining nine items from the original 17-item HRSD were condensed into four symptom categories, both to minimize participant burden and to reflect what is known about the relations among these items from factor analyses of the HRSD (Shafer, 2006). The three insomnia items (initial, middle, and delayed) were combined to represent "Sleep" symptoms. Two of the anxiety items (anxiety – psychic and anxiety – somatic) were combined to represent "Anxiety" symptoms. Two items, somatic symptoms – general, and genital symptoms (also referred to as libido on other versions of the HRSD), were combined to represent "Somatic" symptoms. Finally, "loss of weight" and "somatic symptoms – gastrointestinal" were combined to represent "Appetite & Weight" symptoms.

To summarize, participants were asked to rate the importance of reductions in these nine symptom categories: Depressed Mood, Guilt, Suicide, Sleep, Work, Retardation, Anxiety, Somatic, and Appetite & Weight. The construction of the nine symptom categories can also be seen in Table S1.

We created descriptions for each symptom category that reflected the six possible reductions that could occur given a pre-treatment score of 3 (to 2, to 1, or to 0), 2 (to 1 or to 0), or 1 (to 0). We did not present descriptions reflecting the highest possible pre-treatment score (i.e., 4) for any of the symptoms because such scores are uncommon, even in populations of patients with severe forms of depression (see Evans et al., 2004). See Table S2–S7 for the exact wordings of the symptom change descriptions (9 symptoms by 6 pre/post combinations) presented to participants. Note that the order of the nine symptoms presented within each set was randomized.

For each of the vignettes presented to participants, their task was to imagine that they had experienced the improvement represented in the vignette and to select the word or phrase that best answered the question, "How important or meaningful was this change?" Their options were: "Not at all"; "Slightly"; "Moderately"; or "Very" (which we coded, respectively, as 1, 2, 3, or 4). Each participant was randomized to encounter three sets of vignettes which represented three of the six possible pre/post combinations. Each set presented symptom change descriptions for all nine symptoms, with the pre- to post-treatment score (e.g., from 3 to 1) held constant within a set. The order of the nine symptoms presented within each set was randomized. Note that participants were not given the numeric scores associated with any of the descriptions.

Participants

The institutional review board approved all procedures. The study was posted on MQ's Take Part in Research platform, where United Kingdom (UK) citizens could fill out the survey online. MQ is a registered charity in England that focuses on transforming mental health through research. At the beginning of their encounter with the survey, participants were presented with a consent form and ensured that all responses would be anonymous. Only participants who answered that they had received treatment for depression were asked to continue. At the end of the survey, participants had the option to provide comments and to enter a raffle to win a £50 Amazon gift card.

Statistical Analyses

In order to generate estimates of the main effects of the symptoms and pre/post combinations on importance ratings, we created a mixed-effects model using the lme4 package in R (Bates et al., 2014), with importance rating specified as the dependent variable. The nine symptoms (Depressed Mood, Guilt, Suicide, Sleep, Work, Retardation, Anxiety, Somatic, and Appetite & Weight) and the six pre/post combinations (3 to 0, 3 to 2, 3 to 1, 2 to 0, 2 to 1, and 1 to 0) were entered as categorical fixed-effect predictors, and participant was entered as a random effect.

Categorical predictors were entered into the model using unweighted effect coding. Compared to dummy coding, where a specific reference group is

selected, unweighted effect coding uses the unweighted mean of all the group means as the reference point (Cohen et al., 2013; IDRE, 2021). Thus, coefficients can be interpreted in relation to the grand mean and, in this case, adjusted for the random effect of participant (Cohen et al., 2013). Effect coding is especially useful for providing interpretable estimates from models that include interactions. Specifically, significant interaction terms reflect pairs (symptom by pre/post combination) that participants reported as significantly more or less important than what was accounted for by mean level ratings for both the symptom and the pre/post combination (Aguinis, 2004). We also applied the general linear hypothesis test function from the "multcomp" package (Bretz et al., 2016) to this model to generate estimates and to perform pairwise tests of the average importance rating for each of the 54 pairs of symptom and pre/post combination. These estimates are mathematically equivalent to adding the effect-coded coefficients for symptom and pre/post combination for each respective pair. Because of the large number of interactions (54 possible pairs of symptom and pre/post combination), we only interpreted interaction terms if the significance test met an adjusted threshold of $p < 0.0009$, which we arrived at by dividing 0.05 by 54, or the number of tests (i.e., a Bonferroni correction; Bland & Altman, 1995).

Since we were interested in whether there were reliable differences in participants' judgments as a function of the symptom or the pre/post combination, we conducted two Type III Analyses of Variance, one with symptom as a factor, and one with pre/post combination as a factor, using Satterthwaite's method (Luke, 2017). If a significant overall effect of symptom (or pre/post combination) was detected, we explored the differences between each pair of symptoms (or each pair of pre/post combinations) using the Tukey method, which controlled for the family-wise Type I error rate. The general linear hypothesis test function from the "multcomp" package (Bretz et al., 2016) in R was used to implement the Tukey method (1949).

Results

Sample Characteristics

A total of 2,676 responses was obtained from 158 participants. Table I presents the sample characteristics, specifically age, gender, participants' current depression status, type of treatment, as well as the number of sessions of treatment they had received. Note that the present sample comprised a very high percentage of females (94.9%), even when considering that the majority of UK patients with depression treated through the NHS are female (65%; Clark, 2018).

Table I. Demographics.

Characteristics	
Age	34.8 ± 12.7
Female	94.9%
Self-reported currently depressed?	86.7%
Number of treatment sessions[a]	20
Received both antidepressants and psychotherapy	79.7%
Received only antidepressants	15.2%
Received only psychotherapy	5.1%
If received psychotherapy, what type?	
Cognitive behavioral therapy	69.0%
Psychodynamic therapy	22.8%
Behavioral activation	5.7%
Counseling	15.8%
Dialectical behavioral therapy	6.3%

Note. Statistics reported are in percentages (n/N) for categorical variables and mean ± SD for continuous variables.
[a]Number of treatment sessions is reported as a median.

Differences in Symptoms

A significant effect of symptom was obtained (F [8, 2808.4] = 9.22, $p < 0.001$), indicating that judgments across the nine symptoms differed. Estimates of the importance ratings were, in order from most to least important: Anxiety, Suicide, Depressed Mood, Work, Guilt, Somatic, Sleep, Appetite & Weight, and Retardation. Significant differences, revealed by the Tukey method (1949), indicated that reductions in Anxiety were judged as more important than reduction in four of the other eight symptoms (Somatic, Sleep, Appetite & Weight, Retardation). Reductions in Suicide and Depressed Mood were judged more important than reductions in Sleep, Appetite & Weight, and Retardation. Reductions in Work and Guilt were judged more important than reductions in Appetite & Weight and Retardation (see Figure 1). See Table S8 for standardized regression coefficients, which characterize the differences in importance, controlling for pre/post combination, between each symptom and each of the other symptoms.

Differences in Pre/Post Combinations

A significant effect of pre/post combination was obtained (F [5, 2904.92] = 255.97, $p < 0.001$). Differences were significant in 13 of the 15 pairwise comparisons (all $ps < 0.05$). Estimates of the mean importance ratings of each of the six pre/post combinations, as well as significance tests (Tukey,

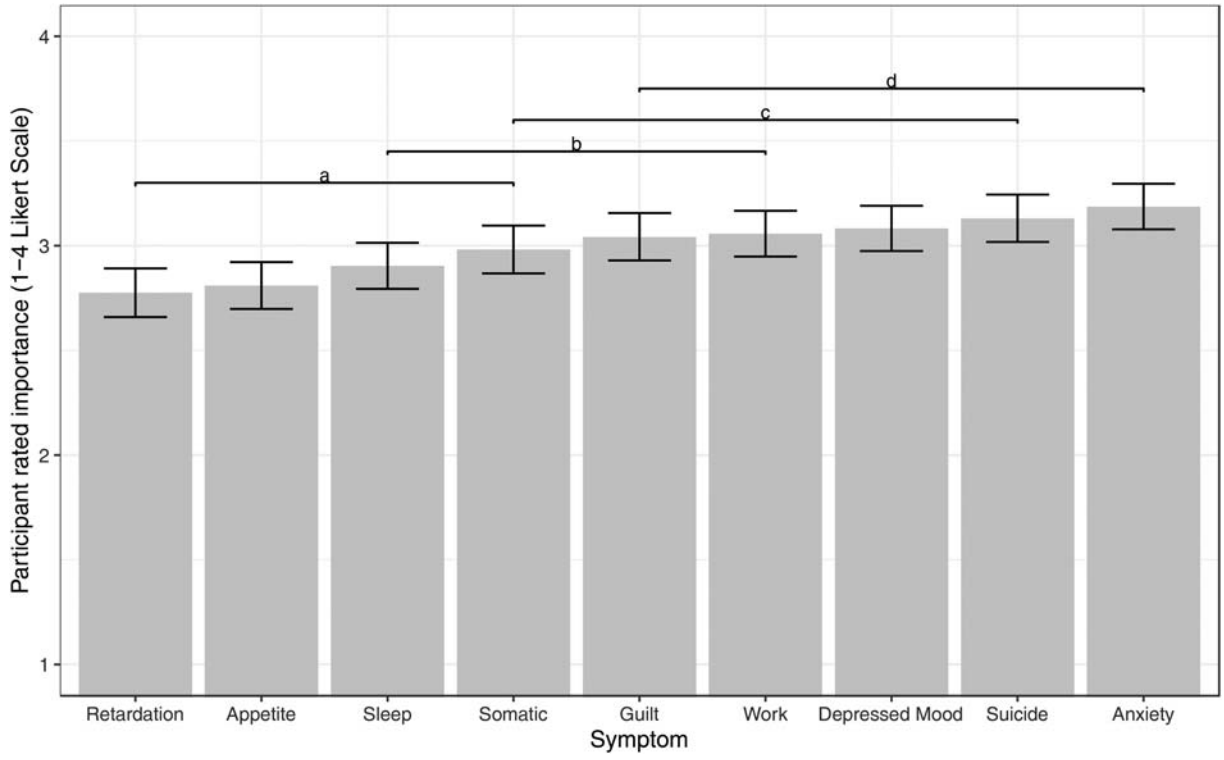

Figure 1. Modeled mean importance ratings for reductions in the nine symptoms.
Note: Symptom categories not in the same group (e.g., group a) are significantly different from each other according to Tukey pairwise contrasts at $p < 0.05$. Error bars represent standard errors.
Figure 1 shows that reductions in Anxiety are judged significantly more important than 4 of the 8 other symptoms (Somatic, Sleep, Appetite & Weight, and Retardation). Reductions in Suicide and Depressed Mood are judged more important than Sleep, Appetite & Weight, and Retardation. Reductions in Work and Guilt are judged more important than Appetite & Weight and Retardation.
Reductions in the more somatic symptoms (group a) were judged as not significantly different from each other, and reductions in the more psychological or cognitive symptoms (group d) were judged as not significantly different from each other.

1949) between the 15 possible pairs, are shown in Figure 2. Judgments of reductions from 3 to 0 and 2 to 0 were judged as the most important, and they did not differ from each other ($p = 0.96$). The next highest in importance was the reduction from 1 to 0, followed by the reduction from 3 to 1. The lowest mean importance ratings were given to reductions from 2 to 1 and 3 to 2, which did not differ from each other ($p = 0.47$).

Symptom by Pre/Post Combination Interactions

Estimates of mean importance ratings for each of the 54 cells are given in Table II (a visualization of these estimates is shown in Figure S1). At the $p < 0.05$ level, 18 of the 54 interactions were significant. Of these, eight met the more stringent threshold ($p < 0.0009$). Participants judged six pairings of symptoms and pre/post combinations as more important than expected: Guilt (2 to 1), Somatic (3 to 2), Suicide (3 to 2), Anxiety (2 to 1), Retardation (1 to 0), and Appetite (2 to 1). Participants judged two pairings of symptoms and pre/post combinations as less important than expected: Work (3 to 2) and Suicide (2 to 1).

Note that changes on Suicide were judged to be more important than the model-derived expectation for the change from 3 to 2, and less important for the change from 2 to 1. Moreover, the mean for a change from 3 to 2 on Suicide (2.90) was the highest of the nine symptoms (mean for all 3 to 2 changes = 2.27), whereas the mean for a change from 2 to 1 on Suicide (1.97) was the lowest of the nine symptoms (mean for all 2 to 1 changes = 2.37). Both estimates represented reductions of one point on the HRSD, yet they differed by nearly a full point on the 1 to 4 scale of importance.

Discussion

With ratings of the importance of changes on nine symptoms represented by the HRSD from former or currently depressed patients, we found that: (a) changes on these symptoms were valued differently; (b) changes reflecting higher numeric reductions

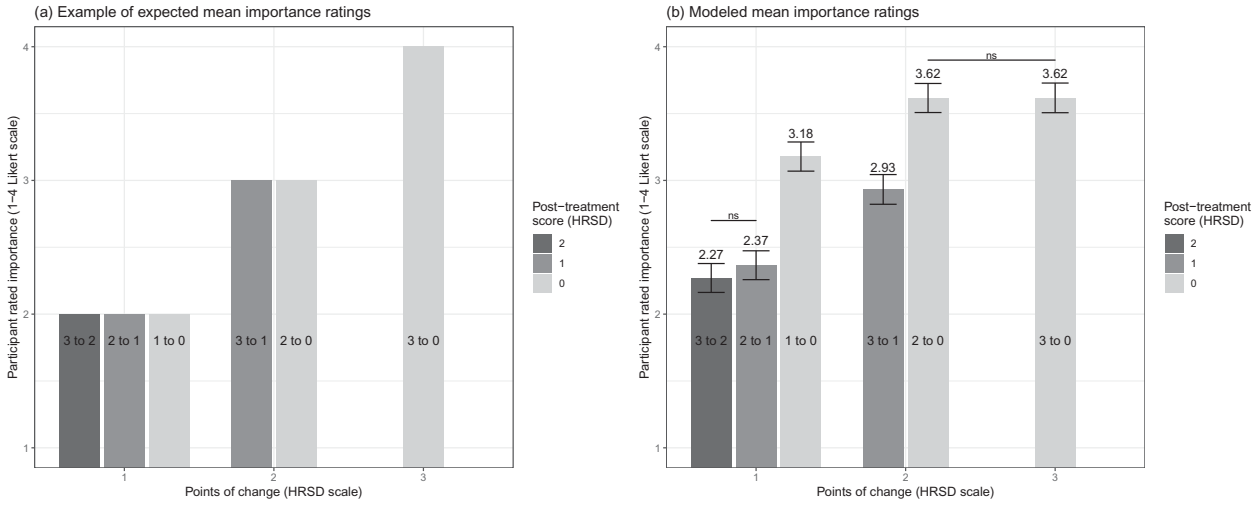

Figure 2. Example of expected juxtaposed with the modeled mean importance ratings for the six pre/post combinations.
Note: (a) is one example of the expected mean importance ratings if judgments of pre- to post-treatment reductions within symptoms approximated the patterns expected from an interval scale. We would expect higher numeric reductions to be judged as more important than lower numeric reductions, and equivalent points of change to be judged as equally important, irrespective of the pre-treatment score. (b) shows the actual modeled mean importance ratings. Bars that do not share "ns" are significantly different from each other according to Tukey pairwise contrasts at $p < 0.05$. Error bars represent standard errors. Higher numeric reductions were judged as more important than those of lower numeric reductions for most comparisons. Post-treatment scores of 0 were especially valued, even when considering the magnitude of reduction.

were not always judged as more important than those reflecting lower reductions; and (c) changes reflecting the same numeric reduction were valued differently depending on the level of severity at post-treatment.

Differences in Symptoms

Gibbons et al. (1993) concluded from their research that eight symptoms (anxiety – psychic, anxiety – somatic, suicide, depressed mood, work and

Table II. Modeled mean importance ratings, stratified by symptom and pre/post combination.

	1 point of reduction			2 points of reduction		3 points of reduction	
Symptom	1 to 0	2 to 1	3 to 2	2 to 0	3 to 1	3 to 0	**Modeled overall**
Anxiety	3.26 (2.94–3.59)	2.78 (2.55–3.01)	2.05 (1.81–2.29)	3.90 (3.61–4.19)	3.15 (2.89–3.41)	3.97 (3.67–4.28)	3.19 (2.97–3.40)
Suicide	3.22 (2.99–3.45)	1.97 (1.68–2.26)	2.90 (2.56–3.23)	3.63 (3.38–3.89)	3.14 (2.77–3.52)	3.92 (3.59–4.25)	3.13 (2.91–3.35)
Depressed mood	3.38 (3.05–3.71)	2.43 (2.15–2.7)	2.30 (2.06–2.53)	3.75 (3.51–4)	2.76 (2.49–3.02)	3.88 (3.63–4.13)	3.08 (2.87–3.29)
Work	3.26 (2.94–3.59)	2.63 (2.38–2.87)	1.89 (1.66–2.13)	3.73 (3.43–4.03)	3.06 (2.81–3.31)	3.77 (3.47–4.07)	3.06 (2.84–3.27)
Guilt	3.19 (2.91–3.48)	2.71 (2.49–2.94)	2.06 (1.73–2.4)	3.73 (3.45–4)	2.98 (2.61–3.35)	3.58 (3.25–3.9)	3.04 (2.82–3.26)
Somatic	3.11 (2.78–3.44)	2.21 (1.84–2.57)	2.63 (2.38–2.88)	3.54 (3.19–3.89)	3.00 (2.7–3.3)	3.40 (3.09–3.72)	2.98 (2.76–3.21)
Sleep	3.15 (2.82–3.48)	1.98 (1.68–2.29)	2.27 (2.03–2.51)	3.62 (3.34–3.89)	2.79 (2.52–3.07)	3.61 (3.34–3.89)	2.90 (2.69–3.12)
Appetite	2.80 (2.56–3.05)	2.42 (2.19–2.65)	2.19 (1.85–2.53)	3.25 (2.97–3.52)	3.09 (2.72–3.46)	3.11 (2.78–3.44)	2.81 (2.59–3.03)
Retardation	3.22 (3.00–3.45)	2.16 (1.79–2.52)	2.15 (1.81–2.48)	3.40 (3.05–3.75)	2.42 (2.05–2.79)	3.31 (2.98–3.64)	2.78 (2.55–3.00)
Modeled overall	3.18 (2.96–3.39)	2.37 (2.15–2.58)	2.27 (2.06–2.48)	3.62 (3.40–3.83)	2.93 (2.71–3.15)	3.62 (3.40–3.83)	

Note. Confidence intervals represent modeled 95% confidence intervals.

interests, guilt, agitation, and genital symptoms) defined a unidimensional index of global depression severity, and that most somatic or vegetative symptom items did not contribute to this definition. More recently, Demyttenaere et al. (2015) reported that physicians and patients ranked changes in items that described somatic symptoms as the least important in 'being cured from depression.' We found a similar pattern. Reductions in the more psychological or cognitive symptoms were judged as more important (Anxiety, Suicide, Depressed Mood, Work, Guilt) than reductions in the more somatic or vegetative symptoms (Somatic, Sleep, Appetite & Weight, and Retardation). Furthermore, judgments of reductions in the somatic symptoms were not significantly different from each other, nor were judgments of reductions in the cognitive symptoms different from each other.

Anxiety is not included as a symptom in the DSM-5 diagnosis of MDD (American Psychiatric Association, 2013). It is therefore of note that reductions in anxiety received the highest average ratings. Future studies should replicate this finding and examine if this finding, as well as the others we report here, extends to other measures of depression severity. Note that one specifier of a diagnosis of MDD "indicates MDD episodes associated with anxious distress" (Hasin et al., 2018, p. 337). Indeed, this specifier commonly follows a diagnosis of MDD – 75% of participants in the National Epidemiologic Survey on Alcohol and Related Conditions III met criteria for lifetime MDD followed by an anxious/distressed specifier (Hasin et al.). In addition, participants with the anxious/distressed specifier have reported greater symptom severity, poorer functioning, and poorer coping ability (Zimmerman et al., 2019).

Fried and Nesse (2015), applying a network analysis to a large number of depressive symptoms to elucidate interconnections among symptoms of depression, found the highest centrality scores for the following nodes: anxiety, diminished interest/pleasure, and sad mood. Although Fried et al.'s network was constructed with single assessments from each patient, the theory behind network analysis would predict that changes in more central nodes would influence changes in other nodes in the network more than would changes in less central nodes. The present findings are consistent, then, with Fried et al.'s, as well as those from Zimmerman et al. (2019), in the context of symptom changes. Indeed, reductions in these three symptoms and suicide were judged to be more important than any of the other five symptoms we examined.

Differences in Pre/Post Combinations

As expected, participants tended to value more highly descriptions that reflected higher numeric reductions. Two of the three combinations that reflected reductions of one point (specifically, reductions from 3 to 2 and 2 to 1) yielded very similar average importance ratings and were the two lowest among the six combinations presented to participants. However, reductions from 1 to 0, the other combination that reflected a one-point reduction, were judged as more important than either of the other one-point reductions, suggesting that the resolution of a symptom is seen as especially valuable. Indeed, the average importance rating for reductions from 1 to 0 was higher than those obtained for reductions from 3 to 1, a clear departure from the pattern expected if higher numeric reductions are valued more than lower ones. More evidence of the high value placed on symptom resolution comes from the observation that reductions from 2 to 0 were rated as substantially more important than reductions from 3 to 1.

Reductions from 3 to 0 could not be compared to any other reduction of three points (e.g., 4 to 1), due primarily to the features of the HRSD. However, it is notable that the estimates we obtained for the average importance ratings of reductions from 3 to 0 and 2 to 0 did not differ from each other and, indeed, were the same to the second decimal place. We considered three possible accounts of this surprising pattern. First, participants might have judged as similar in intensity descriptions of phrases associated with pre-treatment scores of 3 and 2. This is ruled out by the fact that changes from 3 to 1 were judged to be more important than changes from 2 to 1. Another possibility, not ruled out by our data, is that high ratings for changes from 2 to 0 left no room for participants to provide yet higher ratings for descriptions of changes from 3 to 0. A third possibility is that the resolution of a symptom is especially highly valued, beyond the magnitude of the reduction it represents. This account is supported by our finding that changes from 1 to 0 were more highly valued than changes from 3 to 1, as well as both of the other one-point changes represented in our stimulus set.

Symptom by Pre/Post Combination Interactions

Of the eight significant interactions from tests of symptoms by pre/post combination pairings, the two whose coefficients reflected the greatest departure from expectation concerned reductions in

Suicide. Although reductions in Suicide were judged, overall, as second in importance only to reductions in Anxiety, the estimate for a change of 2 to 1 in Suicide was not only the lowest of all pairings from 2 to 1, but also the lowest of all the 54 pairings. In contrast, the estimate for a change of 3 to 2 in Suicide (also a reduction of one point) was substantially higher than expected, and nearly as high as the average rating across symptoms for reductions from 3 to 1 (a reduction of two points). The descriptions of each level for Suicide, which closely adhere to those of the original and modern versions of the HRSD (Hamilton, 1960; Williams, 1988), may explain this phenomenon. The change from 3 to 2 was represented thus: from "had suicidal ideas and sometimes a specific suicide plan" to "sometimes wished that he (or she) were dead, but is not considering taking his (or her) own life." Our findings suggest that participants placed a high value on the shift, represented in the change from 3 to 2 on this item, from active to passive suicidal ideation. The change from 2 to 1, also a one-point change, yielded an estimate of importance from our participants just below "Slightly." The wording, reflecting a change from 2 ("sometimes wished that he (or she) were dead, but was not considering taking his (or her) own life") to 1 ("sometimes thinks that life is not worth living, but he (or she) does not think about suicide"), does appear to be subtle, as neither indicates a current interest in suicide per se'.

Implications

The present research suggests that properties of the HRSD may limit its utility to evaluate some reductions in depression severity. For example, if Patient A and Patient B both experienced a change from a pre-treatment score of 27 to a post-treatment score of 9, the HRSD would suggest these changes to be equivalent in importance. However, Patient B may have had more symptoms reduced to a post-treatment score of 0 compared to Patient A, and/or experienced more reductions in cognitive than somatic symptoms. The present study would suggest that Patient B would judge their change to be higher in importance than Patient A, despite the equivalent point reduction and overall level of severity at post-treatment.

Similar properties may extend to other measures of depression severity, as well as to instruments that are used to assess changes in severity during treatment in other mental disorders. We chose the HRSD as the basis for this work because of its coherent descriptions of symptoms at various levels of severity. Future studies should first examine whether similar results can be found in other commonly used depression measures, such as the Beck Depression Inventory-II (BDI-II; Beck et al., 1996). A measure that is becoming very popular, the Patient Health Questionnaire-9 (PHQ-9; Kroenke et al., 2001), uses the same anchor for each symptom, where the anchor references the frequency of the presence of the symptom over the previous two weeks ("Not at all," "Several days," "More than half the days," "Nearly every day"). It would be interesting to determine whether patterns of participants' judgments of reductions in intensity more closely align with properties of interval scaling if the descriptions reflect frequency, as in the PHQ-9, rather than descriptions of symptom severity, as in the HRSD. However, it would still be important to discover whether symptom resolution, *per se*, is more highly valued than would be expected on the basis of numerical reductions, and whether changes on some symptoms are more highly valued than changes on others.

Limitations

We assessed only former or current patients who have received treatment for depression. Clinicians constitute another stakeholder group whose judgments of importance would be valuable to obtain. While clinicians and patients work together to evaluate depression severity, there may be differences in clinicians' perspectives. Future studies might also include evaluations of various degrees of worsening of symptoms, as well as improvements. Future studies should also include a more balanced gender distribution.

In this study, vignettes were used to prompt patients' judgments. While studies employing vignettes can provide an interpretation of the real world (Hughes, 1998), it may be worthwhile to query patients about changes in symptoms they experience during a course of treatment. However, this procedure could be susceptible to recall bias – patients would have to consider their changes in symptoms from a period of four months ago. Furthermore, responses to vignettes can closely resemble responses to real-life experiences, if the vignettes appear to be both relevant and real to participants (Finch, 1987; Rahman, 1996).

Acknowledgements

We would like to thank Dr. Lois Gelfand, Dr. Zachary Cohen, and Mr. Akash Wasil for their helpful comments on a version of this manuscript.

Supplemental data

Supplemental data for this article can be accessed https://doi.org/10.1080/10503307.2021.1938731.

ORCID

Thomas T. Kim http://orcid.org/0000-0003-2412-6647

References

Aguinis, H. (2004). *Regression analysis for categorical moderators*. Guilford Press.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Bagby, R. M., Ryder, A. G., Schuller, D. R., & Marshall, M. B. (2004). The Hamilton Depression Rating Scale: Has the gold standard become a lead weight? *American Journal of Psychiatry*, *161*(12), 2163–2177. https://doi.org/10.1176/appi.ajp.161.12.2163

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.*5823.

Bech, P., Allerup, P., Reisby, N., & Gram, L. F. (1984). Assessment of symptom change from improvement curves on the Hamilton depression scale in trials with antidepressants. *Psychopharmacology*, *84*(2), 276–281. https://doi.org/10.1007/BF00427459

Bech, P., Gram, L. F., Dein, E., Jacobsen, O., Vitger, J., & Bolwig, T. G. (1975). Quantitative rating of depressive states. *Acta Psychiatrica Scandinavica*, *51*(3), 161–170. https://doi.org/10.1111/j.1600-0447.1975.tb00002.x

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. (1996). Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *Journal of Personality Assessment*, *67*(3), 588–597. https://doi.org/10.1207/s15327752jpa6703_13

Bland, J. M., & Altman, D. G. (1995). Statistics notes: Multiple significance tests: The Bonferroni method. *BMJ*, *310*(6973), 170. https://doi.org/10.1136/bmj.310.6973.170

Bretz, F., Hothorn, T., & Westfall, P. (2016). *Multiple comparisons using R*. CRC Press.

Carmody, T. J., Rush, A. J., Bernstein, I., Warden, D., Brannan, S., Burnham, D., Woo, A., & Trivedi, M. H. (2006). The Montgomery Asberg and the Hamilton ratings of depression: A comparison of measures. *European Neuropsychopharmacology*, *16*(8), 601–611. https://doi.org/10.1016/j.euroneuro.2006.04.008

Clark, D. M. (2018). Realizing the mass public benefit of evidence-based psychological therapies: The IAPT program. *Annual Review of Clinical Psychology*, *14*(1), 159–183. https://doi.org/10.1146/annurev-clinpsy-050817-084833

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Demyttenaere, K., Donneau, A. F., Albert, A., Ansseau, M., Constant, E., & van Heeringen, K. (2015). What is important in being cured from depression? Discordance between physicians and patients (1). *Journal of Affective Disorders*, *174*, 390–396. https://doi.org/10.1016/j.jad.2014.12.004

DeRubeis, R. J., Hollon, S. D., Amsterdam, J. D., Shelton, R. C., Young, P. R., Salomon, R. M., O'Reardon, J. P., Lovett, M. L., Gladis, M. M., Brown, L. L., & Gallop, R. (2005). Cognitive therapy vs medications in the treatment of moderate to severe depression. *Archives of General Psychiatry*, *62*(4), 409–416. https://doi.org/10.1001/archpsyc.62.4.409

Evans, K. R., Sills, T., DeBrota, D. J., Gelwicks, S., Engelhardt, N., & Santor, D. (2004). An item response analysis of the Hamilton Depression Rating Scale using shared data from two pharmaceutical companies. *Journal of Psychiatric Research*, *38*(3), 275–284. https://doi.org/10.1016/j.jpsychires.2003.11.003

Faries, D., Herrera, J., Rayamajhi, J., DeBrota, D., Demitrack, M., & Potter, W. Z. (2000). The responsiveness of the Hamilton Depression Rating Scale. *Journal of Psychiatric Research*, *34*(1), 3–10. https://doi.org/10.1016/S0022-3956(99)00037-0

Finch, J. (1987). The vignette technique in survey research. *Sociology*, *21*(1), 105–114. https://doi.org/10.1177/0038038587021001008

Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*(1), 72. https://doi.org/10.1186/s12916-015-0325-4

Gibbons, R. D., Clark, D. C., & Kupfer, D. J. (1993). Exactly what does the Hamilton Depression Rating Scale measure? *Journal of Psychiatric Research*, *27*(3), 259–273. https://doi.org/10.1016/0022-3956(93)90037-3

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery & Psychiatry*, *23*(1), 56–62. https://doi.org/10.1136/jnnp.23.1.56

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, *71*(1), 105–131. https://doi.org/10.3102/00346543071001105

Hasin, D. S., Sarvet, A. L., Meyers, J. L., Saha, T. D., Ruan, W. J., Stohl, M., & Grant, B. F. (2018). Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the United States. *JAMA Psychiatry*, *75*(4), 336–346. https://doi.org/10.1001/jamapsychiatry.2017.4602

Hollon, S. D., DeRubeis, R. J., Fawcett, J., Amsterdam, J. D., Shelton, R. C., Zajecka, J., Young, P. R., & Gallop, R. (2014). Effect of cognitive therapy with antidepressant medications vs antidepressants alone on the rate of recovery in major depressive disorder: A randomized clinical trial. *JAMA Psychiatry*, *71*(10), 1157–1164. https://doi.org/10.1001/jamapsychiatry.2014.1054

Hughes, R. (1998). Considering the vignette technique and its application to a study of drug injecting and HIV risk and safer behaviour. *Sociology of Health & Illness*, *20*(3), 381–400. https://doi.org/10.1111/1467-9566.00107

IDRE. (2021). *Coding systems for categorical variables in regression analysis*. UCLA: Statistical Consulting Group. https://stats.idre.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis/.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Rahman, N. (1996). Caregivers' sensitivity to conflict: The use of the vignette methodology. *Journal of Elder Abuse & Neglect*, *8*(1), 35–47. https://doi.org/10.1300/J084v08n01_02

Shafer, A. B. (2006). Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of Clinical Psychology*, *62*(1), 123–146. https://doi.org/10.1002/jclp.20213

Tukey, J. W. (1949). Moments of random group size distributions. *The Annals of Mathematical Statistics*, *20*(4), 523–539. https://doi.org/10.1214/aoms/1177729945

Williams, J. B. (1988). A structured interview guide for the Hamilton Depression Rating Scale. *Archives of General Psychiatry*, *45*(8), 742–747. https://doi.org/10.1001/archpsyc.1988.01800320058007

Zimmerman, M., Martin, J., McGonigal, P., Harris, L., Kerr, S., Balling, C., Kiefer, R., Stanton, K., & Dalrymple, K. (2019). Validity of the DSM-5 anxious distress specifier for major depressive disorder. *Depression and Anxiety*, *36*(1), 31–38. https://doi.org/10.1002/da.22837

Zimmerman, M., Posternak, M. A., & Chelminski, I. (2005). Is it time to replace the Hamilton Depression Rating Scale as the primary outcome measure in treatment studies of depression? *Journal of Clinical Psychopharmacology*, *25*(2), 105–110. https://doi.org/10.1097/01.jcp.0000155824.59585.46