HHS Public Access

Author manuscript

J Affect Disord. Author manuscript; available in PMC 2017 August 01.

Published in final edited form as:

J Affect Disord. 2016 August ; 200: 111–118. doi:10.1016/j.jad.2016.01.051.

A Protocol for the Hamilton Rating Scale for Depression: Item Scoring Rules, Rater Training, and Outcome Accuracy with Data on its Application in a Clinical Trial

Kelly J. Rohan, Ph.D.[1,*], **Jennifer N. Rough, B.A.**[1], **Maggie Evans, B.A.**[1], **Sheau-Yan Ho, B.A.**[1], **Jonah Meyerhoff, B.A.**[1], **Lorinda M. Roberts, M.A.**[1], and **Pamela M. Vacek, Ph.D.**[2]

[1]Department of Psychological Science, University of Vermont, Burlington, VT

[2]Department of Medical Biostatistics, University of Vermont College of Medicine, Burlington, VT

Abstract

Background—We present a fully articulated protocol for the Hamilton Rating Scale for Depression (HAM-D), including item scoring rules, rater training procedures, and a data management algorithm to increase accuracy of scores prior to outcome analyses. The latter involves identifying potentially inaccurate scores as interviews with discrepancies between two independent raters on the basis of either scores (≥ 5-point difference) or meeting threshold for depression recurrence status, a long-term treatment outcome with public health significance. Discrepancies are resolved by assigning two new raters, identifying items with disagreement per an algorithm, and reaching consensus on the most accurate scores for those items.

Methods—These methods were applied in a clinical trial where the primary outcome was the Structured Interview Guide for the Hamilton Rating Scale for Depression—Seasonal Affective Disorder version (SIGH-SAD), which includes the 21-item HAM-D and 8 items assessing atypical symptoms. 177 seasonally depressed adult patients were enrolled and interviewed at 10 time points across treatment and the 2-year followup interval for a total of 1,589 completed interviews with 1,535 (96.6%) archived.

Results—Inter-rater reliability ranged from ICCs of .923 to .967. Only 86 (5.6%) interviews met criteria for a between-rater discrepancy. HAM-D items “Depressed Mood,” “Work and Activities,” “Middle Insomnia,” and “Hypochondriasis” and Atypical items “Fatigability” and “Hypersomnia” contributed most to discrepancies.

Limitations—Generalizability beyond well-trained, experienced raters in a clinical trial is unknown.

Conclusions—Researchers might want to consider adopting this protocol in part or full. Clinicians might want to tailor it to their needs.

[*]Department of Psychological Science, University of Vermont, John Dewey Hall, 2 Colchester Avenue, Burlington, VT 05405-0134. Phone: (802) 656-0798, FAX: (802) 656-8783, ; Email: kelly.rohan@uvm.edu

Keywords

Hamilton Rating Scale; Depression assessment; Rater training; Scoring rules; Inter-rater reliability

Introduction

The Hamilton Rating Scale for Depression (HAM-D; Hamilton, 1960) is one of the longest standing, most widely used measures of depression severity in research and clinical practice. Originally designed to measure symptom severity in depressed inpatients, the 17-item HAM-D has evolved over the past 50 plus years into 11 modified versions that have been administered to various patient populations in an array of psychiatric, medical, and other research settings (Williams, 2001).

Although the HAM-D has been referred to as the "gold standard" for measuring depression severity, the measure is limited by scoring difficulties and psychometric weaknesses. In a review of the HAM-D, Bagby and colleagues (2004) examined the psychometric properties of the 17-item version across 70 studies, including reliability, item-response characteristics, and validity of the measure. Results indicated adequate reliability (internal, inter-rater, and retest reliability) and validity (convergent, discriminant, and predictive validity). However, the measure demonstrated poor item-level inter-rater reliability, test-retest reliability, and content validity. Bagby and colleagues examined internal reliability using Chronbach's alpha. They found alphas ranging from .46–.92. In eight of the 12 studies reporting Chronbach's alphas, internal reliability coefficients were less than or equal to .76. Bagby et al. (2004) concluded that the total scale score is multidimensional and its clinical meaning is unclear. Additionally, they found problems in the scaling of particular items (e.g., depressed mood, feelings of guilt, hypochondriasis). Another meta-analytic review concluded that some HAM-D items show poor or marginally acceptable internal consistency, particularly the insight item (Trajković et al., 2011).

Bagby and colleagues ultimately recommended the development of a new scale, but also suggested ways to improve the HAM-D. Suggestions included revising the content and rating scale of items to address the psychometric problems and developing clear interview prompts and scoring guidelines. Consistent with that recommendation, a multi-site study aimed at improving the inter-reliability of the 17-item HAM-D in primary care (Morriss, Leese, Chatwin, & Baldwin, 2008) developed item-by-item scoring rules for lay interviewers to use as a means to reduce inter-individual clinical judgment in the ratings. The overall intraclass correlation (ICC) was .947 with a standard deviation of 1.25, comparable to previous studies that relied on inexperienced raters. Morriss et al. (2008) emphasized that scoring rules were critical in yielding high inter-rater reliability.

Other attempts have been made to address the aforementioned critiques and improve the HAM-D. Structured interview guides, including Williams' Structured Interview Guide for the Hamilton Depression Rating Scale (SIGH-D; Williams, 1988), were developed to improve item reliability and facilitate rater training. The SIGH-D provides parenthetical qualifications in order to provide more consistent anchor points across raters. The creators of the HAM-D recently attempted to overhaul the measure, citing poor item reliability. The

overhauled measure, called the GRID-HAMD (Williams et al., 2008), addresses poor item reliability by creating separate item anchors for symptom intensity and symptom frequency. These anchors are placed along a vertical and horizontal grid that yields a single cell which contains a score (of 1–4) for any given item. These, and other modifications to the HAM-D, result in a new measure with more reliable items and simpler administration (Williams et al., 2008).

The Structured Interview Guide for the Hamilton Rating Scale for Depression–Seasonal Affective Disorder version (SIGH-SAD, Williams et al., 1992) is comprised of the 21-item HAM-D and 8 items assessing atypical symptoms of depression (e.g., hyperphagia, hypersomnia), which are not part of the original scale and are common in certain subtypes of depression, including seasonal depression. The SIGH-SAD is the standard measure of winter seasonal affective disorder (SAD) severity and is widely used in SAD research. Given that the SIGH-SAD is comprised of HAM-D items, the inter-rater reliability of the SIGH-SAD is also of interest as it may present similar psychometric issues for the assessment of depressed patients with atypical symptoms. However, the rater training requirements, item scoring methods, and inter-rater reliability statistics are not widely available for the SIGH-SAD. The few clinical trials that have employed multiple raters and reported the ICCs between SIGH-SAD raters report adequate reliability, ICC = .95 (Lam et al., 2006; Terman et al., 1998). Given the paucity of data, it remains unknown whether these reliability coefficients are typical of most SAD trials and what training protocols and scoring rules are required to obtain high agreement.

Although Morriss and colleagues' (2008) HAM-D item scoring guidelines are applicable to 17 of the 29 total items on the SIGH-SAD, many of their rules were not clearly defined and all rules focused on distinguishing a score of one from zero or two, thereby not informing scoring decisions for the full range of scores. Morriss et al. (2008) justified their lack of attention to ratings above two on any item by stating that symptoms in that range are quite rare in primary care. Furthermore, Morriss et al. (2008) did not provide scoring rules for the 12 items included on the SIGH-SAD that are beyond those on the 17-item HAM-D. Continued common problems in using the HAM-D that also apply to the SIGH-SAD include not citing whether a specific structured interview guide was used, providing no description of rater training in the methods, and wide variability in rater training protocols. For these reasons, it is desirable to disseminate a comprehensive protocols that might inform research and practice using the HAM-D.

Here, we share the methodology our group has adopted to address the aforementioned psychometric flaws of the HAM-D in the context of our program of research testing the efficacy of SAD treatments. We use the 29-item SIGH-SAD version of the HAM-D, but our methods can be applied to other versions of the HAM-D contained within the SIGH-SAD (e.g., the 21-item and 17-item HAM-D). First, we outline clear, comprehensive guidelines for scoring each item on the SIGH-SAD, noting where they differ from those proposed by Morriss et al. (2008) for the 17-item HAM-D. Second, we describe the structured protocol we use to train beginning SIGH-SAD raters and to prevent rater drift over time. Third, we articulate a data analytic approach to increase accuracy of ratings prior to outcome analyses. The protocol involves identifying interviews that meet criteria for significant between-rater

discrepancy and a procedure for resolving the discrepancy to estimate the most accurate score for analysis. Fourth, we present data from a recently completed clinical trial with a 2-year followup interval that enrolled 177 SAD patients, used the SIGH-SAD as the primary outcome measure, and followed the approach detailed in this paper. Specifically, we present inter-rater reliability at each time point (i.e., baseline, weekly during 6-weeks of acute treatment, and at followups the next summer, next winter, and second winter). We also present frequency data on identified between-rater discrepancies per our algorithm at each time point and the specific items that most commonly contributed to those discrepancies. We conclude with general recommendations for future work using the SIGH-SAD and HAM-D.

Item Scoring Rules

The following section proceeds item-by-item according to the item numbering system of the scale, using the prefix "H" for HAM-D items and "A" for atypical subscale items, followed by the name of the item as it appears on the scale. Per the scale instructions, the assessment timeframe for all items is over the past week and the comparison for ratings is current behavior vs. "when feeling OK." We interpret the latter to indicate when euthymic and in, the case of SAD, during the summer. For all items, a score of zero (0) indicates the absence of that particular symptom. Our administration method involves following the probing questions verbatim from the structured interview guide for the Hamilton Depression Rating Scale (SIGH-D; Williams, 1998) in the order provided, with a few exceptions where noted (see items H5, H8), and also using the probing questions for the eight atypical items provided by Williams et al. (1992). As noted by Bagby et al. (2004), the HAM-D has problems with item scaling and content, as some items assess frequency whereas others assess severity. Our scoring guidelines use the scale as designed without altering it, but strive to increase inter-rater reliability despite these limitations. The following guidelines illustrate the conventions we use to rate each item. It is also worth noting that if the respondent is experiencing a health condition (e.g., cold, flu, arthritis) at interview, we do not attempt to parse out whether responses are related to illness or depression because such distinctions are inherently complicated by qualitative overlap in somatic symptoms of depression vs. physical ailments (with one exception on H9). As a general rule, we rate all symptoms that are qualitatively consistent with depression at interview, even if possibly attributable to a physical condition.

H1. Depressed Mood (sadness, hopelessness, helplessness, worthlessness)

This item is specific to depressed mood, defined by our guidelines as any emotional state commensurate with sadness such as those named in the item label as well as down, low, blue, melancholic, dysphoric, teary, etc. Other types of negative mood (e.g., anxious, tense, angry, irritable) might be expressed but are not coded on H1, as they are captured by other items. The ratings distinguish verbal and nonverbal expressions of sadness. A score of (1) is given if depressed mood is verbally endorsed only upon direct questioning (i.e., an affirmative response to "Have you been feeling depressed, down, sad, hopeless, helpless, or worthless?"). We score verbal reports of depressed mood as (2) if a depressive descriptor is expressed spontaneously in response to either of the two open-ended questions that precede H1: "How have you been feeling since then?" and "What's your mood been like this past

week?" Nonverbal expressions of depressed mood (e.g., crying, sullen posture, frowning) during the interview receive a score of (3). In our experience, nonverbal expressions most often manifest during the H1 probes, but can sometimes emerge later in the course of the interview. In that case, we return to item H1 and factor them into the rating. A score of (4) "virtually only" indicates the presence of both nonverbal expression and spontaneous reports of depressed mood (i.e., criteria for ratings of 2 and 3 are both met).

H2. Work and Activities

This item assesses diminished time and interest in the domains of work and leisure activities/hobbies. As this item considers work "in or out of the home," we consider engagement in household chores and family caregiving the respondent is responsible for when rating primary homemakers and employed individuals alike. We issue a score of (1) for responses indicating relatively normal interest and engagement, with some vague or mild difficulty (e.g., thoughts of fatigue related to work/activities). A score of (2) is given when the response indicates a clear decrease in interest, while maintaining regular time/frequency in activities (e.g., "I have to push myself"). Regardless of interest level, any noticeable decrease in productivity or time spent in work or leisure activities warrants a (3). Therefore, the distinction between (2) vs. (3) is decreased interest vs. decreased frequency/time in activities compared to when euthymic. We follow the anchor for scoring a (4), "stopped working because of illness."

A1. Social Withdrawal

The scoring for this item is less ambiguous as the scoring options are directly read as a forced-choice item to respondents who endorse "No" to "Have you been as social as when you feel well?" However, even if the response is "Yes," we still assess interest level to distinguish between a (0) and (1). A score of (1) indicates social activities that are otherwise normal but accompanied by decreased interest, whereas a (0) indicates absence of withdrawal in activity or loss of interest.

H3. Genital Symptoms

This item assesses loss of libido. The scoring for H3 is a clear severity rating. However, the probes and rating scale force the respondent to choose whether the change is 'mild' ("a little less") or 'severe' ("a lot less") compared to when euthymic. We code all 'moderate' responses as 'mild.' Morriss et al. (2008) suggested rating this item over the past month. However, we chose to maintain consistency with the 'past week' timeframe of the measure in rating sexual interest.

H4. Somatic Symptoms Gastrointestinal

This item assesses decrease in appetite. A score of (1) is given for individuals who notice a decrease in appetite but continue to eat regularly. A score of (2) is indicated if the individual has skipped meals, notices that they have been eating less when they do not have someone urging them to eat, or has used medications for stomach or intestinal problems in the past week.

H5. Loss of Weight

This item assesses weight loss due to depression, not including effortful weight loss (e.g., dieting). The B set of scoring criteria involve actual weight changes measured on a scale and are self-explanatory. Rather than weighing our participants, we use the A set of scoring criteria. We distinguish between ratings of (1) and (2) based on the respondent's certainty of weight loss. A score of (2) is coded when the respondent clearly affirms that he/she lost weight within the past week or that his/her clothes felt looser this week. A score of (1) indicates some ambiguity in their response (e.g., "I think I have, but I am not sure"). It is unclear why the initial probe provided for this item is "Have you lost any weight since you started feeling depressed or down," as this timeframe would capture weight loss prior to and within the past week. Importantly, the remaining probes and the rating stems clearly focus on the past week, which is the reference timeframe for rating. Therefore, we bypass the initial probe as it is confusing to the respondent and rater alike and lead with the second probe, "Did you lose any weight this last week?"

A2. Weight Gain

This item assesses weight gain due to depression. The same scoring principles from H5 are also applied to A2.

A3. Appetite Increase

This item assesses the extent of appetite increase over the past week. If the respondent endorses greater appetite than normal (i.e., when euthymic), his or her response to the second probe dictates the rating for wanting to eat a little more (1), somewhat more (2), or much more (3) than usual.

A4. Increased Eating

This item assesses an increase in actual food consumption. As in A3, if increased eating is present, the respondent chooses their rating for eating a little more (1), somewhat more (2), or much more (3) than normal.

A5. Carbohydrate Craving or Eating

This item assesses a change in food preference towards more carbohydrates, reflected by increased craving and/or consumption of starches or sugars. Following an affirmative response to the initial probe, the respondent selects their rating for more (1), much more (2), or irresistible (3) carbohydrate craving or eating. The remaining probes are for coding purposes only and are not tallied in the total score (e.g., mainly a preference for starches, sweets, or both; increased craving, eating, or both; time of day). We have found utility in asking one of these supplemental prompts ("Which specific foods have you been craving?") in order to determine whether or not the respondent is aware of what constitutes carbohydrates. If not, the interviewer briefly explains what carbohydrates are with a few examples and re-administers the questions.

Preface to the Insomnia Items (H6, H7, H8)

Three items (H6, H7, H8) assess insomnia during the initial, middle, and latter segments of the sleep episode. Consistent across all of these items, an episode of insomnia is defined as a wakeful period of more than 30 minutes during the time when one intends to be asleep. Morriss et al. (2008) used this same 30-minute cutpoint. The scale loosely defines these three segments as "the beginning of the night," "in the middle of the night," and "in the early hours of morning," respectively, providing little guidance to the rater. Therefore, as a convention, we divide the individual's daily sleep length up into thirds to determine whether any insomnia disclosed counts as early (H6), middle (H7), or late (H8). Our scoring rules that follow for H6, H7, and H8 anchor ratings based on cutpoints governed by the frequency of problems sleeping throughout the week. Our approach differs from Morriss et al. (2008)'s rule that "One disturbed night's sleep may be ignored unless it is severe. Delay or disruption of sleep for more than 30 min is required on an occasional or regular basis to score a 1 or 2" (p. 207).

H6. Early (Initial) Insomnia

The H6 scoring stems operationally define early insomnia as taking longer than 30 minutes to initially fall asleep. The rating stems indicate that the scores differ on the basis of frequency, but do not provide a cutpoint. Therefore, we issued a score of (1) for initial insomnia on 1–4 nights and a score of (2) for 5 or more nights. In our experience, first-time respondents may require some direction in that this item asks how long it takes to fall asleep from the time that they are intending to begin sleep, which is not necessarily from the time they get in bed (i.e., if they are doing other activities in bed at night such as reading or if they make a habit of falling asleep at night on the couch).

H7. Middle Insomnia

In contrast to H6, the scoring stems for H7 focus on severity rather than frequency. Therefore, we rate middle insomnia as present even if it occurred only one night of the week. Otherwise restful sleep that involves waking during the night to use the bathroom and falling right back to sleep is rated as a (0). The endorsement of either perceived restless/disturbed sleep or spontaneous awakenings followed by a return to sleep within 30 minutes is coded as (1). We give a score of (2) if the respondent got out of bed during the night because they could not sleep or lay awake in bed unable to fall back asleep for more than 30 minutes. In the event that the respondent endorses interrupted sleep due to an external stimulus (e.g., fire alarm, child, pet, sudden noise, etc.), we rate a (0) as long as sleep was otherwise restful and he/she was able to fall back asleep within 30 minutes after returning to bed and a (2) if it took 30 minutes or longer to fall back asleep.

H8. Late (Terminal) Insomnia

A key aspect of rating this item is identifying the respondent's euthymic rise time, assessed by the last probe, and comparing it to the current wake time in the past week. If the respondent is confused by the probes, we use an alternative question that might be helpful: "In the last week, did you wake up earlier than you intended to?" If respondents awoke to an alarm clock or reported that they spontaneously awoke at their intended rise time, the item is

scored (0). Respondents who slept later than they intended to are also rated as (0) as hypersomnia is captured by the next item (A6). For early morning awakening(s), a score of (1) is given if the respondent was able to fall back asleep within 30 minutes and a score of (2) denotes an inability to fall back asleep within 30 minutes or rising earlier than intended (by > 30 minutes).

A6. Hypersomnia

This atypical subscale item is useful as the HAM-D items only assess insomnia. The key to rating this item is to compare the current vs. euthymic hours of sleep/day (including naps). This item involves a calculation of time asleep, a different approach than in items H6–8, which involve identifying periods of insomnia. Rather than simply relying on the initial probe ("Have you been sleeping more than usual this past week?"), we have found that the secondary probes are necessary to determine sleep length, including the optional probes shown in (). In our experience, it is essential to inquire about naps and weekend sleep, as some individuals who do not spontaneously endorse the initial prompt to this question do endorse naps and/or sleeping more on weekends when asked specifically. Our approach is to sum a respondent's total number of hours of sleep (including naps) in the past week and divide by 7. This number yields a daily average of hours of sleep that is then compared to their average normal or euthymic sleep length. The rating scale is clearly articulated for the increase in sleep length and we interpret it literally (without rounding): (0) no or less than 1 hour, (1) at least 1 hour, (2) at least 2 hours, (3) at least 3 hours, and (4) 4 or more hours. The language of this item assumes a regular sleep schedule; however, in our experience, some respondents report erratic sleep schedules. If that is the case, it may be necessary to calculate sleep length for each of the past 7 days.

H9. Somatic Symptoms General

We rate ambiguous answers (e.g., when the respondent was unsure or unclear about the presence of symptoms) as (1). We score any definite indication of any one or more of the somatic symptoms listed (e.g., fatigue/loss of energy, aches, pains, heaviness) as (2), consistent with the stem "any clear-cut symptom." We rate aches or pains related to a diagnosed acute infection or known injury (e.g., pain related to a broken bone, sprain, or bacterial infection) as (0). However, we rate aches and pains possibly related to underlying, chronic conditions (e.g., arthritis, fibromyalgia) as a (1) or (2) depending on clarity. Of note, our system for H9 differs from that of Morriss et al. (2008), which simply makes the distinction between nonspontaneous (1) and spontaneous (2) reports of somatic symptoms (e.g., low energy, muscle fatigue).

A7. Fatigability

The ratings are based on the pervasiveness (frequency, duration) of fatigability, if endorsed in the previous item, H9. Two of the scoring stems provide specific criteria (e.g., a 2 is at least one hour/day on at least three days, and a 1 is fatigue that is less frequent than in 2). We interpret a (3) "much of the time on most days" as at least half of the day on at least 5 days and a (4) "almost all the time" as normal energy levels for no more than 2 hours per day every day.

H10. Feelings of Guilt

This item is rated as (1) if there are general thoughts of putting oneself down, doing things wrong, or letting others down. A score of (1) includes statements of self-reproach about not getting things done at home or work due to current depressive symptoms. We code this item as (2) if the respondent endorses feeling guilty about events that happened a long time ago. An example of a (2) is expressing self-blame over past mistakes or deeds in a self-blaming way (e.g., "I should have done things differently when…"). Responses that directly characterized their depression as a form of punishment for past wrongdoings are rated as (3). We do not score this item as (3) if the respondent simply acknowledges their own possible role in maintaining depressive symptoms (e.g., "I should be doing things to improve how I feel"), which we rate as a (1). Our scoring differs from Morriss et al. (2008), which rates H10 based on "appropriateness" of guilt and pervasiveness.

H11. Suicide

A score of (1) is provided if thoughts that life is not worth living are endorsed. We code a score of (2) if the respondent reported thinking he or she would be better off dead or endorsed passive thoughts about his/her own death, but without a specified plan of action or intention to act. We give a rating of (3) if the person endorsed making plans or thinking about taking action to hurt oneself and a (4) if the person actually made attempts at suicide in the past week. Our system for ratings of (1) and (2) is commensurate with Morriss et al. (2008).

H12. Anxiety Psychic

We rated this item as (1) if the participant endorsed feeling more tense, irritable, argumentative or impatient as compared to normal. This item was rated as (2) if the participant reported worrying about little things they do not ordinarily worry about. Note that the probes ask for examples of the types of "little things" worried about. Here, we find that it is important to distinguish momentary worries about everyday things that are consistent with a (2) (e.g., where to go, what to do, what to wear, what to eat). A (2) is warranted if the participant endorsed worrying about serious matters that would similarly elicit anxiety during a euthymic mood episode (e.g., major financial stressors, relationship strain with partner). Ratings of (3) or (4) were provided on the basis of clinical presentation per the scoring stems provided.

H13. Anxiety Somatic

In addition to the probes provided, we ask the respondent to identify and then rate the worst (i.e., most bothersome) symptom endorsed from the list provided as "mild" (1), "moderate" (2), "severe" (3), or "incapacitating" (4). It is potentially confusing that the probes include questions to try and determine whether the symptom is due to depression, an illness, or a medication, but the item instructions state "BUT RATE SYMPTOMS ANYWAY." We interpret this literally and do not make efforts to tease out what is causing the symptom or whether it waxes and wanes with mood; we simply rate its severity. Our system for H12 and H13 differs from the Morriss et al. (2008) guidelines in that if both H12 and H13 were endorsed, they rated only the most distressing form of anxiety.

H14. Hypochondriasis

We rate this item as (1) if the respondent endorses an increase in bodily self-absorption (i.e., more thoughts about his/her health or how his/her body is working). We also scored (1) for any thoughts expressed about the health-related implications of current depressive symptoms (feeling fatigued, gaining/losing weight, etc. could be bad for my health). This is distinguished from generally ruminating about one's depression, which would be a (0) in the absence of health-related thoughts. We rate this item as (2) if the respondent endorsed excessively thinking about the possibility of becoming sick or ill with a disease, cold, or flu. This item is coded as (3) if the respondent reported asking for help with things because they believed that they were too sick or ill to do them alone, or if they believed that they might become sick or ill if they did them. A score of (3) is not warranted for respondents who asked for help simply because they felt too tired or depressed to do something alone; this explanation would be rated as (0) on this item because there are several other items that capture fatigue and other depressive symptoms vs. this item's specific focus on illness fears. We rate the item as (4) if the participant described symptoms indicative of hypochondriacal delusions. Our guidelines for ratings of (1) or (2) on H14 are commensurate with Morriss et al. (2008).

H15. Insight, H16. Agitation, H17. Retardation

These are observational items based on thoughts, behavior, and speech during the interview. We use the anchors provided for each scoring stem and do not feel they require further clarification. This is in line with Morriss et al. (2008)'s guidelines.

H18. Diurnal Variation Type A

If the respondent endorsed diurnal variation in response to the initial probe (i.e., morning type/feels worst before going to sleep or evening type/feels worse after waking), the secondary probes are necessary to rate the degree. We rate "a little bit worse" or "medium/moderate" responses as (1 "mild") and "a lot worse" responses as (2 "severe").

A8. Diurnal Variation Type B

This item measures regularly occurring afternoon or evening slumps in energy and/or mood. We define "regular" as at least 4 out of 7 days in order to justify a rating greater than (0). According to the scale, a slump requires recovery (i.e., a return to the level present throughout the rest of the day) at least one hour before going to sleep for the night. Therefore, slumps in mood or energy that persisted or increased in intensity until bedtime receive a rating of (0). It is essential to ask the secondary probes that assess this and for the interviewer, not the respondent, to determine whether or not there is a true slump. A slump that results in a nap is counted as a slump as long as the person wakes up in a "recovered" state for an hour before going to sleep for the night. If a slump is deemed present, the respondent verbally selects its intensity as mild (1), moderate (2), or severe (3). In-between responses are rounded downward.

H19. Depersonalization and Derealization

If any symptoms are endorsed in response to the initial probe, we ask about the severity of the symptom to generate a rating [i.e., mild (1), moderate (2), severe (3), or incapacitating (4)]. However, "spacey feelings" in the absence of other symptoms were uniformly rated as mild (1), regardless of the self-reported severity. Additionally, we did not rate any disclosed cognitive symptoms of depression, such as non-dissociative forgetfulness or difficulties in concentration, as these constructs are qualitatively distinct from the content of this item.

H20. Paranoid Symptoms

We rate paranoid suspicions without ideas of being directly targeted as (1). We define ideas and delusions of reference (i.e., ratings of (2) and (3), respectively) as beliefs that irrelevant, unrelated, or innocuous things in the world refer to oneself directly or have special personal significance. We inquire regarding the degree of conviction in the reported belief in order to clarify between a rating of (2) or (3), with an idea being less- and a delusion being more firmly-held and met with resistance upon questioning.

H21. Obsessional and Compulsive Symptoms

We use the definitions of obsessions and compulsions provided in DSM-5 (APA, 2013) as a reference for coding this item. We do not include reported depressive ruminative thinking (i.e., repetitive thoughts about one's depressive symptoms and their causes and consequences) or paranoid suspicions from H20 in this item. We differentiate between a (1) "mild" and (2) "severe" based upon the number of symptoms reported, their frequency, the amount of time involved in each, and the overall impairment or distress associated with the obsessions or compulsions.

Interview Rating Process

All subjects gave consent to taping all study interviews. The parent study used an assessor-blind design. At each study visit, the SIGH-SAD was administered live by the primary rater and archived as a digital recording, consistent with the standard practice of video or audiotaping interviews in clinical trials. Consistent with Hamilton's (1960) recommendation for two raters, a second rater later listened to and rated each archived SIGH-SAD. Both primary and secondary raters were drawn from a general rater pool, such that all raters served in both roles. The study spanned eight consecutive winters, including six winters of recruitment/treatment and two winters to complete remaining followups. Therefore, the rater pool changed each year, depending on laboratory staff. The grand rater pool (N = 34), included 1 PI, 1 Project Coordinator, 6 graduate students, and 26 undergraduate students. Other than the PI, no one had prior experience with the HAM-D before working on the study.

Protocol for Training Clinical Raters

The following two-pronged training protocol was repeated each year of recruitment prior to enrolling any new subjects in the next cohort. In an effort to limit rater drift, every member of our team doing SIGH-SAD ratings for the upcoming year was required to participate in

annual training, no matter how experienced. First, the PI led two or three training sessions with the study team to discuss each of the 29 SIGH-SAD items in detail and the nuances of scoring them per our scoring rules described above. All raters had access to the item-by-item scoring rules throughout the study. Second, the raters practiced rating at least three audiotapes of SIGH-SADs from past studies and discussed their ratings in a group session lead by the PI. New trainees were additionally required to observe a veteran rater perform a live SIGH-SAD in the study, independently record their ratings, and discuss them with the interviewer after the session. To become an independent rater, a new trainee had to perform a mock SIGH-SAD interview on the PI (role-playing a depressed patient) with good flow and proficiency and obtain item ratings that corresponded well to the PI's judgments, defined as item score disagreement on ≤ 3 items and a difference in total score ≤ 5 points.

Our training protocol is similar to other approaches that utilize both a didactic and applied training process prior to rater administration of the instrument (Koback, Lipsitz, & Feiger, 2003; Morriss et al., 2008). Some studies reference the use of didactic trainings incorporating lectures, viewing videotaped interviews, and/or scoring practice prior to administration of the instrument (Moberg et al., 2001; Muller & Dragicevic, 2003). Our protocol also actively involves the rater and requires accurate administration and scoring of the instrument before independent performance. Morriss and colleagues (2008) who describe a process in which interviewers watch videotapes of patients undergoing the HAM-D, observe interviews with live patients, and subsequently administer interviews that are audiotaped and reviewed for feedback. Our protocol is similar in its multi-step approach, but is unique in providing ongoing training and review of scoring rules, particularly in the context of a multi-year study.

Protocol for Ensuring Accuracy of Scores

The following protocol was developed to increase the accuracy of SIGH-SAD scores for data analysis. The project statistician identified any SIGH-SAD administration with (1) a 5-point or greater discrepancy in total score (minus items H16 and H17, which require a direct observation of the patient) between the original and second rater and/or (2) a between-rater discrepancy on whether SIGH-SAD recurrence criteria[1] were met at a followup interview. In the absence of a consensus on what the smallest clinically meaningful difference between SIGH-SAD scores is, the threshold of 5-points was selected as an unacceptable split between raters because a 5-point difference is inarguably clinically meaningful. A difference of opinion on depression recurrence status at followup was another criterion for a between-rater discrepancy, regardless of split in scores, because recurrence was the primary outcome in this trial. For interviews identified as rater-discrepant based on either criterion, two new blinded raters were selected from our team to re-rate those particular interviews. The statistician then examined ratings from all four raters of each discrepant interview to identify SIGH-SAD items where the original rater disagreed with at least two of the other three raters. Subsequently, the two new raters discussed responses to those particular items on each interview and tried to reach a consensus on the most accurate score. If a consensus

[1]Recurrence was defined using Terman, Terman, and Rafferty's (1990) criteria for current winter depression, i.e., SIGH-SAD score ≥ 20 (including atypical symptom subscale score ≥ 5).

could not be reached, the plan was for the PI to be consulted about what was said in the interview (keeping the PI blind to subject identity, and, therefore, group assignment) to make a final decision about the most accurate score for that item. However, we did not have to resort to this level of oversight to resolve any discrepancy. Any corrected SIGH-SAD scores were used in the primary and secondary data analyses. However, inter-rater reliability statistics reflected agreement between the original (uncorrected) rater and the second rater.

Clinical Trial Data

Our group applied these methods in the context of an R01-level NIMH-funded clinical trial with the SIGH-SAD as the primary outcome. Details of the trial are beyond the scope of this paper. The clinical trial protocol is archived elsewhere (Rohan et al., 2013), the sample baseline characteristics and participant flow are presented in Rohan et al. (2015, in press), and treatment efficacy results are published for the acute treatment (Rohan et al., 2015) and followup (Rohan et al., in press) phases. In brief, 177 community adults aged 18 or over with Major Depression, Recurrent with Seasonal Pattern were enrolled at our University-based research lab specializing in SAD, randomized to light therapy or SAD-tailored cognitive-behavioral therapy (CBT-SAD), and interviewed at 10 time points: pre-treatment, weekly across the 6-weeks treatment phase, and at followups the following summer, next winter (i.e., the next wholly new winter after the winter of study treatment), and two winters following enrollment. Therefore, the followup interval was approximately 2 years, depending on when treatment was completed (range = 23–26 months).

A total of 1,589 interviews were completed in the trial, and 1,535 of them were successfully archived on audiotape (96.6%) and rated by a second rater. Table 1 displays inter-rater reliability statistics as intra-class correlations (ICCs) for SIGH-SAD interviews at each time point. Inter-rater reliability was consistently very good and ranged from ICCs of .923 to . 967. Table 2 displays the number of SIGH-SAD interviews that met criteria for between-rater discrepancies at each time point based on a difference in total score or in recurrence status. Only 86 interviews met either criterion for a between-rater discrepancy (5.6% of interviews archived). For the three followup timepoints that detect between-rater discrepancies based on point-split OR differential recurrence status, only one interview satisfied both sets of criteria at summer and at second winter followup. At next winter followup, there were four rater-discrepant interviews that met both sets of criteria. Specifically, 4/12 of interviews identified as rater-discrepant on scores also differed by rater on recurrence status and 4/10 interviews identified as rater-discrepant on recurrence status also had rater-discrepant scores. This pattern suggests that the two sets of criteria for between-rater discrepancies largely capture unique (i.e., non-overlapping) rater discrepancies and both should be retained. For the subset of interviews flagged as discrepant, the magnitude of the mean difference in scores between raters was relatively consistent across the treatment phase and at summer (range = 5.0–6.2). The magnitude of the difference between raters in scores on discrepant interviews was lower at the two winter followups, reflecting that most of the interviews flagged as discrepant on recurrence status were not discrepant on scores.

Table 3 displays the specific items that received different ratings from Rater 1 vs. the other three raters in the subset of interviews identified as rater-discrepant. The items that contributed most frequently to discrepancies (i.e., on 20 or more interviews) were HAM-D items “Depressed Mood,” “Work and Activities,” “Middle Insomnia,” and “Hypochondriasis” and Atypical items “Fatigability” and “Hypersomnia,” suggesting that these items are particularly problematic to rate.

Conclusions

To our knowledge, this is the first comprehensive protocol for using the HAM-D from rater training to administration/scoring to data analysis. Our approach is three-pronged. First, consistent with the recommendations of Bagby’s et al. (2004) seminal review of the HAM-D, we proposed item scoring rules in an effort to improve upon the measure’s psychometric properties as an alternative to Morriss et al’s (2008) guidelines for the 17-item HAM-D. Second, we delineated our procedures for training beginning raters and preventing rater drift among all raters, including experienced ones. Third, we shared a data management algorithm for detecting potentially inaccurate interviews and resolving rater discrepancies prior to primary outcome analyses. The latter algorithm can be used to identify specific items that may be problematic to rate, thereby requiring greater emphasis in training.

Using data from a clinical trial with longitudinal followup, we reported relatively high inter-rater reliability (ICCs = .923 to .967), a relatively low frequency (5.6%) of interviews identified as rater-discrepant (i.e., potentially inaccurate), and successful resolution of all discrepancies using these methods. Although our trial used the seasonal affective disorder version of the HAM-D (SIGH-SAD), the 17- and 21-item versions of the HAM-D are contained within it. In fact, the SIGH-SAD has some advantages over the HAM-D for depression assessment and research in general, beyond SAD, because it captures a broader range of depressive symptoms, including the atypical symptoms. Clinical trials researchers using the HAM-D or SIGH-SAD might want to consider adopting our protocol in part or full. Practicing clinicians might want to consider tailoring our methods to their particular needs. For example, clinicians could potentially benefit from using the scoring rules we developed or using our methods to train staff to evaluate patient outcomes over treatment.

Limitations

It is unknown whether results using our algorithm generalize beyond relatively well-trained, experienced raters working on a clinical trial in a research setting. As Williams (1988) noted as a limitation with any protocol where one rater performs the interviewer and a second rater rates the same interview, the two ratings lack independence and taping likely increases the inter-rater reliability. We conceptualized these training and surveillance methods as a means of quality assurance, including maximizing our inter-rater reliability, preventing rater drift, and increasing our measurement accuracy. However, we did not include experimental conditions to test the efficacy of our methods (e.g., randomly assign raters to participate in training or not, or to have access to the item scoring rules or not). We also recognize that there are several parts to this protocol, and it remains unclear which aspects are necessary to achieve these results. Our procedure of assigning two additional raters for interviews

identified as discrepant may not be feasible for all clinical trials. Nonetheless, we chose to share our articulated protocol to provide one example of a systematic approach to an otherwise notoriously problematic measure. It is possible that other groups have developed their own unpublished HAM-D protocols with advantages over ours.

References

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. fifth. Arlington, VA: American Psychiatric Publishing; 2013.

Bagby RM, Ryder AG, Schuller DR, Marshall MB. The Hamilton Depression Rating Scale: Has the gold standard become a lead weight? Am. J. Psychiat. 2004; 161:2163–2177. [PubMed: 15569884]

Hamilton M. A rating scale for depression. J. Neurol. Neurosur. Ps. 1960; 23:56–62.

Kobak KA, Lipsitz JD, Feiger A. Development of a standardized training program for the Hamilton Depression Scale using internet-based technologies: results from a pilot study. J. Psychiat. Res. 2003; 37:509–515. [PubMed: 14563382]

Lam R, Levitt A, Levitan R, Enns M, Morehouse R, Michalak E, Tam E. The Can-SAD study: A randomized controlled trial of the effectiveness of light therapy and fluoxetine in patients with winter seasonal affective disorder. Am. J. Psychiat. 2006; 163:805–812. [PubMed: 16648320]

Moberg PJ, Lazarus LW, Mesholam RI, Bilker W, Chuy IL, Neyman I, Markvart V. Comparison of the standard and structured interview guide for the Hamilton Rating Scale in depressed geriatric inpatients. Am. J. Geriat. Psychiat. 2001; 9:35–40.

Morriss R, Leese M, Chatwin J, Baldwin D. Inter-rater reliability of the Hamilton Depression Rating Scale as a diagnostic and outcome measure of depression in primary care. J. Affect. Disorders. 2008; 111:204–213. [PubMed: 18374987]

Muller MJ, Dragicevic A. Standardized rater training for the Hamilton Rating Scale (HAMD-17) in psychiatric novices. J. Affect. Disorders. 2003; 77:65–69. [PubMed: 14550936]

Rohan KJ, Evans M, Mahon JN, Sitnikov L, Ho S, Nillni YI, Postolache TT, Vacek PM. Cognitive-behavioral therapy vs. light therapy for preventing winter depression recurrence: Study protocol for a randomized controlled trial. TRIALS. 2013; 14:82. [PubMed: 23514124]

Rohan KJ, Mahon JN, Evans M, Ho S, Meyerhoff J, Postolache TT, Vacek PM. Randomized trial of cognitive-behavioral therapy vs. light therapy for seasonal affective disorder: Acute outcomes. Am. J. Psychiat. 2015; 172:862–869. [PubMed: 25859764]

Rohan KJ, Meyerhoff J, Ho S, Evans M, Postolache TT, Vacek PM. Outcomes one and two winters following cognitive-behavioral therapy or light therapy for seasonal affective disorder. Am. J. Psychiat. in press.

Terman M, Terman JS, Rafferty B. Experimental design and measures of success in the treatment of winter depression by bright light. Psychopharmacol. Bull. 1990; 26:505–510. [PubMed: 2087543]

Terman M, Terman JS, Ross DC. A controlled trial of timed bright light and negative air ionization for treatment of winter depression. Arch. Gen. Psychiat. 1998; 55:875–882. [PubMed: 9783557]

Trajković G, Starčević V, Latas M, Leštarević M, Ille T, Bukumirić Z, Marinković J. Reliability of the Hamilton Rating Scale for Depression: A meta-analysis over a period of 49 years. Psychiat. Res. 2011; 189:1–9.

Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. Arch. Gen. Psychiat. 1988; 45:742–747. [PubMed: 3395203]

Williams JBW. Standardizing the Hamilton Depression Rating Scale: Past, present, and future. Eur. Arch. Psy. Clin. N. 2001; 251:6–12.

Williams JB, Kobak KA, Bech P, Engelhardt N, Evans K, Lipsitz J, Olin J, Pearson J, Kalali A. The GRID-HAMD: Standardization of the Hamilton Depression Rating Scale. Int. Clin. Psychopharm. 2008; 23:120–129.

Williams, JBW.; Link, MJ.; Rosenthal, NE.; Amira, L.; Terman, M. Structured Interview Guide for the Hamilton Depression Rating Scale—Seasonal Affective Disorder Version (SIGH-SAD). New York, NY: New York State Psychiatric Institute; 1992.

Highlights

- We share a comprehensive protocol for the Hamilton Rating Scale for Depression.
- The protocol includes item scoring rules and rater training procedures.
- We present an algorithm to detect inaccurate interviews and resolve rater discrepancies.
- This protocol was applied in a longitudinal clinical trial of 177 depressed patients.
- These methods yielded high inter-rater reliably and few between-rater discrepancies.

Table 1

Inter-rater reliability on the SIGH-SAD interviews at each time point

Time Point	Inter-Rater Reliability (ICC)
Pre-Treatment	0.923
Week 1	0.958
Week 2	0.965
Week 3	0.950
Week 4	0.967
Week 5	0.962
Post-Treatment	0.961
Summer	0.963
Next Winter	0.965
Second Winter	0.967

Note. SIGH-SAD = Structured Interview Guide for the Hamilton Depression Rating Scale—Seasonal Affective Disorder Version.

Table 2

Between-rater discrepancies on SIGH-SAD interviews at each time point

Time Point	No. interviews completed/No. participants eligible	No. interviews with Rater 2[a] /No. interviews completed	No. interviews with discrepant scores[b]	No. interviews with discrepant recurrence status[c]	Mean discrepancy between raters (SD)[d]
Pre-Treatment	177/177 (100%)	171/177 (96.6%)	6 (3.5%)	N/A	6.2 (1.0)
Week 1	154/177 (87%)	149/154 (96.8%)	6 (4.0%)	N/A	6.2 (1.3)
Week 2	153/177 (86.4%)	146/153 (95.4%)	9 (6.2%)	N/A	5.2 (0.4)
Week 3	154/177 (87%)	148/154 (96.1%)	1 (0.7%)	N/A	6.0 (N/A)
Week 4	149/177 (84.2%)	144/149 (96.6%)	4 (2.8%)	N/A	5.3 (0.5)
Week 5	149/177 (84.2%)	141/149 (94.6%)	3 (2.1%)	N/A	5.7 (1.2)
Post-Treatment	173/177 (97.7%)	168/173 (97.1%)	5 (3.0%)	N/A	5.0 (0.0)
Summer[e]	143/153 (93.5%)	140/143 (97.9%)	4 (2.9%)	2 (1.4%)	5.6 (1.9)
Next Winter	169/177 (95.5%)	165/169 (97.6%)	12 (7.3%)	10 (6.1%)	4.4 (2.3)
Second Winter	168/177 (94.9%)	163/168 (97.0%)	12 (7.4%)	12 (7.4%)	3.7 (2.1)

[a] Represents the number of interviews that were successfully recorded and, therefore, available for a second rater to verify.

[b] Interviews with a 5 or more point difference between Rater 1 and Rater 2 on total SIGH-SAD score. The denominator for the percentages is the number of interviews with a Rater 2 from the prior column.

[c] Interviews where Rater 1 and Rater 2 differed on whether SIGH-SAD recurrence criteria were met at a followup visit.

[d] The absolute value of the mean difference between Rater 1 and Rater 2 in total SIGH-SAD scores for the interviews identified as discrepant based on discrepant scores and/or discrepant recurrence status.

[e] 24 subjects participated during the initial (feasibility) year, which did not include a summer followup, leaving 153 total participants eligible for summer followup.

Table 3

SIGH-SAD items with rater discrepancies on the 86 interviews identified as discrepant

Scale item number (symptom assessed)	No. discrepant interviews with rater differences on item (% of discrepant interviews with rater differences on item)
H1 (Depressed Mood)	34 (39.5%)
H2 (Work and Activities)	31 (36.0%)
A1 (Social Withdrawal)	4 (4.6%)
H3 (Genital Symptoms)	4 (4.6%)
H4 (Somatic Symptoms Gastrointestinal)	10 (11.6%)
H5A (Weight Loss)	3 (3.5%)
A2 (Weight Gain)	13 (15.1%)
A3 (Appetite Increase)	3 (3.5%)
A4 (Increased Eating)	2 (2.3%)
A5A (Carbohydrate Craving or Eating)	2 (2.3%)
H6 (Early Insomnia)	17 (19.8%)
H7 (Middle Insomnia)	31 (36.0%)
H8 (Late Insomnia)	18 (20.9%)
A6 (Hypersomnia)	29 (33.7%)
H9 (Somatic Symptoms General)	27 (31.4%)
A7 (Fatigability)	27 (31.4%)
H10 (Guilt)	12 (14.0%)
H11 (Suicide)	2 (2.3%)
H12 (Anxiety Psychic)	9 (10.5%)
H13 (Anxiety Somatic)	10 (11.6%)
H14 (Hypochondriasis)	35 (40.7”%)
H15 (Insight)	7 (8.1%)
H18 (Diurnal Variation Type A)	7 (8.1%)
A8A (Diurnal Variation Type B)	5 (5.8%)
H19 (Depersonalization, Derealization)	5 (5.8%)
H20 (Paranoid Symptoms)	1 (1.2%)
H21 (Obsessive and Compulsive Symptoms)	4 (4.6%)

Note. Our accuracy protocol involves identifying interviews as discrepant between Raters 1 and 2 on the basis of a split in scores and/or recurrence status and subsequently assigning two new independent raters to rate those interviews in their entirety. This table displays specific items that differed between the original rater and at least two of the other three raters on the 86 interviews identified as discrepant. See Table 1 for the criteria used to define a between-rater discrepancy on an interview.