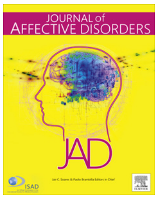
Journal of Affective Disorders 200 (2016) 111–118

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad



Research paper

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[a,*], Jennifer N. Rough[a], Maggie Evans[a], Sheau-Yan Ho[a], Jonah Meyerhoff[a], Lorinda M. Roberts[a], Pamela M. Vacek[b]

[a] *Department of Psychological Science, University of Vermont, Burlington, VT, United States*
[b] *Department of Medical Biostatistics, University of Vermont College of Medicine, Burlington, VT, United States*

ARTICLE INFO

Article history:
Received 11 August 2015
Received in revised form
14 December 2015
Accepted 24 January 2016
Available online 20 April 2016

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

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 1589 completed interviews with 1535 (96.6%) archived.

Results: Inter-rater reliability ranged from ICCs of .923–.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.

1. 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 et al. (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 to .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,

[*] Correspondence to: Department of Psychological Science, University of Vermont, John Dewey Hall, 2 Colchester Avenue, Burlington, VT 05405-0134, United States.

E-mail address: kelly.rohan@uvm.edu (K.J. Rohan).

http://dx.doi.org/10.1016/j.jad.2016.01.051
0165-0327/© 2016 Elsevier B.V. All rights reserved.

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 et al., 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 et al.'s (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 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.

2. 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, 1988) 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