Acta Psychiatrica Scandinavica

WILEY

Acta Psychiatrica Scandinavica

ORIGINAL ARTICLE OPEN ACCESS

Improving the Accuracy and Reliability of Ratings on the Hamilton Depression Rating Scale via a Video-Based Training Program

Pernille Kølbæk[1,2] | Botilla Dalsgaard Jensen[2] | Erik Roj Larsen[3] | Søren Dinesen Østergaard[1,2]

[1]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark | [2]Department of Affective Disorders, Aarhus University Hospital—Psychiatry, Aarhus, Denmark | [3]Translational Neuropsychiatry Unit, Department of Clinical Medicine, Aarhus University, Aarhus, Denmark

Correspondence: Søren Dinesen Østergaard (soeoes@rm.dk)

Received: 13 May 2025 | **Revised:** 14 August 2025 | **Accepted:** 18 August 2025

Funding: The authors received no specific funding for this work.

Keywords: depression | education | psychiatric status rating scales | psychopathology | staff development

ABSTRACT

Introduction: The clinician-rated 17-item Hamilton Depression Rating Scale (HAM-D17) allows for a systematic severity assessment of depressive symptoms. Applying the HAM-D17 in clinical practice requires that staff members' ratings on the HAM-D17 are accurate and reliable. Here, we aimed to investigate whether such accuracy and reliability can be achieved through a brief video-based training program.

Methods: One-hundred-and-ten psychiatric hospital staff members (psychologists, medical doctors, nurses, health care workers, physio-/occupational therapists, and social workers) performed baseline HAM-D17 ratings after watching a videotaped patient interview. Subsequently, a theoretical introduction video was displayed, followed by five successive videotaped patient interviews. After watching each interview, individual ratings were conducted before a video providing the gold standard rating was displayed. Accuracy was estimated by calculating the proportion of participants whose ratings did not display a deviation from the gold standard of > 1 point on all individual HAM-D17 items and > 6 points on the HAM-D17 total score. Reliability was calculated using Gwet's agreement coefficient (AC1).

Results: At baseline and after the sixth rating session, 43% versus 70% of the staff members, respectively, rated within the acceptable deviation of the gold standard ($p < 0.001$). At the HAM-D17 item level, baseline reliability indices were highest for item 6 (Late Insomnia) and lowest for item 14 (Sexual Interest) (AC1 = 0.97 vs. 0.47), but both improved following training (AC1 = 0.99 vs. 0.84 at the sixth rating session).

Conclusions: Most staff members conducted accurate and reliable HAM-D17 ratings after participating in a brief video-based training program.

1 | Introduction

Monitoring of symptom severity is an essential part of the treatment of depression. Ideally, symptom assessment should be carried out by means of rating scales to ensure an objective and systematic approach [1, 2]. Routine use of rating scales in clinical practice allows for measurement-based care, which refers to the systematic use of rating scales to guide clinical decision-making [3]. Measurement-based care has shown promise in the treatment of major depression by shortening the time to both treatment response to and remission [4–7].

Summary

- Significant outcomes
  - Accuracy and reliability of staff ratings on the 17-item Hamilton Depression Rating Scale was relatively low prior to training.
  - Accuracy and reliability of staff ratings on the 17-item Hamilton Depression Rating Scale was substantially improved following participation in a video-based training program.
- Limitations
  - The participants' interview technique was not evaluated.
  - Persistence of the training effect was not tested.

The clinician-rated 17-item Hamilton Depression Rating Scale (HAM-D17) is one of the most widely used rating scales to measure the severity of depressive symptoms. It contains a six-item subscale covering core depression symptoms (HAM-D6), which possesses strong psychometric properties, including unidimensionality [8, 9] and high sensitivity to the separation of the effects of antidepressants and placebo [10–13]. The remaining 11 items assess symptoms that are often associated with depression and have high clinical relevance, such as insomnia and suicidal thoughts. However, these 11 items are less specific for depression and may pick up on common side effects of the most widely used antidepressant drugs [14], which is problematic from a psychometric perspective [15, 16].

A prerequisite for the utility of the HAM-D17 and the HAM-D6 subscale is that clinical staff can produce ratings that are both reliable (i.e., reproducible) and accurate (i.e., in agreement with a gold standard). Achieving such reliability and validity may be facilitated by providing clinicians with a standardized introduction to the scale, including information about scoring conventions, the psychopathology covered by the items, and the process of collecting data for the purpose of rating via a clinical interview and observation of the patient [17–21].

Studies evaluating the effect of training staff in HAM-D17 rating have obtained promising results. However, they were limited by including only inexperienced raters or lacking reproducibility due to live instruction and discussion (which impedes consistency and scalability, making it difficult to standardize across different settings and instructors), inadequate randomization of patient interviews, small sample sizes (i.e., a limited number of raters being tested or patients involved in the training program) [19], or using interviews of actors instead of patients [17, 19, 22–24]. To overcome these limitations, we created a video-based HAM-D17 rater training program including a theoretical introduction, authentic patient interviews, and gold standard rating rationales.

2 | Aims of the Study

The primary aim of the study was to assess potential improvement in the reliability and accuracy of ratings produced by staff members with diverse professional backgrounds and different levels of experience during the training process. The secondary aim was to identify staff characteristics associated with rating skills at baseline. Finally, the staff members' self-perceived rating skills before and after the training program—as well as their level of satisfaction with it—were evaluated.

3 | Methods

The study was designed based on an analogue video-based training program for the six-item Positive and Negative Syndrome Scale (PANSS-6) [25].

3.1 | Setting and Participants

Clinical staff members from the Psychiatric Services (outpatient clinics and inpatient units) of the Central Denmark Region were invited to participate in the study. Data collection took place between May 2021 and October 2022. All participating staff members provided informed written consent prior to participation. The data were collected, processed, and stored in accordance with the European Union General Data Protection Regulation.

3.2 | Study Procedure

The overall study procedure is illustrated in Figure 1, and the individual training components are described in detail below.

Participants were divided into eight groups. Group composition was based on logistical considerations, such as scheduling, room capacity, and availability of participants. All groups followed the same structured training protocol; however, the order of patient interview videos and corresponding gold standard rating videos was randomized across groups to reduce potential order effects and learning biases.

Training was conducted on-site over three consecutive days within a single week. On Day 1, participants first completed a pre-training questionnaire assessing prior experience with and confidence in using the HAM-D17 (see Pre- and Post-Training Questionnaires). Immediately afterward, they watched and rated a baseline patient interview video without any prior instruction. This served as a measure of baseline scoring skills.

Next, participants viewed a theoretical introduction video, which provided a structured overview of the HAM-D17, including item definitions, scoring rules, and common pitfalls. Later the same day, they watched and rated a second patient interview video, followed by the relevant gold standard rating video.

On Days 2 and 3, participants viewed and independently rated four additional patient interviews (two per day), each followed by the relevant gold standard rating video. These gold standard rating videos presented item-level scores along with brief

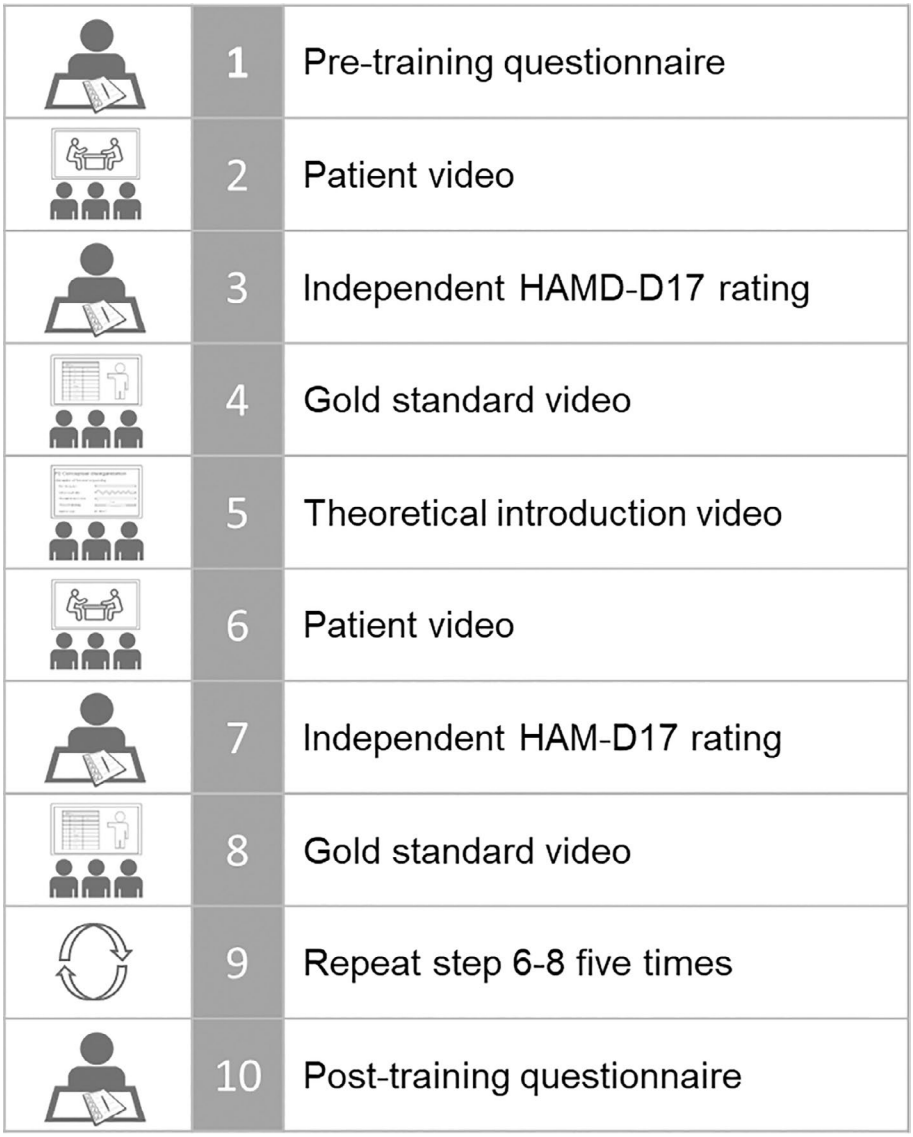

FIGURE 1 | Overview of the study procedure. Modified from Kølbæk et al. [20].

justifications, offering insight into expert reasoning and scoring criteria.

Throughout the training, participants always completed their ratings before viewing the gold standard video to preserve the integrity of their independent assessments.

At the end of the final day, participants completed a post-training questionnaire designed to evaluate changes in confidence, perceived competence, and satisfaction with the training program.

Further details on the three main components—the theoretical introduction, patient interviews, and expert rating videos—are provided in the following section.

3.3 | The Video-Based Training Program

As shown in Figure 1, the training program consisted of three components:

i. A video-based theoretical introduction using a script written by P.K., E.R.L., and S.D.Ø. This 60-min video covers scoring conventions for HAM-D17 ratings, the psychopathology covered by the items included in the HAM-D17, and the principles of conducting a semi-structured interview. In addition, several examples of applying the scoring conventions for the HAM-D17 are presented. To support replicability, the slide deck and accompanying script for the theoretical introduction can be obtained by contacting the first author.

ii. Display of video recordings of HAM-D17 interviews of inpatients and outpatients from the Department of Affective Disorders at Aarhus University Hospital, conducted by trained clinicians using a semi-structured interview guide for the HAM-D17 [26]. Eligible patients had a diagnosis of unipolar depression (ICD-10 codes F32 or F33) as determined by their treating psychiatrist and were at least 18 years old. Patients were considered ineligible if they were under the influence of alcohol or illicit drugs at the time of the interview (assessed by the

recruiting clinician), were treated using coercive measures, or had comorbid organic mental disorder or intellectual disability (IQ < 70—as per clinical judgment by the referring clinician). All patients provided written informed consent for the video recordings to be used for research and educational purposes. A total of 12 patient interviews were recorded for potential use in the training program. From these, six were selected based on two primary criteria: (1) coverage of a broad range of symptom severity across HAM-D17 items to ensure sufficient variance in clinical presentation, and (2) acceptable technical quality, including clear audio and visual recordings to facilitate accurate scoring. The selected videos were intended to expose raters to varied clinical profiles while maintaining a consistently high standard of recording quality. The mean duration of the patient videos was 27 min (SD = 10).

iii. Display of videos presenting the gold standard ratings, including item-by-item scoring rationales, for each of the six HAM-D17 patient interviews. The gold standard scores were established by a group of eight trained raters who first rated the videos independently, followed by a structured group discussion to reach consensus on each item. In cases of disagreement, expert input was provided by P.K. and E.R.L. The gold standard videos also included brief comments on interview technique when relevant, offering additional pedagogical value. The mean duration of the gold standard videos was 13 min (SD = 3).

3.4 | The HAM-D17

The participants rated the patient video interviews using a version of the HAM-D17, which contains specific anchoring points for the scoring of each item. The reference period for rating is the previous 3 days, except for the rating of items 8 (Psychomotor Retardation), 9 (Agitation), and 17 (Weight Loss). Items 8 and 9 are rated based on observation during the interview, and item 17 is rated based on changes since the baseline assessment. Eight items of the HAM-D17 are rated on a scale from 0 to 2, and nine items are rated on a scale from 0 to 4, yielding a total score ranging from 0 to 52. We used a structured version of the HAM-D17 based on the rating criteria introduced by Bech et al. [27], combined with the Danish modified version of the semi-structured interview guide originally developed by Williams [28], as adapted by Bech et al. This version is commonly referred to as the ABC-model, a term promoted by Bech [26], in which the 17 HAM-D items are grouped into: Group A: Core symptoms of depression (the HAM-D6), Group B: Stress-related arousal symptoms, and Group C: Suicidal thoughts and lack of insight. In accordance with this model, the clinical interviews used in our training videos were structured to begin with Group B items (e.g., sleep, appetite), proceed to Group A (core affective symptoms), and conclude with Group C (more sensitive topics), thereby reflecting a clinically intuitive and patient-centered interview flow. The HAM-D17 is described in greater detail in a previous study that employed this version of the scale [29].

3.5 | Pre- and Post-Training Questionnaires

The pre-training questionnaire consisted of questions concerning the participants' demographic and clinical background (i.e., their age, sex, education, department of current employment, years of clinical experience, research experience, and the number of HAM-D17 assessments conducted in clinical practice), their level of interest in the psychopathology of depression (rated on a scale ranging from 1 [no interest] to 10 [highest possible interest]), and their self-perceived skills in HAM-D17 rating (rated on a scale ranging from 1 [not at all competent] to 10 [maximally competent]). In the post-training questionnaire, the participants again rated their self-perceived skills in HAM-D17 ratings and recorded their satisfaction with the training program (on a scale from 1 [extremely unsatisfied] to 10 [extremely satisfied]). Finally, the participants were invited to leave comments regarding the theoretical introduction video, the patient interviews, the gold standard videos, and other aspects of the training program.

3.6 | Statistical Analysis

In the following, rating accuracy refers to the degree of agreement with gold standard ratings, while inter-rater reliability refers to the agreement among raters. The randomization of video presentation order was implemented to minimize the influence of the order on the accuracy and reliability estimates. However, this approach presented statistical challenges as participants rated different videos at each stage of the training program. Nonetheless, we considered randomization to be crucial and also used this approach in the analogue study of the video-based PANSS-6 training program [25]. The reliability of ratings was assessed after each of the six rating sessions by means of Gwet's agreement coefficient (AC1) [30]. The validity of ratings was also evaluated after each rating session by calculating the proportion of participants whose ratings did not display a deviation > 1 point from the gold standard on all individual HAM-D17 items and > 6 points on the gold standard HAM-17 total score. The item level threshold is commonly used in HAM-D17 training and inter-rater reliability context [24, 31], whereas the total score threshold mirrors clinical trial definitions of minimal detectable change on HAM-D17 (≈7 points) and corresponds to the typical placebo response magnitude [32]. Changes in the accuracy of ratings during the training program were evaluated by comparing the proportion of participants whose scores met these criteria after each post-baseline patient video using the McNemar test.

Logistic regression analysis was employed to identify participant characteristics associated with baseline rating skills. Specifically, we first conducted a univariable logistic regression analysis to examine the association between participant characteristics (i.e., their age, sex, education, department of current employment, years of clinical experience, research experience, and the number of HAM-D17 assessments conducted in clinical practice), their level of interest in the psychopathology of depression, and their self-perceived competence in conducting HAM-D17 ratings and their performance at baseline in terms of conducting an accurate HAM-D17 rating. For

continuous variables, a median-split was used. Subsequently, covariates with a p-value <0.1 in this initial test were taken forward for multivariable logistic regression analysis.

Differences in the self-perceived rating skills of staff members before and after the training program were examined using the paired t-test.

For all analyses, two-tailed p-values <0.05 were considered statistically significant. The analyses were carried out using STATA version 18.5.

4 | Results

4.1 | Participant Characteristics

The median age of the participants ($n=110$) was 41 years (interquartile range [IQR = 24]). Most of the participants were female (81%), and 55% worked in outpatient clinics. The professional backgrounds of participants were as follows: psychologists (13%), nurses (56%), medical doctors (10%), and other professions (health care workers, physio-/occupational therapists, and social workers) (21%). In total, 44% were employed at Aarhus University Hospital—Psychiatry, and the remainder at regional psychiatric hospitals. The median time since education completion was 8 years (IQR = 17) and the median experience with depression treatment was 5 years (IQR = 14). The median number of HAM-D17 ratings conducted prior to the training was 10 (IQR = 49). A total of 15% of the participants had research experience. The median score for interest in the psychopathology of depression measured on a scale from 1 (no interest) to 10 (highest possible interest) was 8 (IQR = 3).

4.2 | Reliability of Ratings

Table 1 illustrates the reliability of ratings for the HAM-D17 items, which ranged from 0.47 (item 14—Sexual Interest) to 0.97 (item 6—Late Insomnia) at baseline, and from 0.74 (item 13—Somatic Symptoms) to 0.99 (item 6—Late Insomnia) at the endpoint (rating of the sixth patient video interview).

4.3 | Accuracy of Ratings

Table 2 lists the proportion of participants with accurate ratings. When compared to the baseline rating, the proportion of the participants' ratings meeting this cut-off was statistically significantly improved (McNemar test, $p<0.01$) after the third, fifth, and sixth patient videos.

Table 3 lists the results of the analyses investigating the association between staff characteristics and baseline rating accuracy. The only characteristic that was statistically significantly associated with accurate HAM-D17 rating was working in an inpatient unit (univariable logistic regression analysis odds ratio = 0.44 (95% CI: 0.20–0.97)). Consequently, a multivariable analysis was not conducted.

4.3.1 | Self-Perceived Post-Training Skills and Satisfaction With Training

The median level of the participants' self-perceived skills in conducting the HAM-D17 assessment was 5 (IQR = 4) before and 8 (IQR = 1) after participating in the video-based training program (Figure 2). The median level of satisfaction with the video-based training (on a scale from 1 (extremely unsatisfied) to 10 (extremely satisfied)) was 7 (IQR = 2). Written feedback from the participants in the post-training questionnaire is summarized and categorized in Table S1. In brief, the most frequent (46%–75%) comments regarding the theoretical introduction video and the patient and expert videos were generally positive statements such as "interesting," "good," and "educational." The most frequent (32%) overall feedback comment was the request for dialogue/discussion or live instructions.

5 | Discussion

We found that a video-based training program focusing on HAM-D17 rating resulted in a substantial and statistically significant improvement in the accuracy and reliability of HAM-D17 ratings conducted by staff members with varying professional backgrounds and experience. The only participant characteristic that was statistically significantly associated with accurate HAM-D17 rating at baseline was employment within an outpatient unit.

The HAM-D17 is widely used in both research and clinical practice, including in Denmark, where one of the national quality indicators for psychiatric treatment of patients with depression includes the rating of patients on the HAM-D17 within specified timeframes [33]. Thus, it was surprising to observe the relatively poor baseline rating skill in this study, as fewer than half of the participating staff members rated the HAM-D17 adequately. The findings from the "other" staff group—who had more heterogeneous backgrounds and typically lacked formal clinical training—highlight an important limitation of the video-based training approach. Despite completing the same training as other participants, only 48% of this group met the predefined accuracy criteria at the final assessment. Although this proportion corresponds to the baseline accuracy of medical doctors and psychologists, it also underscores that while video-based training can support skill development, some subgroups of staff may require additional training.

Furthermore, we were surprised to find that levels of experience and professional background were not associated with HAM-D17 rating skills at baseline. These results are in contrast with those from our analogue study on the effect of a video-based training program in PANSS-6 rating [25], where a significantly larger proportion of psychologists and medical doctors rated adequately at baseline compared to other professions. A potential explanation may lie in the fact that systematic monitoring of psychotic symptoms has not yet been fully implemented into clinical practice in the hospital system under study. This situation may lead psychologists and medical doctors—who have studied psychopathology during

TABLE 1 | Reliability of HAM-D17 items (Gwet's agreement coefficient) after participants' ratings of each patient video (randomized order).

	1st day of training			2nd day of training		3rd day of training	
	First (baseline)		Second	Third	Fourth	Fifth	Sixth
HAM1—Depressed mood	0.88 (0.83; 0.93)	Theoretical introduction	0.95 (0.93; 0.97)	0.93 (0.90; 0.95)	0.92 (0.88; 0.95)	0.93 (0.91; 0.96)	0.95 (0.92; 0.97)
HAM2—Guilt	0.84 (0.77; 0.91)		0.92 (0.88; 0.96)	0.91 (0.88; 0.94)	0.90 (0.86; 0.94)	0.94 (0.91; 0.97)	0.93 (0.90; 0.96)
HAM7—Work and activities	0.81 (0.76; 0.86)		0.86 (0.83; 0.89)	0.89 (0.87; 0.92)	0.81 (0.75; 0.87)	0.86 (0.81; 0.90)	0.87 (0.83; 0.91)
HAM8—Retardation	0.71 (0.65; 0.77)		0.78 (0.73; 0.83)	0.81 (0.76; 0.85)	0.82 (0.78; 0.87)	0.85 (0.82; 0.89)	0.85 (0.81; 0.89)
HAM10—Anxiety, psychic	0.87 (0.82; 0.91)		0.89 (0.85; 0.94)	0.86 (0.81; 0.91)	0.91 (0.88; 0.95)	0.92 (0.89; 0.95)	0.86 (0.82; 0.91)
HAM13—Somatic symptoms general	0.61 (0.51; 0.70)		0.54 (0.42; 0.67)	0.68 (0.59; 0.77)	0.65 (0.55; 0.75)	0.75 (0.69; 0.81)	0.74 (0.66; 0.82)
HAM4—Insomnia early	0.77 (0.69; 0.84)		0.95 (0.92; 0.99)	0.89 (0.83; 0.94)	0.84 (0.77; 0.91)	0.94 (0.90; 0.98)	0.93 (0.88; 0.97)
HAM5—Insomnia middle	0.88 (0.81; 0.94)		0.89 (0.83; 0.95)	0.90 (0.84; 0.95)	0.92 (0.87; 0.98)	0.89 (0.83; 0.96)	0.91 (0.85; 0.96)
HAM6—Insomnia late	0.97 (0.95; 1.00)		0.94 (0.88; 0.99)	0.92 (0.86; 0.99)	0.92 (0.86; 0.99)	0.87 (0.79; 0.95)	0.99 (0.96; 1.00)
HAM9—Agitation	0.91 (0.88; 0.93)		0.93 (0.91; 0.95)	0.94 (0.91; 0.97)	0.94 (0.91; 0.96)	0.89 (0.86; 0.93)	0.90 (0.87; 0.94)
HAM11—Anxiety, somatic	0.85 (0.80; 0.90)		0.80 (0.75; 0.85)	0.91 (0.88; 0.94)	0.89 (0.84; 0.93)	0.85 (0.80; 0.90)	0.85 (0.80; 0.90)
HAM12—Loss of appetite	0.76 (0.66; 0.87)		0.91 (0.84; 0.98)	0.90 (0.84; 0.96)	0.90 (0.84; 0.97)	0.97 (0.95; 1.00)	0.97 (0.94; 1.00)
HAM14—Sexual interest	0.47 (0.29; 0.64)		0.82 (0.72; 0.91)	0.79 (0.68; 0.90)	0.87 (0.79; 0.95)	0.77 (0.66; 0.89)	0.84 (0.75; 0.93)
HAM15—Hypochondriasis	0.96 (0.95; 0.98)		0.98 (0.97; 0.99)	0.98 (0.97; 0.99)	0.97 (0.95; 0.98)	0.98 (0.97; 0.99)	0.97 (0.95; 0.99)
HAM17—Loss of weight	0.89 (0.82; 0.97)		0.93 (0.87; 0.99)	0.93 (0.87; 0.98)	0.91 (0.84; 0.98)	0.90 (0.84; 0.97)	0.98 (0.95; 1.00)
HAM3—Suicidality	0.93 (0.90; 0.97)		0.92 (0.88; 0.97)	0.98 (0.96; 1.00)	0.97 (0.96; 0.99)	0.97 (0.95; 0.99)	0.97 (0.95; 1.00)
HAM16—Insight	0.80 (0.72; 0.89)		0.95 (0.92; 0.99)	0.94 (0.89; 0.98)	0.86 (0.79; 0.93)	0.99 (0.97; 1.00)	0.96 (0.92; 1.00)
HAM-D6—Subscale score	0.80 (0.74; 0.87)		0.89 (0.85; 0.92)	0.90 (0.87; 0.93)	0.91 (0.87; 0.94)	0.89 (0.86; 0.92)	0.90 (0.87; 0.93)
HAM-D17—Total score	0.84 (0.79; 0.88)		0.91 (0.88; 0.94)	0.94 (0.92; 0.96)	0.92 (0.89; 0.95)	0.91 (0.88; 0.94)	0.89 (0.85; 0.93)

Note: The item order follows the ABC version of the Hamilton Depression Rating Scale developed by Per Bech [26].

TABLE 2 | The proportion of participant ratings with ≤1 point deviation from the gold standard on individual HAMD-17 item scores and ≤6 point deviation from the gold standard on the HAM-D17.

	Psychologists and medical doctors[a]	Nurses	Other	All participants
Rating of video	***n* = 25**	***n* = 62**	***n* = 23[b]**	***n* = 110**
First (baseline)	48%	45%	30%	43%
Theoretical introduction video				
Second	52%	62%	35%	54%
Third	68%	65%	61%	65%*
Fourth	68%	54%	52%	57%
Fifth	52%	82%	52%	69%**
Sixth	83%	74%	48%	70%**

[a]Psychologist, psychiatrists, and medical doctors.
[b]Health care workers, occupational therapist, and physiotherapist.
*McNemar test, $p < 0.01$.
**McNemar test, $p < 0.001$.

their education—to administer the most accurate ratings of psychotic symptoms. In contrast, nurses—who conduct HAM-D17 ratings with a similar frequency as psychologists and medical doctors—are well acquainted with the psychopathology of depression. The only participant characteristic associated with accurate HAM-D17 ratings at baseline was employment within an outpatient unit. A potential explanation for this finding is that outpatient staff may have more experience with longitudinal assessments and that outpatients present a broader range of symptoms, enabling staff to develop finer calibration and sensitivity to different symptom levels. Moreover, outpatient staff are likely more accustomed to basing their ratings on patient interviews—which is very similar to the video-based training we test (i.e., ratings of patient interviews)—whereas inpatient staff may rely more on observable symptoms.

The reliability of the HAM-D17 ratings improved considerably after the theoretical introduction. Hence, providing staff with basic information on the theoretical principles of assessing the psychopathology in question according to the severity anchors should be a cornerstone in standardizing HAM-D17 ratings within clinical practice. It is worth noting that the item that initially showed the lowest reliability at baseline (item 14: Sexual Interest) demonstrated a substantial increase from 0.47 [95% CI: 0.29; 0.64] to 0.82 [95% CI: 0.72; 0.91] after the theoretical introduction video. The initial low reliability for this item has also been found in other studies [19], one plausible explanation may be that—given its sensitive nature—it is often skipped during training and supervision out of consideration for the patient. However, this practice does not only compromise the reliability of HAM-D17 ratings but also perpetuates the stigma surrounding the evaluation of an important symptom of depression and a common side effect of antidepressant treatment, negatively impacting quality of life [34]. Therefore, we strongly recommend that clinicians prioritize training in the assessment of reduced sexual interest, just as they do for other symptoms of sensitive nature (e.g., suicidal ideation).

The second lowest reliability at baseline was found for item 13 (Somatic Symptoms). This has also been observed elsewhere [24, 28] and may be attributed to differing understanding of what psychopathology this item covers (i.e., tiredness, loss of physical energy, fatigue, and muscular aches and pains). Thus, other somatic symptoms should be rated elsewhere [35, 36]. Likewise, the principle of not rating symptoms due to other causes clearly unrelated to depression is particularly relevant for this item. Specifically, the presence of, for example, acute infections, muscular ache/fatigue following physical activity/sports, and other clearly unrelated reasons for tiredness and muscular aches should not be rated here [35, 36].

While scoring reliability improved significantly for some items (e.g., item 4—Early Insomnia and item 16—Loss of Insight) after the theoretical introduction, most items appeared to require further training. Importantly, in addition to item 13 (Somatic Symptoms), another HAM-D6 subscale item (item 10—Psychomotor Retardation) initially displayed low reliability, in line with other studies [28, 37]. Although it did improve following the theoretical introduction, from 0.71 [95% CI: 0.65; 0.77] to 0.78 [95% CI: 0.73; 0.83], the highest reliability was achieved by the end of the training program with an AC1 of 0.85 [95% CI: 0.81; 0.89]. This pattern may be attributed to the observational nature of the assessment of this item. Hence, the display of a range of symptom severities is likely required to calibrate the assessment of severity by staff members. A similar tendency was seen for the reliability of the other HAM-D17 item that is rated based on observation (item 9—Psychomotor Agitation), for which other studies have also reported poor reliability [19, 28, 37]. The rating of such symptoms can easily be trained in clinical practice since they are based solely on observation and do not require an interview.

Lastly, staff satisfaction with the video-based training was high, and the subjective rating skills showed a statistically significant increase from baseline to endpoint. Notably, about one-third requested that the video-based training be

TABLE 3 | Univariable logistic regression analysis.

	Univariate	
	OR (CI)	***p***
Sex (ref: male)		
Female	0.99 (0.38; 2.60)	0.989
Age (ref: < 41 years)		
≥ 41	1.01 (0.47; 2.15)	0.978
Education (ref: other health workers)		
Psychologists & doctors	2.11 (0.65; 6.90)	0.217
Nurses	1.88 (0.68; 5.22)	0.224
Department (ref: Regional Psychiatric Hospital West)		
Aarhus University Hospital Psychiatry	1.00 (0.40; 2.52)	1.00
Regional Psychiatric Hospital Centrum + Randers	2.18 (0.79; 6.02)	0.133
Place of employment (ref: outpatient unit)		
Inpatient unit	0.44 (0.20; 0.97)	0.042
Years of clinical experience (ref: 0–1 years)		
2–5 years	1.36 (0.47; 3.96)	0.574
> 5 years	2.18 (0.86; 5.55)	0.100
Number of HAM-D17 conducted in clinical practice (ref: = 0)		
1–10 ratings	0.47 (0.15; 1.50)	0.203
> 10 ratings	0.99 (0.37; 2.69)	0.990
Interest in psychopathology of depression (ref: < 8)		
≥ 8	0.90 (0.40; 2.01)	0.800
Attitude toward video training (ref: prefers face-to-face)		
Video based training is a good supplement	1.37 (0.52; 3.62)	0.530
Self-perceived competence (ref: 1–3)		
4–6	1.11 (0.44; 2.79)	0.826
> 6	1.26 (0.49; 3.26)	0.629

Note: Association between participant characteristics and providing accurate Hamilton ratings (≤ 1 point deviation from the gold standard on individual HAMD-17 item scores and ≤ 6 point deviation from the gold standard on the HAM-D17 total score) at baseline.

supplemented by live instructions and discussion, which could potentially have further increased the beneficial effects of training. Moreover, the use of real patients enhances ecological validity and improves the generalizability of findings, as it exposes raters to the complex and heterogeneous symptom presentations of depression that are often difficult to replicate using actors. This approach more closely reflects the complexity encountered in real-world clinical assessments and may better prepare raters for routine practice. While only six patient interviews were used in the training program, this number was sufficient to produce a measurable and statistically significant improvement in scoring accuracy across sessions. These improvements suggest that even a brief but structured exposure to diverse clinical presentations can yield meaningful learning gains. However, the optimal number of training videos may depend on individual learning trajectories. Ideally, a future adaptive training program could tailor the number and type of videos to each participant's needs, ensuring progressive difficulty while maintaining broad coverage of symptom types and severities. This video-based training lends itself well to replication in other settings and professional groups. We encourage future studies to explore the feasibility and effectiveness of such standardized training across a broader range of rating scales and clinical populations.

The following limitations should be considered when interpreting the results of this study. First, it would be ideal to test the training program in a randomized controlled design to isolate the effects of the intervention from potential learning effects (e.g., repeated exposure to rating tasks without theoretical instruction or gold standard rating videos). Additionally, participants were assigned to groups based on logistical feasibility

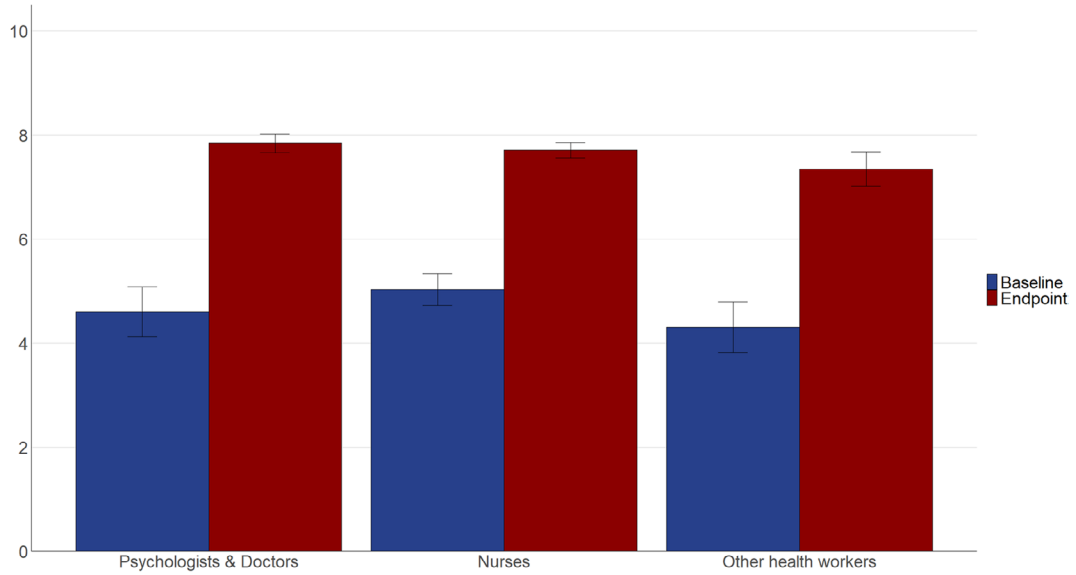

FIGURE 2 | Subjective skills in conducting Hamilton rating before and after participation in the video-based training rated on a scale ranging from 1 (not at all competent) to 10 (maximally competent). Error bars represent standard errors. All *p*-values for baseline–endpoint subgroup comparisons were <0.001.

rather than randomization, which may have resulted in an uneven distribution of professional backgrounds and other demographic and clinical characteristics across groups. To minimize potential confounding due to case difficulty, the order of patient interview videos was randomized across groups. This design choice aimed to distribute easier and more challenging cases evenly, thereby improving the assessment of the training effect, at the cost of direct within-subject comparisons.

Second, while the study focused exclusively on evaluating HAM-D17 rating skills, the raters' skills in conducting semi-structured interviews and effectively probing for necessary information are equally important. This point is particularly relevant given the suboptimal quality of interviews in clinical trials [38].

Third, the extent to which the effect of training persists over time should be assessed to evaluate whether raters drift in their assessments [21]. Applying the consistency checks (i.e., flags indicating consistency/inconsistency ratings) proposed by Rabinowitz et al. [39] as part of a systematic quality monitoring system may guide the required frequency and focus points for re-training.

Fourth, the participants in this study were rather heterogeneous, with approximately one-fifth belonging to professions whose primary tasks do not typically involve HAM-D17 rating (e.g., healthcare workers, physiotherapists, occupational therapists, and social workers). However, we included these participants to establish a common language for communication regarding a patient population that is supported in different capacities by each profession.

Fifth, it may be argued that the cut-off point for rating accuracy in this study was too permissive, especially for items rated on a 3-point scale; this criterion was augmented by a cut-off on the total score to mitigate potential systematic bias in assessing symptom severity.

Sixth, inter-rater reliability was assessed using Gwet's agreement coefficient (AC1), which, although less commonly used than Cohen's kappa (κ) or the Intraclass Correlation Coefficient (ICC), offers methodological advantages in certain contexts. Specifically, Gwet's AC1 is less affected by imbalanced score distributions—a frequent issue in clinical rating scales such as the HAM-D17, where the prevalence of certain symptom ratings may be high or low. In such cases, traditional measures like κ or ICC can yield paradoxically low values even when agreement is substantial [40–42]. By contrast, Gwet's AC1 provides a more stable and robust estimate of agreement under these conditions, making it well-suited for our study design involving multiple raters and real patients.

Seventh, the training was conducted over three consecutive days within a single week, during which participants viewed two patient videos and two expert-rated videos per day, in addition to a theoretical introduction video on the first day. The cumulative cognitive load and repeated exposure to similar tasks may have introduced fatigue effects, potentially affecting attention and rating accuracy—especially toward the end of each day.

Eighth, while the video-based modality proved effective in enhancing inter-rater reliability, accuracy of ratings, and subjective skills in conducting HAM-D17 assessments, one potential drawback of this approach is the absence of opportunities to ask an instructor questions. Conversely, providing theoretical introductions and gold standard videos instead of relying on discussions ensured that all participants were equipped with the same fundamental rating principles and understood the rationale behind the gold standard scores. Additionally, it is worth noting that the video-based approach facilitates the training of a large number of staff members without being limited by the availability of expert raters or patients willing to be interviewed. In fact, based on the results of this study, completing the video-based training in HAM-D17 has become mandatory for all newly employed staff treating depression in the Central Denmark Region.

Our findings show that education and professional experience alone do not guarantee proficiency in applying rating scales like the HAM-D17 and underscore the need for ongoing training and periodic skill assessments as part of quality assurance in mental health services. Video-based training proved effective in enhancing both subjective and objective rating skills. Rater training should be complemented by supervision in interview techniques to ensure that relevant information for rating is obtained.

Author Contributions

P.K. and S.D.Ø. designed the study. The theoretical introduction to HAM-D17 rating was prepared by P.K., E.R.L., and S.D.Ø. P.K. and B.D.J. undertook the statistical analyses. All authors contributed to the interpretation of the results. P.K. drafted the first version of the manuscript, which was subsequently revised critically for important intellectual content by the remaining authors. All authors approved the final version of the manuscript prior to submission.

Acknowledgments

The authors are grateful to the participating patients and staff members and to Cecilie Wolf from the Psychiatric Services of the Central Denmark Region, who recorded and edited all video material. S.D.Ø. reports funding from Independent Research Fund Denmark (grant numbers: 7016-00048B and 2096-00055A), the Lundbeck Foundation (grant numbers: R358-2020-2341 and R344-2020-1073), the Danish Cancer Society (grant number: R283-A16461), the Central Denmark Region Fund for Strengthening of Health Science (grant number: 1-36-72-4-20), and the Danish Agency for Digitization Investment Fund for New Technologies (grant number: 2020-6720). These funders played no role in the design or conduct of the study; collection, management, analysis, or interpretation of the data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

Ethics Statement

All participating staff members provided informed consent. The patients contributing to the video interviews consented to the recordings being used for education and research. The data were collected, processed, and stored in accordance with the European Union General Data Protection Regulation.

Conflicts of Interest

S.D.Ø. received the 2020 Lundbeck Foundation Young Investigator Prize. S.D.Ø. owns/has owned units of mutual funds with stock tickers DKIGI, IAIMWC, SPIC25KL, and WEKAFKI, and owns/has owned units of exchange traded funds with stock tickers BATE, TRET, QDV5, QDVH, QDVE, SADM, IQQH, USPY, EXH2, 2B76, IS4S, OM3X, and EUNL. The other authors declare no conflicts of interest.

Data Availability Statement

Participants did not provide consent for sharing of their individual-level data beyond the scope of the original research objective. Therefore, the data from this study cannot be shared.

References

1. P. Bech, *Clinical Psychometrics* (Wiley, 2012), https://doi.org/10.1002/9781118511800.

2. C. U. Correll, T. Kishimoto, J. Nielsen, and J. M. Kane, "Quantifying Clinical Relevance in the Treatment of Schizophrenia," *Clinical Therapeutics* 33, no. 12 (2011): B16–B39, https://doi.org/10.1016/j.clinthera.2011.11.016.

3. M. H. Trivedi, A. J. Rush, M. L. Crismon, et al., "Clinical Results for Patients With Major Depressive Disorder in the Texas Medication Algorithm Project," *Archives of General Psychiatry* 61, no. 7 (2004): 669–680, https://doi.org/10.1001/archpsyc.61.7.669.

4. T. Guo, Y. T. Xiang, L. Xiao, et al., "Measurement-Based Care Versus Standard Care for Major Depression: A Randomized Controlled Trial With Blind Raters," *American Journal of Psychiatry* 172, no. 10 (2015): 1004–1013, https://doi.org/10.1176/appi.ajp.2015.14050652.

5. M. L. Crismon, M. Trivedi, T. A. Pigott, et al., "The Texas Medication Algorithm Project: Report of the Texas Consensus Conference Panel on Medication Treatment of Major Depressive Disorder," *Journal of Clinical Psychiatry* 60, no. 3 (1999): 142–156.

6. R. Ricken, K. Wiethoff, T. Reinhold, et al., "Algorithm-Guided Treatment of Depression Reduces Treatment Costs—Results From the Randomized Controlled German Algorithm Project (GAPII)," *Journal of Affective Disorders* 134, no. 1–3 (2011): 249–256, https://doi.org/10.1016/j.jad.2011.05.053.

7. M. H. Trivedi, A. J. Rush, S. R. Wisniewski, et al., "Evaluation of Outcomes With Citalopram for Depression Using Measurement-Based Care in STAR*D: Implications for Clinical Practice," *American Journal of Psychiatry* 163, no. 1 (2006): 28–40, https://doi.org/10.1176/appi.ajp.163.1.28.

8. R. W. Licht, S. Qvitzau, P. Allerup, and P. Bech, "Validation of the Bech-Rafaelsen Melancholia Scale and the Hamilton Depression Scale in Patients With Major Depression; Is the Total Score a Valid Measure of Illness Severity?," *Acta Psychiatrica Scandinavica* 111, no. 2 (2005): 144–149, https://doi.org/10.1111/j.1600-0447.2004.00440.x.

9. P. Bech, L. F. Gram, E. Dein, O. Jacobsen, J. Vitger, and T. G. Bolwig, "Quantitative Rating of Depressive States," *Acta Psychiatrica Scandinavica* 51, no. 3 (1975): 161–170, https://doi.org/10.1111/j.1600-0447.1975.tb00002.x.

10. F. Hieronymus, J. F. Emilsson, S. Nilsson, and E. Eriksson, "Consistent Superiority of Selective Serotonin Reuptake Inhibitors Over Placebo in Reducing Depressed Mood in Patients With Major Depression," *Molecular Psychiatry* 21, no. 4 (2016): 523–530, https://doi.org/10.1038/mp.2015.53.

11. F. Hieronymus, A. Lisinski, S. Nilsson, and E. Eriksson, "Influence of Baseline Severity on the Effects of SSRIs in Depression: An Item-Based, Patient-Level Post-Hoc Analysis," *Lancet Psychiatry* 6, no. 9 (2019): 745–752, https://doi.org/10.1016/s2215-0366(19)30216-0.

12. S. D. Østergaard, P. Bech, and K. W. Miskowiak, "Fewer Study Participants Needed to Demonstrate Superior Antidepressant Efficacy When Using the Hamilton Melancholia Subscale (HAM-D_6) as Outcome Measure," *Journal of Affective Disorders* 190 (2016): 842–845, https://doi.org/10.1016/j.jad.2014.10.047.

13. P. Bech, E. Paykel, L. Sireling, and J. Yiend, "Rating Scales in General Practice Depression: Psychometric Analyses of the Clinical Interview for Depression and the Hamilton Rating Scale," *Journal of Affective Disorders* 171 (2015): 68–73, https://doi.org/10.1016/j.jad.2014.09.013.

14. F. Hieronymus, A. Lisinski, E. Eriksson, and S. D. Østergaard, "Do Side Effects of Antidepressants Impact Efficacy Estimates Based on the Hamilton Depression Rating Scale? A Pooled Patient-Level Analysis," *Translational Psychiatry* 11, no. 27 (2021): 249, https://doi.org/10.1038/s41398-021-01364-0.

15. P. Bech, "Is the Antidepressive Effect of Second-Generation Antidepressants a Myth?," *Psychological Medicine* 40, no. 2 (2010): 181–186, https://doi.org/10.1017/s0033291709006102.

16. S. D. Østergaard, "Do Not Blame the SSRIs: Blame the Hamilton Depression Rating Scale," *Acta Neuropsychiatrica* 30, no. 5 (2018): 241–243, https://doi.org/10.1017/neu.2017.6.

17. K. A. Kobak, J. D. Lipsitz, and A. Feiger, "Development of a Standardized Training Program for the Hamilton Depression Scale Using

Internet-Based Technologies: Results From a Pilot Study," *Journal of Psychiatric Research* 37, no. 6 (2003): 509–515, https://doi.org/10.1016/s0022-3956(03)00056-6.

18. K. A. Kobak, M. G. Opler, and N. Engelhardt, "PANSS Rater Training Using Internet and Videoconference: Results From a Pilot Study," *Schizophrenia Research* 92, no. 1–3 (2007): 63–67, https://doi.org/10.1016/j.schres.2007.01.011.

19. M. J. Müller and A. Dragicevic, "Standardized Rater Training for the Hamilton Depression Rating Scale (HAMD-17) in Psychiatric Novices," *Journal of Affective Disorders* 77, no. 1 (2003): 65–69, https://doi.org/10.1016/s0165-0327(02)00097-6.

20. P. Kølbæk, D. Dines, J. Hansen, et al., "Standardized Training in the Rating of the Six-Item Positive and Negative Syndrome Scale (PANSS-6)," 2020.

21. B. H. Mulsant, K. B. Kastango, J. Rosen, R. A. Stone, S. Mazumdar, and B. G. Pollock, "Interrater Reliability in Clinical Trials of Depressive Disorders," *American Journal of Psychiatry* 159, no. 9 (2002): 1598–1600, https://doi.org/10.1176/appi.ajp.159.9.1598.

22. J. Rosen, B. H. Mulsant, P. Marino, C. Groening, R. C. Young, and D. Fox, "Web-Based Training and Interrater Reliability Testing for Scoring the Hamilton Depression Rating Scale," *Psychiatry Research* 161, no. 1 (2008): 126–130, https://doi.org/10.1016/j.psychres.2008.03.001.

23. K. A. Kobak, J. D. Lipsitz, J. B. Williams, N. Engelhardt, and K. M. Bellew, "A New Approach to Rater Training and Certification in a Multicenter Clinical Trial," *Journal of Clinical Psychopharmacology* 25, no. 5 (2005): 407–412, https://doi.org/10.1097/01.jcp.0000177666.35016.a0.

24. S. Wagner, I. Helmreich, K. Lieb, and A. Tadić, "Standardized Rater Training for the Hamilton Depression Rating Scale ($HAMD_{17}$) and the Inventory of Depressive Symptoms (IDS(C30))," *Psychopathology* 44, no. 1 (2011): 68–70, https://doi.org/10.1159/000318162.

25. P. Kølbæk, D. Dines, J. Hansen, et al., "Standardized Training in the Rating of the Six-Item Positive and Negative Syndrome Scale (PANSS-6)," *Schizophrenia Research* 228 (2021): 438–446, https://doi.org/10.1016/j.schres.2020.12.044.

26. P. Bech, "The ABC Profile of the HAM-D17," *Revista Brasileira de Psiquiatria (São Paulo, Brazil: 1999)* 33 (2011): 109–110, https://doi.org/10.1590/S1516-44462011000200001.

27. P. Bech, M. Kastrup, and O. J. Rafaelsen, "Mini-Compendium of Rating Scales for States of Anxiety Depression Mania Schizophrenia With Corresponding DSM-III Syndromes," *Acta Psychiatrica Scandinavica. Supplementum* 326 (1986): 1–37.

28. J. B. Williams, "A Structured Interview Guide for the Hamilton Depression Rating Scale," *Archives of General Psychiatry* 45, no. 8 (1988): 742–747, https://doi.org/10.1001/archpsyc.1988.01800320058007.

29. P. Kølbæk, C. W. Nielsen, C. W. Buus, et al., "Clinical Validation of the Self-Reported 6-Item Hamilton Depression Rating Scale (HAM-D6-SR) Among Inpatients," *Journal of Affective Disorders* 354 (2024): 765–772, https://doi.org/10.1016/j.jad.2024.03.014.

30. K. L. Gwet, "Computing Inter-Rater Reliability and Its Variance in the Presence of High Agreement," *British Journal of Mathematical and Statistical Psychology* 61, no. Pt 1 (2008): 29–48, https://doi.org/10.1348/000711006X126600.

31. M. Sajatovic, R. Gaur, C. Tatsuoka, et al., "Rater Training for a Multi-Site, International Clinical Trial: What Mood Symptoms May Be Most Difficult to Rate?," *Psychopharmacology Bulletin* 44, no. 3 (2011): 5–14.

32. M. P. Hengartner, J. C. Jakobsen, A. Sørensen, and M. Plöderl, "Efficacy of New-Generation Antidepressants Assessed With the Montgomery-Asberg Depression Rating Scale, the Gold Standard Clinician Rating Scale: A Meta-Analysis of Randomised Placebo-Controlled Trials," *PLoS One* 15, no. 2 (2020): e0229381, https://doi.org/10.1371/journal.pone.0229381.

33. P. Videbech and A. Deleuran, "The Danish Depression Database," *Clinical Epidemiology* 8 (2016): 475–478, https://doi.org/10.2147/clep.S100298.

34. E. Atlantis and T. Sullivan, "Bidirectional Association Between Depression and Sexual Dysfunction: A Systematic Review and Meta-Analysis," *Journal of Sexual Medicine* 9, no. 6 (2012): 1497–1507, https://doi.org/10.1111/j.1743-6109.2012.02709.x.

35. J. B. Williams, K. A. Kobak, P. Bech, et al., "The GRID-HAMD: Standardization of the Hamilton Depression Rating Scale," *International Clinical Psychopharmacology* 23, no. 3 (2008): 120–129, https://doi.org/10.1097/YIC.0b013e3282f948f5.

36. K. J. Rohan, J. N. Rough, M. Evans, et al., "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," *Journal of Affective Disorders* 200 (2016): 111–118, https://doi.org/10.1016/j.jad.2016.01.051.

37. L. P. Rehm and M. W. O'Hara, "Item Characteristics of the Hamilton Rating Scale for Depression," *Journal of Psychiatric Research* 19, no. 1 (1985): 31–41, https://doi.org/10.1016/0022-3956(85)90066-4.

38. N. Engelhardt, A. D. Feiger, K. O. Cogger, et al., "Rating the Raters: Assessing the Quality of Hamilton Rating Scale for Depression Clinical Interviews in Two Industry-Sponsored Clinical Drug Trials," *Journal of Clinical Psychopharmacology* 26, no. 1 (2006): 71–74, https://doi.org/10.1097/01.jcp.0000194621.61868.7c.

39. J. Rabinowitz, N. R. Schooler, A. Anderson, et al., "Consistency Checks to Improve Measurement With the Positive and Negative Syndrome Scale (PANSS)," *Schizophrenia Research* 190 (2017): 74–76, https://doi.org/10.1016/j.schres.2017.03.017.

40. H. C. de Vet, L. B. Mokkink, C. B. Terwee, O. S. Hoekstra, and D. L. Knol, "Clinicians Are Right Not to Like Cohen's κ," *BMJ (Clinical Research Ed.)* 346 (2013): f2125, https://doi.org/10.1136/bmj.f2125.

41. S. Mehta, R. F. Bastero-Caballero, Y. Sun, et al., "Performance of Intraclass Correlation Coefficient (ICC) as a Reliability Index Under Various Distributions in Scale Reliability Studies," *Statistics in Medicine* 37, no. 18 (2018): 2734–2752, https://doi.org/10.1002/sim.7679.

42. N. Wongpakaran, T. Wongpakaran, D. Wedding, and K. L. Gwet, "A Comparison of Cohen's Kappa and Gwet's AC1 When Calculating Inter-Rater Reliability Coefficients: A Study Conducted With Personality Disorder Samples," *BMC Medical Research Methodology* 13 (2013): 61, https://doi.org/10.1186/1471-2288-13-61.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Table S1:** Written feedback in the post-training questionnaire categorized according to subject.