Published in final edited form as:

Psychiatry Res. 2008 October 30; 161(1): 126–130. doi:10.1016/j.psychres.2008.03.001.

Web-based training and interrater reliability testing for scoring the Hamilton Depression Rating Scale

Jules Rosen^{a,d}, Benoit H. Mulsant^{a,b}, Patricia Marino^c, Christopher Groening^d, Robert C. Young^c, and Debra Fox^e

^aDepartment of Psychiatry, University of Pittsburgh School of Medicine, PA

^bGeriatric Mental Health Program, Centre for Addiction and Mental Health, and University of Toronto

^cDepartment of Psychiatry, Weill Medical College of Cornell University

^dKatz Graduate School of Business, University of Pittsburgh

^eFox Learning Systems, Inc; Pittsburgh, PA

Abstract

Despite the importance of establishing shared scoring conventions and assessing interrater reliability in clinical trials in psychiatry, these elements are often overlooked. Obstacles to rater training and reliability testing include logistic difficulties in providing live training sessions, or mailing videotapes of patients to multiple sites and collecting the data for analysis. To address some of these obstacles, a Web-based interactive video system was developed. It uses actors of diverse ages, gender and race to train raters how to score the Hamilton Depression Rating Scale and to assess interrater reliability. This system was tested with a group of experienced and novice raters within a single site. It was subsequently used to train raters of a federally-funded multi-center clinical trial on scoring conventions and to test their interrater reliability. The advantages and limitations of using interactive video technology to improve the quality of clinical trials are discussed.

1. Introduction

1.1. Background

In clinical trials, the reliability of the data collected ultimately determines the validity of the studies' conclusions (Kobak et al., 1996). In psychiatry, the primary outcome measures often depend on interviewers' skills in eliciting information, as well as their interpretations of the subjects' responses (Kobak et al., 2005a). When multiple raters are used in a clinical trial, differences between raters in terms of interviewing technique and scoring criteria introduce variability that can distort the outcome measures (Muller and Szegedi, 2002; Bourin et al., 2004). Despite the importance of statistically establishing raters' reliability, review of the literature suggests that this issue is often ignored in clinical trials, including those of depression treatment (Mulsant et al., 2002). This is especially problematic in multi-center trials in which groups of raters are geographically dispersed, may change over time, and that may recruit patients over several years.

Corresponding Author: Jules Rosen MD, Professor of Psychiatry, University of Pittsburgh School of Medicine, 3811 O'Hara St., Pittsburgh, PA 15213, 412 246 5900; 412 586 9300 (fax), rosenji@upmc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

We have previously reported that videotapes of professional actors using scripted interviews of the Hamilton Depression Rating Scale (HDRS) could not be distinguished from the videotapes of the actual patients when scored by experienced raters (Rosen et al., 2004). Building on the findings of that study, we developed a Web-based system using professional actors both to train raters on scoring the HDRS using shared scoring conventions and to assess interrater reliability. This report describes: 1) the development of the system, 2) a study of HDRS scoring-tutorial and reliability testing with both naïve and experienced raters, and 3) results of a field test of this system in a multi-site NIMH-funded study.

1.2. Development and Description of Web-based System

The web-based system consists of three components: 1) a scoring tutorial program, 2) a reliability testing program, and 3) an administrative program. In order to use this system, users must have a high speed internet connection with “Flash” plug-in for the internet browser. To develop the scoring tutorial and reliability testing programs, informed consent was obtained to video-record 21 HDRS interviews of seven patients participating in an NIMH-funded study of depression at initiation of treatment, in mid-treatment, and in partial or full remission. The semi-structured interview utilized for this project is based on the published interview by Williams et al. (Williams, 1988), and has been previously used in depression trials in the U.S. (Mulsant et al., 1999; Tew, Jr. et al., 1999;; Sackeim et al., 2000; Sackeim et al., 2001; Gildengers et al., 2005; Feske et al., 2004; Reynolds, III et al., 2006; Dombrowski et al., 2006). The scores of these 21 interviews ranged from below 10 (absence of depression), 11–20 (mild to moderate depression), 21–29 (severe depression), and greater than 30 (very severe depression, including psychosis) as each patient was followed through the course of his or her treatment. The videotaped interviews of the patients were transcribed yielding 21 scripts, which were modified to remove all information that might identify the actual patients. In order to create realistic portrayals of different stages of depression in diverse populations, three male and three female actors were recruited to portray young, mid-life, and elderly adults. One of the male and one of the female actors were African-American. Each actor recorded 9 or 10 scripts that were slightly modified to be age and gender appropriate for the actor (for instance a reference to a child may be changed to a reference to a grandchild).

Ten of the scripts were used to create the tutorial program designed to train raters on scoring conventions. The scoring tutorial program provides video vignettes for every possible score of 28 HDRS items. For item-scores not represented by actual interviews, the scripts were modified by changing either the intensity or frequency to move the score into a more or into a less severe rating. In the tutorial mode, trainees have the option of watching every vignette for each question in the order of increasing severity. Alternatively, they can watch them in random order. While the rater is observing the interview in the tutorial mode, the scoring guidelines are presented in text format in a box below the video for reference. In the tutorial mode, raters assign scores, and the system informs them when their scores differ from the scores assigned by two expert psychiatrist /raters. These raters (JR and BMH) have more than 20 years of cumulative experience administering and scoring the HDRS.

Following completion of the tutorial, the raters are directed by the system to the reliability testing program. The testing program was created with the 11 scripts that were not used in the tutorial program. To test interrater reliability, raters are presented with 6 of the HDRS interviews representing a full range of severities of depression. As in the tutorial mode, while raters watch the interview, the scoring guideline corresponding to the item that is being probed by the interviewer is presented in text format below the video stream. After raters select a score for a particular item, the system progresses to the next questions. Raters have the opportunity to go back and review any question and their score until they have scored all the items and “lock-in” their scores at the end of the testing session. Once raters complete a particular

interview and lock their scores, these scores are stored in a database and are available to calculate interrater reliability. All raters who are associated with a given study complete the reliability testing mode with the same 6 interviews. Repeat testing to assess rater drift over time can be accomplished with an alternate set of interviews.

The system is designed to provide scoring tutorials and reliability testing using the 17-, 24- or 28- item versions of the HDRS. The scoring conventions used for the first 17 items are based on the published conventions of the 17-item “Grid-Hamilton” that provides a single score for each item based on both the intensity and frequency of depressive symptoms (Kalali et al., 2002). The scoring convention used for items 18–28 were adapted by two of the authors (JR and BHM) to be congruent with the Grid-Hamilton scoring conventions.

The administrative program is designed to perform several functions. The overall administrator of a clinical trial can identify the sites participating in the study and designate for each site a site-coordinator. The overall administrator also specifies the version of the HDRS to be used for training and reliability testing (i.e., 17-, 24-, or 28-item version). In multi-site studies, the site coordinators enter the names and ID numbers of raters at each research site. Sites and raters can be added or removed during the course of a clinical trial. A database stores the test scores of each rater. Intra-class correlations (ICC) coefficients are calculated for raters participating in a particular study or by site according to the formula described by Shrout and Fleiss (Shrout and Fleiss, 1979) who described calculations based on one of 3 main cases depending on the assignment of judges. Our study follows Case 3, where “each target is rated by each of the same k judges, who are the only judges of interest” (p. 421).

2. Methods

2.1. Study 1: Initial Evaluation

Research raters were recruited from the research programs of the Department of Psychiatry at the University of Pittsburgh School of Medicine to conduct an initial evaluation of the web-based system prior to finalization of the system and actual field testing. All participants were research raters in one of the psychiatry research programs. All of them had received previous training on at least one rating instrument using classroom instructions and videotapes to establish reliability. However, some had not been trained to administer and score the HDRS. Regardless of their prior experience with the HDRS, they were required to complete the tutorial program prior to reliability testing. ICC's were calculated for the entire group and for three subgroups based on prior experience with the HDRS: 1) naive raters with no prior experience with the HDRS, 2) experienced raters who had administered the HDRS less than 150 times, and 3) highly experienced raters who had administered the HDRS 150 times or more.

Raters were also asked to keep track of, and to report the amount of time and the number of sessions needed to complete the tutorial program and the reliability testing program.

2.2. Study 2: Field Trial

To further evaluate the system, a field trial was completed. In this study, six sites involved in a NIMH-funded multi-site study of late-life mood disorders used this system to train raters on shared scoring conventions and to assess interrater reliability.

For both Study 1 and Study 2, the scoring-tutorial and the reliability testing were to be completed within a two week window. Within that time frame, raters were permitted flexibility in terms of how much time they spent with the system and the number of sessions needed to complete the tutorial and the testing. For Study 1, the 28-item version of the HDRS was used; for Study 2, the 17-item version of the HDRS was used.

3) Results

3.1) Study 1: Single site study

Of the 17 raters who participated in this study, 7 were naive, 3 experienced, and 7 experts. The mean age was 42.3 years (range: 22–60). One rater was male; one rater was an African American woman; the remaining raters were Caucasian women.

Based on self-reports, the tutorial was completed in a mean of 1.8 hours (range: 1 – 2.5 hrs.) in 2.5 sessions (range 1–4). The mean number of hours to complete the reliability testing was 3.3 (range 2.5 – 5) in 2.6 separate sessions (range: 1 – 4).

The ICC for the naive, experienced and expert subgroups were 0.94, 0.93, and 0.96, respectively. The ICC calculated for the entire group was 0.95.

3.2) Study 2: Multi-site study

Of the 13 participating raters, 10 were female. One woman was Asian and one Hispanic and the remaining were Caucasian. The mean age was 34.3 years (range: 23–58). All participants completed the tutorial before going on to the test mode. The ICC was 0.98 for this group, and no outliers were identified. There were no problems accessing the web-site, completing the interactive tutorial and testing, and recovery of ICC data.

3.3) Individual interviews / items

Each of the 11 testing interviews used in studies 1 and 2 were individually assessed. Table 1, below, describes characteristics of these interviews.

4.) Discussion

The interrater reliabilities were excellent for both Study 1 and Study 2. Establishing rater reliability in studies of depression treatment is critically important, however, most studies do not report on rater training or reliability measures (Mulsant et al., 2002). In typical industry-supported clinical trials, investigators meetings convene to provide instruction to raters and investigators on the proper use of the various instruments. However, rigorous assessments of rater reliability rarely occur at these meetings or at any following time. The practical importance of interrater reliability has been established as essential to reducing variability in multi-site trials (Small et al., 1996). In that report, inadequate rater training and the absence of a measure of interrater reliability was shown to skew the results.

The relatively high ICCs for all groups of raters participating in this study are consistent with interrater reliability described in several studies with the HDRS that used traditional videotapes. In a large multi-center study, the ICC for conventionally-trained raters on scoring the HDRS was 0.97 (Sackeim et al., 2001), and in a single center study with multiple raters, the ICC for conventionally trained raters was 0.95 (Feske et al., 2004).

Although not supported as a training tool in all settings (Sanchez et al., 1995), videotapes of patients have been used to train raters and establish reliability in some clinical trials (Andreasen et al., 1982; Muller et al., 1998; Muller and Wetzel, 1998; Muller and Dragicevic, 2003). Limitations to this technique include the logistical support needed to mail videotapes to all raters at multiple sites, the inability to interact with video-based training, and the additional burden of mailing paper scoring sheets to a data management center and entering the data.

The computer-based system provides interactive training that is continuously available through any computer that has high speed connectivity to the Internet. The integrated database provides ICC calculations and report generation without the additional work of mailing or faxing data

and entering the data by hand. New raters and new sites can be added over time and the ICC's can be calculated with the group. Finally, rater drift can be assessed.

It is important to note that the Web-based system described in the current manuscript addresses rater training only in regard to scoring conventions. The equally important component of training raters on clinical interview skills required for the administration of the HDRS was not addressed in this study. The importance and effectiveness of providing rater interview training for the HDRS with a Web-based instrument has been previously demonstrated (Kobak et al., 2003; Kobak et al., 2005a; Kobak et al., 2005b; Targum, 2006; Kobak et al., 2006; Jeglic et al., 2007). Additional limitations to this study include the relatively small sample size and the fact that all of the "naïve" raters were experienced clinicians or raters that used different assessment instruments. Finally, the use of videotapes may artificially inflate estimates of reliability by reducing the information variance that would result if each rater interviewed the patient independently (Spitzer and Williams, 1980).

In conclusion, the current study evaluated a web-based system of interactive scoring training and reliability testing in a group of raters using the HDRS as a single site and multiple site study. The ICC's calculated support the effectiveness of this system without the additional logistical burden involved in with the use of videotapes.

Acknowledgments

This work was sponsored in part by the National Institute of Health (MH061639, MH069430, MH062565, MH067028, MH068847, HS011976).

References

- Andreasen NC, McDonald-Scott P, Grove WM, Keller MB, Shapiro RW, Hirschfeld RM. Assessment of reliability in multicenter collaborative research with a videotape approach. *American Journal of Psychiatry* 1982;139(7):876–882. [PubMed: 7091404]
- Bourin M, Deplanque D, Zins-Ritter M. Mean Deviation of Inter-rater Scoring (MDIS): a simple tool for introducing conformity into groups of clinical investigators. *Int.Clin.Psychopharmacol* 2004;19(4): 209–213. [PubMed: 15201567]
- Dombrowski AY, Blakesley-Ball RE, Mulsant BH, Mazumdar S, Houck PR, Szanto K, Reynolds CF III. Speed of improvement in sleep disturbance and anxiety compared with core mood symptoms during acute treatment of depression in old age. *American Journal of Geriatric Psychiatry* 2006;14(6):550–554. [PubMed: 16731725]
- Feske U, Mulsant BH, Pilkonis PA, Soloff P, Dolata D, Sackeim HA, Haskett RF. Clinical outcome of ECT in patients with major depression and comorbid borderline personality disorder. *American Journal of Psychiatry* 2004;161(11):2073–2080. [PubMed: 15514409]
- Gildengers AG, Houck PR, Mulsant BH, Dew MA, Aizenstein HJ, Jones BL, Greenhouse J, Pollock BG, Reynolds CF III. Trajectories of treatment response in late-life depression: psychosocial and clinical correlates. *Journal of Clinical Psychopharmacology* 2005;25(4 Suppl 1):S8–S13. [PubMed: 16027561]
- Jeglic E, Kobak KA, Engelhardt N, Williams JB, Lipsitz JD, Salvucci D, Bryson H, Bellew K. A novel approach to rater training and certification in multinational trials. *Int.Clin.Psychopharmacol* 2007;22(4):187–191. [PubMed: 17519640]
- Kalali A, Williams JB, Koback KA, Lipsitz J, Engelhardt N, Evans K, Olin J, Pearson J, Rothman M, Bech P. The new GRID HAM-D: pilot testing and international field trials. *Int J Neuropsychopharmacol* 2002;5:S147–S148.
- Kobak KA, Engelhardt N, Lipsitz JD. Enriched rater training using Internet based technologies: a comparison to traditional rater training in a multi-site depression trial. *J.Psychiatr.Res* 2006;40(3): 192–199. [PubMed: 16197959]

- Kobak KA, Feiger AD, Lipsitz JD. Interview quality and signal detection in clinical trials. *American Journal of Psychiatry* 2005a;162(3):628. [PubMed: 15741493]
- Kobak KA, Greist JJ, Jefferson JW, Katzelnick DJ. Computer-administered clinical rating scales. A review. *Psychopharmacology* 1996;127(4):291–301. [PubMed: 8923563]
- 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.Psychiatr.Res* 2003;37(6):509–515. [PubMed: 14563382]
- Kobak KA, Lipsitz JD, Williams JB, Engelhardt N, Bellew KM. A new approach to rater training and certification in a multicenter clinical trial. *Journal of Clinical Psychopharmacology* 2005b;25(5):407–412. [PubMed: 16160614]
- Muller MJ, Dragicevic A. Standardized rater training for the Hamilton Depression Rating Scale (HAMD-17) in psychiatric novices. *J.Affect.Disord* 2003;77(1):65–69. [PubMed: 14550936]
- Muller MJ, Rossbach W, Dannigkeit P, Muller-Siecheneder F, Szegedi A, Wetzel H. Evaluation of standardized rater training for the Positive and Negative Syndrome Scale (PANSS). *Schizophr.Res* 1998;32(3):151–160. [PubMed: 9720120]
- Muller MJ, Szegedi A. Effects of interrater reliability of psychopathologic assessment on power and sample size calculations in clinical trials. *Journal of Clinical Psychopharmacology* 2002;22(3):318–325. [PubMed: 12006903]
- Muller MJ, Wetzel H. Improvement of inter-rater reliability of PANSS items and subscales by a standardized rater training. *Acta Psychiatrica Scandinavica* 1998;98(2):135–139. [PubMed: 9718240]
- Mulsant BH, Kastango KB, Rosen J, Stone RA, Mazumdar S, Pollock BG. Interrater reliability in clinical trials of depressive disorders. *American Journal of Psychiatry* 2002;159(9):1598–1600. [PubMed: 12202285]
- Mulsant BH, Pollock BG, Nebes RD, Miller MD, Little JT, Stack J, Houck PR, Bensasi S, Mazumdar S, Reynolds CF III. A double-blind randomized comparison of nortriptyline and paroxetine in the treatment of late-life depression: 6-week outcome. *Journal of Clinical Psychiatry* 1999;60:16–20. [PubMed: 10513853]
- Reynolds CF III, Dew MA, Pollock BG, Mulsant BH, Frank E, Miller MD, Houck PR, Mazumdar S, Butters MA, Stack JA, Schlernitzauer MA, Whyte EM, Gildengers A, Karp J, Lenze E, Szanto K, Bensasi S, Kupfer DJ. Maintenance treatment of major depression in old age. *N.Engl.J.Med* 2006;354(11):1130–1138. [PubMed: 16540613]
- Rosen J, Mulsant BH, Bruce ML, Mittal V, Fox D. Actors' portrayals of depression to test interrater reliability in clinical trials. *American Journal of Psychiatry* 2004;161(10):1909–1911. [PubMed: 15465990]
- Sackeim HA, Haskett RF, Mulsant BH, Thase ME, Mann JJ, Pettinati HM, Greenberg RM, Crowe RR, Cooper TB, Prudic J. Continuation pharmacotherapy in the prevention of relapse following electroconvulsive therapy: a randomized controlled trial. *Journal of the American Medical Association* 2001;285(10):1299–1307. [PubMed: 11255384]
- Sackeim HA, Prudic J, Devanand DP, Nobler MS, Lisanby SH, Peyser S, Fitzsimons L, Moody BJ, Clark J. A prospective, randomized, double-blind comparison of bilateral and right unilateral electroconvulsive therapy at different stimulus intensities. *Archives of General Psychiatry* 2000;57(5):425–434. [PubMed: 10807482]
- Sanchez LE, Adams PB, Uysal S, Hallin A, Campbell M, Small AM. A comparison of live and videotape ratings: clomipramine and haloperidol in autism. *Psychopharmacology Bulletin* 1995;31(2):371–378. [PubMed: 7491394]
- Shrout PE, Fleiss JL. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin* 1979;2:420–428. [PubMed: 18839484]
- Small GW, Schneider LS, Hamilton SH, Bystritsky A, Meyers BS, Nemeroff CB. Site variability in a multisite geriatric depression trial. *International Journal of Geriatric Psychiatry* 1996;11:1089–1095.
- Spitzer, RL.; Williams, JBW. *Classification in Psychiatry*. Baltimore: Williams & Wilkins; 1980.
- Targum SD. Evaluating rater competency for CNS clinical trials. *Journal of Clinical Psychopharmacology* 2006;26(3):308–310. [PubMed: 16702896]

- Tew JD Jr, Mulsant BH, Haskett RF, Prudic J, Thase ME, Crowe RR, Dolata D, Begley AE, Reynolds CF III, Sackeim HA. Acute efficacy of ECT in the treatment of major depression in the old-old. *American Journal of Psychiatry* 1999;156(12):1865–1870. [PubMed: 10588398]
- Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. *Archives of General Psychiatry* 1988;45:742–747. [PubMed: 3395203]

Table 1

Interview*	Interview										
	1	2	3	4	5	6	7	8	9	10	11
Min	18	4	7	0	24	3	28	13	0	32	10
Max	26	15	18	6	32	8	35	19	2	42	17
Median	21	13	15	4	28	6	31	16	0	38	13
Mean	21.56	12.24	14.53	3.59	28.18	5.65	31.23	16.62	0.62	37.23	12.92
Std. Dev.	2.23	2.73	2.58	1.87	2.40	1.66	2.31	1.61	0.77	3.47	2.29
1st Quartile	18	4	7	0	24	3	28	13	0	32	10
2nd Quartile	20	12	14	3	27	5	30	16	0	34	11
3rd Quartile	21	13	15	4	28	6	31	16	0	38	13
4th Quartile	22	14	16	5	30	7	33	18	1	38	14
End of 4th Quartile	26	15	18	6	32	8	35	19	2	42	17
ICC	0.994	0.982	0.984	0.961	0.991	0.975	0.995	0.992	0.735	0.984	0.986

* Interviews 1–6 were used in Study 1; Interviews 4 and 7–11 were used in study 2