# Face-to-face vs. Remote Assessment of Inter-rater Reliability on the HAMD

**Kenneth A. Kobak[1], Ph.D., Janet B.W. Williams[2,3], D.S.W., Nina Engelhardt[4], Ph.D.**

[1]MedAvante, Inc., [2]Columbia University, [3]NY State Psychiatric Institute, [4]Consultant, Indianapolis, IN

**BACKGROUND.** Inter-rater reliability (IRR) has been recognized as an important methodological factor that may contribute to failed trials (Muller & Szegedi, 2002). The sheer number of raters involved in large, multicenter trials and the lack of standardized scoring conventions presents a formidable challenge to train and calibrate raters across sites. The most common approach for evaluating IRR in clinical trials involves observation of videotapes. However, this approach artificially inflates reliability estimates by reducing the "information variance" that occurs when two raters independently interview the same patient (Spitzer & Williams, 1980). A more rigorous approach involves two raters independently evaluating the same patient, blind to each other's scores. Implementing this approach in multi-site trials is limited by the diverse location of raters. One way of accomplishing this is through the use of videoconferencing. In order to evaluate the potential impact of remote evaluation on IRR, we compared IRR obtained via videoconference to IRR obtained using face-to-face interviews. In addition, a "hybrid" model was examined where one of the interviews was done face-to-face and one was done remotely by videoconference.

**METHOD.** Four raters at three different locations (Madison, WI (2), New York City, and Indianapolis) participated in the study. Prior to assessing IRR on the HAMD, raters were trained using didactic and applied methods. Following training, assessment of IRR was done using all pair-wise combinations of raters. Each paired rater independently conducted an interview with the same patient, who was at a third, central location. Raters were blind to the other rater's score. Following completion of both interviews, raters compared scores and discussed discrepant items. If the total score difference was more than 2 points, then the rater pair would conduct further interviews until they were within two points. In order to compare ICC obtained by videoconference to ICC obtained by two face-to-face interviews, data from a second sample were analyzed. In this sample, two cohorts were used: IRR assessed using two independent face-to-face interviews (n=21); and IRR assessed using one face-to-face interview and one remote interview (n=21). All interviews utilized a modified version of the SIGH-D (Williams, 1988).

**RESULTS.** The ICCs were high for each of the three methods of assessment (Table 1). The 95% confidence intervals for all three methods indicate no significant difference between the modalities. In addition, there was no significant difference in the mean HAMD scores between raters using any of the three approaches, with the mean difference less than one point in all cases (Table 2). Agreement on an item level for the total cohort was high for most items.

**CONCLUSION.** Videoconferencing can be used as a tool to calibrate raters at diverse sites. There appears to be no significant difference in ICC using remote methods of calibration compared to traditional face-to-face methods. Hybrid models will allow calibration of raters using face-to-face interviews with remote raters interviewing the same patient with videoconferencing. With good applied training, high levels of calibration can be obtained, even using rigorous assessment methodology of independent interviews.

***Table 1. ICC by Assessment Methodology***

| | ICC | 95% C.I. | F value | P value |
|---|---|---|---|---|
| Face vs. Face (n=21) | .93 | .8401, .9711 | 27.9008 | .0000 |
| Face vs. Video (n=21) | .88 | .7372, .9503 | 16.0306 | .0000 |
| Video vs. Video (n=22) | .90 | .7703, .9553 | 18.2902 | .0001 |
| Total Sample (n=64) | .91 | .8528, .9428 | 20.6733 | .0000 |

***Table 2. Mean Differences by ICC Interview Method***

| GROUP | Mean (SD) |
|---|---|
| Two Face-to-Face (n=21) | |
| HAMD1 | 17.62 (6.61) |
| HAMD2 | 18.48 (6.32) |
| Difference | -.857 |
| t | -1.702 |
| p | .104 |
| Face-to-Face vs. Video (n=21) | |
| HAMD1 (face-to-face) | 18.38 (5.28) |
| HAMD2 (video) | 18.24 (6.00) |
| Difference | -.14 |
| t | .234 |
| p | .817 |
| Video vs. Video (n=22) | |
| HAMD1 (n=22) | 15.36 (5.67) |
| HAMD2 (n=22) | 15.72 (4.81) |
| Difference | -.36 |
| t | .704 |
| p | .489 |

***Table 3. Intra-class correlation on HAMD Items, Total Sample (N=64)***

| HAMD ITEM | ICC |
|---|---|
| 1. Mood | .81 |
| 2. Guilt | .42 |
| 3. Suicide | .93 |
| 4. Initial Insomnia | .84 |
| 5. Middle Insomnia | .90 |
| 6. Late Insomnia | .67 |
| 7. Work/Interest | .61 |
| 8. Retardation | .58 |
| 9. Agitation | .23 |
| 10. Psychic Anxiety | .60 |
| 11. Somatic Anxiety | .75 |
| 12. Appetite | .87 |
| 13. Energy (Somatic General) | .67 |
| 14. Sex | .84 |
| 15. Hypochondriasis | .65 |
| 16. Loss of Weight | .94 |
| 17. Insight | 0 |
| **Total Score** | **.91** |