

The Relationship Between Interview Quality and Scoring Discrepancy: Application of the Rater Applied Performance Scale

Popp, D¹, Detke, M^{1,2}, Williams, J.B.W.^{1,3}

¹ MedAvante, Inc., ² Indiana University School of Medicine, ³ College of Physicians and Surgeons, Columbia University

ABSTRACT

Introduction: Variable quality of interviews may be an important contributor to the high number of failed CNS clinical trials. Recent advances in the measurement of interviewer skill allow for an examination of the effects of interview quality on signal detection¹. Kobak and colleagues demonstrated that higher quality baseline interviews were more likely to distinguish an effective drug from placebo, suggesting that CNS clinical trials could benefit from monitoring of interview quality². Brown and colleagues demonstrated that interview quality improved over the number of reviews as part of an ongoing quality control program in two clinical trials³. The current study seeks to identify which areas of interview quality are most closely related to scoring discrepancies between site raters and calibrated, blinded quality reviewers. While scoring accuracy (i.e. valid, reliable scores) cannot be determined without unblinded study results, it is hypothesized that scoring agreement should be high when interview quality is assessed as high. It is noted that information variance is reduced, and thus inter-rater reliability increased because the rater dyads are scoring one interview, the site rater's interview.

Methods: Data were pooled from two ongoing clinical trials of Major Depressive Disorder (MDD). Site rater assessments were audio taped and uploaded to a central server. A group of calibrated, blinded independent clinician reviewers scored a subset of downloaded site rater assessments selected according to an a priori algorithm. The reviewers also rated interviews on a 1 to 4 scale on 5 domains of the Rater Applied Performance Scale (RAPS): Adherence, Follow Up, Clarification, Neutrality, and Rapport. Independent reviewers' scoring of the scale items and the RAPS were completed and locked prior to revealing site rater scores.

Results: 1057 interviews across 2 MDD studies were analyzed to examine relationships between RAPS ratings and scoring discrepancy. As the two studies relied on different scales, all total scores were standardized to z-scores ($M = 0, SD = 1$). Scoring discrepancy was computed as the absolute difference between standardized site rater scores and standardized reviewer scores.

50.4% of the interviews were categorized as Good or Excellent quality (average of 5 RAPS domains ≥ 3) and 49.6% had an average RAPS score of less than 3.0. Scoring discrepancies were significantly greater at inclusion visits (visits during which entry criteria were applied; $n = 624$) than all other visits ($n = 423$), $t(1055) = 2.98, p < .01$. However, differences between site raters and quality reviewers were statistically significant at both screening/baseline ($t(633) = 15.90, p < .001$) and all other visits ($t(422) = 8.68, p < .001$).

For baseline and screening visits during which cut-off scores were applied, scoring discrepancy was significantly correlated with all 5 individual RAPS domains and overall RAPS score (computed as the rater's average score across all 5 items) (all r 's $> .15$, all p 's $< .05$). In order to examine the predictive relationships of each individual RAPS domain, a multiple regression was performed demonstrating that 3 domains of the RAPS significantly predicted scoring discrepancy: Adherence ($\beta = -.109, p < .05$), Follow Up ($\beta = -.143, p < .001$), and Clarification ($\beta = -.127, p < .01$). Higher scores on each of these domains predicted lower scoring discrepancies.

At follow up visits, scoring discrepancy was significantly correlated with the overall RAPS average and all 5 of the RAPS domains (all r 's $> .14$, all p 's $< .05$). Multiple regression predicting scoring discrepancy from the 5 RAPS domains demonstrates that Clarification significantly predicts scoring discrepancies ($\beta = -.132, p < .05$).

Conclusions: Perfectly accurate (i.e., valid, reliable) scores may not be attainable when interview quality is less than excellent, and poor scoring agreement may predict limited detection of efficacy or other signals. The current study demonstrates that lower interview quality is associated with greater scoring discrepancies of site rater interviews in two CNS trials' quality control programs. The RAPS is a useful tool in determining which interviews are of lower quality and identifying which domains are important for rater training and remediation to minimize scoring discrepancies, as part of a quality control program.

INTRODUCTION

- Quality of assessment interview may be an important contributor to the high number of failed CNS clinical trials.
- Lipsitz and colleagues developed the Rater Applied Performance Scale (RAPS) to measure interviewer skills in CNS clinical trials¹.
 - Kobak and colleagues examined data from a multi-site depression trial and demonstrated significant drug-placebo separation in subjects whose baseline interviews were of higher quality. There was no significant drug-placebo separation when lower quality baseline interviews were included in the analysis².
- Brown and colleagues assessed site rater interview quality as part of ongoing quality control programs in two MDD clinical trials. Interview quality improved over the number of reviews interviewers received suggesting that CNS clinical trials could benefit from ongoing monitoring of interview quality³.
- The current study seeks to identify which areas of interview quality are most closely related to scoring discrepancies between site raters and calibrated, blinded independent clinician reviewers in a continuing quality control program.
- While scoring accuracy (i.e., valid, reliable scores) cannot be determined without unblinded study results, it is hypothesized that scoring agreement should be high when interview quality is assessed as high. It is noted that information variance is reduced, and thus inter-rater reliability increased because the rater dyads are scoring one interview, the site rater's interview.

METHODS

Overview

Data were pooled from two ongoing clinical trials of Major Depressive Disorder (MDD). Site rater assessments were reviewed and **feedback was provided** to raters by quality reviewers as part of a continuing quality control program.

Site raters audiotaped their MADRS or HAM-D assessments. A group of calibrated, blinded independent clinician reviewers listened to a subset of taped site rater assessments selected according to an a priori algorithm. Reviewers rated the site rater interviews on 5 domains of the Rater Applied Performance Scale (RAPS): Adherence, Follow Up, Clarification, Neutrality, and Rapport. Scores on each domain of the RAPS range from 1 (unsatisfactory) to 4 (excellent).

The quality reviewers also scored the site raters' interviews using the same standards and conventions agreed to at the rater training session. All interviews were conducted by site raters. Blinded quality reviewers listened to those interviews and scored patients according to the site rater interview.

Quality reviewers provided detailed **feedback** on interview quality and scoring conventions to site raters after each assessment was reviewed. Raters received between 1 and 29 independent reviews and associated feedback throughout the course of the study ($M = 8.28, SD = 6.98$).

Scoring discrepancy is defined here as the difference on scale total score between the site rater and the quality reviewer based solely on the site rater interview. It is possible that blinded independent interviews would produce different levels of agreement. However, most quality control programs rely on audio or videotaped review of site rater assessments and not separately conducted independent interviews. Often, results of scoring discrepancies between site rater and quality reviewer are used to determine whether site raters require additional rater training or remediation. As such, it is important to understand which domains of interview quality are most predictive of this type of discrepancy in order to tailor future quality control programs.

As the two studies relied on different scales, total scores on the MADRS (Study 1) and HAM-D (Study 2) were standardized to z-scores ($M = 0, SD = 1$).

Subset Analyzed:

- A total of 1,057 interviews across 2 MDD studies were analyzed to examine relationships between RAPS ratings and scoring discrepancy on site rater interviews.

Site Raters:

- Study 1:** 106 site raters performing 997 MADRS assessments on subjects.
- Study 2:** 21 site raters performing 60 HAM-D assessments on subjects.
- In these studies, the site raters were selected by sponsors to interview subjects.
- All site raters received rater training prior to the start of the study by the reviewers.

Reviewers:

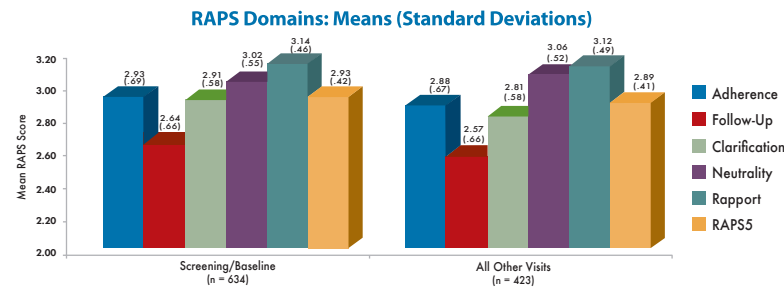
- Study 1:** 18 calibrated, blinded quality reviewers observed 997 MADRS assessments.
- Study 2:** 5 calibrated, blinded quality reviewers observed 60 HAM-D assessments.
- Based on independent interviews, calibrated quality reviewers using the same calibration and training methods as those used here have demonstrated interrater reliability to be^{4,5}:
 - MADRS Scoring ICC = .93 | HAM-D Scoring ICC = .93

REFERENCES

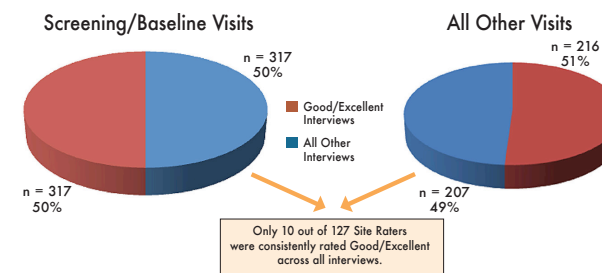
- Lipsitz, J., Kobak, K. A., Feiger, A., Sikich, D., Moroz, G., & Engelhardt, A. (2003). The Rater Applied Performance Scale (RAPS): Development and reliability. *Psychiatry Research*, 124, 147 - 155.
- Kobak, K. A., Feiger, F., & Lipsitz, J. (2007). Impact of interview quality on signal detection in clinical trials. *American Journal of Psychiatry*, 162, 628.
- Brown, B., De Santi, S., Detke, M., Brown, J., & Williams, J. B. W. (2010). Assessing interview quality and scoring accuracy in clinical trials with continuous quality control (CQC). Poster presented at the National Clinical Drug Evaluation Annual Meeting, Boca Raton, FL.
- Williams, J.B.W., & Kobak, K. A. (2008). Development and reliability of the SIGMA: A structured interview guide for the Montgomery-Asberg Depression Rating Scale (MADRS). *British Journal of Psychiatry*, 192, 52-58.
- Kobak, K.A., & Williams, J.B.W. (2006). Development and reliability of a combined Hamilton depression, anxiety, and atypical symptoms scale. American Psychiatric Association, 159th Annual Meeting, Toronto, CA.

RESULTS

- Scores on the 5 RAPS domains were averaged to create an overall RAPS score (RAPSS).



- Interviews were categorized into the following groups:
 - Good/Excellent Interviews: Average RAPS ≥ 3
 - All Other Interviews: Average RAPS < 3.0

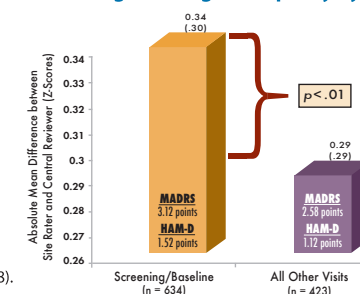


- Scoring discrepancies were computed as the absolute difference between standardized site rater scores and quality reviewer scores of the site rater interview.

- Scoring discrepancies were significantly larger at inclusion visits (visits during which cut-off scores were applied; $n = 624$) than all other visits ($n = 423$), $t(1055) = 2.98, p < .01$.

- At screening/baseline, there was a .34 ($SD = .30$) difference between site rater scores and quality reviewer scores.

Average Scoring Discrepancy by Visit



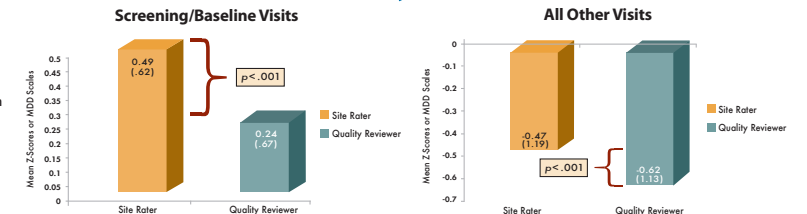
- This equates to a 3.12 point difference in MADRS scores ($M = 30.02, SD = 9.20$) and a 1.52 point difference in HAM-D scores ($M = 20.00, SD = 4.48$).

- At post-baseline, there was a .28 ($SD = .29$) difference between site rater scores and quality reviewer scores.

- This equates to a 2.58 point difference in MADRS scores ($M = 30.02, SD = 9.20$) and a 1.12 difference in HAM-D scores ($M = 20.00, SD = 4.48$).

- Differences between site raters and quality reviewers were statistically significant at both screening/baseline ($t(633) = 15.90, p < .001$) and all other visits ($t(422) = 8.68, p < .001$).

Site Rater and Quality Reviewer Scores



- All 5 RAPS domains were significantly correlated with scoring discrepancies at both screening/baseline and all other visits.

- All correlations suggest that as interview quality goes up, scoring discrepancies goes down.
 - It is possible that in rare situations bad interview quality can result in high scoring agreement particularly when insufficient information was elicited from the patient for the quality reviewer to score the interview adequately.

Correlations with Scoring Discrepancy	Adherence	Follow-Up	Clarification	Neutrality	Rapport	RAPSS
Screening/Baseline	-.26***	-.27***	-.24***	-.21***	-.15***	-.33***
All Other	-.19***	-.17***	-.21***	-.16**	-.14**	-.25***

- RAPS domains were predictive of scoring discrepancies at both screening/baseline and all other visits.
 - In order to examine the predictive relationships of each individual RAPS domain, multiple regressions were performed for each visit type.

- As feedback was given to raters throughout the course of the study after each review, review number was controlled for in these regressions.

Screening/Baseline

- Adherence, Follow-up, and Clarification significantly predicted scoring discrepancies such that as Adherence, Follow-up, and Clarification went up scoring discrepancies went down ($R^2 = .12$).

- Review number was a significant covariate such that as the number of reviews went up scoring discrepancies went down.

All Other Visits

- Clarification significantly predicted scoring discrepancies such that as clarification went up scoring discrepancies went down ($R^2 = .07$).

Screening/Baseline Visits					
Variable	B	SE(B)	β	t	p
Adherence	-.047	.020	-.109	-2.361	.019 **
Follow-Up	-.065	.020	-.143	-3.263	.001 ***
Clarification	-.065	.022	-.127	-2.906	.004 **
Neutrality	-.034	.025	-.062	-1.374	.170
Rapport	.006	.028	.009	.216	.829
Review #	-.004	.002	-.080	-2.057	.040 *

All Other Visits					
Variable	B	SE(B)	β	t	p
Adherence	-.038	.024	-.091	-1.600	.110
Follow-Up	-.021	.024	-.050	-.890	.374
Clarification	-.065	.026	-.132	-2.475	.014 **
Neutrality	-.028	.031	-.051	-.899	.369
Rapport	-.014	.030	-.025	-.450	.653
Review #	-.001	.002	-.026	-.538	.591

* = $p < .05$ ** = $p < .01$ *** = $p < .001$

CONCLUSIONS

- Perfectly accurate (i.e., valid, reliable) scores may not be attainable when interview quality is less than excellent, and poor scoring agreement may predict limited detection of efficacy or other signals.
- Scoring agreement (reliability) and scoring validity are not the same; scores with high agreement (reliability) may be systematically biased. Examples of interviews that may result in good agreement, but questionable validity, include:
 - Interviewers asking leading questions, such as reading the anchors off the scale to the patient.
 - Interviewers not asking follow-up or clarification questions.
- The predictive validity of scoring can only be assessed in the context of evaluation of the treatment-placebo signal detection of known effective treatments.
- The current study demonstrates that lower interview quality is associated with greater scoring discrepancies of site rater interviews in two CNS trials' quality control programs. Thus greater interview quality improves inter-rater reliability.
- The RAPS is a useful tool in determining which interviews are of poor quality and identifying which domains are important for rater training and remediation to reduce scoring discrepancies as part of a quality control program.

AUTHOR DISCLOSURE INFORMATION

D. Popp: MedAvante, Part 1; MedAvante, Part 2; MedAvante, Part 3; MedAvante, Part 5

M. Detke: MedAvante, Eli Lilly, Part 1; MedAvante, Eli Lilly, Part 2; MedAvante, Eli Lilly, Part 3; MedAvante, Part 5

J. Williams: MedAvante, Part 1; MedAvante, Part 2; MedAvante, Part 3; MedAvante, Part 5