The Rater Applied Performance Scale: Evaluating Clinical Interview Skill via Audio Recordings of MADRS Assessments in a Clinical Drug Trial

Engelhardt, N[1], Yavorsky, C[1], McNamara, C[1], Wolanski, K[1], Burger, F[1], Di Clemente, G[1]
[1]Cronos CCS, Inc., Lambertville NJ

Introduction

The quality of clinical interviews in CNS drug trials is frequently overlooked in favor of establishing interrater reliability with passive scoring tasks such as rating patients from a video recorded interview. Audio monitoring of primary outcome clinical interviews has been successfully applied to multi-center psychiatry trials as a way to monitor the quality of outcome data. However, optimal assessment of raters' applied skills requires systematic guidelines so that reviewers can reliably judge a rater's clinical interview skill.

The Rater Applied Performance Scale (RAPS)[1] was developed to provide a systematic and objective assessment of applied rater performance. The RAPS evaluates the clinical interview skill of raters as well as how reliably raters apply scoring criteria. It has been used in structured interview guide development, training, and active monitoring of rater performance in a clinical trial.[2]

We sought to evaluate the level of clinical interview skill of raters in a multicenter clinical drug trial evaluating a compound to treat depression, and to determine if clinical interview skill, as measured by the RAPS, is related to scoring accuracy. We also evaluated the relationship between the severity of illness, as measured by MADRS total score, and RAPS performance.

Methods

The RAPS measures six domains of clinical assessment: Adherence to scale administration guidelines, Follow-up questioning, Clarification of ambiguous responses, Neutrality, Rapport, and scoring Accuracy. Table 1 provides abbreviated definitions of each RAPS Domain.

Table 1: RAPS Domains Definitions

ADHERENCE	Specified instructions for administering the scale are followed, e.g., rater follows sequence of items and skip pattern, assesses appropriate interval, e.g., past week, past 24 hours, etc., reads required text verbatim.
FOLLOW UP	Follow-up questions are asked to elicit sufficient information to score an item accurately. Interview guides may provide questions that serve this purpose, but raters frequently need to add their own.
CLARIFICATION	Clarifying questions are asked if the clinical picture is ambiguous or contradictory. Very general responses often require clarification, e.g., "I feel depressed a lot of the time." The rater clarifies by asking "On how many days? How much of each day?"
NEUTRALITY	A neutral, non-leading style of questioning is maintained. Questions are asked in a way that does not influence or bias the subject to respond in a certain way. , e.g., "How bad has your insomnia been?" versus "How many hours have you been sleeping?"
RAPPORT	A courteous and respectful attitude is conveyed; responsive to the subject's spontaneous verbalizations. Optimal rapport in a research interview differs from rapport during a therapeutic encounter. The rater needs to avoid being overly supportive or talkative and should avoid instilling expectation of positive results.

Each domain is rated as Excellent, Good, Fair, or Unsatisfactory. Table 2 provides abbreviated criteria for judging RAPS performance.

Table 2: RAPS Ratings

UNSATISFACTORY	consistently poor performance or any systematic deviation that would compromise the validity of the assessment, e.g., skipping an item, repeatedly failing to follow up and/or clarify critical information, consistently failing to listen to information that would alter a rating, rushing the subject or repeatedly challenging negative answers.
FAIR	several marked deviations or omissions such as paraphrasing required probes rather than asking verbatim, changing the wording on several questions so that the meaning of the item is slightly altered; making reassuring comments around the subject's hope for treatment; failing to clarify ambiguous responses.
GOOD	less than optimal performance on several items. Taken together, errors do not result in more than 1 or 2 item scores being difficult to verify due to insufficient information. Good is distinguished from Excellent by the frequency and/or significance of errors.
EXCELLENT	a high level of performance throughout, e.g., thorough and consistent clarification of ambiguous information; rapport is neither too therapeutic nor too rigid; asks all necessary follow up questions An Excellent is warranted if the Rater makes a couple of minor, inconsequential, or insignificant errors.

Using domain ratings and pass/fail criteria developed for this analysis, we analyzed individual domain and total RAPS scores of 585 audio recorded interviews of the Structured Interview Guide for the Montgomery-Åsberg Depression Rating Scale (SIGMA) in a randomized, double blind placebo-controlled depression trial. Audio recorded interviews were conducted by remote blinded raters over the telephone. The RAPS was conducted by clinical specialists who were trained and calibrated on the scale by one of the scale authors (Engelhardt).

We also evaluated the relationship of depression severity with RAPS performance in a sample of 550 audio recorded interviews with non-missing value MADRS assessments. Depression severity was represented by total MADRS scores. In one analysis, we split the sample according to the median MADRS total score of 23. In a second analysis we split the sample according to clinically established cutoffs for severity. The sample was split into a low severity range ($\leq$ 12) and a moderate-severe range ($\geq$ 28).

Results

35.4% of raters failed the RAPS evaluation; the majority (64.6%) passed. The MADRS was judged to be scored accurately by 72.6% of the raters. All RAPS domains had a significant relationship to outcome (Pass/Fail) with the exception of Rapport (chi-square score of 4.542, p = .604). Virtually all of the raters (91.5%) received a score of Excellent for this domain. For those raters who failed the RAPS, only 13, or 2.4%, passed scoring Accuracy. Follow-up emerged as the domain that raters had the most difficulty executing: 28% received a Fair or Unsatisfactory rating. Follow-up was associated with scoring accuracy. There was a 72% probability that raters who were Good or Excellent in Follow Up met criteria for passing Accuracy (n = 512). However, for those raters who received Fair or Unsatisfactory in Follow Up (n = 72), the probability of passing Accuracy was 44% and not substantially different than the probability of failing Accuracy (56%).

Relationship between severity of illness and RAPS performance was evaluated in two separate analyses using different MADRS total score cutoffs. RAPS Pass/Fail rates by MADRS total score are show in Table 3. The RAPS pass rate for low severity ($\leq$ 12) was 91.4% while the moderate-severe group ($\geq$ 28) was 77.3%. The difference between the two groups was statistically significant, with z = 3.29, p<.001.

Table 3: RAPS Pass/Fail Rate by MADRS Total Score Severity

MADRS total score	Pass	Fail	n
≤ 23	84.1%	15.9%	277
≥ 23	77.0%	23.0%	287
≤ 12	91.4%	8.6%	116
≥ 28	77.3%	22.7%	172

Conclusion

The majority of remote raters in this sample demonstrated adequate clinical interview skill and scoring accuracy. Remote raters, in addition to being blind to study visit and protocol, received extensive training and regular, ongoing calibration, which may differentiate their rating performance from site raters. Another study found that lower interview quality, as measured by the RAPS, was associated with greater scoring discrepancies between site and remote raters.[3] Raters who engage in appropriate use of follow-up questions to elicit sufficient information tend to score more accurately than raters who do not. Rapport, thought to be critical in mitigating placebo response, was not related to RAPS outcome. Severity of illness affected RAPS performance, with more severe symptomatology negatively impacting RAPS performance. Raters may be less inclined to question subjects who report more severe symptomatology and who may be in greater distress.

References

[1]Lipsitz, J, Kobak, KA, Feiger, A, Sikich, D, Moroz, G, Engelhardt, N. The Rater Applied Performance Scale (RAPS): Development and Reliability. Psychiatry Research, 2004; 127: 147-155.
[2]Rothman, B, Lord-Bessen, J, Sanchez, R, Peters-Strickland, T, Opler M. Performance During "Applied" Training on MADRS/SIGMA as a Predictor of Change in a Global Depression Trial. ISCTM 13th Annual Meeting, 2017, Washington, D.C.
[3]Popp, D, Detke, M, Williams, JBW. The Relationship Between Interview Quality and Scoring Discrepancy: Application of the Rater Applied Performance Scale. ACNP 49th Annual Meeting, 2010, Miami, FL.

All authors are employees of Cronos CCS and report no conflicts of interest.

Corresponding Author: Nina Engelhardt nina.engelhardt@cronosccs.com