



The Rater Applied Performance Scale: development and reliability

Joshua Lipsitz^{a,b,*}, Ken Kobak^{b,c}, Alan Feiger^d, Dawn Sikich^b, Georges Moroz^e,
Nina Engelhardt^f

^aAnxiety Clinic #69, New York State Psychiatric Institute/College of Physicians and Surgeons, Columbia University,
1051 Riverside Drive, New York, NY 10032, USA

^bResearch Training Associates, Madison, WI, USA

^cDean Foundation, Madison, WI, USA

^dColorado University Health Science Center, Denver, CO, USA

^eHoffmann-La Roche Pharmaceuticals, Nutley, NJ, USA

^fIndianapolis School of Medicine, Indianapolis, Indiana, USA

Received 12 July 2003; received in revised form 31 December 2003; accepted 11 March 2004

Abstract

Previous studies of rater performance and interrater reliability have used passive scoring tasks such as rating patients from a videotaped interview. Little is known, however, about how well raters conduct assessments on real patients or how reliably they apply scoring criteria during actual assessment sessions. With growing recognition of the importance of monitoring and review of actual evaluation sessions, there is need for a systematic approach to quantify raters' applied performance. The Rater Applied Performance Scale (RAPS) measures six dimensions of rater performance (adherence, follow-up, clarification, neutrality, rapport, and accuracy) based on reviews of audiotaped or videotaped assessment sessions or on live monitoring of assessment sessions. We tested this new scale by having two reviewers rate 20 Hamilton Depression Scale rating sessions ascertained from a multi-site depression trial. We found good internal consistency for the RAPS. Interrater (i.e. inter-reviewer) reliability was satisfactory for RAPS total score ratings. In addition, RAPS ratings correlated with quantitative measures of scoring accuracy based on independent expert ratings. Preliminary psychometric data suggest that the RAPS may be a valuable tool for quantifying the performance of clinical raters. Potential applications of the RAPS are considered.

© 2004 Elsevier Ireland Ltd. All rights reserved.

Keywords: Rater Applied Performance Scale; Clinical trial; Independent evaluator; Reliability; Depression; Methodology; Psychometrics

1. Introduction

In psychiatry, the success or failure of a clinical trial rests largely in the hands of the clinical rater

(i.e. independent assessor/evaluator). Raters in multi-center trials come from a wide range of backgrounds and disciplines and vary widely with regard to level of experience (Demitrack et al., 1998). There are no accepted standards for prerequisite training or clinical experience for clinical raters. In practice, a rater may begin to administer

*Corresponding author. Tel.: +1-212-543-5417; fax: +1-212-543-6515.

E-mail address: Lipsitz@pi.cpmc.columbia.edu (J. Lipsitz).

a specific rating scale with little or no education about its content, rules of administration, and scoring guidelines. As scientists, regulators and sponsors struggle to explain why half of trials for approved antidepressants fail to separate active drug from placebo (Kahn et al., 2002; Katz et al., 2002), surprisingly little attention has been focused on the raters whose clinical ratings determine the outcomes of these trials.

Among depression trials, only a small percentage of studies provide any information at all about interrater reliability (Mulsant et al., 2002). When interrater reliability is assessed, multiple raters are typically asked to provide ratings based on the same (e.g. videotaped) interview and scores are compared among raters (Andreasen et al., 1982; Baca-Garcia et al., 2001). Demitrack et al. (1998) extended this methodology and categorized individual raters in terms of how closely each rater agreed with the modal score of the entire sample. While this method helps assess agreement with regard to passive understanding of scoring criteria, it does not assess how the rater collects information or how he or she applies scoring criteria in an actual interview setting, i.e. applied performance.

Given concerns about potential for unreliability and systematic bias in clinical assessments (e.g. elevating baseline scores), there is increasing recognition of the need for monitoring of clinical rating sessions, e.g. through review of audiotapes (Klein et al., 2002; Gelenberg, 2002). Audiotape monitoring has already been successfully applied to multicenter depression trials (e.g. Feiger et al., 2001). However, as with assessment of clinical symptoms, optimal assessment of raters' applied performance requires systematic guidelines through which reviewers can reliably judge the strengths and weaknesses of the rater's performance.

In psychotherapy research, rating scales are used to assess and quantify therapist behavior and performance including level of adherence to treatment protocols. Instruments such as the Collaborative Study Psychotherapy Rating Scale (CSPRS; Hollon, 1984) and the Yale Adherence and Competence Scale (YACS; Carroll et al., 2000) assess level of adherence to specific treatment techniques and approaches as well as dimensions of therapist competence (e.g. empathy). In comparative psy-

chotherapy trials, these scales can be used to determine if therapists are providing the type of therapy they are supposed to and if they are doing so in a competent manner. Research suggests that reviewers can reliably rate therapists along these dimensions (e.g. Markowitz et al., 2000). Furthermore, higher ratings on some dimensions predict better treatment response (Frank et al., 1991).

To date, no systematic rating scale has been developed to measure applied performance of clinical raters. We therefore developed the Rater Applied Performance Scale (RAPS) to be used for assessing raters' skills based on actual (audiotaped or videotaped) assessments they conduct. Our first goal for the RAPS was to help quantify judgments as to which raters would be qualified to administer specific scales in a clinical trial. Our second goal was to quantify judgments about ongoing performance of raters and sites during the course of a trial. The current report describes development of this scale based on the dimensions of rater performance that we felt were most salient. We then provide data about internal consistency, interrater reliability, and initial validity of this scale.

2. Methods

Based on the authors' experience with unstructured review of hundreds of audiotaped assessment sessions, we identified the following six salient dimensions of rater performance: (1) adherence; (2) follow-up; (3) clarification; (4) neutrality; (5) rapport and (6) accuracy (see copy of scale in Appendix A).

2.1. Dimensions of rater performance

2.1.1. Adherence

This item assesses the degree to which the rater follows specified guidelines or instructions for administering the assessment scale. For example, most scales that assign ratings based on an assessment interview present items in a specific sequence. The items should generally be administered in that sequence. When using a structured interview guide such as the Structured Interview Guide for Hamilton Depression Scale (Williams, 1988), specific texts are provided for questions

and specific rules apply (e.g. the initial probes should be asked as they appear in the interview guide). The importance of adhering to guidelines is to maximize standardization and thereby decrease information variance. If raters deviate from the prescribed guidelines, e.g. by asking an opening question in a novel way or asking questions out of sequence, they decrease standardization and this may threaten inter-rater reliability.

2.1.2. *Follow-up*

This item assesses appropriate use of follow-up questions to elicit further information. Although interview guides may structure initial inquiry, the goal of a clinical assessment is to elicit enough information so that the rater can make a clinical judgment regarding the presence and severity of symptoms. This usually means going beyond yes/no responses to get a full clinical picture of the symptom, feeling or behavior being assessed. Structured interview guides may provide optional follow-up questions meant to elicit this kind of information and these should be selected or skipped based on the information already provided. However, many situations require the raters to add additional questions of their own (e.g. Tell me more about that. What does that feel like? What have your thoughts been?).

2.1.3. *Clarification*

Answers to open-ended follow-up questions may be vague and difficult to score. For example, when asked how much of the time a symptom has been present during the past week, the patient may reply, 'A lot of the time.' Active clarification is often required to determine more precisely what the patient means (e.g. How many days a week? How much of the time each day?). If the clinical picture is still ambiguous or perhaps the patient has contradicted him/herself, the rater may need to repeat or re-phrase what has been heard in question form (So are you saying that it's hard to get anything done at all while you're at work?).

2.1.4. *Neutrality*

A common pitfall for raters is 'leading the witness.' Raters sometimes ask questions in a way that influences the patient to respond in a certain

direction. For example, during an endpoint assessment with a partially improved depressed patient, one rater responded to a description of persistent insomnia by asking, 'But it's not as bad as it used to be, is it?' Clearly, this question may influence the patient to answer in a more mild direction. As the rater follows up and clarifies, it is sometimes challenging to maintain neutrality and avoid putting words in the patient's mouth.

2.1.5. *Rapport*

The rater should be courteous, respectful and responsive to the patient's spontaneous verbalizations. A stilted or robot-like demeanor or indifference to the patient's current emotional state may lead the patient to shut down and be less forthcoming with information. A common instance of lack of rapport is when the rater reads a question that the patient has just answered in another context without acknowledging that the topic has just been discussed. Optimal rapport in a clinical rating session differs from rapport in other clinical settings such as a therapy or medication treatment session. The rater needs to avoid being too chummy, chatty, or supportive as this may influence the patient, e.g. to report only positive feelings so as not to disappoint a new friend.

2.1.6. *Accuracy of ratings*

Most rating scales have scoring guidelines embedded within item anchors, self-contained within companion guidelines (or scoring conventions), or both. Given the information obtained by the rater, this item rates how accurate the scoring is based on prescribed scoring guidelines. For example, a rater using the Hamilton Depression Scale may be told that the patient has lost weight, but only because the patient has been dieting. If the rater scores this as a symptom, it is a deviation from the prescribed scoring guidelines of the Hamilton Scale and the score is inaccurate. In some cases, conventions for scoring vary from one study to another and 'accuracy' will measure how well the rater adheres to guidelines for a specific study. In some situations, the accurate score may be difficult to determine because the rater has not followed up or clarified sufficiently. Finally, for some items (e.g. psychomotor retardation and agi-

tation) it is difficult to assess accuracy based on review of an audiotape.

2.2. Scale

Each of the six RAPS items described above (adherence, follow-up, clarification, neutrality, rapport, and accuracy) is rated on a four-point scale including 1 (unsatisfactory), 2 (fair), 3 (good) and 4 (excellent). Scores are assigned based on quality and consistency of performance. A score of excellent requires a high level of performance throughout the assessment session (e.g. thorough follow-up on each item for which further information is needed). A score of good indicates less than optimal performance for one or two items (e.g. scoring 1 point higher than is justified on 1–2 items, but good scoring on others). A score of fair is made if there are one or two marked deviations/omissions or frequent minor deviations/omissions (changing the wording subtly on several items of the interview guide so that the meaning of the item is slightly altered). Finally, an unsatisfactory score would be the result of consistently poor performance or any systematic deviation that would clearly compromise the validity of the assessment (e.g. skipping an item, demonstrating that she/he has failed to listen to information that would change a rating, rushing the patient through answers, or repeatedly challenging negative responses).

Each of the six items reflects an independent rating. In addition, a total RAPS score of overall performance is derived from the total of the six items (6–24).

2.3. Procedure

We tested reliability and initial validity of the RAPS scale using 20 audiotaped baseline Hamilton Depression Scale (HAMD) rating sessions conducted as part of a multi-center randomized outpatient depression trial. In the course of monitoring for this trial, we selected the first 20 completely reviewed tapes to be independently reviewed by a second reviewer. Thirteen raters from eight clinical centers conducted the 20 taped sessions. In addition to RAPS ratings for rater performance, each

reviewer assigned HAMD ratings independently based on information from the audiotape. The reviewers rated all HAMD items with the exception of the agitation item and the retardation item, which are rated largely based on visual observation.

2.4. Statistical analyses

2.4.1. Reliability

Internal consistency of the six-item RAPS was assessed using Cronbach's alpha coefficients. Pearson correlations were derived for individual items; two-sided tests for significance were used. Interrater (i.e. inter-reviewer) reliability for RAPS total and RAPS individual items was assessed using a two-way random effects model for intraclass correlation coefficients (ICCs) as described by Shrout and Fleiss (1979).

2.4.2. Validity

As a preliminary examination of validity, RAPS ratings were correlated with quantitative indices of rater agreement/discrepancy with independent reviewer ratings. HAMD scores of the two reviewers were averaged and these scores were compared with the rater's HAMD scores. Difference (discrepancy) scores were derived both for total HAMD (less the agitation and retardation items) and for a sum of the 15 individual HAMD item discrepancies for each reviewed case. For the latter measure, discrepancy scores were derived for each item and the absolute values of these discrepancies were added. Pearson correlations with RAPS items were derived and two-way significance tests were performed.

3. Results

Ratings were made for 20 audiotaped HAMD cases, each rated by two reviewers. Rater HAMD scores (17-item) for the 20 cases ranged from 15 to 26 (mean = 21.5; S.D. = 3.1). Table 1 presents the mean RAPS ratings for the 20 cases. Means for all items approached three, indicating good performance overall in this small sample. Ranges and standard deviations indicate scores across cases for each of the six items. Overall performance

Table 1
Rater Applied Performance Scale item and total scores. ($N=20$ tapes; 2 reviewers/tape)

	Minimum	Maximum	Mean	S.D.
Adherence	1.00	4.00	2.70	0.79
Follow-up	1.00	4.00	2.62	0.95
Clarification	1.00	4.00	2.65	0.80
Neutrality	1.00	4.00	2.75	0.81
Rapport	1.00	4.00	2.67	0.69
Accuracy	1.00	4.00	2.60	0.70
Total six items	9.00	23.00	16.0	3.3

based on the six-item RAPS total score (possible range: 6–24) ranged from 9 to 23.

3.1. Internal consistency

Coefficient alpha for the total RAPS score (items 1–6) was 0.79. Table 2 presents the inter-correlations among specific RAPS items. As shown, item intercorrelations ranged from minimal for some items (0.05 for adherence and rapport) to strongly significant for other items (0.70 for adherence and follow-up).

3.2. Interrater reliability

Table 3 presents the ICCs for interrater (i.e. inter-reviewer) reliability for the RAPS total score and individual RAPS items. The ICC for the total scale score was 0.68, which is adequate. Individual item ICCs ranged from good (0.77 for the adherence item) to poor (0.20 for rapport). Because rapport was unreliable across reviewers in this sample, we also recalculated ICC for the total RAPS score excluding the rapport item. The ICC

Table 2
Inter-correlations among six items of the Rater Applied Performance Scale based on all reviewed cases ($n=20$) by both reviewers ($N=40$)

	Adherence	Follow-up	Clarification	Neutrality	Rapport	Accuracy
Adherence		0.698**	0.558**	0.521**	0.051	0.375*
Follow-up			0.495**	0.541**	0.276	0.418**
Clarification				0.099	0.343*	0.424**
Neutrality					0.217	0.268
Rapport						0.459**

* Significant at 0.05 level, two-tailed.

** Significant at 0.01 level, two-tailed.

Table 3
Interrater reliability for RAPS total and item ratings (reviewers 1 and 2) for 20 reviewed cases

	ICC ^a	95% confidence interval
Total RAPS score	0.68	(0.35–0.86)
1. Adherence	0.77	(0.51–0.90)
2. Follow-up	0.60	(0.22–0.82)
3. Clarification	0.49	(0.08–0.76)
4. Neutrality	0.58	(0.20–0.81)
5. Rapport	0.20	(0.25–0.58)
6. Accuracy	0.49	(0.07–0.76)

^a Intraclass correlation (ICC): two-way measure with raters as random effects.

for this modified RAPS total score was 0.75 (confidence interval 0.48–0.89).

3.3. Validity

Table 4 presents the correlations between RAPS ratings and two quantitative measures of rater discrepancy from expert HAMD ratings. As shown, RAPS total scores correlated significantly with measures of discrepancy derived both on the basis of item and total HAMD scores. Individual items including adherence, follow-up, and clarification also correlated significantly with one or both of the quantitative indices. The RAPS accuracy item correlated with the item-based difference measure but not with the difference measure based on the total HAMD score.

Table 4

Correlations of RAPS scores with quantitative indices of discrepancy from expert HAMD ratings ($N=20$)

	Total score difference ^a $R=$	Total absolute item difference ^b $R=$
RAPS total (six items)	−0.456*	−0.633**
RAPS total (five items) ^c	−0.504*	−0.671**
1. Adherence	−0.507*	−0.621**
2. Follow-up	−0.311	−0.543*
3. Clarification	−0.613**	−0.520*
4. Neutrality	−0.222	−0.438
5. Rapport	0.050	−0.076
6. Accuracy	−0.395	−0.557*

Note: Inverse correlations are in the expected directions (higher RAPS ratings associated with lower discrepancy from expert ratings).

^a Total HAMD score based on 15 items rated by rater and both reviewers. Total of both reviewers was averaged and difference score was derived for each case.

^b A difference score was derived for each of 15 HAMD items by subtracting the rater's score from the average of the two raters' scores for each item. The absolute value of all 15 item-difference scores was then summed to yield a total difference score.

^c The five-item total excluded item 5 (rapport) for which inter-rater reliability was poor.

4. Discussion

Results of this preliminary study suggest that reviewers can reliably assess raters' overall applied performance using audiotaped interviews. Interrater reliability of the RAPS in this sample is comparable to that obtained for ratings scales used to measure therapist adherence and competence (e.g. Carroll et al., 2000). The total RAPS score has good internal consistency and may be a powerful index of overall rater performance. Furthermore, correlations with quantitative indices of scoring discrepancy suggest that higher RAPS scores are associated with better agreement (lower discrepancy) between a rater's HAMD scores and average HAMD ratings by two experienced reviewers.

The magnitude of intercorrelations between specific RAPS items was generally consistent with intuitive expectations. For example, ratings of adherence, follow-up and clarification were fairly highly correlated. We would expect that the more thorough interviewer would adhere to the interview guide, ask more questions, and also clarify responses that are vague. A high correlation of adherence and neutrality is also not surprising since interviewers who adhere to instructions for asking initial standardized questions rather than asking questions in their own way may be less

vulnerable to introducing directional bias. Smaller correlations with the rapport item should be considered in the context of poor interrater reliability for that item in this sample.

Because this was a small initial sample, we do not know if the RAPS ratings presented reflect the quality of overall ratings for the entire multi-center trial. Also, all raters in this trial completed intensive HAMD training prior to participation and were aware of ongoing monitoring. Thus, this sample may be biased toward better performance as compared to raters in other multi-center studies. As a result, the range of RAPS ratings may have been somewhat restricted in the lower range of performance.

Previously, we conducted reliability analyses for the RAPS using three independent reviewers in a sample of 15 taped HAMD assessments (Lipsitz et al., 2003). In that unmonitored trial, a larger proportion of tapes received unsatisfactory RAPS ratings. Interrater reliability was higher in that sample than in the current sample, with item ICCs ranging from 0.69 for reliability to 0.87 for adherence (Lipsitz et al., 2003). The rapport item, which was not reliable in the current sample, achieved satisfactory reliability in that sample ($ICC=0.70$). It is possible that distinctions, especially for subtle skills like maintaining rapport, may be more easily made in the poor range of performance (e.g.

unsatisfactory vs. fair) and less clear in the better range of performance (e.g. excellent vs. good rapport).

Total and item RAPS scores correlated more consistently with the item-based score of rater-reviewer discrepancy than with the total score discrepancy measure. The RAPS accuracy score correlated significantly only with the latter measure. This indicates that reviewers may have based accuracy ratings more on patterns of individual item scoring and less on the total score. It is not surprising that the two discrepancy indices (total score and item score discrepancies) correlated only moderately in this sample ($r=0.58$, $P=0.007$). In practice, a rater may obtain satisfactory agreement on the total HAMD scale even if several item ratings disagree—so long as there is no systematic bias to rate items either higher or lower.

We should note that the validity of HAMD ratings by independent reviewer is necessarily limited by the amount of information originally collected by the rater. The reviewer also lacks access to non-verbal information, which may influence ratings of items other than retardation and agitation. Further research will be needed to determine whether RAPS scores are associated with more meaningful indices of rater performance such as ability to detect active treatment effects. It would also be useful to explore in larger samples whether RAPS scores are correlated with interviewer characteristics such as years of training and experience. Finally, RAPS scores could be compared with more commonly used measures of rater performance such as tests of passive skills (e.g. agreement scores derived from raters scoring the same videotaped interview) to determine whether passive and applied measures correspond.

Data for this study were generated using reviews of the HAMD for which ratings are based primarily on the patients' verbal reports. It may be more difficult to score a rater's performance using other clinical rating scales (e.g. the Positive and Negative Syndrome Scale), that include many items based on direct observation or on indirect assessment from interview behavior. Thus, further research will be needed to determine whether RAPS ratings can be made reliably and meaningfully for other types of assessment instruments.

Although the RAPS may help reviewers to measure rater performance more reliably and systematically, our experience is that the most useful form of feedback for training and calibration of individual raters is a specific description of the problem. For example, "Your questions were very thorough on questions one through four, but on question five the answer was vague and you should have followed up more and gotten more details." Or, "When the patient complained about how bad he felt, you said, 'Do not worry, we'll get you better.' In a research assessment it is important to maintain appropriate rapport and not to be too supportive or encouraging."

Elsewhere we describe an interactive web-based educational system for the HAMD (Kobak et al., 2003). This type of interactive tutorial format can provide a firm knowledge base and also help ascertain information about the rater's knowledge of assessment guidelines and scoring conventions. However, just as it would be unwise to put a new driver behind the wheel based solely on a paper-and-pencil test, it would not make sense to 'qualify' or 'certify' a rater based solely on responses to a verbal test or passive ratings of an interview conducted by someone else.

It is hoped that the RAPS will enable reviewers to make more meaningful assessments of raters' applied performance and communicate on the basis of these ratings. For ongoing monitoring in the course of a study, the RAPS could provide systematic quantitative data on performance of individual raters and research sites. It could also assist in tracking possible changes (improvement or deterioration) in performance over time. The advent of digital recording and encrypted internet communication means that these reviews can be conducted and feedback provided to the rater, investigator, and sponsor with minimal time delay.

As many aspects of multi-center clinical trials (e.g. laboratory studies) come under more centralized control, it is surprising that primary outcome ratings in clinical psychiatry trials have been conducted with so little direct oversight. If applied properly, rigorous standards for rater qualification and ongoing monitoring procedures could help enhance the overall quality and reliability of ratings in clinical trials. Review and feedback based

on applied performance could help calibrate raters to achieve the best possible agreement in assessment of study patients.

Rater Applied Performance Scale (RAPS)

Interview number _____ Interviewer/Rater _____
 Date of interview _____ Date of Review _____
 Reviewer _____

Review based on: audiotape videotape live review

- Adherence:** Adherence to rating scale content and item sequence. If structured interview guide is used, adherence to this guide.

NA	unsatisfactory	fair	good	excellent
----	----------------	------	------	-----------
- Follow-up:** Use of follow-up questions to obtain more information. This includes probes from interview guide and rater's own probes when needed:

NA	unsatisfactory	fair	good	excellent
----	----------------	------	------	-----------
- Clarification:** Use of questions and re-phrasing of what has already been said to clarify ambiguous information:

NA	unsatisfactory	fair	good	excellent
----	----------------	------	------	-----------
- Neutrality:** Use of open-ended questions. Avoiding leading questions:

NA	unsatisfactory	fair	good	excellent
----	----------------	------	------	-----------
- Report:** Maintains appropriate "research rapport" during interview (attentive and respectful, but not overly chummy or therapeutic):

NA	unsatisfactory	fair	good	excellent
If unsatisfactory/fair, circle: Too therapeutic Too Rigid Too chummy				
- Accuracy of scoring:** The rater scores items based on anchor descriptions and other (e.g., manual) guidelines. Ratings are based on positive information obtained.

NA	unsatisfactory	fair	good	excellent
----	----------------	------	------	-----------

Acknowledgments

This research was made possible by grant #N43MH12049 (to Dr Kobak) from the National Institute of Mental health. Audiotapes used for reliability ratings were ascertained during a multi-site depression trial conducted by Hoffmann-La Roche Pharmaceuticals.

References

Andreasen, N.C., McDonald-Scott, P., Grove, W.M., Keller, M.B., Shapiro, R.W., Hirschfeld, R.M., 1982. Assessment of reliability in multicenter collaborative research with a videotaped approach. *American Journal of Psychiatry* 139, 876–882.

- Baca-Garcia, E., Blanco, C., Saiz-Ruiz, J., Rico, F., Diaz-Sastre, C., Cicchetti, D.V., 2001. Assessment of reliability in the clinical evaluation of depressive symptoms among multiple investigators in a multicenter clinical trial. *Psychiatry Research* 102, 163–173.
- Carroll, K.M., Nich, C., Sifry, R.L., Nuro, K.F., Frankforter, T.L., Ball, S.A., Fenton, L., Rounsaville, B.J., 2000. A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. *Drug and Alcohol Dependence* 57, 225–238.
- Demitrack, M.A., Faries, D., Herrera, J.M., Debrot, D., Potter, W.Z., 1998. The problem of measurement error in multisite clinical trials. *Psychopharmacology Bulletin* 34, 19–24.
- Feiger, A.D., Lipsitz, J.D., Kobak, K.A., Evans, K.R., Sills, T., May, 2001. Impact of a Comprehensive HAMD Inter-Rater Reliability Training in a Multi-Site Trial. National Institute of Mental Health, New Clinical Drug Evaluation Unit, 41st Annual Meeting, Phoenix, AZ.
- Frank, E., Kupfer, D.J., Wagner, E.F., McEachran, A.B., Cornes, C., 1991. Efficacy of interpersonal psychotherapy as a maintenance treatment of recurrent depression. Contributing factors. *Archives of General Psychiatry* 48, 1053–1059.
- Gelenberg, A., 2002. Out of the box. *Archives of General Psychiatry* 59, 281.
- Hollon, S.D., 1984. Final Report: System for Rating Psychotherapy Audiotapes. NIMH, US Department of Health and Human Services, Puckville, MD. 37, 509–515.
- Kahn, A., Khan, S., Brown, W.A., 2002. Are placebo controls necessary to test new antidepressants and anxiolytics? *International Journal of Neuropsychopharmacology* 5, 193–197.
- Katz, M.M., Halbreich, U.M., Bowden, C.L., Frazer, A., Pinder, R.M., Rush, A.J., Wheatly, D.P., Lebowitz, B.D., 2002. Enhancing the technology of clinical trials and the trials to evaluate newly developed, targeted antidepressants. *Neuropsychopharmacology* 27, 319–328.
- Klein, D.F., Thase, M.E., Endicott, J.E., Adler, L., Glick, I., Kalali, A., Leventer, S., Mattes, J., Ross, P., Bystritsky, A., 2002. Improving clinical trials: American Society of Clinical Psychopharmacology recommendations. *Archives of Clinical Psychiatry* 59, 272–278.
- Kobak, K.A., Lipsitz, J.D., Feiger, A.D., 2003. 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, 509–515.
- Lipsitz, J.D., Feiger, A.D., Kobak, K.A., Sikich D., Engelhardt, N., 2003. The Rater Applied Performance Scale: development and initial reliability. Poster presented at the National Institute of Mental Health, New Clinical Drug Evaluation Unit, 43rd Annual Meeting, Boca Raton, FL.
- Markowitz, J.C., Spielman, L.A., Scarvalone, P.A., Perry, S.W., 2000. Psychotherapy adherence of therapists treating HIV-positive patients with depressive symptoms. *Journal of Psychotherapy Practice and Research* 9, 75–80.
- Mulsant, B.H., Kastango, M.S., Rosen, J., Stone, R.A., Mazumdar, S., Pollock, B.G., 2002. Interrater reliability in

- clinical trials of depressive disorders. *American Journal of Psychiatry* 159, 1598–1600.
- Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* 86, 420–428.
- Williams, J.B.W., 1988. A structured interview guide for the Hamilton Depression Rating Scale. *Archives of General Psychiatry* 45, 742–747.