Reprinted from JOURNAL OF CLINICAL PSYCHOLOGY, Vol. XXVI, No. 3, 320-323, July 1970

DIFFERENTIAL RELIABILITY IN RATING PSYCHOPATHOLOGY AND GLOBAL IMPROVEMENT*

KARL RICKELS AND KAY HOWARD
University of Pennsylvania

RONALD S. LIPMAN
NIMH, Psychopharmacology Research Branch

LINO COVI AND LEE C. PARK
Johns Hopkins Medical School

AND

EBERHARD H. UHLENHUTH
University of Chicago

INTRODUCTION

The reliability of clinical judgments of varying degrees of complexity has received much attention in recent years. There has, however, been much variability in the reported reliabilities. For example, between-rater correlations have been reported as high as .82 for a 19-point scale of anxiety[2], and as low as .09 for a 4-point scale of emotional lability[3].

During the course of an active ongoing program of psychopharmacological research, we have collected reliability measures for many of our rating instruments. In this clinical note, we wish to report on the reliability of 2 relatively simple rating scales, scales which are often used and which are felt by some to be more reliable than highly complex measures[4].

*This research was supported by USPHS Grants MH-04731-2. It has been conducted as part of a collaborative study carried out jointly by the Psychopharmacology Research Branch of the National Institute of Mental Health, the Johns Hopkins Medical School, the Philadelphia General Hospital, and the University of Pennsylvania (NIMH-PRB Collaborative Outpatient Studies Nos. 2 and 3). The data reported in this paper were collected at the psychiatric outpatient clinics of the Philadelphia General Hospital and the Hospital of the University of Pennsylvania. Reprint requests to Karl Rickels, M.D., 203 Piersol Bldg., University Hospital, 3400 Spruce Street, Philadelphia, Pa. 19104.

Whereas, in an earlier study, we were concerned primarily with agreement between doctor and patient[6], here we are concerned with agreement between doctors in rating psychopathology and treatment respnose.

Method

The data reported herein were collected at the psychiatric outpatient clinics of the Philadelphia General Hospital and the Hospital of the University of Pennsylvania as part of two separate drug trials. Patients were included in both studies if they evidenced neurotic-anxious or mixed anxious-depressive symptomatology; patients evidencing psychotic, organic or sociopathic symptomatology, illiteracy, alcoholism or requiring other psychiatric treatment were excluded. In both studies, individual doctor-patient interviews were observed through one-way mirrors and heard through loud speakers. In all cases, the same observer rated a particular patient on all visits. The main results of these studies have been presented in detail elsewhere[5, 7].

In the first study, ratings of degree of patient psychopathology (neurotic distress) were performed on an 8-point scale (ranging from 1 = no psychopathology to 8 = extreme psychopathology) at the initial visit and again after one week. Ratings of patient global improvement were made after one week on a 7-point scale (ranging from "very much better" [+ 3] to "no change" [0] to "very much worse" [− 3]). These ratings were made by 4 treating doctors, psychiatric residents, and by 2 observers, staff psychiatrists, each observing 2 of the treating doctors.

In the second study, patients were evaluated at an initial visit, at 2 weeks and again after 4 weeks by their treating doctors, experienced psychiatrists rather than residents, and by the observers, again staff psychiatrists. There were 8 treating doctors and 3 observers in this study. Psychopathology (8-point scale, as above) was rated by both observer and doctor only on the first visit, and global improvement (7-point scale, as above) during the past 2 weeks was rated on the second and third visits.

Results

Table 1 presents correlations between observer and doctor ratings of psychopathology, change in psychopathology rating from visit 1 to visit 2, and global improvement for all doctor-observer pairs pooled for the first study. Observations were pooled to obtain a more generalizable estimate of reliability, in that all of our studies involve more than one doctor. Furthermore, no consistent pattern of differences in reliability among doctor-observer pairs was observed.

Table 1. Correlations Between Staff-Level Observer and Psychiatric Resident Ratings of Psychopathology and Global Improvement in a 1 Week Drug Study

Measure	N	r	p
Initial Psychopathology	123	.55	$<.001$
Psychopathology at 1 Week	109	.31	$<.001$
Change in Psychopathology	104	.07	NS
Global Improvement	110	.74	$<.001$

Table 2 presents correlations between observer and doctor ratings of psychopathology and of the various global improvement ratings of the second study. Also included in Table 2 are the percent deviations between observer and doctor ratings which are no greater than 1 point (including exact agreement).

TABLE 2. CORRELATIONS BETWEEN STAFF-LEVEL OBSERVER AND EXPERIENCED PRACTICING PSYCHIATRIST RATINGS OF PSYCHOPATHOLOGY AND GLOBAL IMPROVEMENT IN A 4 WEEK DRUG STUDY

Measure	N	r	p	% Absolute Deviations $\leq$* 1 Scale Point
Initial Pathology	225	.16	$<.05$	48.4
Improvement at 2 Weeks	174	.73	$<.001$	94.3
Improvement at 4 Weeks	141	.62	$<.001$	90.8

* $\leq$ denotes "equal or less"

DISCUSSION

The correlations from the first study of .31 and .55 between observer and doctor ratings of psychopathology, although significant, are only of moderate size. We examined the correlations between changes in observer and doctor psychopathology ratings, to determine whether, even though there might be differences between doctors and observers in absolute psychopathology ratings due to individual bias, changes in ratings from the first to the second visit might be similar, with the individual bias in absolute ratings, in a sense, removed. This did not prove to be the case, there being virtually no correlation between changes from the first to the second visit in observer and doctor psychopathology ratings. It should be noted that since the first and second visit ratings of psychopathology were made independently, without knowledge of the previous rating, the global improvement rating could not influence the size of the change in psychopathology rating.

The between-rater correlation for psychopathology of .16, found in the second study, although significant, is quite low, particularly as observers and doctors were of more comparable levels of training than in the first study. Furthermore, for nearly half of the patients, the raters disagreed by more than 1 scale point, a criterion accepted by Beck[1] as indicating good clinical agreement on a 4-point scale of depression. If 2 point differences were used as an agreement criterion, which might be more appropriate for a longer scale, then "agreement" was reached for 80% of the patients. One possible explanation is that the doctors, who were private psychiatrists brought into the hospitals for this study, were accustomed to a different patient population and were therefore using different criteria or evaluative processes to make their judgments than the observers, experienced clinic psychiatrists.

The correlations between doctor and observer ratings of global improvement were quite respectable in both studies, indicating good inter-rater correspondence in rating treatment response. The high percentage of agreement by 1 scale point or better found in the second study supports this conclusion.

It is interesting that there are higher between-rater correlations for global improvement ratings than for psychopathology ratings. One possible explanation may be that these represent different types of procedures. In evaluating psychopathology, both doctors and observers are comparing the patient against a set of internal criteria, and furthermore are required to estimate global neurotic pathology, which is a somewhat diffuse concept. In judging global improvement, however, the patient's initial appearance serves as a criterion and furthermore, the patient frequently volunteers or is asked how much better he is, providing a definite statement as a basis for a rating, which is not so for psychopathology. This is supported by the fact that patients' ratings of global improvement on the same scale tend to agree with both doctor (r = .65) and observer (r = .66) ratings, confirming our earlier results[6]. It is also true that in a number of our clinical drug studies where several doctors are involved, global ratings of psychopathology

have been less sensitive in detecting drug-placebo differences than other criterion measures. Thus, it seems that the higher reliability found for global improvement ratings may be due to the presence of greater definition and to greater reliance of both treating physician and observer on the patient's own verbal report.

Summary

Correlations between observer and doctor ratings of psychopathology (8-point scale) and global improvement (7-point scale) from 2 psychiatric drug studies were reported. The reliability of psychopathology ratings was found to be low, but the ratings of global improvement were quite reliable. It was suggested that global neurotic psychopathology is a relatively diffuse concept, and therefore difficult to assess reliably. In rating global improvement, however, the patient's initial appearance serves as a clinical criterion. Furthermore, the patient frequently verbalizes how improved he feels, thus providing more structure, which leads to greater agreement between raters in rating global improvement.

References

1. Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. and Erbaugh, J. An inventory for measuring depression. *Arch. gen. Psychiat.*, 1961, *4*, 561-571.
2. Hamburg, D. A., Sabshin, M. A., Board, F. A., Grinker, R. R., Korchin, S. J., Basowitz, H., Heath, H. and Persky, H. Classification and rating of emotional experiences. *Arch. Neurol. Psychiat.*, 1958, *79*, 415-426.
3. Jones, N. F. and Kahn, M. W. Dimensions and consistency of clinical judgment as related to the judges' level of training. *J. nerv. ment. Dis.*, 1966, *142*, 19-24.
4. Lipman, R. S., Cole, J. O., Park, L. C. and Rickels, K. Sensitivity of Symptom and Non-symptom-Focused Criteria of Outpatient Drug Efficacy. *Amer. J. Psychiat.*, 1965, *122*, 24-27.
5. Lipman, R. S., Park, L. C. and Rickels, K. Paradoxical influence of a therapeutic side-effect interpretation. *Arch. gen. Psychiat.*, 1966, *15*, 462-474.
6. Park, L. C., Uhlenhuth, E. H., Lipman, R. S., Rickels, K. and Fisher, S. A comparison of doctor and patient improvement ratings in a drug (Meprobamate) trial. *Brit. J. Psychiat.*, 1965, *3*, 535-540.
7. Rickels, K., Lipman, R. S., Park, L. C., Covi, L., Uhlenhuth, E. H. and Mock, J. Drug, doctor warmth and clinic setting in the symptomatic response to pharmacotherapy. (In preparation.)