APPENDIX A-2

June, 1952
L. Luborsky/df

SUPPLEMENT

to

Write-Up of Progress of the Health-Sickness Rating Scale
of March, 1952

Reliability Studies of Health-Sickness Rating Scale

I. Introduction:

The best rating scale for our purposes is one which can give perfect agreement by different judges and agreement by single judges with themselves from time to time on the same patient (assuming the patient hasn't changed in the meantime). Practically speaking the best we can do is use those members of the staff who know a patient very well and approximately to the same extent (preferably w ith the same kind of personal contact). Only two of the situations locally approximate this: psychotherapy supervisors and the psychotherapist, and, second, the group control situation. The results of these are reported more fully in this supplement.

Some fairly adequate reliability situations haven't yet been exploited:
(1) The section evaluations of patients offer a situation where several members know a patient over a period of a week or several weeks and have seen the patient personally. Also among the members of the team are representatives of other specialties. We should know how the psychologist's ratings, based largely on the tests, agrees with those of other members of the team. This knowledge is important to us mainly in relation to the retest study of patients in psychotherapy in which we hope the psychologist doing the retests will make independent health ratings of the initial tests as well as the tests after treatment.

(2) Repeat reliability of ratings by the same judge suffers from the major defect that the patient (we hope) does change. The therapist also remembers his earlier rating so that they are not independent assessments of the patient. It would be of interest nevertheless to know how often the therapist changes his mind even when he considers that the patient hasn't changed. We might also find some useful information about our scale through asking therapists for their "range of undertainty," that is, after giving their ratings to say that their range of uncertainty "could go as low as ___ and as high as ____."

II. Summary of Reliability Data Available:

A. Dr. Watterson's analysis of supervisors' vs. therapists' ratings and members of group control vs. therapists' and control leader's ratings.

B. Dr. Aronson's analysis of data from two staff case conferences, Dr. Epstein's and Dr. Lythgoe's

C. Miscellaneous reliability studies:

(1) Raters relatively untrained both in psychology and in the use of the scale: K.U. students in Dr. Bergman's class in Psychotherapy.

(2) Raters trained in psychiatry but relatively untrained in the use of the scale:

(a) D.A.P. therapists rating case descriptions.

(b) Ratings and rankings of our sample case descriptions by other staff members, i.e., outside the Committee (six psychologists.) (Results not yet analyzed.)

III. General Results and Conclusions From All Reliability Studies:

We had hoped at the outset to construct a 100 point scale which would give us interrater reliability within 10 points for most raters. We can conclude from our results so far that we have achieved approximately this. In Dr. Watterson's study the discrepancies in rating the patients' condition at the beginning of treatment, present state, and predictive end state indicate that approximately two-thirds of all the paired ratings were less than 10 points apart. A sample correlation: The correlation coefficient between therapists and observors in rating the state at the beginning of the treatment is , which is significant at the level of significance, for thirty patients.

There is also close agreement in the study between therapists and observers in estimating the amount of change from the start to the present point of treatment. (24% of the observors agree perfectly, 77% agree within 5 points more or less of each other). There is somewhat less agreement in predicting the amount of further change from the beginning point. (Only about 33% are within 5 points more or less of each other).

In Dr. Aronson's study of ratings by the audience of therapists of case presentations a similarly encouraging level of reliability emerges: 50% of the raters fall within less than 10 points and 75% less than 20 points difference from each other in rating the beginning stage and end stage. (Incidentally, we had made a 100 point scale but in practice the entire range of points isn't used. In Dr. Watterson's study about 45 to 55 points were used.)

The main difference between Watterson's and Aronson's reliability situation was one of degree of intimate knowledge of the observors. The audience of observors had only a two-hour case presentation on which to base their ratings. Also they may not have been as well trained in the use of the scale. It isn't surprising then that their reliability is not as great as the therapists-supervisors situation. (But I could argue the other way too. Since the therapist and supervisor know the patient more intimately they have enough basis for making up their own minds and

conceivably could differ more than in a situation where the presenter knows the patient and the others largely have to take his word for the description.)

Some other findings from the reliability studies:

(1) The effect on ratings of type of relationship with the patient:

Watterson found that the therapist tended to predict slightly more change than the observers. We have no basis for judging who is more right. The greater previous experience of the observers may be partly responsible but even more likely is the greater personal involvement of the therapists. Otherwise in rating the patients' present and beginning state they are (as stated above) very close together.

(2) The effect on ratings of amount of knowledge of the patient:

In Aronson's case presentation situation those who knew the patient best, at least in Dr. Epstein's case, were clustered very closely together, but I can't tell about Dr. Lythgoe's case. This is very slight evidence from which to conclude that greater knowledge of the patient gives greater reliability of ratings.

(3) The rating of present state or beginning state vs. predicted end state:

There is evidence from Watterson's study that one does best in rating the present state and not so well in rating a predicted state.

(4) Ratings of beginning state and present state vs. ratings of end state:

The main data here is from Aronson's case conference situation where we find that it is easier to rate the state at the beginning of treatment than at the end of treatment. This finding may have to do with the more exact description of the case in the beginning state than at the end; to [illegible] the general optimism about change through psychotherapy of patients; or ones belief in the therapeutic powers of the therapist. We have also noticed in our Committee a tendency to be more sure of the beginning state than of the end state ratings, as if we take seriously what the patient is like at the beginning while at the end we look him over more carefully to be sure that what we see is what we see and not an artifact of the therapy.

(5) The effect of training in the use of the scale:

In general in the rating of sample cases and in the case presentation situation, the committee members rated lower than others. I am sure there are other factors that may have produced this result than practice in the use of the scale—but that might be one. In the case conference situation there are no differences between the rating of the two cases that can be attributed to practice effects.

(6) Individual differences in judges in their ratings:

We have noticed that some judges have a tendency to rate high both beginning and end state, some to rate low both beginning and end state. We

haven't gone further in the study of such individual differences in conservatism-optimism.

(7) The effect of training in psychiatry on reliability of ratings:

The only evidence we have is from Bergman's K.U. students in a class in Psychotherapy. These raters were relatively untrained both in psychiatry and in the use the scale. At any rate, as one would expect, these students ranked and rated the sample cases more divergently than we or the other staff members here at the Clinic.

(8) The comparison of ratings based on psychological tests vs. clinical case descriptions:

We have only the datum from Dr. Epstein's case conference presentation in which the psychologists rated the case slightly higher than the group. We may guess that the tests didn't show up the very high degree of social disorganization in this patient which was weighted more heavily by those who heard only the case description. This is in the direction we expected: that test criteria would be more heavily based on the usual test evidences of personality disorganization which is only one of our seven criteria.

A general conclusion: In view of the difficulty we have experienced in agreeing on and communicating our criteria for "health" and how the criteria should be combined to obtain a rating, it is all the more interesting that we were able to obtain so much agreement. Health appears to be a concept (like ego strength) which judges have a feeling of knowing, often with considerable certainty, and as experience with the scale piles up we demonstrate that (with slight instruction in the use of the scale) most judges estimate a patient's health with fairly good agreement.