RELIABILITY, CHANCE, AND FANTASY IN INTER-JUDGE AGREEMENT AMONG CLINICIANS *

William A. Hunt, Franklyn N. Arnhoff, and John W. Cotton

Northwestern University

While this paper presents research data its primary purpose is pedagogical. Reliability is the Achilles heel of those clinical disciplines employing the intuitive judgmental process as an operating technique and hence it is of tremendous interest to olinical psychologists. Whenever a clinical study yields data which may bear upon the factor of reliability, such reliability is eagerly surveyed and reported, and always can count on a fascinated if not necessarily enthusiastic audience. For our purposes here, reliability will be defined as inter-judge agreement, in a judgmental situation closely approaching actual operating clinical practice.

Our data were obtained from a previously reported study in which 60 clinicians were given a group of 10 schizophrenic responses to items from the Wechsler-Bellevue and Terman-Binet vocabulary tests (2). They were then asked to rate each of the responses for the severity of the disorder in the thinking processes exhibited using an 11-point scale. The subjects were 60 professional clinicians with four years or more on-the-job professional experience. As a measure of reliability or inter-judge agreement we correlated the rank order of the 10 stimuli for each judge with that of the rank order assigned by averaging the judgments of all 60 clinicians. While there is some contamination here, since each judge contributed to the group average, the proportion of 1-60 renders this negligible.

Brevity and economy in reporting usually dictate the use of some single measure of reliability, which in this case might well be the equivalent of an average $\underline{r}$. For our purposes, however, it seems wiser pedagogically to present a complete table of all 60 $\underline{r}$s. This appears in Table I. Inspection immediately shows the wide range of $\underline{r}$s from + .02 to + .93, with a modal clustering in the 60's. This might be viewed as representing a true value in the 60's with an error distribution about this point, or it might be viewed as a continuum of ability with individual clinicians distributed upon it. Actually, it must be both but it seems safe to assume that differences in ability are at least in part responsible for the distribution, and that in terms of the ability to make reliable judgments in the sense used here,

---

* This study is part of a larger project continuing under ONR contract 7onr-450(11) with Northwestern University. The opinions expressed, however, are those of the individual authors and do not represent the opinions or policy of the Naval service. The present article has been accepted for publication in the Journal of Clinical Psychology.

clinicians vary tremendously. There would seem to be "good" clinicians and "bad" clinicians. While this fact is implicitly recognized among clinicians and occasionally reported in the literature (5), it is seldom taken account of in either experimental designs involving clinical judgment (4) or in actual clinical practice utilizing such judgments. Certainly this wide range of ability is concealed by the use of any single measure.

To illustrate this, let us use such a single measure. We select Alexander's r* as a measure of the average $\bar{r}$ between pairs of judges (1). When such an overall measure is applied to our data it comes out +.33. It is an honest measure and statistically justified, but in this case it conceals some very important information concerning the range of ability among clinicians, a fact which is evident if we consult Table I.

So far we have been considering "reliability." Let us now consider "chance." Since 60 clinicians is an unusually large sample for this type of study, we may feel secure. Most studies use many fewer subjects. Suppose we had had only 20 subjects in our group. We can attempt to answer this conjecture by splitting our group of 60 randomly into three groups of 20 clinicians each. When we do this and apply the same measure, we find the following average $\bar{r}$s: +.19, +.51, +.26. The range here is noticeable. In fact the $\bar{r}$ of +.51 is so far out of line as to establish the lack of homogeneity in this sampling, although it was achieved in random fashion. In terms of our total sample of 60 it is evident that a value of +.19 would underestimate "typical" clinical ability and +.51 would overestimate it. This demonstrates the part that chance (in sampling) may play in measuring inter-judge agreement among clinicians.

Now let us go one step further. Let us assume either that our data are not amenable to treatment by the method of rank order correlation, or that such a stodgy, commonplace measure seems too pedestrian for our use. Then let us indulge in a little status-connected statistical and logical interpretation. We may say that unless the clinicians were showing some agreement between themselves the items judged would not be statistically significantly distinguished one from the other. A measure of the significance with which the items are discriminated would then be closely related to inter-judge agreement and might serve as an indirect measure thereof. This assumption is sensible and supposedly the resulting measure might be informative provided we kept firmly in mind how it had been derived. For such a purpose we decided to use Hoyt's $\bar{r}$ which gives a measure of the reliability of the average judgments for the several items (3). When applied to our data, it gives an $\bar{r}$ of +.97. Its use for our purposes represents fantasy in inter-judge agreement among clinicians.

Table 1

Rank Order Correlations of Each Judge's Ratings with Average Rating of 60 Judges

.02	.25	.41	.49	.59	.64	.68	.73	.82	.86
.08	.27	.42	.50	.59	.64	.68	.74	.83	.86
.13	.37	.42	.52	.60	.64	.68	.74	.84	.89
.21	.38	.45	.54	.61	.66	.70	.75	.84	.92
.22	.40	.45	.57	.61	.66	.71	.76	.85	.93
.23	.41	.46	.58	.62	.67	.71	.77	.86	.93

REFERENCES

1. Alexander, H. W. The estimation of reliability when several trials are available. Psychometrika, 1947, 12, 79-99.

2. Arnhoff, F. N. Some factors affecting the unreliability of clinical judgments. J. clin. Psychol. In press.

3. Hoyt, C. J. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

4. Kelly, E. L., and Fiske, D. W. The prediction of performance in clinical psychology. Ann Arbor: University of Michigan Press, 1951.

5. Klehr, H. Clinical intuition and test scores as a basis for diagnosis. J. consult. Psychol., 1949, 13, 34-38.