April 1971

METHOD FOR A PROGRAM OF RESEARCH

INTO THE THEORY OF POSITIVE DISINTEGRATION

Introductory Inferential Notes and Associations

by

Lawrence F. Spreng, M.A.

PART I. INTERJUDGE RELIABILITY AND DEVELOPMENTAL ASSESSMENT

Preliminary Considerations

Consistently in the psychological literature one always finds "interjudge reliability" or "interjudge agreement" discussed in these research situations where the technique of measurement is such that judges are required to apply a set of implicit and/or explicit criteria in order to make a decision regarding the labels or numerical values to be assigned to a set of data or stimuli. In other words we are involved here with the situation where the instrument of measurement (nominal, ordinal, interval or ratio) is an individual together with a set of criteria (explicit and/or implicit rules). The measurement process in this case constitutes an individual (judge) interacting with a set of criteria and a set of data, e.g. subject responses, in order to arrive at judgements matches between the criteria and the data.

In our case we have the following units:

1. A set of criteria, derived from the Theory of Positive Disintegration, which prescribe what the relationships between subject's responses and the levels of development are. These criteria may be conceptualized as rules for deciding what position on a continuum of development a subject's responses are indicative of (e.g. dynamisms and levels of emotional and instinctive functions).

2. Judges who use the criteria.

3. Subjects' responses in various response modes, e.g. responses to verbal stimuli, biographical responses, neurological responses, Rorschach responses and perhaps interview responses.

Interjudge reliability or agreement is analogous to the split-half reliability of a questionnaire or test. In both cases we are concerned with the question of whether the measuring instrument is reliable. That is we are concerned with the question of whether the measuring instrument consistently yields the same values when applied to a set of data. In the case of judges-using-criteria as the measuring instrument we are concerned with their ability to yield the same values consistently to a set of data. The split-half reliability of a test may be likened to the agreement between judges (do the judges represent the same instrument?) for any given type of response data. Test-retest reliability may be likened to the consistency of any particular judge-using-criteria in assigning values to the same set of response data at different points in time.

Interjudge agreement or reliability, if found to be high (or perfect), indicates that different judges or individuals can, using roughly implicit and/or explicit criteria derived from the Theory of Positive Disintegration, apply these criteria to any given type of response data and can arrive at the same values, e.g. levels of development. Thus, for a given type of response data, e.g. responses to verbal stimuli, interjudge agreement is a necessary but not sufficient condition for the demonstration that judges-using-criteria can validly apply these criteria.

The use of only one judge instead of at least two is unacceptable since only one judge may be assigning values on bases other than the criteria derived from the theory. With only one judge he, the judge, is inseparably confounded with the type of response data being considered. More than one judge (at least three I believe) is required for the measurement of development in any particular type of response data in order to demonstrate the objectivity of measuring development for that particular type of response data. This is the scientific definition of objectivity, i.e. consensus, i.e. inter-subjective empiricism.

Let us consider a case in point. If three judges had high agreement in terms of scoring verbal stimuli responses for level of development and if this agreement was valid, i.e. was derived from the criterion of the Theory of Positive Disintegration, then we could say that these judges were measuring more than just their individual perceptions of a verbal stimulus protocol. If these judges then scored a set of new data (1/3 for each judge) we will be inclined to believe that we had a measure of development based on the verbal stimulus responses made for a set of subjects. Now if we take another response mode like the neurological data and if only one judge were to score these data we would not know whether or not the neurological data represented developmental assessment based on neurological responses as a technique or if the "neurological" data merely represented the responses of this one judge that had nothing really to do with the neurological responses.

When we then correlate the developmental levels of subjects on verbal stimuli responses (objectively scored through a process of interjudge assessors) with the neurological data we could not claim or assert that level of development assessed in verbal stimuli responses is highly associated with level of development assessed in neurological responses given that we had a high correlation!

Why? Because we do not know whether or not the judge who scored the neurologicals was really using neurological data to make his decisions. We are in this case in a state of ignorance!

Let us consider another case in point. Several judges score verbal stimuli and we have not assessed interjudge agreement. We then correlate the levels of development on verbal stimuli with any other measure (assume that this other measure is perfectly reliable and valid). We find a low correlation. Now this correlation could reflect a low association between level of development using the technique of verbal stimuli-response scoring and the other measure. However, thellow correlation could also be due to low reliability on the verbal stimulus response technique due to low interjudge reliability. In other words it could be the case that the judges who scored the response to verbal stimuli didn't use the same criteria in the same way. Again we don't know, we are in a state of ignorance. We are unable to make clear inferences!

Interjudge Agreement and Goals

Interjudge agreement or reliability is what we need if we are to achieve the development of developmental assessment. for different types of subject data. We need it because it is demonstrative of objectivity. We need it because it demonstrates that the criteria of the Theory of Positive Disintegration can be objectively applied to the behavior and expression of individuals. We need it because only through it can we ever hope to have a substantial group of persons equiped to assess and diagnose the development of personality we are concerned with.

Program of Research

I will now outline what to do, how to assess interjudge reliability, how to develop a training program for judges, and how to apply judger scoring techniques to the investigation of the relationships between levels of development using various types of data.

A. For any particular type of response data, e.g. responses to verbal stimuli or biographical data, collect 30 cases that seem to represent a fair range of development.

B. Take at least three judges familiar with the theory and criteria.

C. Have each judge score the 30 cases with preferably the criteria in hand.

D. Insure that the judges make no marks on the case material.

E. Insure that the judges do not communicate with each other at all during the period of scoring the 30 cases.

F. Assess interjudge agreement at an interval level in this way:

1. Compute the Pearson Product Moment correlation between each and every pair of judges.
2. Take noteeof the Y intercepts along the way.
3. Average these correlates using a z' transformation.
4. This average correlation is the index of interjudge agreement at an interval level.
5. The y intercepts will indicate whether any particular judge soores a little higher or lower than the other judges even though he agrees in the distances between the 30 cases.
6. If the y's appear to be high or low then compute a univariate ANOVA using the conservative number of degrees of freedom (since we have a repeated measure situation) followed by Duncan's multiple range test.

G. Assess interjudge agreement at an ordinal level using Kendall's W with the correction 6or ties.

H. If interjudge agreement is high ($>$ .8) you are finished if "G" yields no differences.

I. If interjudge agreement is low ($<$ .8) take the same 30 cases and have the judges discuss their disagreements and resolve the reasons why they differed.

J. Then, take another 30 cases and continue repeating the process until the agreement is high or give up.

Giving up implies that judges can't agree on scoring levels of development for the response made being considered. Once interjudge agreement has been established have the judges score another 30 cases. These cases should reveal a comparable level of agreement. Now we have, given high interjudge agreement, two sets of 30 cases for which our core of judges agree.

When a new person is to be trained as ajjudge at a later date, he may be trained verbally by the core judges and tested on a number of sample cases. Then he may be interrelated with the core group on the first set of 30 cases. If his agreement compares to the core judges he passes the test. If not he is retrained and tested in the second set of 30 cases.

When the core of judges developed for each type of response data are the same individuals, e.g. when Jim, Betty and Matilda agree onsscoring verbal stimuli responses and agree on scoring biographies, etc., a special procedure must be used to do the <u>actual scoring</u> of data for research purposes.

Let us say we have several types of data, e.g. verbal stimuli responses and biographies. Now each judge <u>must</u>--

1. Not know the scores on any other measure for the subjects data he is scoring on a particular type of data (double or triple blind).
2. Score a different set of subjects in one type of data than he does on any other type of data. This means that if Jim scores subjects 1-30 on verbal stimuli, Betty scores subjects 31-60 and Matilda scores subjects 61-90, <u>then</u> for biographies Jim scores either subjects 31-60 or 61-90, Betty scores subjects 1-30 or 61-90 depending on which group Jim scored, etc.

Points number 1 and 2 insure a double or triple blind and

and insure that there will be no judge carry-over affect that will artificially inflate the correlation between verbal stimuli and biographies.

TO BE CONTINUED

AND REVISED--A PREVIEW

In the future I will outline a program for a rating format, studies in the processes of judgment used in assessment levels of development and the development of a rating manual and test for use in other centers of research.