NEGATIVE RELIABILTY: THE IGNORED RULE

Dick J. Bierman

Round-table: Reliability and other ignored issues in Parapsychology

PA-Convention 1980 Reykjavic, Iceland.

see also *RIP* 1980, pp14-15

One of the ways to measure reliability is to divide the total data set into two parts and compare the resulting data-sets.

It so happens that, though there have been hardly any reliabilty figures published in Parapsychological journals, there have been experiments in which the data set was for other reasons split into two parts and compared. Reasons for those data splits ranged from the practical one of sharing the workload of the analyzers to the recently proposed Edinburgh split where one part of the data is used as a pilot and the other part as confirmation. One could even consider the classical pilot-confirmation paradigm as a data-split procedure meant to show reliability of results..

Let us then consider some of the available data, starting with the results of the classical pilot / confirmation paradigm.

It is no secret that not infrequently scoring directions reverse from pilot to confirmation. The use of direction independent statistics and 2-tailed testing should not be allowed to obscure the fact that we have here instances of negative reliability. One might object that pilot and confirmation study are not identical and therefore do not have to result in comparable datasets.

However if we look at the other evidence the situation becomes even more discouraging. In several studies I formally introduced a data-split procedure in the design. This was done to examine the effect of psi entering on a more global level than on that of the individual subjects (a higher hierarchical level in Kennedy's nomenclature). In most of these studies I found significant differences between the 2 sets: In our present discussion this means that these studies also showed negative reliability.

There is one other well known series of studies in which data-sets were split and checked or analyzed by different persons. These are the Feather & Brier studies. Those too showed strong negative reliability of the runscores. Recently also Carl Sargent reported comparable analyzer effects which he however regarded as of minor importance.

In retrospect the well-known Fisk-West studies in which differences were observed with regard to datasets where targets were prepared by different experimenters should be re-evaluated in the light of the analyzer or checker effect. Personal communication with West revealed that indeed the two experimenters did not only prepare their own sets of targets but also analyzed the related 'own' datasets.

How can we interpret these findings of negative reliability. Note that if the differences between the data-sets are significant this is at least an indication that psi entered into the data somewhere.

A negative reliability then seems to imply that we wrongly choose the unit of analysis. In differential process oriented research the unit of analysis is often taken to be the subject. This is, of course, based upon the assumption that the subject has something to do with the significant results. I suggest that these negative reliability figures may simply be telling us that this assumption is incorrect. Indeed the Observational theories would indicate that a proper unit of analysis is the 'observer' or the 'observers' of the results and not the subject per se.

The traditional OT's of Schmidt and Walker however allow for an infinite number of observers which would yield un reliability as the invariable rule. I talked this morning about a more limited model which does include multiple observers but converges rapidly. This model suggests that practically reliability can be obtained if we take as the unit of analysis scores related to the subject and experimenter and maybe a few subsequent observers. According to the OT's a subject in an ESP experiment is simply a complex RNG. The actual psi comes in at a later time when feedback is given. Then the psi-source is triggered and time-displaced PK effect is excerted on his/her earlier guessing.

Let us consider as an example a Ganzfeld study. The Ganzfeld stimulation induces randomness in the subject. Later, after the session, the target is revealed and the psi-source is triggered 'backwards in time' causing the subject to have experienced that particular imagery during the GF-stimulation.

Now it happens that feed-back in the GF research is given to the experimenter as well as to the subject for each trial. Therefore the trial is a suitable unit of analysis for these experiments. Variations in the psi-strength of subject and experimenter are averaged over a number of trials and I expect these experiments to show more reliable results even if analyzed one-tailed.