Comment to J. Rojahn

# Validity and Reliability of Data from Naturalistic Observational Studies – Problems and Alternatives

J. H. Ellgring[1]

Max-Planck-Institut für Psychiatrie, München

The systematic observation of behaviour has become an important tool in behaviourally oriented therapies. It is central to develop criteria for the general methodology in this field to evaluate different tools and their user's skillfulness. In his paper, Rojahn proposes a two-step procedure to construct and evaluate observational methods.

1. Development of reliable behavioural categories with inter-observer agreement of more than 90 % as a criterion.
2. Assessment of intra-observer reliability over a period of time, which has to exceed 80 %.

This procedure is reasonable, and without regard to the numerical values a necessary but not a sufficient one. Rojahn is surely right in stating that inter-observer agreement has not proven to be a collective measure for both the accuracy of the behaviour coding system and the competency of observers' performance. It is nearly a truism to state that reliability is a necessary but not a sufficient criterion for accurate observation. Obviously it is not advisable" . . . using the obsolete inter-observer reliability as the only proof for accurate data . . ." However, one should drop the term "obsolete", a classification which cannot be justified by simply stating it.
The specific procedure proposed in the article, is illustrated in a flow-chart. It is quite unusual that this flow-chart contains a lot of definitions, besides some temporal or logical flow. It might be summarized as follows:

1. Revise each single behavioural category unless inter-observer agreement of more than 90 % is achieved – and you will get a coding system with internally valid categories.
2. Take a category as a basis for data-collection only if the intra-observer index during application is higher than 80 %.

As the main point, Rojahn stresses the concept of observer-drift: A change in the manner in which observers apply definitions of behavioural categories over time. The drift corresponds to intra-observer agreement. He claims that the index of intra-observer agreement sufficiently indicates the observers' performance, mainly changes in the observers' performance over time.

---

[1] Requests for reprints should be sent to: Heiner Ellgring, Max-Planck-Institut für Psychiatrie, Kraepelinstraße 10, D-8000 München 40.

It is interesting to know to what extent "observer drifts" are actually to be expected and how they can be distinguished from behavioural changes and behavioural instability. Some empirical evidence is needed here. Taking the intra-observer reliabilities from Rojahn et al. (1978, p. 189), there is not much drift. An actual run like that outlined in Figure 2 would be of great help in understanding the computational procedures intended.

However, it has been known since Bessel and Wundt, that intra-individual observations can be consistent and stable but differ inter-individually. Intra-observer agreement is a valuable information and inter-observer reliability is so too.

Medley and Mitzel (1963, pp. 253 – 254) distinguished between three aspects or types of reliability: To obtain inter-observer reliability or observer agreement, different persons observe the same behaviour at the same time. The stability coefficient is assessed by having the same observer observe at different times. The reliability coefficient is assessed by having different observers observe at different times. Each of these measures gives different informations on the quality of observation.

The terms for and the procedures of inter- and intra-observer reliability seem to be most widespread. It is doubtful if renaming of inter-observer reliability as internal validity, as Rojahn does by simple re-interpretation of those different concepts, adds to clarification.

Without comment, Rojahn uses the numerical values of 90 % for inter-observer agreement and of 80 % for intra-observer agreement.

These algebraic values are obviously arbitrary and their computational procedures are not specified. When using some kind of rating-scales as means of behaviour description, a correlation may indicate agreement as well. Looking at categorial data, percentage of agreement can be used. On the other hand, correlational methods are available to account for chance (cf. Fleiss, 1971). Lower values can be accepted when clinically important behaviours are difficult to observe.

Rojahn's numerical values mainly suggest a higher intra-person variability than inter-person variability of observations as the norm. The intra-observer agreement of 80 % indicates that Rojahn expects a single person to be more unstable in his observational skills and to show more intra-individual changes (when interpreted as observer drift) as the group of observers and their inter-individual variability. That would be a most interesting effect, in contrast to common empirical knowledge and worth-while to be pursued empirically.

It is obvious that even if a group of observers shares a common view on aggressive acts in children, this is not necessarily so for a new observer. He can then be trained and share their definition. But as the old common observation of channels on the planet Mars not necessarily indicates that it is a valid observation, some external validation in the observation of aggressive behaviour is needed.

To summarize: There seems to be no advantage in evading the concept of inter-observer reliability by renaming it as internal validity of categories. The same applies to the interpretation of intra-observer reliability mainly as an indicator of

observer drift. It is not sufficient to rely on both indices as the main criteria to evaluate observational methods only.
When discussing observational methods and criteria to assess their usefulness and quality, it should be taken into account:

1. It is not useful to develop general coding-systems for all observational tasks. Microscopes, pocket-lens, glasses, and long distance binoculars are technical aids for different observational tasks. They are suited for different kinds of objects which have to be observed. A general physician will be content with his simple microscope rather than taking the trouble of using a mighty electron-microscope. Similarly, a practitioner in behaviour modification might want to have only a quick and rough overview to check if his treatments have an effect. A comparatively rough observational method would be given preference over a differentiated, elaborated, but time-consuming procedure.
First and foremost, applied behaviour analysis needs handy tools. Development of these tools ought to be improved and the armamentarium enlarged.
2. Different criteria for applicability and quality of observational instruments and observers' performance ought to be used depending on the specific task.
To amplify on the microscope-example a little bit: The general physician might be interested in a microscope of small size, low weight, easy to handle, with a magnifying power that is completely insufficient for the medical researcher. Nevertheless, both would not recommend their own instrument to the other, but rather use their own for their own purposes.
Looking at the practitioner in behaviour therapy, some discussion on criteria is needed. These criteria have to be oriented towards the requirements of daily therapy. The practitioner needs e.g. some self-control methods to indicate his quality of observing to him. An observer drift as postulated by Rojahn is of importance here. The researcher, who wants to publish and recommend a new therapy or re-evaluate an old one has to meet higher standards regarding reliability and validity of his observations.
3. When observational methods are used as scientific instruments in order to assess behavioural changes, a variety of information is needed to evaluate their quality.

It is advisable to maintain the assessment of inter-observer reliability as an indicator of person-independent observation, clarity of category definition etc. It is a global measure which is affected by different factors. Besides stating percentage of agreement, other measures can be used and different minimal criteria values can be defined depending on the specific purpose, the kind of behaviour etc.
If possible, not only internal, face or content validity should be investigated. There are plenty of ways to do so that are discussed in the literature. Correlations with some other, independently assessed variables are obviously one way. Correlations between subjective measures of the patient and observational data is another. Still another approach might be used in therapeutic studies: If videotapes are observed by persons unaware of temporal order of sessions, substantial relation between state of therapy (baseline, training, etc.) and observed behaviours are an indirect

proof of validity. Then the observational data can be regarded as valid with respect to therapeutic procedures and changes induced by them.
Relying on two indices only as Rojahn suggests, neglects a lot of valuable information. By demanding high standards in very specific aspects of quality, other aspects of equal or greater importance are disregarded.
The goal of systematic observation should be to achieve a versatile repertoire of tools for observation, to evaluate their quality according to a variety of criteria and to improve their users' skillfulness.

**References**

Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 1971, *76*, 378–382.

Medley, D. M., & Mitzel, H. E. Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

Rojahn, J. Validity and reliability of data from naturalistic observational studies – Problems and alternatives. *Behavioural Analysis and Modification*, 1978, *3*, 296–305.

Rojahn, J., Mulick, J. A., McCoy, D., & Schroeder, S. R. Setting effects, adaptive clothing, and the modification of head-banging and self-restraint in two profoundly retarded adults. *Behavioural Analysis and Modification*, 1978, *2*, 185–196.