Analytic Intervention Rating System (AIRS): a rating system for psychoanalytic interventions

Guenther Klug & Dorothea Huber

Since Strachey published his seminal paper in 1934 on the central role of transference interpretation as the mutative interpretation, a great deal of work has been done to empirically verify his hypothesis, for example by Malan (1976), Marziali (198o, 1984), Luborsky et al. (1979), Silberschatz et al. (1986), Piper et al. (1986, 1991), Mc Cullough et al. (1991) and Crits-Christoph (1993) to mention only the most important contributions for example. Henry et al. (1994) reviewed the literature and pointed out the need for further research because of the many open questions about such a central issue of psychodynamic psychotherapy. In our Munich Psychotherapy Study (MPS) we are studying the process-outcome-link with the transference interpretation being regarded as the most important intervention of the psychotherapist.

In this connection we developed a system for rating the interventions of the psychoanalyst: The Analytic Intervention Rating System (AIRS). The key features of the AIRS are as follows:

> a) it is based on a psychoanalytic theory of technique (and is not pantheoretical, as are all of the systems intended for use in comparing different psychotherapies).
>
> b) it is, in Russell and Stiles´ typology (1979), intersubjective; that is, it is descriptive of syntactically implied and other relationships between the communicator and the recipient.
>
> c) it uses a pragmatic coding strategy, describing the characteristics of the communicator by inferring the communicator´s intent (as opposed to the classical coding strategy, describing the characteristics of the text, with only minimal inference).
>
> d) it has mutually exclusive categories, that is there is only one way to classify a given intervention.
>
> e) The categories within the system are exhaustive, that is, all relevant interventions can be put into one of the categories.
>
> f) it pays special attention to one intervention, the interpretation, differentiating it along a temporal line from the "there and then" to the "here and now".
>
> g) it pays special attention to the context of one type of intervention, the interpretation.
>
> h) it is especially suitable for evaluating long-term treatments.

From this list of key features it is clear that the existing rating systems, e.g. those of Strupp (1957), Stiles (1979), Hill (1978), Piper (1984), Stuhr (1984) and Gaston (1988) to name only the best known, were not appropriate for our purposes.

I will now give a brief overview of the intervention categories. The manual includes a precise definition of each type of intervention and some typical examples.

I. Type of Intervention

A. Silence

> 30 seconds of silence after last utterance by analysand

B. Phonetic Interventions

Sounds indicating affirmation or astonishment

Sounds indicating confrontation, disapproval or negation

C. Explorative Interventions

Unspecific and specific inquiries

Encouragement for self-exploration

D. Clarifying Interventions

Differentiations

Amplifications

Recapitulations

E. Confrontational Interventions

Pointing out omissions

Pointing out contradictions

Direct contradiction

F. Interpretation-like Interventions

Mirroring

Implicit and explicit accentuation

Affective reinforcement

Combining conscious material

Combining conscious material with the analyst

G. Interpretations

Reconstructions

There-and-then interpretations

Current extra-transference interpretations

Not-current transference interpretations

Current transference interpretations

H. Structuring Interventions

Making arrangements and providing information about place, time, setting and theory of treatment

Information concerning the role of analysand

Information concerning the role of analyst

I. Directive Interventions

Requests concerning activity within the session

Requests concerning activity outside the session

Suggestions and advice

J. Formal Interventions

Saying hello and good-bye, using set social phrases

Saying kind words

Saying unkind words

K. Not elsewhere classifiable

In addition, each **interpretation** had to be evaluated on the following dimensions using a 5-point Likert scale:

II. Depth of interpretation (from too trivial to too deep)

III. Style of interpretation (from stereotypic to overadaptive)

IV. Correctness of interpretation (from completely wrong to completely correct)

V. Timing of interpretation (from much too early to much too late)

VI. Enlightening enhancement of interpretation (from too confrontational to too supportive).

We made a first attempt to establish content validity by having the AIRS evaluated by 5 psychoanalysts who had a completed training at an approved psychoanalytic institute and had many years of psychoanalytic experience. Their suggestions led to some improvements in the rating system, above all to clearer definitions.

Interrater reliability was assessed with the aid of three raters (B, W and K), all of whom had completed training at an approved psychoanalytic institute and had many years of psychoanalytic experience.

Rater training consisted of a careful study of the manual before the raters got together and then assessment of three 10-minute segments of an audiotaped session, which had also been transcribed, followed by an extensive discussion of each rating; total training time was between 4 and 5 hours.

Two pairs of sessions of a psychoanalysis that was totally audiotaped (the first author was the therapist) were transcribed according to the rules suggested by Mergenthaler (1986). The whole text spoken by the therapist was segmented by the first author; the units were chosen by following principles of clinical meaningfulness more than of automatic exactness; interrater reliability for segmentation was not assessed because it was done by only one person.

Three raters independently judged a total of 192 interventions by the therapist. Interrater reliability was assessed for the types of intervention (which were scored in a nominal category system) by computing kappa (Cohen 1960) for all three pairs of raters.* Kappa for pair B / W was .45; for pair B / K .50 and for pair W / K .51. Since the total number of interpretations was too small, the level of agreement between the pairs of raters for the dimensions of interpretation (depth, style, etc.) could not be computed.

In a first attempt to establish construct validity, the first author determined two "good hours" and two "bad hours" using a retroreport (Meyer 1981) he had made after each approximately 3 years ago; a "good hour" and a "bad hour" were close together in the course of psychoanalysis but the pairs came from different phases of the psychoanalysis; two of the judges (B and W) were blind to this assessment. The prediction was:

a) in the "good hour" there are more interpretations than in the "bad hour" and b) in the "good hour" the dimensions of the interpretation (depth, style, correctness, timing and enlightening enhancement) are scored better than in the "bad hour".

In order to infer construct validity the measure should operate in the predicted way.

As already stated, a statistical analysis could not be performed because of the limited number of judgements. A merely descriptive statistic showed that there was a clear tendency towards more identical rating of interpretations in the "good hour" and a tendency towards a better score on the dimensions of the interpretation. Moreover, though not predicted and therefore no measure of construct validity but in line with the theory of psychoanalytic technique, in the "good hour" the raters reached a higher level of agreement in scoring an interpretation than in the "bad hour". We also computed a chi-square test for all interventions and found no significant difference between "good hour" and "bad hour" for two judges (B and K) and a significant difference at the 10% level for the third judge (W).

We would like to thank Dr. Dipl. psych. G. Henrich, for his contribution to this paper.

Our first reaction to the unacceptably low interrater reliability (according to Landis and Koch (1977) a kappa of .41 to .60 is only moderate) was to omit some of the subcategories of the AIRS and to combine others under a new heading after a systematic analysis of the divergent scores on the item level. The modifications were:

B. Accompanying Interventions (instead of **Phonetic Interventions)**

Sounds indicating affirmation or astonishment

Paraphrasing

E. Confrontational Interventions

Sounds indicating confrontation, disapproval or negation

Indirect contradiction

Direct contradiction

F. Interpretationlike Interventions

Whole category deleted.

A new reliability and validity study was then conducted with this altered version of AIRS but with the same raters. Meanwhile two years had passed since the first evaluation, and we hoped that they all had forgotten the first round.

The second round with the new version of the AIRS was in May 1997. Kappa for pair B/W was now .66; kappa for pair B/K was .59; and kappa for pair W/ K was .60. Again the level of agreement between the pairs of raters for the dimensions of interpretation (depth, style, etc.) could not be computed because of the small number of interpretations.

Construct validity was again assessed by comparing "good hours" and "bad hours". On a descriptive level - because of the small number of judgements - we again found more ratings of interpretations in the "good hour" and a tendency towards better scores on the dimensions of the interpretation. In the "bad hour" either no interpretation was scored (hour 3) or there was a low level of agreement between the judges (hour 1), with one judge scoring interpretations frequently and the others scoring none. The chi-square test for all interventions showed significant differences between the "good hour" and the "bad hour" for all judges: for judge B at the 5% level, for judge K at the 10% level and for judge W at the 10% level; on the subcategory level C.0 (= unspecific and specific inquiries) and E.1 (= indirect contradiction) differed significantly at the 10% level between "good hours" and "bad hours".

We will try to refrain from anticipating what we hope will be a lively discussion but simply have to give some interpretations of the results - in a presentation dealing with interpretations we simply have to.

Clearly better interrater reliability was achieved with the new version of the AIRS. This was the result of deleting the category Interpretationlike interventions and replacing it with the subcategory paraphrasing, as a comparison between the distributions of frequency of interventions in the first and second investigations suggests. Five subcategories (mirroring, implicit and explicit accentuation, affective reinforcement, combining conscious material and combining conscious material with the analyst) were reduced to one subcategory (paraphrasing); it is obvious that this operation reduces the possibilities for disagreement between the judges, and it yielded a better inter-rater reliability, but what got lost are the interventions aiming at the preconscious of the analysand or, in other words, the different interventions that make the preconscious conscious are lost and in this regard clinical validity is reduced.

Now some comments on the validity study. For sure we would not claim that our instrument can distinguish between a "good hour" and a "bad hour" since the sessions were evaluated by simple clinical impression by one analyst, who moreover was one of the raters and therefore not blind before the rating procedure. But there must be a difference between the different pairs of sessions anyway and the AIRS is evidently able to reproduce it. Looking at the differences at the category level, it makes sense clinically that in the "bad hour" the analyst asks more and interprets less than in the "good hour".

I will stop now in order to have a "good half an hour" of discussion. Thank you for your patience and for your attention.