Observation and judgment in psychology: Assessing agreement among markings of behavioral events

BRUNO THOMANN

University of Zurich, Zurich, Switzerland

In a sequence from a videotaped dialogue between two persons (e.g., psychotherapist and patient), observers are given the task of searching for, marking, and commenting on conspicuous events. The observers receive no information on where to look for conspicuous events and no information on what events are worth marking. Moreover, nothing is known about the observers' marking preferences. Marking is therefore very spontaneous and subjective. This marking freedom leads to the problem that a superimposition of all observers' markings yields complicated configurations of mutually overlapping intervals. To address this difficulty, a new definition of the concept of marking agreement is proposed that is based upon relational considerations. According to this definition, a set of *agreeing marking* subsets of various sizes can be found in marking configurations. A significance concept is proposed to allow comparison of different configurations, and standardization on the basis of random configurations is performed using Monte Carlo simulations.

In the field of psychological judgment formation, one of the main areas of interest is the fundamental relationship between the *judgment* about a person and the individual *observations* made by the person judging. Persons being judged may be patients and therapists in a counseling context, subjects participating in research projects on judgment formation, or any other persons. Observers may be therapists, psychological experts, raters, or subjects.

Experience with the observation of video-recorded case material by several judges has shown that differences at the level of overall judgments are often associated with significant differences at the level of observations. Observers can agree on their overall judgments even if they arrive at them on the basis of different observations; similarly, they can observe the same individual event and deem it relevant to their judgment but nevertheless reach different overall judgments. Psychological practice shares this particular circumstance with everyday life experience. To gain empirical access to this fundamental issue, the concepts discussed below may offer some insight.

Guided Observing and Judging

Judgments in psychological practice and research are mostly of a distinctly integrative nature; to a greater or lesser extent, judgments integrate the speech content and patterns of behavior observed over a longer time period rather than being based explicitly upon individual obser-

vations made within this time period. Empirically oriented research on judgment formation generally makes use of predetermined judgment categories (diagnostic systems, questionnaires, assessment scales, content analytical classification of unstructured expressions of judgment, and so on); less frequently, post hoc judgment categories are used. Integrative judgments of this kind, however, are not considered in this paper. Instead, the focus here is on "local" judgments that are tied directly to individual observations. These are explained in more detail below.

Observations and associated observational methods differ as to whether they are implemented for a search task, an assessment task, or a combination of both.

Search tasks are used frequently in psychology. In recordings of psychotherapeutic sessions, dialogues, and interaction research conducted in various settings, for example, the observer typically searches for the occurrence of certain words, formulations, and turns of phrase, for the occurrence of certain facial and body expressions, for certain patterns of interaction, and so on. It is essential that an observer who is to carry out a direct search task in this manner receive advance instruction as to where to search and what to look for; that is, there must be prior agreement as to the exact nature of the searched-for events.

Assessment tasks, too, are naturally frequent in psychology. Where assessment involves the observation of video-recorded sequences, the material is usually segmented into more or less rigid units of observation for this purpose. For conversations, segments are commonly chosen to coincide with phrases, entire grammatical sentences, or contributions to the conversation as marked at

Correspondence should be addressed to B. Thomann, University of Zurich, Department of Clinical Psychology, Schmelzbergstrasse 40, CH-8044 Zurich, Switzerland (e-mail: thomann@klipsy.unizh.ch).

each end by a change in speaker. For studying nonverbal action, the rigid units frequently defined are single video frames or preset time intervals. Observers must then judge each of these segments in their entirety according to predetermined criteria, which means, for example, that each individual contribution by each speaker, no matter how long, is judged as a whole. This is essentially the same in studies observers and their assessment behavior are the object of research.

Strict guidelines of this kind greatly simplify the conception of empirical designs and are typical for the majority of empirically oriented studies in psychological judgment formation. For this reason, software packages for data collection and analysis developed commercially or at universities are as a rule geared toward guidelines of this kind.

Unguided Spontaneous Observing and Judging

A different situation arises when there are no such preset guidelines. In the following considerations, a hypothetical situation is presented that includes all problems typical of unguided observation procedures. This hypothetical situation is characterized as follows:

1. The observers' task is to search the material for any conspicuous event that they deem relevant for forming their judgment.
2. Observers do not have any information whatsoever on the location and nature of such conspicuous events—if they exist.
3. There is no information available on any possible observational or judgmental predilections of the observers that could be used to validate their ratings.

This situation is sometimes found in psychoanalytical contexts when new and unfamiliar material is assessed. Here, judgment formation depends strongly on ideas, associations, assumptions, attributions, and preliminary "local" suppositions and speculations that the therapist forms immediately on the basis of single observations and continues to compile into an overall judgment. See, for example, Sandler, Dare, and Holder (1992).

The same situation arises, however, in other psychological disciplines concerned with judgment formation. It always occurs when studies in assessment, attributions, and so on are based not only on standardized questionnaires, rating catalogues, and specially prepared "test" sequences, but also on examination of the subjects' individual observations that have led to a certain assessment.

Of course, this prototypical situation never appears in reality in the pure and strict form defined above. Neither in psychological practice nor in everyday life do we find pure, unguided, spontaneous observing and judging, and it is hardly something to strive for. Observing and forming judgments are always based on theories, whether explicit or implicit. The context in which observations and judgments are made, the type of material being assessed, and the observer's professional point of view are some of the main influencing variables.

The procedure here will be to construct a case that is "pure" from a methodological perspective to permit investigation of the fundamental methodological problems. This pure case is thus a kind of worst-case scenario—again from a methodological, but also perhaps a psychological, perspective. This worst-case scenario will serve as a reference point for real cases, in that they will differ from it and from each other in the type and degree of additional information that is available. This way of proceeding is thus somewhat similar to the mathematical principle of solving a particular problem by initially formulating a "general" solution and then embedding the practical cases as particular solutions under individual framework conditions (the additional information available).

The Basic Problem in Unguided Marking: Agreement

In the design of empirical research projects based on unguided, spontaneous observing, a fundamental problem arises, as illustrated by the following example, which will serve as the prototypical case: Observers are asked to mark individually all the places in a video-recorded sequence of a clinical dialogue between patient and therapist that seem relevant or noteworthy to them and to comment on them, as described above. When all markings of all observers are subsequently superimposed, we can expect to find *marking configurations* like the one shown in Figure 1.

The example was constructed for illustrative purposes. The somewhat chaotic appearance of these configurations is characterized by the following:

1. Each observer produces a different number of markings.
2. There is a free "choice" of marking points on the time axis.
3. The markings vary in length.
4. There is multiple reciprocal overlapping of the markings.

The actual appearance of a configuration—the degree of chaos in its appearance—is of course strongly dependent upon the degree of the effects of any of the influencing variables mentioned above on unguided spontaneous marking in the concrete case. We might expect or wish to think that markings by experts would result in less chaotic configurations. So far, studies have revealed that the opposite is the case. Particularly in the psychoanalytical context, configurations such as the one just presented are found frequently, even where the observers are all experts with similar training backgrounds.

Graphical representation of overlapping markings can take various forms:

1. Each observer may be represented by its own row, or number of rows, if some of the observer's own markings overlap each other. This can occur, but it will not be discussed here (see Thomann, 2001).
2. The markings are sorted in such a way that a minimum of rows is required for a representation that is free of intersections, as shown in Figure 1.

Figure 1. Superimposition of the markings of several observers. The observed material is represented by a time axis. Markings are intervals with two time addresses for the beginning and end of the event that they record. See text for further explanations.

There are advantages and disadvantages to each of the two forms of graphical representation. Method 1 allows a quick view of the originator of a marking and also of the number of markings by observers, but in large configurations, the total picture becomes very unclear. Unavoidably, large areas of empty places show up that make it difficult to gain a visual impression of the (local) distribution of density of the markings. Method 2 is an improvement in this regard, but the observers are represented less well (the markings have to be given an observer number, as in Figures 2 and 3). On-screen computer representations, on the other hand, can easily be programmed

to highlight the markings of a particular observer of interest. The sorting principle visualized in the Method 2 representation is, moreover, prerequisite to establishing the nuclei of a configuration, which will be described below.

To illustrate the problem, two empirically obtained configurations of some 500 and 400 markings, respectively, by a total of 17 observers are shown in Figures 2 and 3, which will be discussed below.

It is obvious that if such configurations of markings are to be allowed in empirical studies, an adequate definition of the agreement of overlapping markings is a central issue.

Figure 2. Configuration obtained from 509 markings of verbal events by 17 observers. The material contains 504 words and pauses in a conversational sequence of about 4 min. Here the units of the x-axis represent the words and pauses. The number above a marking denotes the observer. The number below the marking is an observer-related code made up of session and serial number: For instance, 8 (above) and 412 (below) is the code for the 12th marking made by Observer 8 in the fourth observation hour (fourth session). The markings are sorted—using a special algorithm not described here—in a manner that requires a minimum of rows for representation without intersections.

Figure 3. Configuration obtained from 409 markings of nonverbal events by 17 observers. The same material as in Figure 2, 4 min in length, is divided into some 1,150 units of 1/5 sec each (x-axis) to mark nonverbal events. Further details as in Figure 2.

A Definition of Agreement

When dealing with the issue of agreement or nonagreement between observations, raters, and so on, we immediately think of well-known agreement measures, such as Cohen's κ (Cohen, 1960), its extension to more than two raters (e.g., Fleiss, 1971), and the quite useful weighted κ (Cohen, 1968). For an overview, see Bortz, Lienert, and Boehnke (1990). There is also an interesting further development that takes entropy measures into consideration (see von Rennenkampff, 2000).

A look at Figure 4, however, reveals the basic difficulties entailed in the use of such measures. Apparently obvious *metric* criteria such as the formation of intersection or union sets for defining agreement (Figures 4a–4c) cannot succeed. The reason lies in the basal overlapping pattern in Figure 4d, with its different and complex variations, as can be seen in the upper areas of Figures 2 and 3. If there were no overlapping pattern of this kind, or if such overlapping patterns were prevented through cur-

tailoring marking freedom, there would be markings that still overlap but show only *one* shared average area (Figure 4e). In that case, appropriate metric criteria could be defined for the application of κ coefficients.

κ coefficients are also used in cluster analysis to investigate agreement (see, e.g., Breckenridge, 1989). Thus we could attempt to deal with configurations like those in Figures 2 and 3 through nondisjunctive overlapping cluster building (for overlapping clusters, see, e.g., Collins & Dent, 1988) or through so-called fuzzy cluster analysis (see, e.g., Miyamoto, 1990). The markings to be clustered would be seen as intervals in the form of linearly ordered sets of points (in this case, for instance, points are words or video frames). However, it would be difficult to find criteria for an overlapping or nonoverlapping cluster formation of such intervals. Once again, the main difficulty is the basic pattern shown in Figure 4d.

For this reason, the definition we propose for marking agreement goes back to a lower formal level in that it dis-

Figure 4. Problematic metric definition of agreement (see text for explanations).

Figure 5. Definition of nuclei. For further explications, see text.

penses with metric criteria. By applying purely relational concepts—based on the mathematical theory of so-called interval graphs (see, e.g., Fishburn, 1985; Hanlon, 1982; Korte, 1987)—the markings are taken as elements of the set of all superimposed markings. Overlapping is conceived of as a binary relation in the product set of the markings.

The following definition for agreement of markings is proposed: *A subset of markings in a set of markings is declared to be in agreement if (1) each marking in the subset overlaps with every other marking of the subset, and (2) the markings in the subset are not distinguishable from each other regarding their overlapping relation to the remaining markings of the configuration. Such subsets shall be called nuclei* (Figure 5).

In the example shown, we find three nuclei according to this definition (the markings shaded in gray). The second condition in the definition is not as apparent intuitively as Condition 1. It means that in a configuration, either *all* markings in a nucleus overlap with any marking outside the nucleus or *all* markings in a nucleus do *not* overlap with *any* marking outside the nucleus. For this reason, for instance, the subset made up of Markings 6, 7, 8, 9, and 10 cannot form a nucleus, for whereas Condition 1 is fulfilled, Condition 2 is not. If we examine here Marking 5 (outside the nucleus), we see that Markings 6, 7, and 8 overlap with Marking 5, but Markings 9 and 10 do not. That is, the markings of the subset do not behave—as required by Condition 2—*identically* to all markings lying outside the subset.

The ratio of nuclei markings to all markings of the configuration defines an *absolute degree of agreement*. The three nuclei shown in Figure 5 contain 7 of the total of 12 markings, which results in an absolute degree of agreement of about .58.

The determination of nuclei in extensive configurations is not a trivial matter; the number and size of nuclei are not ascertainable through visual inspection and cannot be estimated. Thus 340, or 67%, of the total of 509 markings in Figure 2 are nuclei markings, as opposed to just 181, or 44%, of the total of 409 markings in Figure 3. This

means that the absolute degree of agreement of .67 in Figure 2 is significantly greater than that in Figure 3 (.44), which is not necessarily visually apparent. An algorithm for determining the distribution of nuclei in any given configuration can be found in Thomann (2001). Note, however, that this degree of agreement is concerned with a purely formalistic agreement of the markings that is not necessarily accompanied by agreement of associated comments (see below).

Measure of Agreement

In order to be able to compare different configurations through the application of generally used statistical tests, we propose the use of a concept of statistical significance in lieu of absolute values—that is, a referencing of empirical configurations to random configurations. Suppose that a configuration of n markings contains k markings in z nuclei. We define *degree of organization* (DO) as a measure that can take on values between -1 and $+1$ (Figure 6). A value of -1 signifies the absence of nuclei in the configuration ($k = 0$, no agreement). The value of $+1$ is reached when all markings in the configuration are organized in nuclei ($k = n$, maximum agreement). A degree of 0 means that an empirical configuration contains a number of nuclei markings equal to the average number of markings that would occur in a random configuration ($k = k_{\text{r}}$, with k_{r} = the mean of the number of markings in nuclei that occur by chance).

A random distribution of nuclei is created by generating a random distribution of n markings by a random generator. The starting points of the n markings are randomly distributed over the material to be observed, for—according to definition—there is no a priori knowledge of noteworthy places in the material that would justify a weighting of the local distribution of random markings. In contrast, it makes sense for the length of the markings to show the same distribution as the empirical markings in order to function as a meaningful reference. In the empirical examples in Figures 2 and 3, for instance, the length of the markings is approximately Poisson distributed. A random generator weighted by these

Figure 6. Degree of organization (DO). For explanations, see text.

parameters produces configurations that contain an average of $k_- = 78.1$, or about 15%, nucleus markings for verbal behavior and $k_- = 80.5$, or about 20%, for non-verbal behavior. More will be said about these possibly unexpectedly high values in the next section.

There are various possibilities of establishing the course of the curve between the limits of -1 and $+1$. The two subintervals, $[0 \dots k_-]$ and $[k_- \dots n]$, maybe viewed separately and each given a linear course. Some alternative suggestions are discussed in Thomann (2001), such as measures based on information theory. The DO^+ and DO^- values associated with an empirical k value are expressed as a fraction of the maximum and minimum DO values, respectively:

$$DO^+ = (k - k_-)/(n - k_-) \quad (\text{for } k > k_-)$$

$$DO^- = (k - k_-)/k_- \quad (\text{for } k < k_-)$$

With the values from Figure 2 of $k = 340$, $k_- = 78.1$, and $n = 509$, the resulting value of DO is .61, whereas in the case of Figure 3, $DO = .31$. Comparison of these values with the absolute degrees of agreement of .67 and .44 mentioned above illustrates clearly the effect of the standardization. The contrast between agreement as it appears visually and agreement as empirically derived is even greater for Figure 3. Notice that the number of nuclei z does not enter into DO, but rather only the total number of nucleus markings k . The DO thus gives no information on the *distribution* of nucleic size in the configuration. Agreeing markings can be organized in a few large or in many small nuclei, both of which can lead to the same DO, but to different interpretations. A small number of large nuclei can be interpreted as the result of a strongly “unifying” influence of the material on the observers. Many small nuclei also signify agreement. However, this can be interpreted to mean that although a great many places in the material are marked by a collective of observers, the collectives are relatively small. In other words, the observers form only “local” alliances.

In analogy to DO, a measure could be introduced with which the empirical number of nuclei z in a configuration could be referenced to the number of nuclei z_- arising by chance. Such a “degree of concentration” also

having a range of values between -1 and $+1$ would be positive—for instance, if there are fewer nuclei in the empirical configuration than would occur in the random configuration (i.e., a strong concentration of k in a small number of nuclei)—and negative in the opposite case. For a detailed discussion of this, see Thomann (2001).

Monte Carlo Simulation of Random Configurations

The k distributions resulting from random configurations are determined through Monte Carlo simulations. Figure 7 shows the relationship between the number of markings n and the number and size of emerging nuclei for two different random distributions of marking lengths. In Figure 7A, the marking lengths are randomly and uniformly distributed within the interval $[1 \dots 200]$ units, whereas in Figure 7B they are randomly and uniformly distributed within the interval $[0 \dots 20]$ units. In both cases the material length is taken to be 1,000 units, within which the starting points of the markings are randomly and uniformly distributed. The diagrams on the left show in both cases a snapshot of $n = 100$ markings distributed by the random generator and sorted in the manner described above. Noteworthy on the one hand is the “resistance” to the development of strong nuclei with an increasing number of markings n , and on the other hand the strong dependency of nuclei formation on the length of the markings. With $n = 200$ long markings (Figure 7A), about $k = 23$ nuclei markings, or roughly 11%, are formed; with short markings (Figure 7B), however, this is about 30%. This dependence and the (possibly unexpectedly high) values of random agreement underline the importance of using a significance concept for empirical configurations instead of absolute values.

This also applies to the global relationship among number of markings n , their distribution of length, and the number of emerging nucleic markings k . This relationship is a rather complex one, and it yields—possibly again unexpected—local maxima and minima (Figure 8).

Some Remarks Regarding DO

DO and Cohen’s κ in nuclei. DO is obviously a weak measure relative to Cohen’s κ and measures that are based upon metric criteria. To utilize the power of κ in spite of this, κ could be calculated for nucleus markings only. However, it is easy to see that in configurations such as those in Figures 2 and 3, very small κ values would result that would lie far below the values of $\kappa > .7$ that are typically demanded in psychological research. In the single nucleus, there is as a rule only a small subset of observers represented (and from nucleus to nucleus a different subset), so that pairwise comparisons between all of the observers involved in the configuration naturally yield very low values of agreement.

DO and additional information. Nevertheless, in configurations such as those in Figures 2 and 3, there is agreement and structure that differ from completely random configurations. The DO represents that minimal

Figure 7. Relationship between number of markings n and number and size of developing nuclei. For explanations, see text.

amount of agreement information that can be derived from such configurations if no additional information is available or—for whatever reasons—additional information should not be utilized. In practical cases, such information, as noted, is available to a greater or lesser extent, such as, for example, category-forming marking objects (facial expressions, hand movements, and the like) in markings of nonverbal events. Asking for such information from the observer or taking account of such information given can be utilized for finding variants of, and a more precise, DO. The subdivision of markings in a “verbal” and a “nonverbal” group according to Figures 2 and 3 does in fact take account of additional information that the observer was asked to give and that led to two different DOs. Since all such conceivable variations

of DO start from a mutually shared “ideal” DO, there is better comparability among them.

DO range. The two configurations in Figures 2 and 3 show DOs of .61 and .31. Is this high or low? There is no way to tell as yet. A comparison with traditional κ value ranges is not very helpful, because the DO is about another form of agreement entirely. Just as for Cohen’s κ , some empirically established figures are needed, according to which, say, an increase in DO from .5 to .6 could be compared with an increase from .8 to .9 in an appropriate manner. The same holds with regard to determining validity criteria as well, which have been developed gradually in individual social scientific areas for Cohen’s κ and, for example, for correlation coefficients as well.

Figure 8. The number of markings in nuclei (k) as a function of the number of markings (n) and their maximum length. The lengths of the markings are randomly and uniformly distributed between 1 unit and the given maximum length (x -axis). Length of the material: 1,000 units.

Studying Growing Configurations

A possible strength of the DO may lie in its suitability for studying *growing* configurations. The starting point for consideration of growing configurations is the following question: What is the appropriate time span for working with a given set of material, and how many observers should be employed? To study this issue, the 17 observers already mentioned were asked to mark and comment on the material again and again in a total of six sessions of 45 min each. Figure 9a shows the growth of the markings with an increasing number of sessions and an increasing number of observers in the form of a matrix. Figure 9b shows the development of the corresponding DO.

Matrix 9a reads as follows: The first cell at top left contains, with $n = 5.41$, the number of markings made by 1 observer following *one* session as the average of the markings of all 17 observers. The cell immediately below that cell contains, with $n = 10.82$, the average number of markings made by all possible pairs (here there are 136 possible pairs) of the 17 observers. The cell to the right of the first cell contains, with $n = 11.00$, the total number of markings made by 1 average observer following two sessions (Sessions 1 and 2 taken together). This continues vertically in an analogous fashion with all observer trios, quartets, and so on, and horizontally for the accumulation of further sessions.

The matrix represents the total of $n = 509$ verbal markings shown in Figure 2 (cell at bottom right). It shows quite steady growth of n in both directions, which of course arises through the calculated averages mentioned above. In analogous fashion, Figure 9b reveals the changes—also mainly growth—in the average DO associated with

each average n . The value of .0528 for 2 observers and one session is thus the mean of all DOs of the 136 possible pairs of observers.

The initial question can now be reformulated in a new way and answered: Given a limit of 12 working hours in all, for instance, should 12 observers work 1 h each or 2 observers 6 h each? Both cases result in approximately the same configuration size (marked cells in Figure 9a), but they show substantial differences in agreement ($DO_{12 \times 1} = .41$ vs. $DO_{2 \times 6} = .11$, marked cells in Figure 9b). The answers to questions of this kind are thus scalable. In the present example, the result can be interpreted as follows: If a high level of agreement is of primary interest, then many observers should work for relatively short time periods; if, however, the objective is to obtain the greatest possible variety of different observations, the strategy of having few observers work over longer periods of time should be applied.

Further Applications

The concept of agreement outlined above opens up a number of new possibilities for research.

Agreement of markings versus agreement of the accompanying comments. On the basis of the assumption that comments—in the form as mentioned above, for example—are categorized in a suitable content-analytical manner, a configuration can be examined to determine whether or not an agreement of markings is associated with an agreement of corresponding comments. For instance, we can ask whether there is stronger agreement among commentaries in strong than in weak nuclei.

Studies on the marking behavior of observers. The definition of DO does not require that all observers make

Markings							DO						
obs. acc.	Sessions accumulated						obs. acc.	Sessions accumulated					
	1	2	3	4	5	6		1	2	3	4	5	6
1	5.41	11.00	16.59	21.35	26.24	29.94	1	-	-	-	-	-	-
2	10.82	22.00	33.18	42.71	52.42	59.88	2	.0528	.0935	.0900	.1066	.1165	.1139
3	16.24	33.00	49.76	64.06	78.71	89.82	3	.0978	.1686	.1737	.2002	.2262	.2188
4	21.91	44.29	66.32	85.92	105.47	120.49	4	.1488	.2219	.2357	.2771	.3034	.2944
5	27.42	55.46	83.39	107.92	132.40	151.02	5	.1938	.2749	.2991	.3435	.3681	.3572
6	32.63	66.19	99.71	128.51	157.88	180.22	6	.2288	.3230	.3557	.3931	.4155	.4038
7	37.98	77.16	116.29	149.71	183.89	209.79	7	.2604	.3634	.3976	.4313	.4525	.4400
8	43.39	88.13	132.88	171.14	210.16	239.71	8	.2929	.3949	.4390	.4685	.4845	.4700
9	48.77	99.10	149.44	192.50	236.37	269.61	9	.3270	.4282	.4758	.4975	.5108	.4955
10	54.30	110.28	166.17	214.08	262.89	299.91	10	.3567	.4574	.5064	.5209	.5310	.5160
11	59.78	121.37	182.89	235.72	289.40	330.06	11	.3838	.4831	.5335	.5416	.5499	.5344
12	64.93	131.98	198.80	255.72	314.43	359.17	12	.4076	.5057	.5580	.5540	.5636	.5518
13	70.55	143.35	216.23	278.84	342.02	389.76	13	.4303	.5283	.5771	.5769	.5823	.5656
14	75.76	154.00	232.24	298.94	367.29	419.18	14	.4504	.5444	.5952	.5873	.5937	.5789
15	81.18	165.00	248.82	320.29	393.53	449.12	15	.4715	.5604	.6093	.6012	.6057	.5897
16	86.59	176.00	265.41	341.65	419.76	479.06	16	.4914	.5743	.6217	.6127	.6154	.5992
17	92.00	187.00	282.00	363.00	446.00	509.00	17	.5106	.5864	.6299	.6236	.6250	.6078

(a)

(b)

Figure 9. Growing configurations. Matrix of accumulation of observers and observing sessions. DO, degree of organization.

the same number of markings. Therefore, the following question can be investigated: Do observers who mark considerably more events than others also mark more idiosyncratically? That is, do they produce a disproportionately greater share of markings that do not form nuclei, or do they simply agree with a correspondingly greater number of other observers? Individual observers in a collective of observers can be differentiated with regard to whether their markings lie mainly within nuclei or mainly outside nuclei. The former could be said to show "mainstream" marking behavior, whereas the marking behavior of the latter might be described as "idiosyncratic" or "exotic."

Temporal effects. How do the agreement measures change through the course of a longer observation period? Are the differences in marking among observers neutralized over the course of a given period of observation, due to the fact that observers tend to mark the same places in the material but in a differing temporal sequence? Or does a longer period of observation lead to an increase in marking differences because, after having reached agreement on markings of eye-catching places, individual preferences take on greater significance?

Comparing different observer categories. Observers that are a priori assigned to various categories (such as observers having a particular therapeutic orientation, experts as opposed to laypersons, or observers with or without prior information about the material to be investigated) can be differentiated according to the observa-

tions that they find relevant and also according to the degree of interrater agreement within their category. In this way, for example, experts could stand out by virtue of the fact that they not only mark more "relevant" phenomena than laypersons (however *relevant* is to be defined) but also show higher intermarker agreement in this among themselves. A further application is conceivable in the field of training for observation and judging, in that teachers' markings could be compared with students' markings.

REFERENCES

- BORTZ, J., LIENERT, G. A., & BOEHNKE, K. (1990). *Verteilungsfreie Methoden in der Biostatistik*. Heidelberg: Springer-Verlag.
- BRECKENRIDGE, J. N. (1989). Replication cluster analysis: Method, consistency, and validity. *Multivariate Behavioral Research*, **24**, 147-161.
- COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement*, **20**, 37-46.
- COHEN, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213-220.
- COLLINS, L. A., & DENT, C. W. (1988). Omega: A general formulation of the Rand-Index of cluster recovery suitable for non-disjoint solutions. *Multivariate Behavioral Research*, **23**, 231-242.
- FISHBURN, P. C. (1985). *Interval orders and interval graphs*. New York: Wiley.
- FLEISS, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**, 378-382.
- HANLON, P. (1982). Counting interval graphs. *Transactions of the American Mathematical Society*, **272**, 383-426.
- KORTE, N. (1987). *Intervallgraphen und MPQ-Bäume*. Unpublished doctoral dissertation, Aachen University of Technology.

- MIYAMOTO, S. (1990). *Fuzzy sets in information retrieval and cluster analysis*. Dordrecht: Kluwer.
- SANDLER, J., DARE, C., & HOLDER, A. (1992). *The patient and the analyst: The basis of the psychoanalytic process*. London: Karnac.
- THOMANN, B. (2001). *Iteratives Signieren und Beurteilen audiovisuellen Materials*. Frankfurt: Verlag für Akademische Schriften.
- VON RENNENKAMPFF, J. (2000). *Estimation of theoretical characteristics of a new measure of agreement for multiple categorical ratings*. Unpublished master's thesis, University of Zurich.

(Manuscript received October 25, 2000;
revision accepted for publication June 16, 2001.)