Psychology Science, Volume 47, 2005 (3/4), p. 315-325

Psychotherapy research based on Two-Sample CFA and Markov-Chain Analysis

JOACHIM KRAUTH[1]

Abstract

A combination of configural frequency analysis (CFA) and time-series analysis is proposed for studying the effects of psychotherapy. By means of Two-sample CFA the multivariate behavior patterns observed in clients for each session are reduced to a zero-one criterion. For the resulting binary sequences a first-order Markov dependence is assumed. Using results on intercalary independence and the truncation property of Markov processes we can test nonparametrically for therapy effects in single-case studies. Procedures for combining and comparing the effects of independent single-case studies are described. The performance of a single-case analysis is demonstrated for an empirical data set from therapy research.

Key words: Psychotherapy research, Two-sample CFA, single-case studies, time-series analysis

[1] Prof. Dr. J. Krauth, Department of Experimental Psychology, University of Düsseldorf, D-40225 Düsseldorf, Germany; email: J.Krauth@uni-duesseldorf.de

1. Introduction

For several reasons it is difficult to show that psychotherapies have positive effects. One problem is that it is quite often not clear which kind of effect the researcher is looking for. In Krauth (1983a) we made a difference between the measurement of effect, of success and of change. In most cases, the therapist, either in cooperation with the client or alone, formulates the target of the therapy, and if both therapist and client agree at the end of the therapy that this object has been achieved it is claimed that the therapy was successful. It is obvious that this kind of measuring the success of a therapy is very subjective. Often a multivariate criterion of success is used and the therapy is claimed to be successful if at least some of the components of the criterion seem to be improved. In most cases it is neither asked whether the criterion is realistic, e.g. whether it could be achieved by a normal subject, nor whether it reflects the behavior of a normal subject. For these reasons it is often only claimed at the end of a therapy that the therapy has had a positive effect, i.e. the state of the client seems to be improved in comparison with the beginning of the therapy even if this observed effect may be in no relation to a criterion of success which had been formulated before the therapy started. In practice, even this very weak criterion of a positive outcome is often not applicable and at least the therapist may be often inclined to formulate that the therapy has been useful for the client because it has caused a certain change of behavior and this is considered as a kind of success by itself.

In my opinion, neither an observed success nor an observed effect nor an observed change of behavior can justify post hoc the performance of a therapy. Even if we could perform a controlled randomized double-blind study – which is not really possible in psychotherapy – in order to measure the effectiveness of a therapy on the basis of a criterion of success, effect or change as described above, it cannot be ruled out that a given therapy had no effect at all or even a harmful effect. The reason for this is obvious: Anything may happen if subjective or arbitrary measures of success (or effect or change) are used.

In the following we propose a procedure by which at least some difficulties connected with the considered problem are addressed, e.g. the selection of an objective multivariate criterion and the dependence of observations in longitudinal studies.

2. Selection of a criterion

The first step is the selection of an appropriate measuring device or test by which the behavior of a client can be measured at given points of time. The result of a measurement can be a multivariate vector of measurements where we assume that the dimension of the vector (related to one point of time) is not too high (a maximum of 6 components might be appropriate). For example, we might consider the items or subtests of a global rating scale describing the behavior or state of health of the client. Here, it is important that this scale is sensitive to change (Krauth, 1983b; Krauth, 1995, pp. 297 ff.) i.e. is suited to reflect objective changes of the client in an appropriate way. Most psychological scales cannot be used for this purpose because they are optimized only with respect to the measurement at only one point of time. A typical example of a test which obviously is not sensitive to change though it is often used in repeated-measurements designs is the MMPI.

In a second step we need a sample of clients from the same population for which we perform our therapy study and a corresponding sample of normal and healthy subjects which is similar to the first sample with respect to gender, age, and education. To both samples our rating scale is applied. The resulting measurement vectors are analysed by means of a Two-sample CFA (e.g. Krauth, 1993, Chapter 7). As a result we may find configurations which are typical for clients, others which are typical for normal subjects and still others which do not discriminate between the two populations.

Several cases can be distinguished. The worst case occurs if no discriminating types are discovered, i.e. if there do not exist configurations which are either typical for clients or for normal subjects. One reason for this result might be that the sample sizes are too small to detect any types. A second possible reason may be that the potential clients do not differ from normal subjects and therefore do not need any therapy. A third explanation might be that apparently normal subjects in reality need a therapy. However, the most plausible explanation seems to be that the validity of our measuring device is too low and thus differences between normal subjects and potential clients cannot be detected.

The best case occurs if we have one subclass of configurations which are displayed only by clients, a second subclass of configurations which are displayed only by normal subjects, and a third subclass of configurations which are displayed by no subjects at all. Here, the last subclass may be empty. However, in practice it is very unlikely that this best case is ever observed.

In general, a Two-sample CFA will identify some configurations which may serve as discrimination types, i.e. which are typical for exactly one of the two populations while nothing is said with respect to the other configurations. For these other configurations it may occur that they are exhibited only by subjects of one population but in many cases subjects of both populations will exhibit the same configurations. This latter case may also occur for configurations that were identified as types by CFA. Depending on the number of dimensions of the used criterion and the samples sizes the probability of finding configurations that are exhibited by no subjects at all may be quite high.

In a third step we have to partition the set of configurations into two subsets. The configurations of one subset are labeled by 1 and those of the other subset by 0. Here, different definitions are possible:

i) If the main target of a therapy is that clients should adopt the behavior pattern of normal subjects, we assign a 1 to all configurations which are typical for normal subjects or which occur only for normal subjects. To all other configurations we assign a 0.
ii) If the main target of a therapy is that clients should lose their typical client behavior, we assign a 1 to all configurations which are typical for clients or which are exhibited only by clients, and a 0 to all remaining configurations.

These definitions can be modified in more than one respect. E.g., definition (i) could be strengthened by assigning a 1 only to configurations which are identified by CFA to be typical for normal subjects. In a similar way definition (i) could be weakened by assigning a 1 to all those configurations where the majority of subjects which exhibit such a configuration are normal subjects.

3. Single-case analysis

In Section 2 we described in which way a multivariate time series of client behavior during a therapy can be transformed into a binary sequence of 1's (target behavior) and 0's (other behaviors). In case (i) our alternative hypothesis would be that the probability for a 1 is increased by therapy (positive trend) while the null hypothesis would be that this probability does not change or even decreases (no trend or even a negative trend). In case (ii) we would assume under the alternative hypothesis that we have a negative trend while under the null hypothesis we assume that no trend or even a positive trend is exhibited. If we assume for the moment that our binary sequence is a Bernoulli sequence, i.e. that the trials are independent, we can use a one-sided trend test to test for a therapy effect. A nonparametric asymptotically optimum test for this situation is based on the Spearman rank correlation test (Krauth, 1979a) where the pairs are formed by the observed 1's and 0's on the one side and the corresponding points of time on the other side. In Krauth (1988, pp. 155-157) it is described how to perform the exact Spearman rank correlation test. In our particular case, one variable exhibits only two different values (1 and 0). In this case we might compare the sample of points of time connected with behavior 1 with the corresponding sample of points of time connected with behavior 0 by means of the exact Wilcoxon rank-sum test (Krauth, 1988, pp. 49-53).

However, unfortunately the mentioned tests are only valid if our observations are independent and this assumption is rather unrealistic for the single-case studies under consideration. In order to permit at least a certain degree of dependence we assume in the following that our binary sequence can be described by a first-order Markov chain, i.e. we assume that a category of behavior at a given point of time depends only on the category of behavior at the preceding point of time but not on the complete sequence of behavior categories which starts at the beginning of the therapy. In a more formal way we express this by the following equality for conditional probabilities:

$$P(X_T = x_T \mid X_{T-1} = x_{T-1}, X_{T-2} = x_{T-2}, ..., X_1 = x_1) = P(X_T = x_t \mid X_{T-1} = x_{T-1})$$

for all possible combinations of the

$$x_i \in \{0,1\}, i = 1,...,T, \text{ where } T \in \{2,3,...\}.$$

The null hypothesis of no trend in the probabilities is then expressed in that way that we have a binary homogeneous stationary first-order Markov chain. In contrast to the opinion of some authors it is not allowed even in such a restricted situation to use the usual nonparametric tests for independent data (Krauth, 1980a).

We admit, that the assumption of a first-order Markov chain may be too weak to describe the dependence structure of empirical behavior sequences in an adequate way. At least in theory, we can extend the following approach to higher-order Markov chains. For example, in a second-order Markov chain the present category of behavior depends on the behavior categories for the last two preceding points of time. One reason for restricting ourselves to demonstrate the procedure here only for first-order Markov chains is that we propose here very conservative exact tests, and behavior sequences in therapy research are, in general, not

very long. The tests are the more conservative, the higher the order of the Markov chain. Thus, if we do not have very long sequences we cannot expect that our tests will yield significant results in the case that higher-order chains are assumed.

Our approach is based on the properties of „intercalary independence" and the „truncation property" of general Markov processes. These properties were proved by Dufour and Torrès (2000) though intercalary independence has been considered already by Ogawara (1951) who states that U.V. Linnik used it as early as 1949. In Krauth (2005) we consider the special case of a binary homogeneous stationary first-order Markov chain. The results for this special case can be summarized as follows, where we assume that the number (T) of points of time is odd (otherwise we omit the last trial):

(1) Intercalary independence: If we fix the values for the trials with an odd number the random variables for the trials with an even number are conditionally independent.
(2) Truncation property: The conditional distribution of a random variable corresponding to an even trial does not depend on the set of the values for all odd trials but only on the values for the immediate two odd neighbors of the even trial.
(3) Only three different conditional distributions of the random variables for the even trials are possible under the null hypothesis: One for the neighbors (1, 1), one for the neighbors (0, 0), and one for the neighbors (1, 0) or (0, 1).

Thus, if we fix the values for the odd trials the random variables for the even trials can be considered as a sequence of independent variables which is composed of three subsequences of independent identically distributed random variables. Because the values for the odd trials are assumed to be known, we know which and how many of the even trials belong to one of the subsequences. To any of these subsequences we can apply an exact Spearman rank correlation test or the exact Wilcoxon rank-sum test as described above.

Since we can perform three independent tests for identical test problems, we have to decide which kind of strategy we should use in order to utilize the available information in an efficient way. Four obvious strategies can be thought of:

- *Strategy 1:* We perform all three one-sided tests and use an alpha-adjustment, e.g. the Bonferroni or Holm procedure (Krauth, 1988, pp. 36-38). However, considering the independence of the tests and assuming the same value of the significance bound (α) for all tests a more efficient adjustment is given by comparing the three p-values with $1-(1-\alpha)^{1/3}$. This is a consequence of a result given in Krauth (1988, p. 35).
- *Strategy 2:* Since the one-sided test problem is identical for the three independent tests we could consider only the smallest p-value and compare this value with $1-(1-\alpha)^{1/3}$ (Tippett, 1931). Here, again the same significance bound (α) is assumed for all three tests.
- *Strategy 3:* Because we propose to use consistent tests (i.e. the Spearman test or the Wilcoxon test) the power of the test increases with the sample size. Thus, we might perform only one test for the longest subsequence.
- *Strategy 4:* Because we have three independent tests with identical one-sided test problems we can also combine the p-values of the three tests into one p-value which is compared with the significance bound (α). Several procedures of this kind were proposed. For example, we might use the Edgington procedure (Edgington, 1972) where we use the

sum s of the three p-values. If $s \leq 1$ holds, the total p-value is given by $(s^3/6)$, for $1 < s \leq 2$ by $(s^3 - 3(s-1)^3)/6$, and for $2 < s \leq 3$ by $(s^3 - 3(s-1)^3 + 3(s-2)^3)/6$.

For selecting an appropriate strategy the following advice might be of use: If one of the three subsequences is considerably longer than at least one of the two others, Strategy 3 should be used. The reason for this is that in general short subsequences will cause large p-values due to the low power of tests for small sample sizes. If all three subsequences have about the same length, Strategy 4 should be chosen. Strategies 1 and 2 are not recommended.

4. Combination and comparison of the results of samples of independent single case studies

A well-known problem of single-case analysis in therapy research is the short length of the observed time series. Often this may be the reason for the observation that the null hypothesis cannot be rejected by our tests due to the small power of the tests. If we observe such short time-series not only for one client but for a sample of clients who get the same treatment and who cannot influence each other which means, e.g., that they are not participating in the same therapy group, we can increase the efficiency of our tests by pooling the results. If we have obtained the time-series data for n independent clients we might apply Strategy 3 in Section 3 to each client and derive in this way n p-values. These can be combined by means of the Edgington procedure (Edgington, 1972) in the following way: We calculate the sum (s) of the n p-values. For $s \leq 1$ we get the total p-value $p_T = s^n/n!$ where $n! = 1 \times 2 \times \ldots \times n$. For $s > 1$ we calculate p_T by

$$p_T = \frac{s^n}{n!} - \frac{(s-1)^n}{1!(n-1)!} + \frac{(s-2)^n}{2!(n-2)!} - \ldots + (-1)^i \frac{(s-i)^n}{i!(n-i)!}.$$

Here, i denotes the largest integer with $i < s$. For $p_T \leq \alpha$, we can reject the null hypothesis.

If n is a large number and $s > 1$ holds, the Edgington procedure may become numerically unstable because differences of very large numbers have to be computed. This may yield wrong values for p_T (e.g. $p_T < 0$ or $p_T > 1$). Sometimes it helps to replace s by $(n - s)$ and to use then at the end of the calculations $(1 - p_T)$ instead of p_T. Alternatively one might use the procedure proposed by Fisher (Fisher, 1932).

Another problem is the interpretation of a significant result. Without a control group we cannot conclude that a significant trend has been caused by the therapy. We cannot rule out that such an apparent effect might have occured even if no treatment had been applied at all.

Unfortunately, it is difficult in therapy research to split a sample of clients randomly into two subsamples, where one subsample gets the therapy while the other subsample is not treated at all or gets a control condition. Here, a waiting group of clients would be of not much use because it is most probably impossible to observe for this group a time series under similar conditions as for the treatment group. One possible control condition might be a

standard therapy which is to be compared with a new therapy (treatment condition). Sometimes it may be possible to consider a sham treatment in the control group, e.g. an autogenic training or progressive relaxation. These are well-known and generally accepted treatments and few problems are to be expected in contrast, e.g., to the repeated application of a drug which in reality is a placebo.

If we have succeeded in establishing two groups of independent clients who have been randomly assigned to a treatment and a control group, we can observe for each client a time series and derive for this, as explained in Section 3, a *p*-value. We may consider the *p*-values as measures of effect and can compare the two samples of independent *p*-values by means of the exact Wilcoxon rank-sum test (Krauth, 1988, pp. 49-53). Here, it is important that the lengths of the time series in the two samples do not differ because longer time-series may be responsible for smaller *p*-values due to a larger power of the single-case tests. It would be the best if all clients in both groups would underlie exactly the same time pattern.

5. Empirical example

I am very much obliged to my colleague Professor Reinhard Pietrowsky from the clinical section of the Institute of Experimental Psychology at the University of Düsseldorf who made available to me the copy of an original Fear Diary from which only the name of the client had been cut off. In this diary the client had noted for each day during thirteen weeks, i.e. for a total of 91 days, whether he or she had experienced fear.

For each day up to four different fear attacks could be scaled. First, for each fear attack the maximum fear which was experienced was rated on a scale with the eleven ordered categories 0, 1, ..., 10. Second, up to fifteen different symptoms could be enumerated which were coded by the integers 1, 2, ..., 15. Here, the score 15 corresponded to „other symptoms". Third, the location where the fear attack occurred had to be described. Fourth, it was asked whether the fear attack was expected or unexpected. Fifth, it was asked whether the client was accompanied by a person whom he or she trusted or whether this was not the case. Finally, the client had to rate the average fear which he or she experienced during the day apart from the fear attacks mentioned above using again the rating scale with the eleven ordered categories 0, 1, ..., 10.

Thus 6 dependent variables for at most 4 fear attacks were measured each day. By forming classes of similar values for those variables which are based on more than two categories it is possible to generate multivariate fear patterns for each day which allow an easier interpretation than the original data. If we had samples of Fear Diaries from independently responding clients and normal subjects it might be possible to identify by means of a Two-sample CFA fear patterns which are typical exclusively for one of the two populations. As described in Section 2 we could use this information to transform the multivariate time-series of fear responses for each client into a binary sequence which could then be further analysed by the procedure described in Section 3.

Unfortunately, only one Fear Diary was available and therefore it was not possible to use the procedure described in Section 2. In order to be able to illustrate at least the single-case analysis described in Section 3 we coded a day by a 1 if at least one fear attack was reported and by a 0 otherwise. The result of this transformation is given in Table 1. Since we have altogether 91 days which is an odd number of days we can analyse the complete sequence

and it is not necessary to omit the data of the last day. Now, we derive from this sequence three subsequences as described in Section 3. The result is given in Table 2. To explain this procedure consider the first three values 0, 0, 0 for the days Mo, Tu, We of week 1. The first (0) and the third (0) value belong to odd days (1 and 3) while the second value (0) belongs to an even day (2). Therefore, the first value of the (0, 0) subsequence is a 0. Next, we consider the three values 0, 1, 0 for the days We, Th, Fr of week 1. The first (0) and the third (0) value belong to odd days (3 and 5) while the second value (1) belongs to an even day (4). Thus, the second value of the (0, 0) subsequence is a 1. Then, we consider the three values 0, 0, 0 for the days Fr, Sa, Su of week 1. Because the second value is a 0, this is the third value in the subsequence (0, 0). After this we consider the days Su in week 1 and Mo, Tu in week 2 with the values 0, 0, 0, yielding the fourth value (0) of the (0, 0) sequence. The days Tu, We, Th in week 2 follow with values 0, 1, 1 yielding the first value (1) in the (0, 1)/(1, 0) subsequence. In this way we derived the 19 values of the (0, 0) subsequence, the 6 values of the (1, 1) subsequence, and the 20 values of the (0, 1)/(1, 0) subsequence.

Table 1:
Fear Diary of a client for 13 weeks where a 1 denotes a day with at least one fear attack

Week	Mo	Tu	We	Th	Fr	Sa	Su
1	0	0	0	1	0	0	0
2	0	0	1	1	0	1	0
3	0	1	0	1	1	1	0
4	1	0	1	0	1	0	1
5	0	1	1	0	1	1	0
6	0	0	0	0	0	0	0
7	1	0	1	0	1	1	0
8	0	0	0	1	1	0	0
9	1	1	1	1	0	1	0
10	0	1	0	0	0	0	0
11	0	0	1	0	0	0	1
12	0	1	1	0	0	0	0
13	1	0	0	0	0	0	0

Table 2:
Subsequences of behavior at even days for the 3 possible cases (0, 0), (1, 1), and (0, 1)/(1, 0) of odd neighbors

Odd neighbors	Behavior at even days
(0, 0)	0 1 0 0 1 1 1 1 1 0 0 0 0 1 0 0 0 0 0
(1, 1)	0 0 0 0 1 0
(0, 1)/(1, 0)	1 0 1 1 1 1 0 1 0 1 0 1 0 0 0 0 0 1 0 0

We performed exact Wilcoxon rank-sum tests for each subsequence using the program DISFREE (Krauth, 1989). For example, for the subsequence (0, 0) this means that we compared the sample (1, 3, 4, 10, 11, 12, 13, 15, 16, 17, 18, 19) i.e. those points of time which correspond to no fear attack with the sample (2, 5, 6, 7, 8, 9, 14) of points of time corresponding to a fear attack. Our one-sided alternative hypothesis was that the probability of a fear attack decreases with time. The corresponding one-sided p-values were .05992 for subsequence (0, 0), .83333 for subsequence (1, 1), and .01906 for subsequence (0, 1)/(1, 0). If we set $\alpha = .05$ we have $.05/3 = .01667$ and $1-(1-.05)^{1/3} = .01695$. Because of $p(01/10) > .01667$ and $p(01/10) > .01695$ all three procedures mentioned in Strategy 1 of Section 3 (i.e. Bonferroni, Holm, α adjustment for independent tests) fail to give a significant result. The same is true for Strategy 2. Strategy 3 gives a significant result because (0, 1)/(1, 0) is the longest subsequence and we have $p(01/10) < .05$. According to Strategy 4 we calculate

$$(.05992 + .83333 + .01906)^3 / 6 = .12655 > .05,$$

i.e. we get no significant result.

Since the subsequence (0, 1)/(1, 0) with 20 values is considerably longer than the subsequence (1, 1) with only 6 values we should have chosen Strategy 3 according to the advice given at the end of Section 3. Then a significant result would have been obtained.

Since we have only one Fear Diary, the procedures of Section 4 cannot be illustrated here because no results for independent single-case studies are available which could be combined or compared.

6. Discussion

It is well-known that it is difficult and perhaps even impossible to demonstrate the effectiveness of a psychotherapy. For this there exist several reasons: The measurement of success or effect of a therapy is based in general on subjective concepts of client and therapist. Therapy is considered to be rather an art than a scientific method which makes any standardization difficult. It is often not possible to establish even the minimum requirements of experimental research, e.g. randomization and an appropriate control. Statistical evaluation is difficult in view of short time series of varying lengths and missing data.

Here, we described how a multivariate time series describing the behavior of clients sometimes may be condensed into a binary sequence with a minimum loss of relevant information. To achieve this a Two-sample CFA was proposed. Under the possibly problematic assumption that the dependence structure of the binary sequence can be described by a first-order Markov chain we proposed nonparametric tests for testing for a therapy effect. Further, we showed how results from single-case analyses for samples of subjects can be combined and compared.

We do not deny that our proposals cannot really solve the most important problems which prevent a scientific evaluation of the effects of psychotherapies. But we hope that some of our ideas might help researchers in this field.

In the rare case that single-case studies are available for samples of subjects we have previously proposed to perform nonparametric response curve analyses (Krauth, 1973, 1980b).

This approach has also been combined with CFA (Krauth and Lienert, 1978). In this kind of analysis the dependence structure of the longitudinal data is of minor importance because only the trend curves are considered. The assumption of a Markov chain dependence should hold at least approximately for the specific kind of analysis of psychotherapeutic processes in group designs which was discussed in Krauth (1979b).

If we have no group design but only single subjects, assumptions about the dependence structure of the data have to be made. Some models for this kind of assumptions were presented in Krauth (1981). Unfortunately, many authors use analyses which have no justification as has been discussed in Krauth (1986, 1990, 2000b). Only one method by which single-case experiments can be performed and analysed is known that allows a causal interpretation of the results. This was originally invented by Fisher (1966) and was propagated by Edgington (1967; 1995, Chapter 12) and also used and discussed by us (Wurthmann et al., 1996; Krauth, 2000a, Chapter 9). Unfortunately, this approach requires a true double-blind study which is difficult to establish in psychotherapy. Therefore, the approach which we presented here may be sometimes a useful alternative.

References

1. Dufour, J.M. & Torrès, O. (2000). Markovian processes, two-sided autoregressions and finite-sample inference for stationary and nonstationary autoregressive processes. Journal of Econometrics, 98, 255-289.
2. Edgington, E.S. (1967). Statistical inference from N = 1 experiments. Journal of Psychology 65, 195-199.
3. Edgington, E. S. (1972). An additive method for combining probability values from independent experiments. Journal of Psychology 80, 351-363.
4. Edgington, E.S. (1995). Randomization tests. Third edition, revised and expanded. New York: Marcel Dekker.
5. Fisher, R.A. (1932). Statistical methods for research workers. Edinburgh: Oliver and Boyd, 4th edn.
6. Fisher, R.A. (1966). The design of experiments. Eighth edition. Edinburgh: Oliver and Boyd.
7. Krauth, J. (1973). Nichtparametrische Ansätze zur Auswertung von Verlaufskurven. Biometrische Zeitschrift 15, 557-566.
8. Krauth, J. (1979a). Generalized sequential ranks and tests of randomness. Mathematische Operationsforschung und Statistik, Series Statistics, 10, 291-298.
9. Krauth, J. (1979b). Nonparametric interaction analysis for psychotherapeutic processes. Behavioral Analysis and Modification, 3, 98-108.
10. Krauth, J. (1980a). Possible misinterpretations when evaluating psychological time series. Archiv für Psychologie, 133, 139-147.
11. Krauth, J. (1980b). Nonparametric analysis of response curves. Journal of Neuroscience Methods, 2, 239-252.
12. Krauth, J. (1981). Nichtparametrische Ansätze bei Zeitreihenanalysen. In: Janke, W. (Ed.) Beiträge zur Methodik in der differentiellen, diagnostischen und klinischen Psychologie. Festschrift zum 60. Geburtstag von G. A. Lienert. 26-45. Meisenheim: Hain.
13. Krauth, J. (1983a). The present state in the measurement of success and change. In: Lüer, G. (Ed.) Bericht über den 33. Kongreß der Deutschen Gesellschaft für Psychologie, Mainz 1982. 769-776. Göttingen: Hogrefe.

14. Krauth, J. (1983b). Bewertung der Änderungssensitivität von Items. Zeitschrift für Differentielle und Diagnostische Psychologie, 4, 7-28.
15. Krauth, J. (1986). Probleme bei der Auswertung von Einzelfallstudien. Diagnostica, 32, 17-29.
16. Krauth, J. (1988). Distribution-free statistics. An application-oriented approach. Amsterdam: Elsevier.
17. Krauth, J. (1989). DISFREE. Cambridge, U.K., BIOSOFT.
18. Krauth, J. (1990). Comments on the paper by Möller et al. (1989): Problems in single-case evaluation. European Archives of Psychiatry and Neurological Sciences, 239, 391-394.
19. Krauth, J. (1993). Einführung in die Konfigurationsfrequenzanalyse (KFA). Ein multivariates nichtparametrisches Verfahren zum Nachweis und zur Interpretation von Typen und Syndromen. Weinheim/Basel: Beltz, Psychologie-Verlags-Union.
20. Krauth, J. (1995). Testkonstruktion und Testtheorie. Weinheim/Basel: Beltz, Psychologie-Verlags-Union.
21. Krauth, J. (2000a). Experimental design. A handbook and dictionary for medical and behavioral research. Amsterdam: Elsevier.
22. Krauth, J. (2000b). Einzelfalluntersuchungen in der Neuropsychologie – Eine Bemerkung zu Brüggemann, Ruoß und Weber (1999). Zeitschrift für Neuropsychologie, 11, 262-266.
23. Krauth, J. (2005). Test for a change point in Bernoulli trials with dependence. In: Weihs, C., Gaul, W. (Eds.) Classification: the Ubiquitous Challenge, 346-353. Berlin: Springer.
24. Krauth, J., Lienert, G.A. (1978). Nonparametric two-sample comparison of learning curves based on orthogonal polynomials. Psychological Research, 40, 159-171.
25. Ogawara, M. (1951). A note on the test of serial correlation coefficients. Annals of Mathematical Statistics, 22, 115-118.
26. Tippett, L.H. (1931). The methods of statistics. London: Williams and Norgate.
27. Wurthmann, G., Klieser, E., Lehmann, E., Krauth, J. (1996). Single-subject experiments to determine individually differential effects of anxiolytics in generalized anxiety disorder. Neuropsychobiology, 33, 196-201.