REPLICABILITY OF RATING THE CORONARY-PRONE
BEHAVIOUR PATTERN*

BY

C. DAVID JENKINS, Ph.D., RAY H. ROSENMAN, M.D., AND
MEYER FRIEDMAN, M.D.

From the Department of Epidemiology, School of Public Health, University of North Carolina,
Chapel Hill, N.C., and the Harold Brunn Institute, Mt. Zion Hospital and Medical
Center, San Francisco, Calif.

The possible involvement of a specific overt behaviour pattern in the pathogenesis of clinical
coronary heart disease (CHD) has been under careful
study. The characteristics of this overt behaviour
pattern, called Behaviour Pattern A, have been
described previously (Friedman and Rosenman,
1959). In brief, it is characterized by extremes of
competitiveness, striving for achievement, aggres-
siveness (yet sometimes stringently repressed), haste,
impatience, restlessness, hyper-alertness, explosiv-
ness of speech, tenseness of facial musculature, and
feelings of being under the pressure of time and under
the challenge of responsibility. In an earlier study
this pattern was found to be associated with increased
prevalence of CHD, Rosenman, Friedman, Straus,
Wurm, Kositchev, Hahn, and Werthessen (1964).
To test this finding more decisively, a collaborative
prospective study of CHD was initiated in 1960-61
among 3,524 men aged 39–59 years. At the end of a
mean follow-up period of 24 years, subjects classified
at intake as having exhibited Pattern A showed a
considerably higher incidence of new CHD than
those classified prospectively as exhibiting the
converse Pattern B (Rosenman, Friedman, Straus,
Wurm, Jenkins, and Messinger, 1966).

It therefore appeared important to the investigators
to examine the adequacy of their technique for
differentiating Pattern A from Pattern B. The current
study tests the reliability of assessment of the
behaviour patterns, both in terms of agreement
between judges and stability of the rating over time.

The behaviour pattern was assessed by means of
an oral interview designed especially for this
purpose. Detection of Behaviour Pattern A consists
in determining whether a subject is in active and
continuous conflict either with other subjects or with
time. This condition reveals itself through various

---

* This study was supported by U.S. Public Health Service research
grants HE-03429, HE-05121, and HE-10326, and by grants from the
American Heart Association and the Irwin Strasburger Memorial
Medical Foundation of New York.

motor, emotional, and intellectual attributes
described below. The correct classification of a subject
depends more upon the motor and emotional
qualities accompanying his response to specific
questions than the actual content of his answers.

During the personal interviews in the Western
Collaborative Group Study, four general features of
respondent reaction were studied:

(1) Presence of characteristic motor signs.
(2) Degree of drive and ambition.
(3) Degree of past and present competitive,
aggressive, and hostile feelings.
(4) Intensity of the sense of time urgency.

The present inter-judge reliability study is based
on tape-recordings of the original interviews and
hence limits its consideration to the last three of these
categories.

Degree of drive and ambition was assessed by the
style and content of replies to questions such as the
following: Were you in any athletic teams in high
school or in college? Were you captain of any? Did
you participate in other school activities? Did you
hold office in any? Since you began making your
living, have you attended evening classes or taken
other courses to improve your chances of advance-
ment? Are you satisfied with your present job or do
you consciously strive for advancement?

Degree of competitive, aggressive, and hostile
feelings was determined by response to these
questions: Do you think of yourself as hard-driving
and aggressive or as relaxed and easy-going? What
does your wife think? When you play games with
children, do you usually purposely let them win?
(If negative): Do you ever let them win? Is there
much competition in your work? Do you enjoy it?
If you are driving your car on a highway and the car
ahead of you is going too slowly, what do you do
about it? How frequently do you get upset or angry? Do you usually show it, or do you try to keep it to yourself.

Simple affirmative answers do not mark a subject as a Type A. The intensity of feeling revealed in the response is the key, with Type A subjects responding more intensely and abruptly.

The sense of time urgency is particularly acute and widespread in the life-space of the person manifesting Behaviour Pattern A. The following were among the series of questions used in this section: Are there many deadlines in your work? Do you enjoy them? When you have an appointment to meet your wife or a friend, are you punctual? Are you impatient when you see something being done more slowly than you think it ought to be done? Are you irritated if you have to wait for a table at a restaurant? How long will you wait? Do you often have the feeling that time is passing all too quickly each day to get everything done that you want to do?

In addition, two questions were asked in a round-about fashion, with the interviewer intentionally stumbling with the words and hesitating in the completion of the question. The subject with Pattern A typically anticipates the end of the question and begins his response before the question is completed. Verbal behaviour of this sort, called by us speech-hurrying, was also used as an index of time urgency. The present study, then, examines the reliabilities of inferences made from the style and content of responses given in the interview situation described.

**Methods**

(A) **Inter-rater Agreement in Assessment of Behaviour Pattern**

The data employed in this study were drawn from the Western Collaborative Group Study (WCGS) (Rosenman and others, 1964). In the course of this prospective study of CHD a specially-devised interview was given by trained lay interviewers to 3,524 men. All these interviews were tape-recorded to permit a final assessment of behaviour pattern by the physician directing the study (RHR) and also to provide a permanent record available for subsequent deeper analyses by the investigators. Each subject was classified at the time of these initial interviews in 1960–61 as exhibiting Behaviour Pattern A or B. Subjects assessed as Type A were subdivided into those exhibiting the fully-developed (Type A1) or incompletely developed (Type A2) Pattern; and those assessed as Type B were subdivided into those exhibiting the incompletely developed (Type B3) or fully-developed (Type B4) Pattern.

The senior author, a clinical psychologist previously experienced in the principles of assessing the presence of the Behaviour Pattern A, auditioned and rated a series of tape-recorded WCGS interviews, selected randomly within stratifications by age and occupation to assure broad representation.

For the first 25 interviews, general assessments of behaviour type were made on the 4-point scale (described above) by the psychologist and compared with the ratings made by the physician-investigator (RHR) in 1960–61.

For the next fifty interviews each response of each subject was rated as being either in the direction and style of Behaviour Pattern A or in the opposite direction and style, that of Behaviour Pattern B. Items not involving judgement (e.g. “Were you ever captain of an athletic team?”) were eliminated from reliability computations to avoid spurious “padding” of the percentage of agreement. Judgements based on posture or facial expression were also eliminated for obvious reasons. Dichotomous judgements on the remaining 27 items, a total of some 1,350 decisions were compared with the corresponding ratings recorded on tally sheets by the original interviewers at the time of the intake interview.

(B) **Stability of Behaviour Pattern over Time**

In addition to the interviews and ratings given to all subjects at intake into WCGS in 1960–61, a similar interview was administered to all subjects at the first follow-up examination in 1962. These were administered by essentially the same team of lay interviewers and were tape-recorded, and ratings were made by the physician-investigator. The latter was on this occasion influenced to a greater degree by the ratings given by the interviewers.

The team of interviewers varied from three to six women over the course of these 2 years. With respondents assigned to interviewers sequentially by appointment time, it is estimated that 10 to 20 per cent. of the respondents may have been seen by the same interviewer on both occasions. The sheer number of interviews done by each interviewer (600 to 900) and the interval between the interviews of the same respondent (12 to 20 months) would act to reduce the chances of an interviewer recalling the rating she had previously given to a particular respondent. Even if her recollection were correct, she would have no way of knowing whether her judgement had been sustained by the physician-investigator.

The first 1,131 men seen in the 1965 follow-up examination of WCGS who had been interviewed both at intake and in 1962 comprise the sample used for the test-retest reliability estimate. Collation and
comparison of records were performed by electronic data-processing equipment at the Department of Epidemiology, University of North Carolina; 67 subjects were rejected from the comparison because one interview (usually for 1962) had been rated as "undecided" or had not been done.

RESULTS

(A) INTER-RATER AGREEMENT IN ASSESSMENT OF BEHAVIOUR PATTERN

The clinical psychologist (CDJ) listened to the tape recordings of 25 of the original interviews, reaching an independent opinion whether a subject was of Type A or B, and this rating was compared with the original assessment made by the physician-investigator (RHR). The results are presented in Table I. These ratings were found to agree in 21 of the 25 interviews (84 per cent.). Exact replication on the 4-point rating scale (i.e. A1, A2, B3, B4) occurred in sixteen of the 25 interviews (64 per cent.). Where disagreement occurred there was a systematic tendency for the psychologist, the less experienced rater, to tend slightly more in the direction of Pattern A than had the original rater. In all four disagreements on major category the psychologist rated Type A2 where the original assessment had been B3.

Table I
AGREEMENT OF BEHAVIOUR PATTERN RATINGS BETWEEN WCGS INVESTIGATOR AND INDEPENDENT AUDITOR OF TAPE RECORDING

<table>
<thead>
<tr>
<th>Investigator's Rating</th>
<th>Type A</th>
<th>Type B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A1</td>
<td>A2</td>
<td>B3</td>
</tr>
<tr>
<td>Type A</td>
<td>1</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Type B</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>11</td>
<td>9</td>
</tr>
</tbody>
</table>

Percentage of Sample given Identical Rating by Both Raters: 64 per cent.
Percentage of Sample placed in Same Major Category (A or B): 84 per cent.

The item-by-item comparison of 27 items in the interviews of fifty additional subjects yielded the percentage of items agreed upon by the raters as manifesting the style and content of Type A or Type B. Percentage of agreement was tabulated subject by subject, giving fifty 4-fold tables. These are summarized in Fig. 1, each table being represented by an area under the curve located at that point of the abscissa corresponding to the percentage of the 27 rated items categorized identically by both raters. The observed mean rate of agreement of judgements of items per subject was 76.5 per cent. If the frequency of judging the responses to the 27 items as Pattern A or B by each of the two raters be considered fixed and independent (and with the observed overall proportions of judgements prevailing), the distribution of percentage agreement for each subject expected under the null hypothesis would have had a mean of 50.5 per cent. The difference between this and the observed results is highly significant, as may be inferred from Fig. 1. It is of interest that for forty of the fifty subjects, 70 per cent. or more of the items were identically rated by the two raters, whereas under random circumstances such agreement could be expected to occur in only one of fifty subjects.

Merely doing significantly better than chance is not an adequate criterion for reproducibility, however. A scatter diagram comparing the total number of responses judged Type A per subject (Fig. 2, opposite) testifies to the reliability of the entire interview by considering total scores as contrasted with the item-by-item agreement in a dichotomous sense, which was considered in the preceding paragraph. The reliability of the total interview may be particularly important in population surveys because an epidemiologist would be more interested in a total estimate of a person's tendency towards Behaviour Pattern A than in judgements about particular items.

The correlation between the number (and proportion) of responses per interview called Type A by the independent rater and original WCGS interviewers is +0.69. Using Formula 51 (McNemar, 1955, p. 153), a standard error of the number of Type A responses was computed, and its value was found to be 2.75. This indicates that knowing the score
CORONARY-PRONE BEHAVIOUR PATTERN

Tetrachoric Correlation Coefficient: a measure correcting given to a subject by either of these two raters and correcting for any constant difference between raters (such as in this instance, the consistent, slight rightward displacement of scores by the independent rater—see Fig. 2), the score given by the second rater would fall within 2\(\frac{1}{2}\) points of that given by the first rater in 68 per cent. of the cases in large samples. In 95 per cent. of the cases the second rater’s score could be predicted within 5\(\frac{1}{2}\) points (2 standard errors) on this 27-point scale.

(B) STABILITY OF BEHAVIOUR PATTERN OVER TIME

The results of the analysis of stability over time are given in Table II. About 79 per cent. of the men rated A at intake were re-rated A in 1962; and 82 per cent. of the men rated B at intake were re-rated as B. Exactly the same score on the 4-point scale was given to 66.4 per cent. of the men on both occasions. If the tolerance be broadened to include one-point differences in rating, 97.9 per cent. of subjects are included. The use of this concept of Behaviour Pattern to contribute to the prediction of risk to coronary heart disease has been limited to a dichotomous rating: Pattern A (higher risk) or Pattern B (lower risk). In terms of these categories, 80.4 per cent. of the subjects were rated consistently on both occasions.

The stability of this judgement over time can also be expressed as a test–retest correlation coefficient. Product moment correlation is inappropriate, because this variable is represented by dichotomous classification. Tetrachoric correlation is an unbiased approximation of the product moment coefficient, provided that one can assume, as in this case, that the underlying distribution of behaviour pattern is continuous and normally distributed. Casting the data into a 4-fold Table (Type A v. Type B for the two different times) permits the calculation of a tetrachoric correlation coefficient (McNemar, 1955: 197-202). Table II shows that the tetrachoric coefficient of test–retest stability of Behaviour Pattern judgement was found to be +0.82.

The observed differences in the rating of subjects’ behaviour patterns over time derive from many sources. These would include differences between interviewers in the relative emphasis given to the different traits entering into the judgement, different modes of response elicited by the several female interviewers as a function of their own personalities, natural day-to-day variability in subjects’ expression of vigour, and both naturally occurring and planned modifications in style of life over the 12 to 20-month interval between interviews. The relative contribution of each source of variance is unknown, but the total effect of all these influences was sufficient to result in a change of category in only about 20 per cent. of these subjects.

DISCUSSION

These results imply that ratings of Behaviour Pattern A which have been used as a basis for inferring a relationship between behaviour and the prevalence and incidence of CHD are “objective” in character and that these ratings sample a relatively persistent life style. Original tape-recorded interviews were assessed independently for overall ratings, A or B, and agreement in assessment was found for 84 per cent. of the 25 subjects. Furthermore, on an
item-by-item basis, agreement averaged 76.5 per cent. for fifty additional cases.

A similar degree of agreement in utilizing the same interview technique has been reported (Friedman, Hellerstein, Jones, and Eastwood, 1965) in a study of Cleveland attorneys. The oral interview had been given to 107 attorneys and tape-recorded to permit their later concurrent but independent audition and assessment by the two Cleveland investigators and by one of us (RHR). The two Cleveland investigators agreed with RHR in 82.2 and 77.6 per cent. respectively of the 107 interviews. When a 4-point scale of classification was used (i.e. A1, A2, B3, B4), the two Cleveland investigators matched the assessment of the WCGS investigator exactly or within one category in 106 of the 107 subjects. Only two of the total of 214 judgements by the Cleveland investigators differed by as much as two categories from those made by RHR.

The same investigators also studied a sample of 43 businessmen none of whom had heart disease. The ratings of behaviour type made independently by the two Cleveland physicians agreed in 79 per cent. of instances (Friedman and others, 1965).

Keith, Lown, and Stare (1965) studied male hospital patients in the Boston area using this same interview method for the coronary-prone behaviour pattern as one of their variables. The original interviewer (RAK) re-auditioned 100 of his tape-recorded

Fig. 3.—Percentage of inter-judge agreement in selected examples of clinical judgements and diagnoses.

interviews 3 to 18 months later and achieved 74 per cent. agreement on the Type A v. Type B distinction, with 54 per cent. exact agreement on the 4-point scale (A1, A2, B3, B4). This level of replicability is not remarkably different from the findings reported here.

The implications of these observations of inter-rater agreement and of stability over time in rating the coronary-prone behaviour pattern can be best evaluated in comparison with analogous reliability studies of some of the standard diagnostic procedures and clinical judgements. The relevant literature was reviewed, and the results of a number of studies of medical judgements made by qualified diagnosticians are summarized in Fig. 3 in terms of percentages of inter-judge agreement. Because medical diagnosis, electrocardiograms, and chest x rays are expected to change in short intervals of time under treatment, it was thought more appropriate to compare assessment of stability in time of the A–B behaviour type with measures of psychological traits. These findings are summarized in Fig. 4, and the degree of agreement is expressed in terms of the correlation coefficient. The methods used in the comparison studies are described in the references cited in the Figures.

The interview method of assessing the coronary-prone behaviour pattern (Type A v. Type B) compares favourably with other clinical variables in percentage of agreement between qualified raters. Its stability over time is equal to or better than brief psychological tests and most sub-tests of longer standard psychological measures. Its test–retest correlation coefficient is not as high as those of the Full Scale Wechsler–Bellevue or the full Stanford–Binet Tests of Intelligence.

Despite these observed levels of reliability, certain reservations must be made:

(1) The interview method requires interviewers who have been adequately trained and whose standards of assessment have been accurately measured. This is, of course, necessary for any clinical or laboratory procedure. Wide individual differences have been observed in the ability to gain intuitive and practical skill in the assessment of these Behaviour Patterns.

(2) Apart from the problem of training interviewers, the interview procedure itself is time-consuming and expensive in large-scale epidemiological studies.

Efforts to refine further the determination of this behaviour pattern are in process. One approach

---

**Fig. 4.**—Test–retest correlation coefficients in selected examples.

seeks to ascertain the behaviour pattern through a series of objective performance tests derived largely from the field of experimental psychology (Bortner and Rosenman, 1967). Another approach seeks to eliminate examiner judgement and the differential interpersonal impact of examiners by using a self-administered paper-and-pencil test which can be converted directly to punched cards and scored by a computer. This test, the “Jenkins Activity Survey”, underwent initial development and validation against the criterion of behaviour pattern as judged by the standard interview in 1964 and 1965 (Jenkins, Friedman, and Rosenman, 1965; Jenkins, Rosenman, and Friedman, 1967). Validation studies against the criterion of clinical coronary heart disease are in progress, and the test will not be ready for general experimental use until these are completed. Factor analyses and comparisons with biochemical, social, and other psychological variables are being pursued for the purpose of gaining a better understanding of the relation of this behaviour pattern to other parameters in physiological or behavioural science.

SUMMARY

One important attribute of any variable proposed for use in scientific research is its reliability of measurement. The coronary-prone behaviour pattern is a variable which has shown promise as a predictor of risk to coronary heart disease. The reliability of rating this attribute is studied here in terms of general inter-rater agreement whether a subject manifests Behaviour Pattern A or Behaviour Pattern B, inter-rater agreement on whether individual item responses were in the direction and style of Type A or Type B, and stability of the behaviour pattern over time as estimated from the percentage of agreement between two interviews administered the same persons at an interval of 12 to 20 months (usually by different interviewers).

The degree of agreement on the ratings of individual subjects between a clinical psychologist and the physician-investigator of the Western Collaborative Group Study was 84 per cent. The item-by-item comparison for fifty records yielded a mean frequency of 76 per cent. identical ratings as Type A or Type B. The standard error of the mean was 2.75 points, about one-tenth of the possible range of scores. A study of 1,064 of the subjects in the Western Collaborative study showed that 80 per cent. were placed in the same category after an interval of 12 to 20 months.

These approaches to testing the replicability of this interview technique for ascertaining the coronary-prone behaviour pattern indicate that the reliability of these ratings is similar to, and sometimes higher than, that of standard diagnostic procedures in the fields of internal medicine, radiology, psychiatry, and psychological testing.

Efforts are in process further to refine the use of this variable by developing a self-administered, computer-scored questionnaire for rapid determination of the coronary-prone behaviour pattern.

REFERENCES