THE RANDOMIZED MULTIPLE BASELINE EXPERIMENTAL DESIGN: ITS POWER AND A CLINICAL APPLICATION TO THE COGNITIVE MODIFICATION OF DELUSIONS.

Submitted toward the degree of Doctor of Clinical Psychology

University of Leicester, 1998.

Michael John Cliffe

ABSTRACT

THE RANDOMIZED MULTIPLE BASELINE EXPERIMENTAL DESIGN: ITS POWER AND A CLINICAL APPLICATION TO THE COGNITIVE MODIFICATION OF DELUSIONS.

The dissertation describes the first reported application of a small-N experimental design, the randomized baseline experimental design across subjects and behaviours. It is applied to a small scale clinical psychological experiment on the cognitive modification of delusional ideation in four people with a diagnosis of schizophrenia. The data were analysed by a form of randomization test which does not depend on the classical parametric assumptions. The randomization test based on random data permutation gave statistically significant results for the effect of the independent variable (cognitive modification of delusions) on two dependent variables (strength of conviction, and preoccupation) but not on a third dependent variable, amount of distress. Estimates of effect size are provided based on Cohen's d and on the Common Language Effect Size. It presents data on the statistical power of the procedure derived from Monte Carlo power analysis. It provides reviews of the concept of statistical power in applied psychological research, of the concept of effect size, of the use of cognitive modification of delusional ideation and of randomization tests. The results support the feasibility of small-N clinical experiments using the randomized baseline experimental design, analysing the data graphically and by use of randomization tests and designing experiments with the aid of Monte Carlo power analysis.

CONTENTS

INTRODUCTION

This thesis is written in accordance with the principles of the scientist-practitioner model in clinical psychology. Barlow, Hayes and Nelson (1984) describe three roles for the clinical psychologist working with this model:

"In the first role, the practitioner is a consumer of new research findings from research centers, usually new assessment or treatment techniques that he or she will put into practice. In the second role, the practitioner is an evaluator of his or her own interventions using empirical methods that would increase accountability. The third role describes the practitioner as researcher, producing new data from his or her own setting and reporting these data to the scientific community."

Although large scale randomised controlled trials (RCTs) and conventional parametric statistical analyses have an important role in this context, the clinical psychology practitioner whose working context precludes the conducting of large scale RCTs has access to a range of small scale experimental designs that make it feasible to fulfil the second and third roles noted above. For such a practitioner, small scale research designs may have greater applicability than large scale RCTs.

One problem with small scale designs is that at present their statistical base is weak. Analysis of the data has often been solely by means of graphical data analysis (GDA) which carries difficulties in interpretation. The application of conventional parametric statistical methods in this context has been controversial. The concensus is arguably that they are not applicable because of unacceptable violations of the classical parametric assumptions. This thesis proposes a possible solution, in the form of a new type of experimental design that allows the data to be analysed statistically by the use of randomization tests (Edgington, 1995) that do not rely on the classical parametric assumptions.

The thesis is concerned with the concept of statistical power as applied to a small scale experimental design known as the randomized multiple baseline experimental design, when the data resulting from the application of the design are analysed using randomization statistics.

The randomized multiple baseline design was reported by Marascuilo and Busk (1988), who presented hypothetical data from an earlier paper by Wampold and Worsham (1986), with which they illustrated the potential use of the design together with a form of randomization test which they advocated as appropriate for its statistical analysis. There appear to be no reports heretofore of actual implementations of the design.

Attention is drawn to the neglect of the concept of statistical power in applied psychological research (Cohen, 1988). There are no reported data on the power of the experimental design used in the thesis. In order to address this problem, a series of Monte Carlo computer simulations is described which provides data on the power of the procedure under a range of experimentally plausible parameters.

The randomized multiple baseline design, together with the analysis using a randomization test, is implemented in a small scale clinical experiment on the cognitive modification of delusional ideas in four people to whom is attributed a diagnosis of schizophrenia.

OVERVIEW

Chapter 1 introduces the concept of statistical power, reviews the problem of its neglect in the applied psychology literature, points to the problems deriving from low power and makes recommendations on increasing power. It defines the concept of effect size, which is integral to the power concept.

Chapter 2 reviews single-case and small-n experimental designs. It suggests a taxonomy of such designs, examines the problem of generalization and considers the question of graphical versus statistical data analysis with an analysis of the problem of autocorrelation.

Chapter 3 introduces randomization tests. It defines them and gives a numerical example. It considers the question of random versus systematic data permutation, describes randomization test computer programs and considers randomization tests for multiple baseline designs.

Chapter 4 gives an historical review and critique of cognitive behaviour therapy applied to delusional ideation. It describes a clinical experiment on the modification of delusional ideation and the analysis of the resulting data by a randomization test.

Chapter 5 addresses the power of randomization tests applied to randomized multiple baseline designs. It provides tables and graphs derived from computer simulations showing statistical power under a range of experimental parameters.

CHAPTER 1. STATISTICAL POWER

"What behavioral scientist would view with equanimity the question of the probability that his investigation would lead to statistically significant results, i.e., its power?" Cohen, (1969, p. vii)

1.1 INTRODUCTION

The power of a statistical test is the probability that it will correctly reject the null hypothesis (Rossi, 1990; Siegel and Castellan, 1988; Snedecor and Cochran (1989). Power is the complement of beta, the probability of committing a Type II error, i.e. the probability of accepting a false null hypothesis.

Cohen (1962) drew attention to the neglect of the concept of power in the applied psychology literature and illustrated it by a power review of the 1960 volume of the Journal of Abnormal and Social Psychology. He found the mean power to detect medium effect sizes to be 0.48, with a significance level (alpha) = 0.05, two-tailed (effect size will be discussed in Section 1.2 below). Thus the chance of obtaining a significant result approximated that of tossing a head with a fair coin. He attributed this disregard of power to the inaccessibility of a then meager and mathematically difficult literature and attempted to solve the problem with his power handbook "Statistical Power Analysis for the Behavioral Sciences" (SPABS) (Cohen, 1969; 1988). This was intended to make the problem accessible, requiring as background only an introductory psychological statistics course that included significance testing.

The exposition was verbal-intuitive rather than mathematical and was assisted by worked examples from across the spectrum of behavioural science. Further sources of information on statistical power aimed at non-technical readers have become available in recent years, e.g. Kraemer and Thiemann (1987) and Lipsey (1990). Several computer

programs are now available for the determination of power and sample size requirements. These have been reviewed by Goldstein (1989) and by Onghena (1994).

This has apparently not led to an increased use of the power concept in applied behavioural science. Sedlmeier and Gigerenzer (1989) reported a power review of the 1984 volume of the Journal of Abnormal Psychology, 24 years after Cohen's review, titled "Do studies of statistical power have an effect on the power of studies?". They found the answer to be "No". Neither their study nor those they cited (apart from fields in which large sample sizes are common, for example sociology and market research) showed any improvement in power. Cohen (1992) commented "Thus, a quarter century has brought no increase in the probability of obtaining a significant result.". Similar results were obtained by Rossi (1990) in a power review of 6,155 statistical tests in 221 journal articles published in the 1982 volumes of the Journal of Abnormal Psychology, Journal of Consulting and Clinical Psychology, and Journal of Personality and Social Psychology. Cohen (1969) recommended that researchers design their studies so that there is at least an 80% chance of detecting the effect under investigation. In this respect the results of Rossi's (1990) survey are discouraging. The average statistical power exceeded 0.80 only for large effects, and more than a third (35%) of all studies were unable to attain this level of power even for large effects. More than 75% of all studies in the survey failed to achieve power of 0.80 for medium sized effects, and almost half of the studies did not even have a 50% chance of detecting effects of this size. More than 90% of the surveyed studies had less than one chance in three of detecting a small effect.

Hallahan and Rosenthal (1996) observe that failure to consider power in planning and interpreting empirical studies often leads to the drawing of erroneous conclusions, both by overlooking important effects and by prematurely abandoning promising avenues of investigation. Researchers risk wasting time and resources on research that is unlikely to detect an effect that exists, and may mistakenly interpret a non-significant result to

imply that the null hypothesis is true, regardless of a sample's ability to detect an existing non-null effect.

Hallahan and Rosenthal (1996) give an hypothetical illustration as follows. A researcher testing a new treatment randomly assigns N = 20 Ss each to a treatment group and to a comparison group. The treatment group shows a 0.4 standard deviation improvement over the comparison group, but the t test (p = 0.225, two-tailed) fails by the critical 0.05 criterion. He infers that "the difference between treatment and control was not more than would be expected by chance if the groups were identical", and regrets the time spent testing a treatment that "provides no additional benefit".

However, the statistical power of a t test with alpha = 0.05, two-tailed , and N = 20 in each condition to detect a difference of 0.4 standard deviations is only 0.23, with a corresponding probability of committing a type II error equal to (1 - 0.23) = 0.77 (from Cohen's (1988) power tables). Thus if the treatment actually did produce a 0.4 standard deviation gain, fewer than 1 / 4 experiments with this sample size would yield a significant result at the 0.05 level, two-tailed. He had planned a study with little chance of obtaining a significant result, had wrongly concluded that the treatment provided no additional benefit, and had been discouraged from pursuing a promising line of research.

Kazdin and Bass (1989) analysed the power of comparative psychotherapy outcome studies. They found adequate statistical power for treatment versus no-treatment studies, but relatively marginal power for treatment versus treatment and treatment versus active control comparisons. They suggested that the frequently obtained result of no treatment differences in psychotherapy outcome research may well be due to inadequate statistical power. This is an important result given the large amount of effort that has been invested in this field.

In the above discussion the concept of effect size (hereafter referred to as ES) was

referred to without elaboration. ES is addressed in the following section.

1.2 THE CONCEPT OF EFFECT SIZE (ES)

Cohen (1992) suggests that researchers find specifying the ES the most difficult part of power analysis, partly due to the "low level of consciousness of the magnitude of phenomena that characterizes much of psychology". In the Neyman-Pearson approach to statistical inference (see Section 1.4 below) an alternative hypothesis H_i is counterpoised against the null hypothesis H_0. ES is defined as the degree to which H_0 is false, i.e. the discrepancy between H_0 and H_i. The various statistical tests have their own ES indices, all of which are scale free and continuous, ranging upwards from zero, and in all cases the H_0 is that ES = 0.

For example, with the Pearson r, ES is simply the population r, so H_0 posits that r = 0. As a further example, for testing the significance of the departure of a population proportion p from 0.5, the ES index is g = (p - 0.5), so the H_0 is that g = 0. For the tests of the significance of the difference between independent means, correlation coefficients, and proportions, the H_0 is that the difference is zero. Cohen (1992) gives for each of the tests the definition of its ES index. For the case of the difference between two independent means m_a and m_b, the ES index is defined as

$$d = (m_a - m_b) / s , \qquad (1)$$

where s is the within-population standard deviation.

In order to convey the meaning of a given ES index, Cohen (1992) suggests that it is necessary to have some idea of its scale. He proposes, as conventions or operational definitions : small, medium and large values for each, that are at least approximately consistent across the different ES indices. He suggests (Cohen, 1988, p. 13) that ' "Small " effect sizes must not be so small that seeking them amidst the inevitable operation of measurement and experimental bias and lack of fidelity is a bootless task, yet not so large

as to make them fairly perceptible to the naked observational eye...In contrast, large effects must not be defined as so large that their quest by statistical methods is wholly a labor of supererogation, or to use Tukey's delightful term "statistical sanctification"... On the other side, it cannot be defined so as to encroach on a reasonable range of values called medium.' His intention is that medium ES should represent an effect likely to be visible to the naked eye of a careful observer, and states that it has been noted in ES surveys that it approximates the average size of observed effects in various fields. He set small ES to be noticeably smaller than medium but not so small a distance as to be trivial, and set large ES to be the same distance above medium as small was below it. Although the definitions were set subjectively, he states that with some early minor adjustments these conventions have been effectively fixed since the 1977 edition of SPABS and have come into general use. The ES index for the t test of the difference between independent means is d, the difference expressed in units of (i.e. divided by) the within-population standard deviation (Equation 1). For this test, the H_0 is that $d = 0$; and the small, medium and large ESs (or H_is) are d = .20, .50 and .80.

Thus, an operationally defined medium difference between means is half a standard deviation, and as a concrete example, for IQ scores with a population standard deviation of 15, a medium difference between means is 7.5 IQ points. Cohen (1988, pp. 26-27) states that, approximately: small ES is the size of the difference in mean height between 15- and 16- year old girls; a medium ES is the corresponding difference between 14- and 18- year old girls; and a large ES is the corresponding difference between 13- and 18- year old girls.

Lipsey (1990) considered 102 meta-analyses of treatment effectiveness research (mainly in education). He defined the lower 33% of positive mean ESs as "small", the middle 34% as "medium" and the upper 33% as "large". The median ESs for these ranges were 0.15, 0.45 and 0.90, approximately congruent with Cohen's suggested convention.

1.2.1 A COMMON LANGUAGE EFFECT SIZE STATISTIC (CL)

McGraw and Wong (1992) argue that Cohen's ES measure is not usable by someone untutored in statistics, as it is not readily translated into the everyday language used for discussing probabilities. They propose a "common language effect size statistic" (CL) that expresses the probability that a score sampled from one distribution will be greater than a score sampled from another distribution. CL is easily calculated from sample means and variances. They reviewed other approaches such as Rosenthal and Rubin's (1982) binomial effect size display, which will not be further examined here.

Bjorgvinsson and Kerr (1995) suggest that CL can be a useful tool for both statisticians and non-statisticians in judging the true importance of research findings. They give the hypothetical example of a study in which Ss were randomly assigned to groups receiving either a placebo or a drug and were later measured on an outcome variable, giving a CL ES indicator of 0.93. This means that 93 times in 100 a S randomly sampled from the drug group would score higher than a S randomly sampled from the control group. These authors showed that CL can be translated easily into Cohen's d and vice versa. Cohen's ds of 0.2, 0.5 and 0.8 (small, medium and large ESs) correspond to CL ESs of 0.55, 0.64 and 0.71 respectively.

1.2.2. ESTIMATING EFFECT SIZE

In practice the most uncertain part of power analysis involves specifying the expected ES prior to conducting a study. Hallahan and Rosenthal (1996) give recommendations as follows:

1. Consult existing research. Previous research in a given field of investigation may provide a reasonable estimate of the size of the ES that would be expectd in a planned

study. Simple meta-analytic procedures (Rosenthal, 1991) could be used to find the average ES in previous studies.

2. Rely on preliminary data. Pilot research, in addition to providing an opportunity to test and fine-tune experimental procedures, generates data useful in estimating the ES that would be observed in a larger study.

3. Subjective estimation. In the absence of pilot data and previous studies, an educated guess might be appropriate. The value of such a guess is questionable but it may be an enviable situation to be planning a study for which there is no prior information to estimate the size of the ES. The resulting data may be valuable as the first information about the ES of a potentially interesting phenomenon.

4. Cohen's advice. Cohen's (e.g. 1988) suggested benchmarks of small, medium and large ESs may be useful in estimating the expected ES for a planned study. In cases where there is no previous information on which to base an estimate, he suggests that it might be reasonable to expect a small effect because in the absence of previous work the phenomenon of interest is probably not under good experimental control, nor are the available measuring instruments likely to be precise.

5. Cost-benefit analysis. An "implementation threshold ES", or the degree of effectiveness at which an intervention's anticipated benefits would justify its implementation cost, could be determined. This would ensure that a planned study had sufficient power to detect the minimum ES considered important.

1.3 THE POWER FUNCTION

The major determinants of power are (1) the significance level alpha; (2) the sample size n; (3) the effect size (ES). Onghena (1994) defined the power function of a statistical test as the statistical power as a function of ES, for a given alpha and sample size. He noted that in this context power is usually defined irrespective of the truth of the null hypothesis, and thus in broader terms the power is the probability of rejecting the null hypothesis. He further stated that within this broader definition, power is still the

complement of beta, if the null hypothesis is false. If the null hypothesis is true, power is the probability of committing a Type I error, i.e., the probability of rejecting a true null hypothesis, which should not be larger than the significance level alpha (Lehmann, 1986; Silvey, 1975).

Onghena (1994) stated that the relationships between the four parameters (power, alpha, ES and n) could be conceptualized in four dimensions. He presented one two-dimensional representation and two three-dimensional representations regarding the independent-samples t-test, as examples. These three representations are shown below as Figures 1.1, 1.2 and 1.3 (reproduced with permission).

Figure 1.1 (from Onghena, 1994) shows three power functions of independent samples t-tests with alpha = 0.05 and equal sample sizes n1 = n2 = 5, under the classical assumptions of normality, homogeneity of variance, and independence. Onghena notes that for the two-tailed test, power increases as the absolute value of the ES increases. For the upper-tailed test, power increases as the ES increases, i.e., as the mean of the first population increases relative to the mean of the second population. For the lower-tailed test, power increases as the mean of the second population increases relative to the mean of the first population. For both two-tailed and one-tailed tests, power equals alpha if ES is zero. For the one-tailed tests, the probability of committing a Type I error is smaller than or equal to alpha if the ES is compatible with the composite null hypothesis.

Figure 1.2 shows the power of the two-tailed t-test as a function of ES and of n1 (= n2) with alpha = 0.05, under the classical parametric assumptions. Onghena (1994) notes that the "power valley" becomes narrower as n increases. The power functions become steeper for larger sample sizes; or, for a fixed ES, power is an increasing function of the sample size. The minimum power of the surface is 0.05 for an ES of zero for all sample sizes because this is the fixed alpha level.

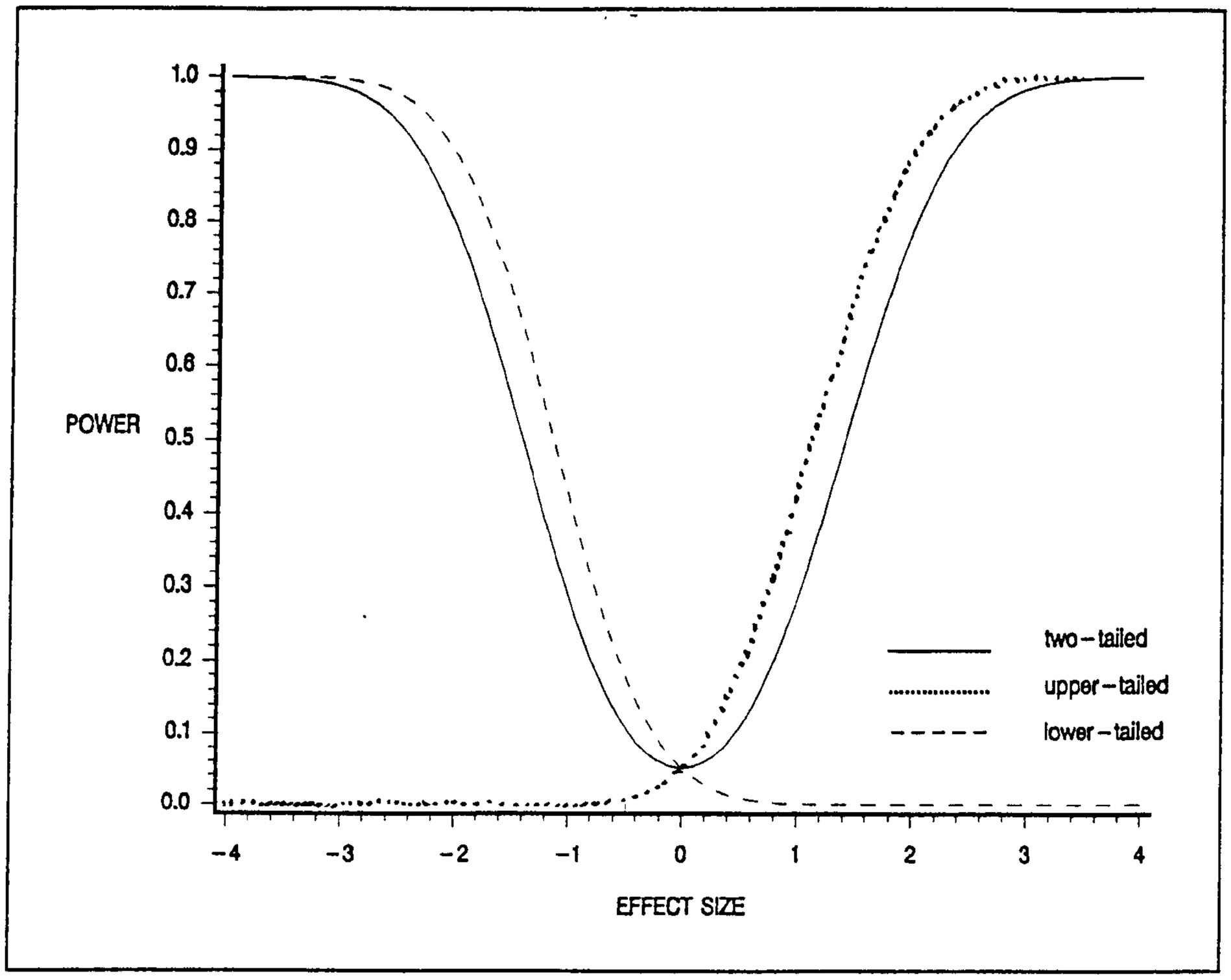


Figure 1.1. Power functions of the 1- and 2- tailed independent samples t tests with alpha = 0.05 and n1 = n2 = 5 under the classical parametric assumptions. The dotted curve shows the power function of the 1- tailed t test with the critical region in the upper tail. The barred curve shows the power function of the 1- tailed t test with the critical region in the lower tail. From Onghena (1994). Used with permission.

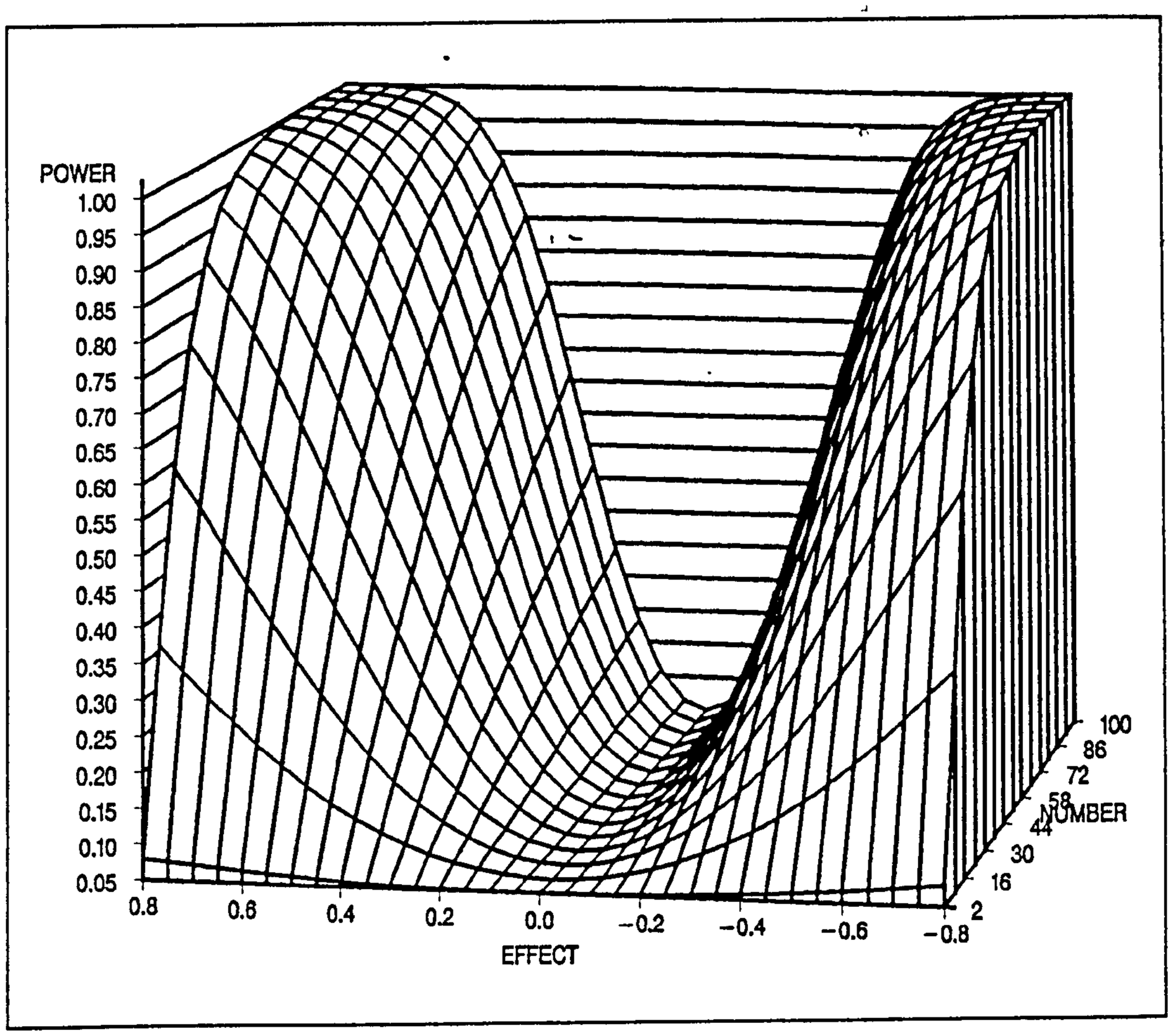


Figure 1.2. Power of the 2- tailed t test with alpha =0.05 as a function of the effect size (EFFECT) and the number of observations (NUMBER, n1 = n2) under the classical parametric assumptions. From Onghena (1994). Used with permission.

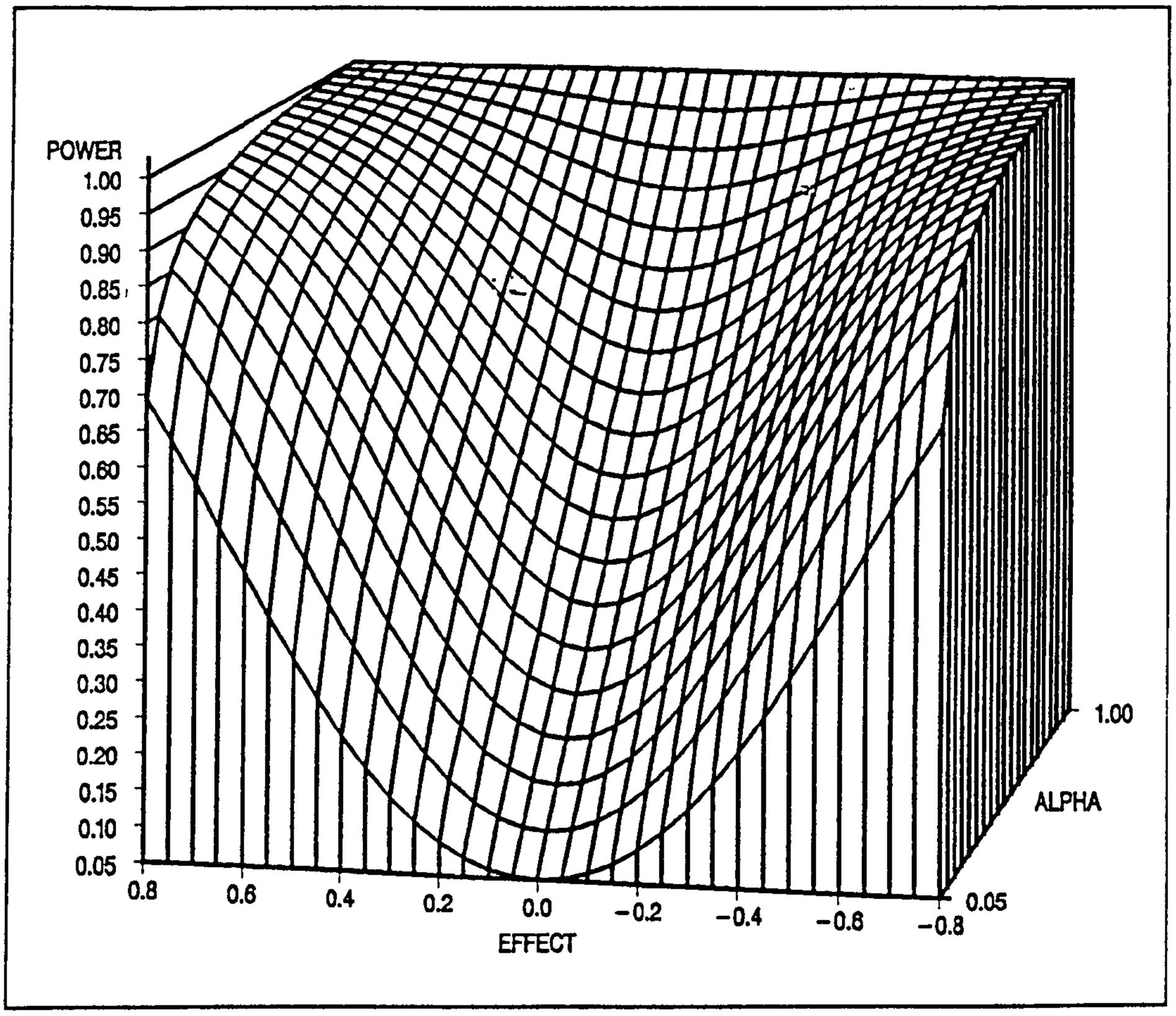


Figure 1.3. Power of the 2- tailed t test with sample sizes of n1 = n2 = 20 as a function of the effect size (EFFECT) and the significance level (ALPHA) under the classical parametric assumptions. From Onghena (1994). Used with permission.

Figure 1.3 shows the power of the two-tailed t-test as a function of ES and of alpha for n1 = n2 = 20, under the classical parametric assumptions. For a fixed n and for each ES, power is an increasing function of alpha. Onghena (1994) states that this "power slide" illustrates the inverse relationship between the probability of Type I errors and the probability of Type II errors.

1.4 THE HISTORICAL BACKGROUND OF POWER

The history of the way in which significance testing was adopted in the psychological field throws light on why psychological researchers have seemed to pay less attention to power and Type II error relative to the null hypothesis and Type I error. According to Gigerenzer and Murray (1987), significance testing came into wide use in psychology during what they termed the "inference revolution" that occurred between 1940 and 1955. There were at that time strong and often vitriolic differences among statisticians about the kind of inferences that could be made from significance tests (Cowles, 1989). Gigerenzer and Murray (1987) describe Sir Ronald Fisher's dispute with Jerzy Neyman and Egon Pearson concerning significance testing and the null hypothesis, which was central to the concept of power. In essence, Fisher's approach to significance testing (e.g.1955, 1956) focussed on testing a null hypothesis whereas the Neyman-Pearson approach (1933) specified both null and alternative hypotheses.

Hallahan and Rosenthal (1966) note that the concepts of power and Type II error are central to Neyman-Pearson but not to Fisher, but that Fisher's views had wider dissemination among psychologists through Snedecor's (1937) influential "Statistical Methods". Psychology ignored the substantial incompatibilities between the two approaches, and assimilated some of Neyman and Pearson's ideas with Fisher's to create a seemingly coherent, seemingly uncontroversial "...single, hybrid theory of which

neither Fisher nor, certainly, Neyman and Pearson would have approved" (Gigerenzer and Murray, 1987, p. 21). Hallahan and Rosenthal (1966) state that although the statistical texts that psychologists used at that time may have mentioned some Neyman-Pearson concepts, e.g. Type II error, they did not attribute these concepts to their founders, nor mention the controversies surrounding them. Thus the way in which psychologists have been taught to analyse data may have given prominence to null hypothesis testing and the avoidance of Type I error at the expense of power analysis and the avoidance of Type II error. This imbalance seems to be reflected in the asymmetry between the probability of Type I and Type II errors in psychological research.

1.5 IMPLICATIONS OF POWER SURVEY RESULTS

Rossi (1990) discussed the implications of low power in psychological research for small, medium and large effect sizes.

1. The case of small effect sizes. Power for small effects was very low (0.17) in his survey. It might be argued that the power of psychological research could not be this low, given that a large proportion of published studies report statistically significant results. However, not only does low power suggest that there may be a large number of Type II errors, but it also suggests the possibility of a proliferation of Type I errors in the research literature. With effect size equal to zero, in the long run significant results will occur at a rate of 5% if alpha is 0.05. Assuming an editorial bias favouring statistically significant results, disproportionately more significant than non-significant results will be published. Because the population effect size is zero in this case, all the published significant results will be Type I errors, despite a Type I error rate of 5%.

Although the above case is contrived, the case with power greater than alpha but still low is not much better. When power is low, the probability of rejecting a true null

hypothesis may be only slightly smaller than the probability of rejecting the null hypothesis when the alternative is true. That is, the ratio of Type I errors to power may be uncomfortably large, indicating that a substantial proportion of all significant results may be due to false rejections of true null hypotheses. Rossi (1990) states that on the basis of his survey, this ratio is 0.05 : 0.17, suggesting one Type I error for approximately every three valid rejections of the null hypothesis, for a "true" Type I error rate of about 0.23. He argues that in this way, low power undermines the confidence that can be placed even in statistically significant results, and that this may well be the legacy of low statistical power for small effects.

2. The case of medium effect sizes. In Rossi's (1990) survey the power to detect medium effect sizes was 0.57. While the problem of increased Type I errors in the published literature is less serious here than in the case of small effect sizes, he argues that a different kind of problem arises when power is marginal, i.e., in the general vicinity of 0.50. An inconsistent pattern of results may be obtained in which some studies yield significant results while others do not. Such a pattern of results may be especially troublesome for research that is directed at a specific problem area and often results in failure to replicate an experimental finding. Rossi (1990) gives the example of the literature on the spontaneous recovery of verbal associations. His analysis suggested that the sample sizes of studies in this area were inadequate to ensure detection of the effect in most studies but were sufficient to guarantee some statistically significant results, and he suggests that it is easy to see how the controversy over the existence of the effect was generated under these circumstances. He states that current texts regard spontaneous recovery as ephemeral, and that the issue was never resolved so much as it was abandoned. He suggests that this may exemplify the legacy of marginal levels of statistical power for medium effects.

3. The case of large effect sizes. Rossi's (1990) survey suggests that if effect sizes in psychological research are large, then power will be somewhat greater than 0.80. He

argues however that it is doubtful that average effect sizes in psychology are large, especially for applied research conducted outside the laboratory. He cites surveys suggesting average effect sizes approximating Cohen's (1977) definition of medium effects, and states that informal observations of the effect sizes reported in published meta-analyses are consistent with this view.

1.6 INCREASING POWER

Given the parameters of sample size, alpha and ES, it is a simple matter to determine the power of a study, e.g. from Cohen's (1988) power tables. An obvious way to increase power is to increase the sample size, but this is merely one of several ways to increase power. Hallahan and Rosenthal (1996) observe that in some cases it may not be possible to increase sample sizes because Ss are rare, difficult to recruit or expensive. In such cases researchers are constrained to work with a small number of Ss, but can achieve reasonable levels of power by other means. Hallahan and Rosenthal suggest ten procedures for increasing power, as follows:

<u>Design</u>

1. Increase sample sizes
2. Administer stronger treatments
3. Avoid restriction of range for dependent variables
4. Standardize experimental procedures
5. Use more reliable measuring instruments
6. Use more homogeneous subject populations
7. Use blocking variables
8. Use repeated measures designs (the ultimate blocking variable)

<u>Analysis</u>

9. Use focussed contrasts rather than omnibus tests

<u>Cumulation</u>

10. Combine results of individual studies.

The power of a study will be affected by any action that has implications for any of the three parameters. Setting alpha to a less stringent level is one possibility. This is referred to as "alpha amelioration" by Rossi (1990) and as "alpha leniency" by Hallahan and Rosenthal (1996). However, this may not be realistic in a world that holds $p < 0.05$ in such high regard. More useful would be steps that increase the observed ES..

Such steps can be seen in terms of the factors that determine ES: (a) the extent to which observations differ as a function of an experimental variable, or the "signal"; and (b) the amount of error variance against which the effect is compared, or the "noise". The ES "d" (formula 1) illustrates this relationship. The numerator, $m_a - m_b$, represents the variability between experimental conditions, and the denominator, s, represents the variability among observations within experimental conditions. Anything that increases between-condition variability, e.g. a stronger experimental manipulation, will increase ES and thereby power. Also it is important to avoid restriction of range. For example, the size of the correlation between exercise and heart rate would probably be smaller in a sample of elite marathon runners than in the general population.

Anything that reduces within-condition variability will increase ES and thereby power. Examples would be efforts to standardize experimental procedures, and the use of more reliable measuring instruments

Between-subject differences are a further source of within-condition variability. One way to reduce subject variance is to use a relatively homogeneous subject population. Another is to use blocking variables, i.e., variables other than the primary independent variable that are also related to the dependent variable. The use of blocking variables increases ES because variance in the dependent variable that is due to the blocking variables is effectively removed from the within-condition variance. Of especial relevance to the present study is Hallahan and Rosenthal's (1996) observation that "Repeated

measures designs are especially powerful because they employ 'the ultimate blocking variable' - the individual S.".

"Student" (1931) offered an early exemplification of the potential to improve power through research design. He argued that an experiment comparing the height and weight of children who received raw or pasteurized milk, with about 5000 children in each condition, could have achieved the same level of power with only 50 sets of identical twins, with one twin being assigned to each condition. This dramatic increase in power would have been achieved because the amount of variance in the height and weight of two identical twins is so much less than that between two randomly chosen children.

The present study provides a similar demonstration of the potential to improve power through research design. It will show that the use of a particular experimental design, provisionally called the "randomized baseline across subjects design" offers a large gain in power in comparison to the conventional randomized controlled trial methodology for measuring treatment intervention effects.

1.7 THE COUNTERNULL HYPOTHESIS

An overemphasis on significance testing at the expense of useful information about the size of effects can lead to two common inference errors (Hallahan and Rosenthal, 1996): (a) interpreting failure to reject the null hypothesis to mean the null is true or that there is no effect; and (b) not distinguishing the statistical significance of a result from its scientific importance. It is good practice to compute and report ES estimates for any effect that is tested, and to provide confidence intervals for effects (Loftus, 1991, 1993). Rosenthal and Rubin (1994) propose the counternull statistic as a way to avoid inference errors of the above kind.

For a given obtained ES, the counternull value of an effect size is the non-null

magnitude of the ES that is supported by just the same amount of evidence as supports the null value of the ES. For example, if a sample ES were d = 0.30, with p = 0.20, then only 1 time in 5 would a sample have an ES as large as d = 0.30 if it were drawn from a larger population where d = 0.00. With p so far from conventional significance, many researchers would conclude that the null was true. However, the counternull specifies the equally likely alternative: a sample ES as small as d = 0.30 would be observed only 1 time in 5 from a population where d = 0.60. That is, the counternull illustrates that populations with d = 0.00 and d = 0.60 are equally likely to produce a sample ES d = 0.30. The counternull is easy to compute. For symmetrically distributed ES statistics (e.g. d), the counternull is simply twice the observed ES minus the null ES:

$$ES_{(counternull)} = 2ES_{(obtained)} - ES_{(null)} \qquad (2)$$

Hallahan and Rosenthal (1996) argue that use of the counternull avoids the errors referred to above. The first error is avoided because the counternull illustrates that it is equally likely that the true population ES is larger than the observed ES as that it is zero, and it avoids the second error because even if the value of the counternull is too small to be scientifically important we will be less tempted to conclude that a result is important merely because it is significant.

CHAPTER 2. SINGLE-CASE AND SMALL-N EXPERIMENTAL DESIGNS

2.1 DEFINITION

A single-case experimental design is an experimental design in which one entity is observed repeatedly during a certain period under different levels of, at least, one independent variable (Barlow and Hersen, 1984; Kazdin, 1982; Kratochwill and Levin 1992; Onghena, 1994). Essential components of this definition (Onghena, 1994) are (1) that only one subject is concerned (single-case) and (2) that there is a manipulation of the independent variable(s) (experimental design). Implications of these components are (3) that the subject is exposed to all levels of the independent variable (within-subject design) and (4) that there are repeated measures or observations (longitudinal or time-series design).

Several other terms have been used more or less synonymously with "single-case designs", including: N = 1 experiments (Davidson and Costello, 1969; Dukes, 1965; Edgington, 1967), N-of-1 randomized controlled trials (Guyatt et al., 1990a, 1990b), idiographic designs (Jones, 1971), intrasubject replication designs (Gentile, Roden and Klein, 1972), intensive designs (Chassan, 1979), and interrupted time-series designs (Cook and Campbell, 1979).

Onghena (1994) states that the aim of a single-case experiment is to find evidence for a causal effect of the independent variable on the dependent variable for the subject under investigation. The focus is therefore on the internal validity of the study (Campbell and Stanley, 1966). In order to maximize the internal validity, the manipulation of the independent variable should be unequivocal, eliminating or controlling any other covariables, and the repeated observations should be sensitive, reliable and valid measurements (Barlow, Hayes and Nelson, 1984). The magnitude of the effect can be assessed by comparing the difference in response under the different conditions to the within-subject variability (within conditions) in the repeated measurements (Barlow and

Hersen, 1984; Kazdin, 1982).

2.2 GENERALIZATION

The results obtained with a single-case experimental design are usually difficult to generalize beyond the single subject who is the focus of the investigation (Onghena, 1994). In order to examine the external validity of the results, single-case designs have to be replicated (Campbell and Stanley, 1966; Cook and Campbell, 1979). Onghena distinguishes 3 types of replication strategies: direct replication, systematic replication, and clinical replication. Direct replication is the replication of the experiment with another subject. Systematic replication is the replication of the experiment under different circumstances (a different experimenter, setting, time of day, etc). Clinical replication is the administration by the same practitioner of a treatment package containing two or more treatment procedures to a series of clients presenting similar combinations of multiple behavioural and emotional problems (Barlow, Hayes and Nelson, 1984; Barlow and Hersen, 1984; Sidman, 1960).

In contradistinction to single-case experimental designs, in group experimental designs the focus is simultaneously on the internal and the external validity. The magnitude of the effect is assessed by comparing the differences in responses under the different conditions to between-subject variability (within conditions). Caution is however necessary in justifying the external validity of group designs by an inferential argument based on a random sampling assumption. Edgington (1966, 1973, 1986,1987) commented on the infrequent use of and the limited relevance of random samples in experimentation. Edgington (1966, 1995) argued that with nonrandom samples, generalization is only possible on nonstatistical, logical grounds. For example, in a clinical context the suitability of a treatment for a given patient can be assessed from the similarity of this patient to other patients for whom it was beneficial (Barlow, Hayes and Nelson, 1984; Barlow and Hersen, 1984). Therefore the external validity of a group

experimental design is in principle not different from the external validity of replicated single-case experimental designs.

Applications of single-case experimental designs have been frequent in the evaluation of treatment effects in clinical psychology (e.g. Kazdin, 1992) and in neuropsychological rehabilitation (Wilson, 1987, 1991). From a theoretical standpoint, single-case experimental designs have been seen as most compatible with behaviourist and operant conditioning paradigms, and so their application has been most prominent in behaviour therapy (Barlow and Hersen, 1984). They have recently been applied in clinical experiments on the cognitive modification of delusions (Chadwick and Lowe, 1990), which will be addressed in Chapter 4.

2.3 A TAXONOMY OF SINGLE-CASE EXPERIMENTAL DESIGNS

Onghena (1994) states that there are three major types of sigle-case experimental designs: phase designs; alternation designs; and multiple baseline designs. In a phase design, comparisons are made within a time-series. In an alternation design, comparisons are made between time-series for the different levels of the independent variable. In a multiple baseline design, comparisons are made both within the time-series and between the different baselines.

2.3.1 PHASE DESIGNS

In phase designs, the entire sequence of repeated measurements is divided into treatment phases and several consecutive measurements are taken in each treatment phase. The most basic phase design is the AB design, in which repeated measurements are taken under control conditions in the first phase (baseline or A phase), and under experimental conditions in the second phase (treatment or B phase).

In the field of behaviour therapy, the necessity of randomization in phase designs has not generally been recognized (Edgington, 1992, p. 134). Having noted that conventional experimental designs are pre-planned with respect to assignment of treatments, Kazdin (1982, p. 263) stressed that this is not so for single-subject behaviour therapy designs: "In single-subject designs, many crucial decisions about the design can be made only as the data are collected. Decisions such as how long baseline data should be collected and when to present or withdraw experimental conditions are made during the investigation itself."

This approach has been referred to as "response-guided experimentation" (Edgington, 1983) because the experimental conditions are adjusted on the basis of responses the subject makes during the experiment. Honig (1966) provides a quotation from Skinner highlighting Skinner's advocacy of response-guided experimentation: "A prior design in which variables are distributed, for example, in a Latin square, may be a severe handicap. When effects on behavior can be immediately observed, it is most efficient to explore relevant variables by manipulating them in an improvised and rapidly changing design."

Edgington (1992) notes that although Skinner's second sentence above seems plausible enough, it should be noted that it begins with "When", not "Because", and that randomized designs and statistical tests are employed in experimental research precisely because "effects on behavior" of a manipulated treatment usually cannot be "immediately observed". Changes in the dependent variable may not be treatment effects at all.

Edgington (1992) states that response-guided experimentation is incompatible with randomization and thus provides no basis for statistical tests. Randomization controls for unknown as well as known sources of confounding, whereas arguments that a research procedure involving nonrandom manipulation is not biased can concern only known sources. In order to strengthen the internal validity of the experiments, Edgington

(1995) proposed phase designs with a predetermined number of measurement times and a randomized phase change.

With the A and B phases as primary units, many variations on the basic phase design are possible, such as ABA or withdrawal designs, ABAB designs and so on. With the addition of other treatments many more variations are possible, such as the ABABACA design and others (Barlow and Hersen, 1984). Such elaborations of the basic AB phase design will not be further explored here. Onghena (1994) provides data on the power of randomization tests applied to AB designs, which is generally low.

2.3.2 ALTERNATION DESIGNS

In alternation designs, the treatments are alternated more rapidly and more frequently than in phase designs, and the "phases" of an alternation design contain only one measurement time, such that the notation ABAABBAB is used to denote an alternation design with 8 treatment times and 2 levels of the independent variable. Onghena (1994) provides data on the power of alternation designs. They will not be further considered here.

2.3.3 MULTIPLE BASELINE DESIGNS

The defining characteristic of multiple baseline designs is that different "targets" are measured simultaneously and the intervention is applied sequentially across targets (Barlow and Hersen, 1984). In behaviour modification experiments, the targets may be different subjects, the same subject in different settings, or different target behaviours of the same subject. This results in, respectively, multiple baseline across subjects, multiple baseline across settings, and multiple baseline across behaviours designs.

Onghena (1994) notes that multiple baseline across subjects designs do not conform

to the definition of a single-case design as given above (Section 2.1). They are discussed in the texts on single-case experimental designs because they originated out of the single-case tradition and because they are structurally equivalent to the single-case multiple baseline designs (across settings or across behaviours).

Because of the simultaneous examination of the subjects, multiple baseline across subject designs are not merely replicated phase designs. This simultaneity allows for a better control of historical confounding factors. If, for example, an intervention is applied to one of the baselines and produces a change in it, while little or no change is observed in the other baselines, it is less likely that other simultaneous events were responsible for the observed change than if this change were observed in an isolated AB design. It is also because of this factor that the graphical analysis of the results obtained from a multiple baseline design consists of both between- and within- baseline comparisons (Hayes, 1981).

It is argued here that in principle there is no logical reason to prevent the application of a multiple baseline design across both subjects and across behaviours. The clinical experiment to be described in Chapter 4 presents such a design, in which 4 subjects are measured on a combined total of 8 behaviours.

2.4 FORMS OF ANALYSIS IN SINGLE-CASE RESEARCH

The application of statistical methods in single-case (N = 1) research has been controversial. In order to overcome objections to the application of such methods, randomization tests (see Chapter 3) have been recommended for N = 1 and small-sample (N > 1) studies by Edgington (1967, 1969a, 1975a,1975b, 1980b, 1980c, 1992, 1995, 1996), by Kratochwill and Levin (1980), and by Levin, Marascuilo and Hubert (1978). In particular, randomization tests have been recommended as a supplement to the visual inspection of graphical data (Edgington, 1967; Gorsuch, 1983; Jones, Vaught and

Weinrott, 1977; Wolery and Billingsley, 1982). In addition, randomization tests have been developed further by Wampold and Worsham (1986) and by Marascuilo and Busk (1988) to allow for the combination of data from replicated designs that still preserves the individual differences of the single-case design. The purpose of combining data from replicated designs is to gain statistical power for rejecting the null hypothesis of no treatment effects.

However, visual inspection (referred to as Graphical Data Analysis or GDA by Onghena (1994) has been recommended as the preferred method for evaluating single-case data (Barlow and Hersen, 1985; Parsonson and Baer, 1992; Sidman, 1960). GDA is the most commonly used method to determine single-case intervention effects. Busk and Marascuilo (1992) report a review of all the articles published in the Journal of Applied Behavior Analysis during 1988. All of the articles using single-case research designs used GDA as the means of data analysis. Statistical methods were however used in the analysis of between-group designs reported in the same journal. Kratochwill and Brody (1978) surveyed four behaviour modification journals (Behavior Therapy, Behaviour Research and Therapy, Journal of Applied Behavior Analysis, and Journal of Behavior Therapy and Experimental Psychiatry). The percentage of all experimental research studies using statistical inference ranged from 18% to 69%. Designs classified as single-case studies that used statistical procedures ranged from 4% to 9%. The results of the above two studies support the view that GDA is the predominant method for analysing single-case data.

Wampold and Furlong (1981b) showed that the reliability of GDA was low. Subjects trained in GDA ignored small intervention effects that could be detected by examining the relative individual-subject variation in the data that is the object of randomization tests.

Further controversy has surrounded the question of the appropriateness of

parametric statistical procedures in this context. Barlow and Hersen (1984) argued that serial dependency and autocorrelation in N = 1 studies invalidated parametric procedures such as t tests and one-way analysis of variance. Counter to this, Kazdin (1984) and Huitema (1985) have recommended the use of parametric procedures under an assumption of no autocorrelation in N = 1 designs. This recommendaton has been challenged by Busk and Marascuilo (1988), Jones, Weinrott and Vaught, (1978), Suen and Ary (1987) and by Toothaker, Banz, Noble, Camp and Davis (1983).

Kazdin (1984) suggested that t and F tests should be preceded by a test of serial dependency, and that parametric procedures would be justified if the tests were not statistically significant. However, showing that an autocorrelation is not statistically significant does not prove that an autocorrelation does not exist (Busk and Marascuilo, 1988). Following the discussion of statistical power in Chapter 1, this is an example of the problems associated with low statistical power. Busk and Marascuilo (1992) state that often the number of observations is too small to determine whether there exists an autocorrelation. In other words, most tests in this context lack statistical power to detect an existing autocorrelation.

2.4.1 GRAPHICAL DATA ANALYSIS (GDA) IN THE ANALYSIS OF N = 1 DATA

In GDA, the researcher is required to look for changes in levels of behaviour or changes in trend in behaviour across an experimental phase change. In practice however it is difficult to detect even pronounced changes reliably (de Prospero and Cohen, 1979; Furlong and Wampold, 1981, 1982; Jones, Weinrott and Vaught, 1978; Wampold and Furlong, 1981b). The study by Wampold and Furlong (1981b) showed that visual inspection has low reliability. Different raters come to different conclusions. A group of graduate students who were trainee counsellors and who had completed a seminar in N = 1 research were compared with a group of graduate students who had had training in

multivariate statistical analysis. The first group, trained in identifying pronounced changes, "appeared to use a scaling heuristic in which they attended to large changes in a time series regardless of the relative variation" (p. 79) and ignored small intervention effects that subjects trained in statistics identified more readily. The results were confirmed by a further study (Furlong and Wampold, 1982). The use of judgemental aids or alternative graphing techniques was suggested by Knapp (1983) and by Bailey (1984) in order to improve ratings based on GDA. They found, however, that GDA was unreliable even with these additional aids. Busk and Marascuilo (1992) stated the need for procedures that can consistently detect the changes in levels of behaviour and or in trends of behaviour if they are real. They suggest that one solution is to supplement GDA with the application of randomization tests to measures of central tendency for changes in levels of behaviour and to measures of slope for changes in trends of behaviour.

2.4.2 THE PROBLEM OF AUTOCORRELATION IN N = 1 RESEARCH

Busk and Marascuilo (1992) state that there are two issues involving autocorrelation in N = 1 research. The first is the effects of the presence of autocorrelations on statistical procedures requiring independence, which they state have been recognized. The second and more controversial issue centres on the question of the existence of autocorrelation or serial dependency in N = 1 data. In N = 1 research the existence of autocorrelation is debated. One view, exemplified by Huitema (1985) is that empirical analyses of N = 1 data have found little or no autocorrelation. The opposed position is that repeated measures on the same individual through time usually are not independent. Busk and Marascuilo (1992) state that there are two aspects of the measurement in N = 1 research that affect serial dependency: 1) the behaviour itself, and 2) the way that the behaviour is assessed. They note that the practice of having serial measurements made by the same observer may introduce structure into the data and hence serial dependency.

Two factors must be considered when trying to measure the magnitude of serial dependency: 1) the precision of the estimate of the coefficient, and 2) the power of the statistical test to determine whether the coefficient is significantly different from zero. Both factors are related to the size of the sample of behaviour. With a small sample, the sample correlation will be quite variable. Busk and Marascuilo illustrate this situation by referring to the data reported by Holtzman (1963) on three functions over 245 successive days on a single schizophrenic patient. The Series A measurements comprised 100 daily observations of : (1) creatinine in the urine; (2) perceptual speed; (3) word association relatedness score. The 100 observations on the three functions can be considered an adequate measurement of the behaviour so that any serial dependency in the three functions would be estimated with precision. The time series for creatinine showed a rapid fluctuation together with an undulating trend. The correlogram showed that the serial correlation fell off sharply in four lags to a trivial value, indicating a small degree of serial correlation. The correlogram for perceptual speed fell off very gradually and continued in a downward trend in the negative direction, indicating a serial correlation. The correlogram for word association relatedness score showed essentially no association. The serial correlations or lags on autocorrelation were: (1) for creatinine 0.60; (2) for perceptual speed 0.75; (3) for word association reletedness,0.25. The first two were statistically significant and the third was not. Estimates of the lag one autocorrelation based on the first 6, 10, 15, 30, and 50 observations were made for each of the three functions. For creatinine the values were 0.22, 0.30, 0.38, 0.36, and 0.69 respectively. For perceptual speed the values were -0.17, 0.04, 0.54, 0.57, and 0.69 respectively. For word association relatedness score the values were -0.19, -0.19, 0.12, 0.31, and 0.38 respectively. The estimates of the lag one autocorrelations for each of the time series based on 100 observations varied widely from those obtained from each of the sample sizes considered. Confidence intervals constructed for each of the functions and each of the estimates covered the autocorrelations based on the 100-observation series, showing that the estimates were not out of range of the original time series. In this example, decisions about the magnitude of the autocorrelations therefore varied with

sample size. The data suggested that trying to infer the "true" or underlying magnitude of autocorrelation is questionable with sample sizes less than 50.

It has been noted above that the statistical power to detect a non-zero autocorrelation is dependent on the size of the time-series. For the function with the non-zero autocorrelation estimate of 0.25, all but the estimate based on 50 observations were nonsignificant. These tests are consistent with the 100-observation series nonsignificant result. For the other two functions with statistically significant autocorrelations, those estimates based on samples of sizes 6 and 10 failed to reject the null hypothesis, indicating lack of power. For the perceptual speed function, the autocorrelation based on the sample size of 15 also failed to reject the null hypothesis. The magnitude of the serial dependency for these two functions was indicative of strong dependency.

The above example provides evidence that serial dependency does exist in behavioural data, that trying to detect it using small samples may result in failure because of low power, and lack of precision leading to erroneous conclusions.

The sample sizes referred to above, i.e. sizes of 6,10, 15, 30 and 50, were chosen to represent typical sizes of behaviour samples based on the results of Busk and Marascuilo (1988), who analysed data from 44 studies from the Journal of Applied Behaviour Analysis from 1975 to 1985. In 101 baseline phases, 47% were from samples of size 6 to 15, 38% were from samples of size 15 to 30, and 15% were from samples of size 30 or more. The following sample sizes from the intervention phase were found: 43%, for 6 to 15; 30% for 15 to 30; and 27% for 30 or more.

Busk and Marascuilo (1988) computed lag one autocorrelations for the data from the 44 studies. There was a total of 248 independent data sets. Of these, 101 were for baseline phases, 125 were for intervention phases, and 22 were for phases beyond the

intervention. They found that many of the N = 1 studies were based on data in which the autocorrelations were greater than zero. In particular, 80% of the autocorrelations were in the range 0.10 to 0.49 for phases of size 5 to 18 observations. Also, 40% of the baseline sets of data yielded autocorrelations greater than 0.25, and in the intervention phases this rose to 59%. Busk and Marascuilo (1988) stated that statistical tests requiring the assumption of independence performed on data from these studies would have an inflated Type I error.

It has been noted above that N = 1 studies are often based on samples of behaviour for which the test for identifying a nonzero autocorrelation as statistically significant has very low power. There is a further problem concerning the points at which observations are made. Busk and Marascuilo (1988) and Holtzman (1963) noted that data observations may be too far apart in time of measurement to detect the autocorrelated nature of the data. Or (see above) the behaviour may have serial dependency but based on few observations, it is not possible precisely to estimate the true dependency. Sharpley (1988) stated that "The process of testing to determine if the level of autocorrelation present in a data series is significant, and then deciding on the basis of the presence or not of a significant autocorrelation whether traditional statistical procedures can be used to test for effects, is unwise, as well as not in keeping with the methodological rigour which requires that data analysis procedures are stipulated prior to data collection."

Because of the difficulty in empirically estimating autocorrelations in behavioural data, it is recommended that N = 1 researchers analyse their data in the absence of assumptions about serial independence. Parametric tests have been proposed as an alternative to GDA (Gentile, Roden and Klein, 1972; Huitema, 1985). Parametric tests can be invalidated by the existence of autocorrelation (Philips, 1983; Scheffe, 1959; Toothaker, Banz, Noble, Camp and Davis, 1983). Autocorrelation poses two problems for parametric tests: (1) because the errors are not independent, the statistical test

overestimates the number of independent sources of information (Kazdin, 1976), and (2) positive autocorrelation can spuriously decrease the error variance and thereby create a liberal bias, whereas negative autocorrelation can spuriously increase error variance and thereby create a negative bias (Phillips, 1983). The latter author showed that serially correlated errors affect ANOVA designs in N = 1 research. Toothaker et al. (1983) showed that even traditional ANOVA F tests modified to allow for autocorrelation have inflated Type I errors in the presence of autocorrelation. Sharpley (1988) showed that a t test can be inflated by 110% when the autocorrelation is only 0.10, and by as much as 435% when the autocorrelation is 0.90. Scheffe (1959) stated that lack of serial independence was the most difficult departure from assumptions with which to deal, in the context of ANOVA. Scheffe (1959) showed that an autocorrelation of 0.30 increased the risk of a Type I error from 0.05 to 0.12; that even with an autocorrelation as low as 0.20, the risk of a Type I error is doubled from 0.05 to 0.10; and that as the autocorrelation increases, the risk of a Type I error increases rapidly.

Busk and Marascuilo (1992) recommend that N = 1 data should be analysed using time-series methods or randomization tests. In the randomization tests used in this dissertation, analysis is based on the treatment of the phase periods as the units of analysis and not the original observation periods. The individual observations are averaged, and this value is used to represent the typical performance in that phase. Marascuilo and Busk (1988) state that averaging helps alleviate the problems of autocorrelated measures associated with single-subject designs.

CHAPTER 3. RANDOMIZATION TESTS

3.1 DEFINITION

"A randomization test is a permutation test based on randomization (random assignment) to test a null hypothesis about treatment effects in a randomized experiment." (Edgington, 1995, p.1).

"A randomization test is defined as a statistical test whose validity is based on the random assignment of units to treatments." (Onghena, 1994, p. 27).

The concept of randomization tests stems from the work of Fisher (1935/1966), Pitman (1937a, 1937b, 1937c), and of Welch (1937). It was further developed by Kempthorne (1952, 1955), and by Edgington (1964, 1966, 1969a, 1969b, 1987b). Onghena (1994) stated that randomization tests only became feasible with the availability of computers and the development of Monte Carlo randomization tests by Dwass (1957) and by Hope (1968).

The test is performed in the following way (Edgington, 1995). A test statistic is computed for the experimental data, then the data are permuted (divided or rearranged) repeatedly and the test statistic is computed for each of the resulting data permutations. Those data permutations, including the one representing the obtained results, constitute the reference set for determining significance. The proportion of data permutations in the reference set that have test statistic values greater than or equal to (or, for certain test statistics, less than or equal to) the value for the experimentally obtained results is the p-value (significance or probability value). For example, if the proportion is 0.02, the p-value is 0.02, and the results are significant at alpha = 0.05 but not at alpha = 0.01. For example (Marascuilo and Busk, 1988) in a single-subject AB design with n_a and n_b observations, a test statistic can be the difference in the means, $d = m_a - m_b$. This

criterion d is computed for all possible assignments of n_a observations to the A phase and n_b to the B phase until the entire distribution of d is found. From this distribution, the significance probability can be found by counting the number of outcomes that equal or exceed the observed d and dividing by the number of all possible assignments.

Edgington (1995) notes that determining significance on the basis of a distribution of test statistics generated by permuting the data is characteristic of all permutation tests, but that it is when the basis for permuting the data is random assignment that a permutation test is called a randomization test.

Edgington's above (1995) definition is broad enough to include procedures called randomization tests that depend on random sampling in addition to randomization. He states however that the modern conceptualization of a randomization test is a permutation test that is based on randomization alone, where it does not matter how the sample is selected. This is the concept of randomization tests used in this dissertation.

3.1.1. A NUMERICAL EXAMPLE

This example is from Edgington (1995). Suppose an experimenter wishes to compare the effectiveness of treatments A and B on reaction time. He expects A to give longer reaction times. Because the requirements of the experimental task are complex he carefully selects 10 suitable subjects. He randomly assigns 5 to each of the treatments and runs the experiment. He conducts an independent t-test, and the t value obtained for the data is 3.450. He is reluctant to determine the significance of t by using t tables because of his method of selecting subjects. Therefore he opts to derive a theoretical distribution of t which does not require the assumption of random sampling. He divides the 10 reaction times in every possible way between treatments A and B with the restriction that each treatment must have 5 reaction times. There are 252 permutations of this kind. For each of the 252 data permutations, t is computed. Ten of the 252

permutations yield a t value as large as 3.450, so the p value is 10/252, or about 0.04. Therefore the results are significant at the 0.05 level.

The logic of this procedure is as follows. The null hypothesis H_0 is that the reaction time for every subject is independent of the treatment assignment. The random assignment of subjects to treatments allowed 252 equally probable ways in which the subjects could be assigned. If H_0 is true, a subject's reaction time would have been the same if the subject had been assigned to the alternative treatment. Given the random assignment of subjects in conjunction with H_0, there are 252 equally probable ways in which the 10 reaction times could have been divided between the two treatments. If H_0 is true, how likely would it be that the random assignment performed in the experiment would yield one of the 10 largest values in the distribution of 252 values? The obtained answer is about 0.04.

3.2 EQUIVALENT TEST STATISTICS

The above example employed the t test. In practice, if data permutations for t tests are ranked from high to low with respect to t, they will always be found to be ranked from high to low with respect to the difference between means $m_a - m_b$. Therefore, the proportion of the data permutations with as large a value of t as the obtained value is the same as the proportion with as large a value of $m_a - m_b$ as the obtained value. Thus t and $m_a - m_b$ are two different test statistics which give the same p value for a randomization test. Therefore one could use the simpler test statistic $m_a - m_b$ to determine significance.

Two test statistics which must give the same p value for a randomization test are defined as "equivalent test statistics" (Edgington, 1995). It can save time to compute a simpler, equivalent test statistic to t, F, r or some other conventional statistic for every data permutation rather than to determine the significance value for the more complex one. Edgington (1995) proposes that two test statistics are equivalent if and only if they

are perfectly monotonically correlated over all data permutations in the set. Expressed in terms of correlation, there will be a perfect positive or negative rank correlation between the values of the two test statistics over the set of data permutations used for determining significance.

3.3 THE RANDOMIZATION TEST NULL HYPOTHESIS

The null hypothesis for a randomization test is that the measurement for each person or other unit that is randomly assigned will be the same under one assignment to treatments as under any alternative assignment that could have resulted from the random assignment procedure. Thus, when the null hypothesis for the randomization test, which is the hypothesis of no differential treatment effect, is true, random assignment of subjects to treatments randomly divides the measurements among the treatments. Each data permutation in the reference set, which functions as a randomization test "significance table", represents the results that would have been obtained for a particular assignment if the null hypothesis is true. To take an example from the statistical literature, Fisher (1935/1966) used the Lady Tasting Tea experiment to introduce the principles of statistical inference. (This is usually referred to as an hypothetical experiment. However Onghena (1994) cites sources providing evidence that it was actually carried out on one Muriel Bishop, a student of algae at Rothamstead who was offended at being offered a cup of tea into which milk had been poured before the tea. The actual results are unknown.).

The experiment was a test to determine whether a lady could tell by taste whether tea was added to milk or milk was added to tea. Eight cups were poured, four in each manner. They were presented to the lady in random order. She was told in advance that four cups were being poured in each way and that she was to decide in which way each cup was poured. In the hypothetical experiment the lady correctly identified all eight cups. There are $8! / 4!\ 4! = 70$ ways in which eight cups can be ordered with respect to

"tea first" or "milk first" when four cups are prepared in each way. Fisher computed the probability of correctly identifying all eight cups, by chance, as 1 / 70 because that is the probability of randomly assigning the cups in an order that would correctly match the statements "tea first" and "milk first" that, under the null hypothesis, the lady would make at the specified times, regardless of the cup she tasted. The null hypothesis tested is that the lady's response ("tea first" or "milk first") at each tasting time is independent of the assignment of cups to tasting times. In other words, that the lady's response is the same at each treatment time as it would have been at that time if the cup had been mixed in the other way.

3.4 VALIDITY CRITERIA FOR RANDOMIZATION TESTS

Detailed discussions of the validity of randomization tests are given by Edgington (1980c; 1995) and by Levin, Marascuilo and Hubert (1978). Edgington (1980c) specified three rules for the valid use of randomization tests: 1) There must be random assignment of treatment times to treatments; 2) the distribution of test statistic values must be based on data divisions that are appropriate for the type of random assignment used; and 3) the test statistic value for a data division must be computed in the same way as it would be computed if that data division represented the obtained results. (p.246)

Edgington (1995) states that the use of a randomization test does not guarantee validity. It is valid only if it it properly conducted. He states that in light of the numerous test statistics and random assignment procedures that can be used with randomization tests, it is essential for the experimenter to know basic rules for the valid execution of such a test. Before dealing with the validity of randomization test procedures, Edgington (1995) specifies a criterion of validity for statistical testing procedures in general. Within the decision-theory model of hypothesis testing, which requires a level of significance to be set in advance of the research, he suggests the criterion:

Decision-theory validity criterion.

A statistical testing procedure is valid if the probability of a Type I error (rejecting H_0 when true) is no greater than alpha, the level of significance, for any alpha.

For example, the practice of determining significance of a one-tailed test in accord with the obtained, rather than the predicted, direction of difference between the means is an invalid procedure because the probability of rejecting H_0 when it is true is greater than alpha. This criterion, implicit in most discussions of the validity of statistical testing procedures, is expressed in terms associated with the decision-theory model: "rejection," "type I error," and "alpha". This might create the impression that only within the Neyman-Pearson decision-theory framework of hypothesis testing can one have a valid test. Edgington (1995) notes that restriction of validity to situations with a fixed level of significance may be suitable for quality control in industry, but not necessarily for scientific experimentation: that interest in pre-set levels of significance is not universal; and that for those experimenters who are interested in using the smallness of a p-value as an indication of the strength of evidence against H_0 (an interpretation of p-values inconsistent with the decision-theory approach), a more general validity criterion is required. He suggests the following operationally equivalent validity criterion that does not use decision-theory terminology:

General validity criterion.

A statistical testing procedure is valid if, under the null hypothesis, the probability of an exact probability or significance value as small as p is no greater than p, for any p.

For example, under the null hypothesis the probability of obtaining a significance value (p-value) as small as 0.05 must be no greater than 0.05, obtaining one as small as 0.03 must be no greater than 0.03, and so on.

The two criteria of validity are equivalent. For any procedure they lead to the same conclusion regarding validity. The general validity criterion can be used by experimenters interested in the decision-theory approach. It is useful also to experimenters who do not set levels of significance in advance but instead use the smallness of the p-value as an indication of the strength of the evidence against H_o. Such experimenters may report their results as significant at the smallest conventional alpha level permitted by their results.

3.5 RANDOM VERSUS SYSTEMATIC DATA PERMUTATION

There are two basic methods of permuting data to compute significance using a randomization test. Onghena (1994) showed that there is considerable lack of agreement on the nomenclature, which has not been standardized. Onghena (1994) refers to the two methods as "exhaustive" and "non-exhaustive". This dissertation will adhere to the usage of Edgington (1995), who refers to the two methods as "systematic data permutation" and "random data permutation". In systematic data permutation, data are permuted systematically (nonrandomly) in determining significance. This is usually done when the number of permutations is small enough to make it practicable.

Random data permutation uses a random sample of all possible data permutations to determine significance. It serves the same function as systematic data permutation with a substantial reduction in the number of permutations that need to be considered. Instead of requiring millions or billions of data permutations, as would be required for the systematic data permutation method for many applications of randomization tests (as exemplified by this dissertation, in which the clinical experiment reported in Chapter 4 has 10^8 possible permutations), the random data permutation method may be effective with as few as 1,000 data permutations (Edgington, 1995).

Should an experimenter decide that the resources needed to deal with all data permutations in the relevant set are too great to be practical, he could use the random permutation method as follows. He performs 999 random data permutations, which is equivalent to selecting 999 data permutations at random from the set that would be used with the systematic method. Under H_o, the data permutation representing the obtained results is also selected randomly from the same set. So, given the truth of H_o, we have 1,000 data permutations that have been selected randomly from the same set, one of which represents the obtained results. Significance is determined as the proportion of the 1,000 test statistic values that are as large as the obtained value. The demonstration that this is a valid procedure for determining significance is as follows (Edgington, 1995). Under H_o the obtained test statistic value (like the 999 associated with the 999 data permutations selected at random) can be regarded as randomly selected from the set of all possible test statistic values, so that we have 1,000 randomly selected values, one of which is the obtained value. If all 1,000 test statistic values were different values, so that they could be ranked from low to high with no ties, under H_o the probability would be 1 / 1000, or 0.001, that the obtained test statistic value would have any specified rank from 1 to 1,000. So the probability that the obtained test statistic value would be the largest of the 1,000 would be 0.001. If some values were identical, the probability could be less than 0.001, but not greater. Given the possibility of ties, the probability is no greater than 0.001 that the obtained test statistic value would be larger than all of the other 999 values. Similarly, the probability is no larger than 0.002 that it would be one of the two largest of the 1,000 values, and so on. In general, when H_o is true, the probability of getting a p-value as small as p is no greater than p, and so the method is valid.

A randomization test using the random data permutation method, although employing only a sample of the possible data permutations, is therefore valid. When H_o is true, the probability of rejecting it at any alpha level is no greater than alpha. If however H_o is false and there is an actual treatment effect, random data permutation is less powerful than systematic data permutation based on the entire set of possible

assignments. Increasing the number of data permutations used with the random data permutation method increases the power of a randomization test employing random data permutation. Edgington (1995) states that using (e.g.) 1,000 data permutations does not give the power provided by several thousand, but that the power is still substantial. He states for example that the probability is 0.99 that an obtained test statistic value that would be judged significant at alpha = 0.01, using systematic data permutation, will be given a p-value no greater than 0.018 by random data permutation with 1,000 random permutations: and that the probability is 0.99 that an obtained test statistic value that would be found significant at alpha = 0.05, from systematic data permutation, will be given a p-value no greater than 0.066 by random data permutation using 1,000 data permutations.

3.6 RANDOMIZATION TEST COMPUTER PROGRAMS.

The amount of computation required for randomization tests made them impractical before the advent of computers. The number of data permutations can be large even for relatively small samples, if systematic data permutation is employed. For example, the small scale clinical experiment described in Chapter 4 generates 10^8 data permutations. Many programs for both systematic and random data permutation methods are available for personal computers. Some, for example Edgington's package of programs called RANDIBM, are available free on the internet. This and other packages of randomization programs with details of how to obtain them are discussed in an Appendix in Edgington (1995).

Because the available programs did not meet the requirements of the present dissertation, special programs relevant to its requirements were written by the author, using QBasic (Perry, 1993) and following the principles described by Edgington (1995). The computer programs used in this dissertation for random data permutation permit the specification of the number of data permutations to be performed. If n permutations are

specified then the programs will treat the obtained results as the first data permutation and randomly permute the data an additional n-1 times.

Edgington (1995) stated that all computer programs for randomization tests must specify the performance of the following operations:

1. Compute an obtained test statistic value, which is the value for the experimental results.
2. Systematically or randomly permute the data.
3. Compare the test statistic value for each data permutation.
4. Compute the significance or probability value. The probability value is the proportion of the test statistic values, including the obtained value, that are as large as (or, where appropriate, as small as) the obtained statistic value.

Figure 3.1 shows a flow chart for randomization test computer programs.

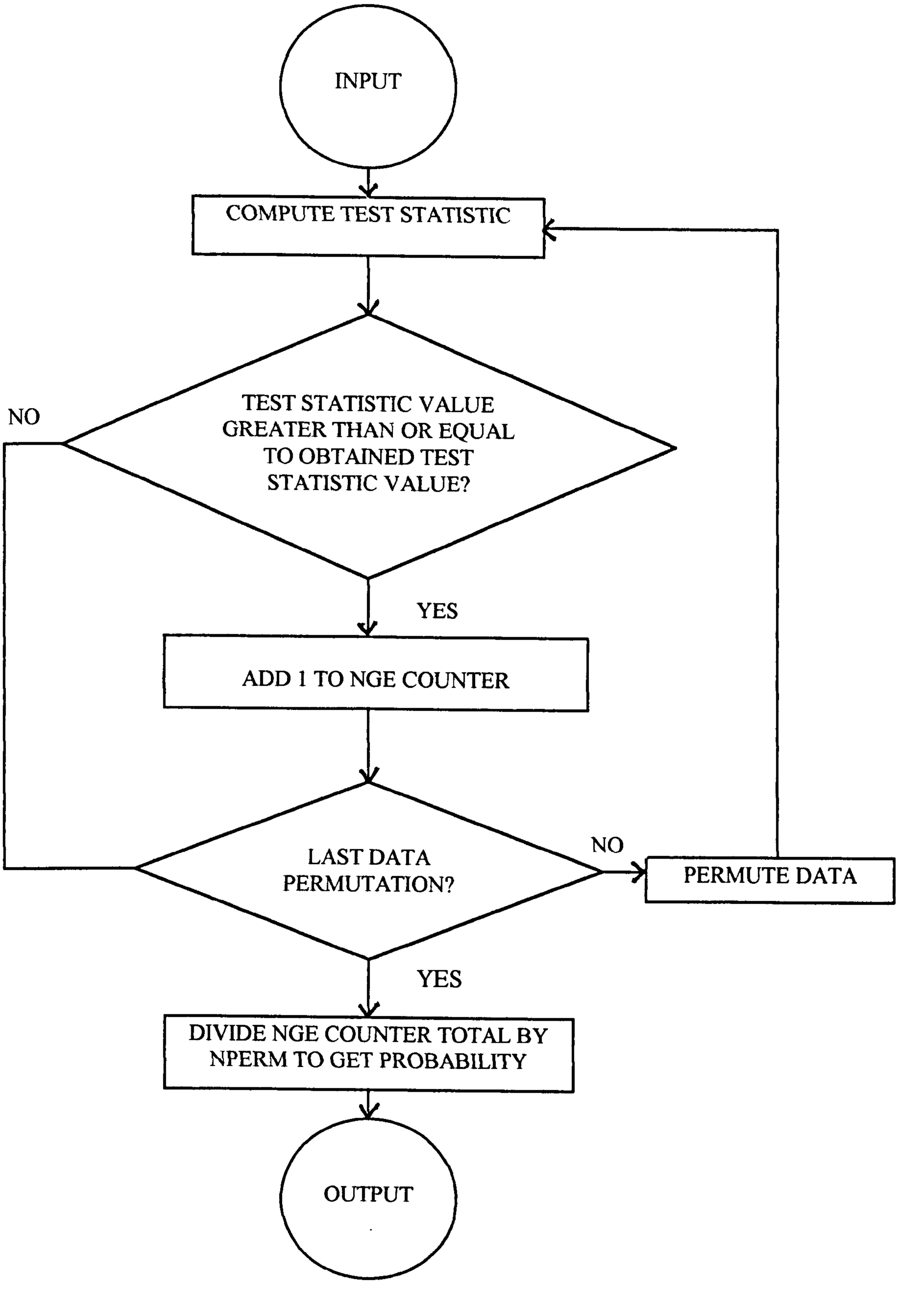


Figure 3.1. Flow chart for randomization test. From Edington (1995). Used with permission.

3.7 RANDOMIZATION TESTS FOR MULTIPLE BASELINE DESIGNS

The traditional, nonrandomised multiple baseline experiment is a replicated AB design (Edgington, 1992). The separate baselines may be data sequences for different subjects, for different behaviours within the same subject, or as in the clinical experiment reported in Chapter 4, a combination of the above. Treatments are introduced at different times for the different baselines in order to reveal associations between intervention and baseline changes.

It is not necessary to have more than one AB baseline to ensure validity when intervention for that baseline is randomized (Edgington, 1992). However, a randomized multiple baseline design can be used to increase the power of AB designs when it is not practical to employ long baselines. This is the logic behind the clinical experiment in Chapter 4. There are many ways in which the AB experiment can be replicated to provide multiple baselines and randomization tests for those multiple baselines. Two such ways are further considered here; the Wampold and Worsham (1986) test and the Marascuilo and Busk (1988) test.

3.7.1 THE WAMPOLD AND WORSHAM TEST

Wampold and Worsham (1986) developed a randomization test for a randomized multiple baseline design that is similar to a conventional, nonrandomized multiple baseline design. Their design, for which they provided hypothetical data, comprised 4 baselines, one for each of 4 subjects. Within each baseline there were 13 possible intervention times. They used only the between baseline feature to develop a randomization test for the multiple baseline design. They fixed the intervention times for the different baselines and then randomly determined which subject takes the earliest intervention, which the next, and so on. They determined a one-tailed test statistic as follows: subtract the mean of the control (pre-experimental or "baseline") measurements

from the mean of the experimental measurements for each baseline and then sum those differences between means over all 4 baselines. This test statistic for the obtained data is then compared to a randomization distribution consisting of test statistics computed for all possible orders in which the subjects could have been subjected to treatment. With 4 subjects there are 4! = 24 possible orders. In their hypothetical experiment the directional randomization test showed that the obtained test statistic was the highest of the 24, giving a p-value of 1 / 24 = 0.042. (Revusky (1967) followed this procedure but applied it to the ranked data, and Wolery and Billingsley (1982) combined Revusky's method with the split-middle technique).

3.7.2 THE MARASCUILO AND BUSK TEST

Marascuilo and Busk (1988) noted that the randomization test of Wampold and Worsham has low power because of the restricted randomization possibilities. They proposed an increase in power by determining the intervention points at random instead of fixing them in advance. This also has the effect of randomly selecting the order in which the subjects are subjected to the treatment. The control over historical sources of confounding is obtained by randomization within a baseline, and the randomization distribution is obtained not only by locating the intervention point of a baseline in the other baselines, but also by locating all possible intervention points in a baseline. To illustrate the greater power of this approach, they assumed that Wampold and Worsham's hypothetical data had been generated by this experimental design. They proposed as a test statistic the difference between the means for the A and the B phase summed over all the 4 baselines. The p-value of this statistic can be derived by comparing it to the reference distribution of this statistic. With 13 possible intervention points in each of 4 baselines, there are 13^4 = 28,561 possible data permutations. Six of the data permutations gave a test statistic value as great as or greater than the obtained test statistic, giving a p-value of 6 / 28,561 = 0.002. (These data were used to test the program marbus.3 presented at Appendix 5). Note that the p-value is considerably

smaller than that obtained by the Wampold and Worsham procedure above. The Marascuilo and Busk procedure is that which is utilised in the clinical experiment described in Chapter 4.

3.7.3 MARASCUILO AND BUSK IN SINGLE-SUBJECT APPLICATIONS

The above randomization tests of Wampold and Worsham (1986) and Marascuilo and Busk (1988) were described as tests of multiple baselines, replicated over subjects. In both papers there was reference to the relevance of their procedures to single-subject multiple baseline designs. Wampold and Worsham considered their procedure applicable to both types of design, shown by their reference to "randomly selecting the order in which the subjects, behaviors, or situations are subjected to the treatment". Conversely, Marascuilo and Busk considered that correlations between behaviours within a subject would tend to render their own procedure inappropriate for single-subject designs:

"One might be tempted to use the proposed methods with multiple baseline designs across behaviors. In most cases, the application cannot be justified because of the correlations that exist between the measures of different behaviors made at the same time (p.23)."

There are two ways in which behaviours in single-subject multiple baselines are likely to be correlated: (1) behaviours covary in the absence of treatment interventions, and (2) a treatment intervention affects more than one behaviour.

Edgington (1992) addressed this problem. He argued that the correlation between behaviours causes difficulty in interpreting significant results, but that the validity of the test is unaffected and that the procedure is valid for application to multiple baseline data from correlated behaviours within a subject. The null hypothesis is that the data for all baselines are the same as they would have been under any possible alternative treatment

intervention. Therefore, if the null hypothesis is rejected, the alternative hypothesis that is accepted is that treatment intervention had an effect on one or more of the baselines. The procedure does not, however, permit the inference of which baseline or baselines were affected. That is, the procedure can be validly applied to single-subject designs with correlated behaviours, but the statistical inference that can be drawn from significant results is not very specific. The implication is simply that somewhere within the configuration of baselines, at least one of the treatment interventions affected at least one of the behaviours. This can still be useful. Adapting an argument by Edgington (1992, p. 153) to the present concern, suppose that the intervention for each baseline were a cognitive approach and that the dependent variable were a measure of strength of conviction in a delusional idea. Then, even if we could not infer which belief responded to the intervention, it might be very useful to have evidence that delusional conviction was influenced by the cognitive intervention, especially if that subject had not previously been known to be responsive in this way.

Two types of correlation were considered: (1) covariation of behaviours in the absence of treatment intervention, and (2) the effect of treatment intervention on more than one bahaviour. In either case, a high correlation between behaviours makes the randomization test less powerful (less likely to detect treatment effects that exist). Edgington (1992) argues that when behaviours covary greatly in the absence of treatment, it is difficult to detect treatment effects that may be small relative to other variation. When there are correlated responses to a single intervention, that is if an intervention on one baseline tends to affect other baselines as well, causing data shifts before or after interventions on those baselines, there may be difficulty in detecting intervention effects. Edgington (1992) argues that these problems may be minimized by selecting interventions likely to have large effects relative to variation in the baseline and by selecting interventions and behaviours to ensure that the intervention for a baseline will primarily affect the behaviour for that baseline rather than the behaviour for other baselines. He states that the procedure can therefore be employed with designs that

minimize the effects of correlated behaviours and correlated effects on behaviours of a baseline intervention in order to make the test sensitive, despite the fact that it is valid even when those effects are not minimized. The above concerns are relevant to the clinical experiment of Chapter 4, though covariations of the kind described above did not appear to present a significant problem.

CHAPTER 4. COGNITIVE BEHAVIOUR THERAPY APPLIED TO PSYCHOTIC DELUSIONS

4.1 HISTORICAL REVIEW

Beck (1952) first reported the application of cognitive therapy in schizophrenia. He encouraged a person with chronic schizophrenia to examine the appearance and behaviour of alleged FBI agents who were visiting his shop, to test his belief that these people had him under surveillance. He was able to narrow down his list of 50 "suspects" to 2-3 possibilities, and stated that he felt he would soon be able to "eliminate them completely". The delusion proved modifiable even though it had been present for 7 years. Hole, Rush and Beck (1979) commented "The combination of tracing the antecedents of the delusion, and helping the patient to test his conclusions systematically, helped him to recognize and to gradually do away with the irrational and rigid belief system.".

Shapiro and Ravenette (1959) reported a preliminary experiment on paranoid delusions in a single case, in which they attempted to scale the intensity of delusional beliefs using Shapiro's (1961) Personal Questionnaire (PQ) technique.

Watts, Powell and Austin (1973) reported the attempted modification of paranoid beliefs in paranoid schizophrenia. They noted that confrontation of such beliefs could result in "psychological reactance", whereby the target beliefs could become more firmly held or even more extreme. They suggested 4 ways to minimize psychological reactance:

1. Target less strongly held beliefs first, unless specific themes bind beliefs together.
2. Avoid direct confrontation, asking the subject merely to consider the facts and arguments discussed with him, and to entertain possible alternative beliefs.
3. Centre discussion not on the belief itself, but on the subject's evidence for it.

4. Encourage the subject to voice the arguments against his own beliefs, even if quite direct questioning is needed to achieve this.

They showed that belief modification with graded re-exposure to avoided social circumstances was successful in reducing the intensity of the beliefs.

Milton, Patwa and Hafner (1978) reported a small-scale study which suggested the efficacy of belief modification through verbal intervention in a proportion of persistently deluded patients.

Hole, Rush and Beck (1979) encouraged 8 delusional inpatients to discuss the nature of their delusional beliefs and the evidence supporting them. These authors defined 4 dimensions for measuring delusions:

1. Conviction.
2. Accommodation (the degree to which a delusion could be modified by external events or incongruities).
3. Pervasiveness (the percentage of the day spent ruminating about delusional concerns, seeking delusional goals, or interpreting experience in terms of delusional systems).
4. Encapsulation (the extent to which a decrease in pervasiveness could occur without an associated decrement in conviction).

These authors found that half the patients showed no change; the remaining half showed reduced pervasiveness, and half of these also showed reduced conviction. They concluded: "We suggest that delusions may function in much the same way as other beliefs and convictions. Delusions may differ from other beliefs only quantitatively with respect to how easily they can be modified by external events.".

Brett-Jones, Garety and Hemsley (1987) studied 9 hospitalized schizophrenic patients,

seen weekly soon after admission. They measured 3 key components of recovery from delusions: strength of conviction in the belief; preoccupation with the belief (that is the amount of time spent thinking about the belief); and the degree to which the belief interfered with the person's daily life. Their results supported a multi-dimensional view of delusions. 7 of the 9 subjects showed fluctuating scores on conviction and preoccupation, with decreases in conviction tending to precede decreases in preoccupation. Correlations between conviction and preoccupation and conviction and interference, for the group as a whole, were not significant (although this may have been due to low power as the sample size was small) suggesting that these components are orthogonal dimensions.

Chadwick and Lowe (1990) reported further evidence for a multi-dimensional view of delusions. Six patients who had held fixed delusional beliefs for 2 or more years were monitored before, during and after 2 psychological interventions; a structured verbal challenge and reality testing. The data for individual clients showed a high degree of desynchrony between conviction, preoccupation and anxiety caused by thinking about the beliefs as the delusions receded.

There has been a modest number of studies reporting attempts to weaken delusions using cognitive techniques, with generally favourable results (Alford, 1986; Alford and Beck,1994; Beck, 1952; Chadwick and Lowe, 1990; Chadwick, Lowe, Horne and Higson, 1994; Fowler and Morley, 1989; Garety, Kuipers, Fowler, Chamberlain and Dunn,1994; Kingdon and Turkington, 1991; Kingdon and Turkington, 1994; Kingdon, Turkington and John, 1994; Hartman and Cashman, 1983; Himadi and Kaiser, 1992; Hole, Rush and Beck, 1979; Johnson, Ross and Mastria, 1977; Lowe and Chadwick, 1990; Milton, Patwa and Hafner, 1978).

Recently reviews and treatment manuals have been published (Birchwood and Tarrier, 1994; Chadwick, Birchwood and Trower, 1996; Fowler, Garety and Kuipers, 1995);

Kingdon and Turkington, 1994).

4.2 CRITIQUE OF CBT APPLIED TO DELUSIONS

Bouchard, Vallieres, Roy and Maziade (1996) reported a critical analysis of cognitive restructuring in the treatment of psychotic symptoms in schizophrenia. They considered 3 elements in evaluating each study: (a) if subjects are reliably diagnosed with schizophrenia with chronic course and severe impairment; (b) if psychotic symptoms are adequately measured; and (c) if designs are methodologically sound. They found that schizophrenia was not reliably diagnosed and that severity was low to moderate. Assessment of psychotic symptoms was satisfactory, but assessment of generalization to other areas was limited. They found that only 5 studies possessed reliable design and were performed with schizophrenia subjects, and that these studies suggested that cognitive restructuring was effective to reduce or eliminate hallucinations or delusions in schizophrenia patients.

When evaluating methodology and research design in the studies they examined, Bouchard et al. (1996) noted that intrasubject designs can provide a powerful experimental methodology to infer the effectiveness of an intervention if they are rigorously applied. Among the important issues with this methodology were the presence of continuous assessment as well as a baseline that is sufficiently long and stable before introduction of the intervention. On the other hand, group designs offered many advantages but they also required more subjects and some level of complexity to be reliable, e.g. the presence of a control condition.

Bouchard et al. found that 3 of the 12 studies using an intrasubject design did not incorporate any baseline, and one study used only one observation as a baseline. Follow-up information was not provided in 5 studies. Many of the intrasubject designs were not very sophisticated, but the following 5 studies were considered to have utilized more

rigorous approaches: Alford (1986); Chadwick and Birchwood (1994) (this study was concerned with hallucinations rather than delusions); Chadwick and Lowe (1994); Fowler and Morley (1989); and Himadi and Kaiser (1992).

Bouchard et al. recommended that the results of the studies by Hole, Rush and Beck (1979) and by Watts, Powell and Austin (1973) should be "interpreted with extreme caution" due to the absence of any baseline and almost no follow-up. They argue that the two group-design studies by Kingdon and Turkington (1991) and by Milton, Patwa and Hafner (1977) should also be considered with the same extreme caution due to the absence of any control condition. They considered the group-design study by Garety et al. (1994) to have been "very well conducted", although assignment to each condition was not random.

Bouchard et al. considered the outcome of cognitive restructuring in the reviewed studies. In general, they found that cognitive restructuring led to a decrease on the measures that were specific to hallucinations or delusions. Treatment of hallucinations was generally less successful than that of delusions. Important methodological considerations restricted the number of studies that could be used to reliably assess outcome. A number of studies were considered to have been rigorous and to have been performed with subjects who were diagnosed with schizophrenia according to "valid diagnostic criteria" (sic): Alford (1986), Chadwick and Birchwood (1994), Chadwick's set of studies on delusions (Chadwick and Lowe, 1990, 1994; Chadwick et al., 1994; Lowe and Chadwick, 1990); Garety et al. (1994), and Himadi and Kaiser (1992). One further study could be considered rigorous but was not performed with schizophrenia subjects (Fowler and Morley, 1989).

However, Bouchard et al. stated that before discarding too rapidly the studies they considered less rigorous, it was essential to recognize that they often represented pioneering work that had helped the field to progress. In this context they cited the work

by Alford, Fleece and Rothblum (1982), Hartman and Cashman (1983), Hole, Rush and Beck (1979), Milton, Patwa and Hafner (1977) and Watts, Powell and Austin (1973).

Although Bouchard et al. examined studies concerned with both hallucinations and delusions, attention will be focussed here solely on delusions. Bouchard et al. found that when considering all the rigorous studies addressing delusions, there was a substantial reduction in conviction, except for one subject who relapsed at follow-up (Alford et al., 1982) and one subject who developed a new delusional belief (Lowe and Chadwick, 1990). They also found a fairly important and positive effect of the intervention on secondary measures such as Scale for the Assessment of Positive Symptoms scores, anxiety or depression. They concluded that in 3 of the 4 rigorous studies on delusions, (Alford, 1986; Chadwick and Lowe, 1994; Himadi and Kaiser, 1992) cognitive restructuring led to a reduction or an elimination of delusions in 12 of the 14 subjects. In the fourth study (Garety et al., 1994), results indicated a reduction in conviction and acting on delusions that was significantly greater than in the waiting list control group.

Recently the efficacy of cognitive behaviour therapy for people with treatment resistant delusions and voices has been demonstrated by large randomized controlled trials, which were reviewed by Fowler, Garety and Kuipers (1998). Published data are presently available only for the trial conducted by Kuipers et al. (1997). Data from two further large RCTs have been reported at conference (Tarrier, 1997; Kingdon, 1997). Fowler et al. (1998) conclude that the evidence from RCTs provides strong support for the use of cognitive-behavioural therapies with people who present with distressing delusions and voices.

Bouchard et al. stated that "it is important to recall that despite the complexity of any intrasubject design, one major limitation is the frequent use of visual inspection, rather than statistical analysis.". This problem has been addressed in Chapter 2, above. They noted that visual inspection is not a reliable technique, citing in support Matyas and

Greenwood (1990). To illustrate the point, they used Crosbie's (1993) ITSA (interrupted time-series analysis) program to perform brief time-series analysis of data from Chadwick and Lowe's (1990) paper for subject number 6 in their Figure 2. Visual inspection of the data on preoccupation with the delusional belief in this subject suggested a reduction which was, however, found to be nonsignificant with ITSA. Bouchard et al. concluded that "Subsequent research should therefore rely more on statistical methods such as time-series analysis wherever possible.".

The conclusion that research should rely more on statistical methods is in line with the arguments of Chapter 2, above. ITSA (Crosbie, 1993) represents one possible form of statistical analysis, which Crosbie claims has adequate statistical power. The use of interrupted time-series analysis alone does not guarantee acceptable experimental design, and it can only be recommended in conjunction with such an acceptable design if it is to be use to detect intervention effects. In addition it is not clear how data from more than one subject can be combined in order to evaluate statistical significance. Because of the above and the necessity to include an element of randomization in an experimental as opposed to a "quasi-experimental" design (Campbell and Stanley, 1966), the use of n = few experimental designs that allow of analysis by randomization tests (see Chapter 3 above) is recommended for small scale clinical research in this field.

The following clinical experiment exemplifies the use of such a design, the randomized multiple baseline design. This is the first reported use of such a design, though the theory has been developed by Marascuilo and Busk (1988). It is essentially a development of the multiple baseline design such as that employed in the study by Chadwick and Lowe (1990) referred to above. The data in the Chadwick and Lowe (1990) study relied on visual inspection of the graphed data without any attempt at statistical evaluation. The randomized baseline design also allows of statistical analysis using randomization tests (see Chapter 3).

4.3 A CLINICAL EXPERIMENT ON DELUSIONAL IDEATION

4.3.1 SUBJECTS.

The subjects were four white British men who were living in a medium-to-low security tertiary intensive therapy psychiatric unit to which they had each been referred from conventional psychiatric wards because their behaviour had been difficult to manage. All had held one or more delusional beliefs for at least two years and all held a diagnosis of schizophrenia by their responsible consultant psychiatrist. All were on stable medication regimes which were not significantly changed during the course of the study. They were aged between 29 and 43, with a mean age of 36. In all cases the duration of the illness was more than 10 years.

Subject 1 "Tom" (all names are disguised to protect anonymity) was 43. His diagnosis was schizophrenia, paranoid type, 295.30 (DSM-IV, American Psychiatric Association, 1994). His medication comprised: Lofepramine 70mg tds; Procyclidine 5mg bd; Lorazepam 2mg IM (PRN). He had for many years held delusional beliefs about sinister groups, usually of foreign origin, who were trying to subvert Great Britain. At the time of the study two beliefs were prominent, both concerning trains. He strongly believed: (Tom1) that local trains he observed were shunting large amounts of gold bullion around the country and (Tom 2) that trains were carrying "Soviet" troops.

Subject 2 "Bill" was 39. His diagnosis was schizophrenia, catatonic type, 295.20 (DSM-IV, American Psychiatric Association, 1994). His medication comprised: Triflouperazine 15mg bd; Carbamazepine 400mg tds; Cyproterone Acetate 50mg bd; Fluphenazine Decanoate 100mg IM every 2 weeks. He had a Wechsler Verbal IQ of 82, Performance IQ of 100 and Full Scale IQ of 89. He believed (Bill 1) that the number "666" had been tattooed onto the back of his head when a child, by a dentist. He thought that this must be the case, else how explain the bad luck that had in his view followed

him throughout his life?

Subject 3 "Pete" was 29. He had a diagnosis of schizophrenia, paranoid type, 295.30 (DSM-IV, American Psychiatric Association, 1994). His medication comprised: Dothiepin 100mg nocte; Propanalol 20mg tds; Procyclidine 5mg bd; Fluphenazine Decanoate 37.5mg IM every 3 weeks. He held two beliefs of a paranoid nature: (Pete1) that ward staff were poisoning his three-weekly Fluphenazine Decanoate injection and (Pete 2) that staff or fellow patients were deliberately contaminating his meals.

Subject 4 "Ken" was 32. He had a diagnosis of schizophrenia, paranoid type, 295.30 (DSM-IV, American Psychiatric Association, 1994). His Wechsler Verbal IQ was reported as 65. The Performance IQ was recorded as 72. There was no record of the Full Scale IQ. His medication comprised: Fluphenazine Decanoate 50mg weekly; Procyclidine 5mg tds; Respiridone 4mg daily. He held three delusional beliefs: (Ken1) that he was a member of the SAS; (Ken 2) that he was being pursued by the IRA; (Ken 3) that he could summon assistance at need by using a special Whitehall telephone number.

4.3.2 METHOD

Measurement of the dependent variable.

Strength of conviction of beliefs was measured by a modified version of Mulhall's (1976, 1978) Personal Questionnaire Rapid Scaling Technique(PQRST) (Cliffe, Possamai and Mulhall, 1995). This yields scores ranging from 0 (zero strength of conviction) to 9 (maximum strength of conviction). This instrument was used because it was found by Cliffe, Possamai and Mulhall (1995) to be much easier to employ with this type of client than traditional Personal Questionnaire methods as described in the pioneering work of Brett-Jones, Garety and Hemsley (1987) and of Garety (1985,

1992). It is easy and quick to administer. Administration follows the instructions in the PQRST Manual (Mulhall, 1978) and scoring is simply by the template provided. Each belief can be assessed in about two minutes. It is idiographic and allows clients to express their beliefs in their own words. Reliability of the results can be assessed by reference to the internal consistency of the answers. It provides an even coverage over the entire continuum and therefore lacks response bias. Measurement of the strength of conviction took place at the end of each session.

In line with the study by Chadwick and Lowe (1990), two further aspects of the subjects' delusional experience were measured; preoccupation (defined as the percentage of time spent thinking about the belief) and the amount of distress or anxiety caused by the belief.. These were both measured using the original unmodified version of Mulhall's (1978) Personal Questionnaire Rapid Scaling Technique. In both cases the measures were retrospective, applying to the week before the assessment.
Measures were taken at the beginning of each session.

Experimental design.

A randomized multiple baseline design (Wampold and Worsham, 1986; Marascuilo and Busk, 1988) was used. For each delusional belief, 19 sessions were allocated. The first 5 sessions were unconditionally reserved as baseline sessions. Treatment could not begin later than session 15, to guarantee a minimum of 5 sessions available for treatment. A random session number between 6 and 15 was selected. Thus for each belief there were 10 possible session numbers that could be designated for the start of the treatment condition. With 8 beliefs, and 10 possible random intervention points for each belief, there are 10^8 possible ways in which the 8 random intervention points could have been assigned. The actual session numbers that were randomly generated by computer were as follows: Tom 1 (7); Tom 2 (15); Bill 1 (11); Pete 1 (6); Pete 2 (14); Ken 1 (6); Ken 2 (10); Ken 3 (15).

The experimental effect was defined as the difference between the mean score in the baseline phase and the treatment phase, summed over the 8 beliefs. Statistical analysis was performed by a non-exhaustive randomization test with 10^5 data permutations.

Because previous work in this field (e.g. Chadwick and Lowe, 1990) had suggested relatively large experimental effects (based on visual inspection of the data as the Effect Sizes were not computed), the significance level alpha was set at $p < 0.01$.

4.3.3. RESULTS

Table 4.1 and Figures 4.1.a, 4.1.b and 4.1.c show the strength of conviction in the 8 delusional beliefs over the 19 sessions, for the baseline phase and the treatment phase, and at follow-up at 1, 3 and 6 months. Figure 4.2 shows the same data with the order of the graphs presented so that the intervention points are in sequence. For 7 of the 8 beliefs, conviction score was a maximum 9 throughout the baseline phase. For one belief (Tom 2) the conviction score varied between 8 and 9 during the baseline phase.

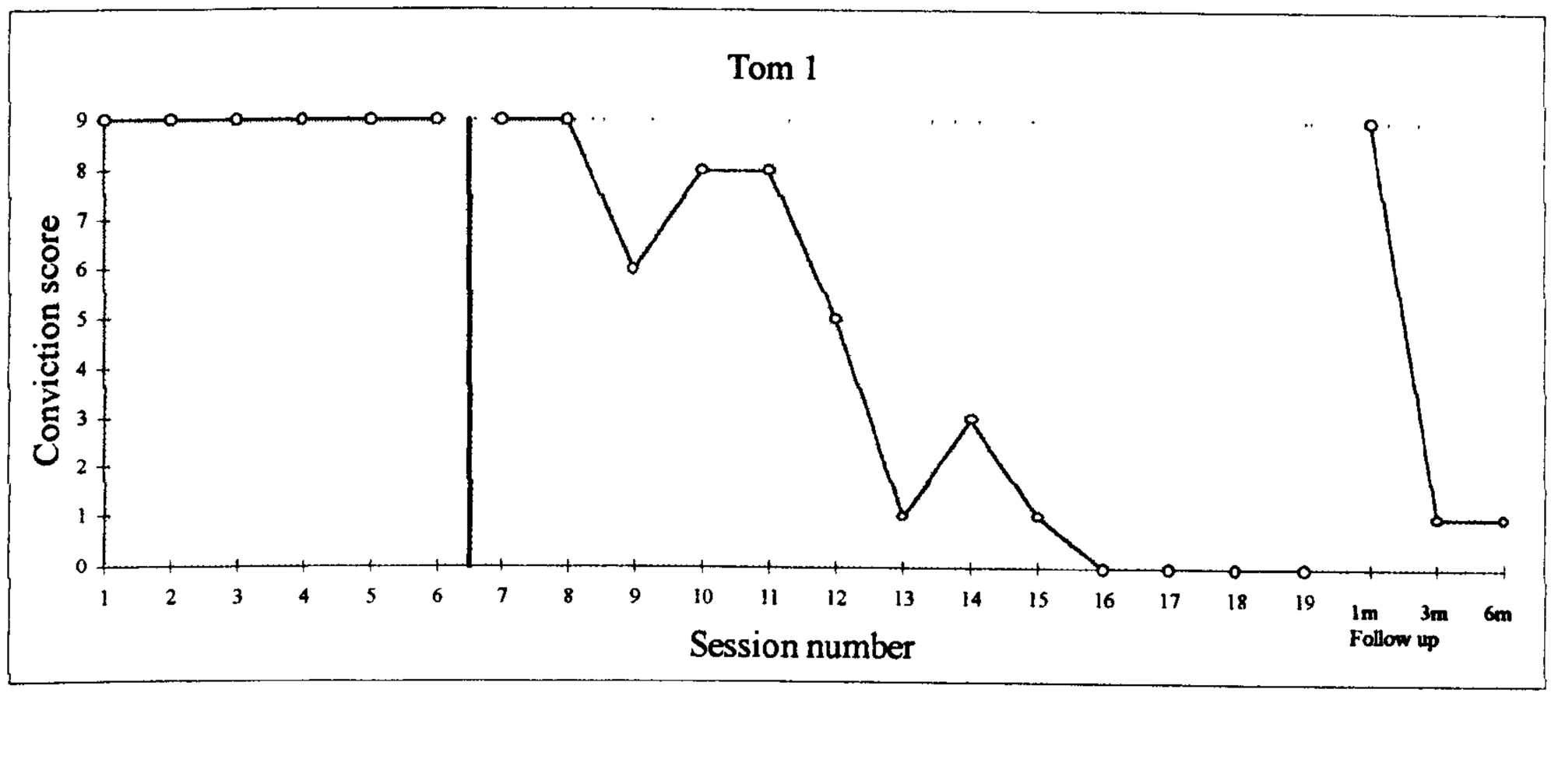
Figures 4.3.a, 4.3.b and 4.3.c show the PQRST preoccupation (PREO) and distress (DIST) scores associated with the 8 delusional beliefs over the 19 sessions, for the baseline phase and the treatment phase, and at follow-up at 1, 3 and 6 months. Figure 4.4 shows the same data with the order of the graphs presented so that the intervention points are in sequence. The effect of the therapeutic intervention on each of the beliefs will be described for each Subject.

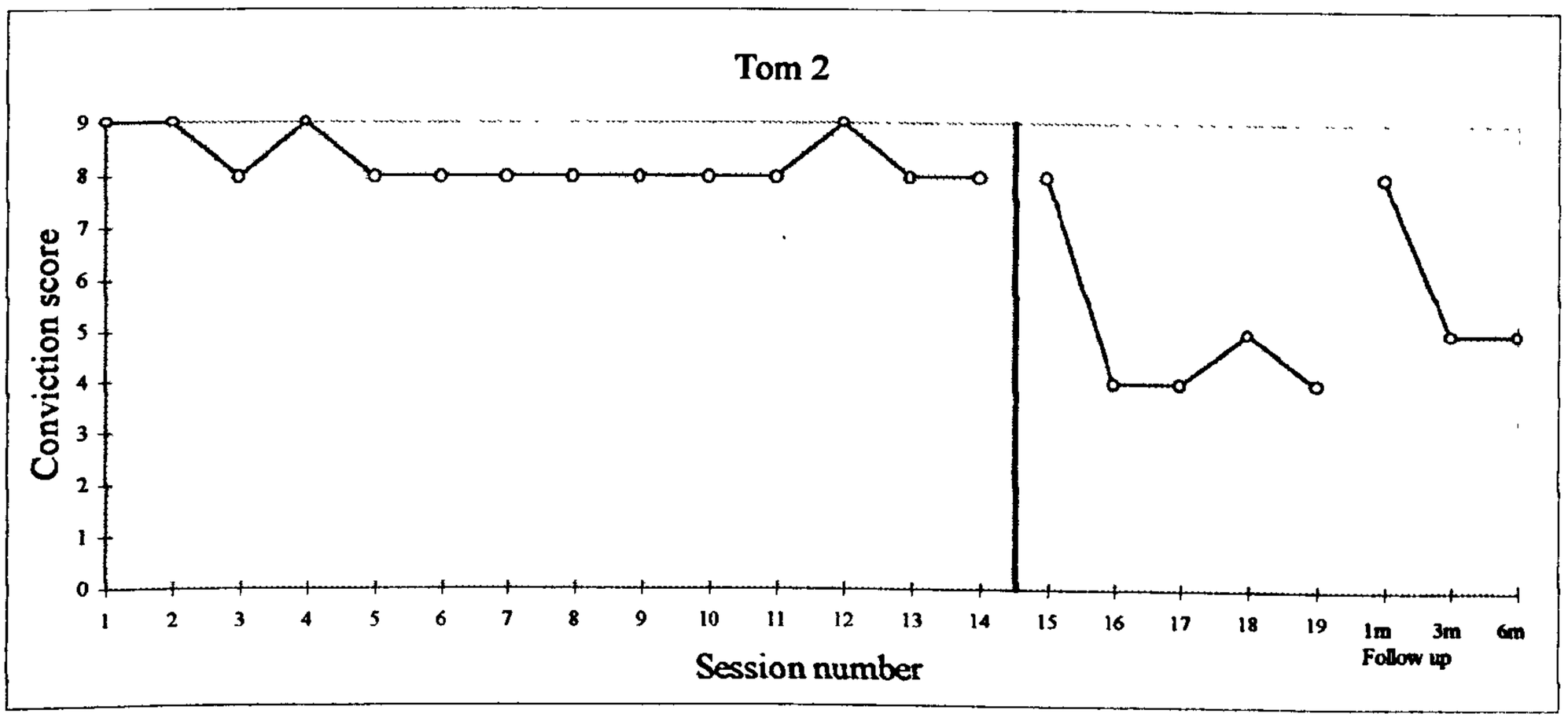
The treatment procedures used followed those outlined by Fowler et al (1998).

	Session	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	F1	F3	F6
BELIEF																							
PETE 1	PHASE	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B	B	B	B	B	F	F	F
	SCORE	9	9	9	9	9	0	0	0	1	0	5	5	5	6	0	0	0	0	0	0	5	5
KEN 1	PHASE	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B	B	B	B	B	F	F	F
	SCORE	9	9	9	9	9	5	8	5	0	6	0	0	0	0	0	0	0	0	0	5	0	0
TOM 1	PHASE	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B	B	B	B	F	F	F
	SCORE	9	9	9	9	9	9	9	9	6	8	8	5	1	3	1	0	0	0	0	9	1	1
KEN 2	PHASE	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B	F	F	F
	SCORE	9	9	9	9	9	9	9	9	9	9	8	8	5	9	8	8	0	1	0	0	8	0
BILL 1	PHASE	A	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	F	F	F
	SCORE	9	9	9	9	9	9	9	9	9	9	0	0	0	0	0	0	0	0	0	0	0	0
PETE 2	PHASE	A	A	A	A	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	F	F	F
	SCORE	9	9	9	9	9	9	9	9	9	9	9	9	9	5	1	0	0	1	1	5	7	5
TOM 2	PHASE	A	A	A	A	A	A	A	A	A	A	A	A	A	A	B	B	B	B	B	F	F	F
	SCORE	9	9	8	9	8	8	8	8	8	8	8	9	8	8	8	4	4	5	4	8	5	5
KEN 3	PHASE	A	A	A	A	A	A	A	A	A	A	A	A	A	A	B	B	B	B	B	F	F	F
	SCORE	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9

Table 4.1

Strength of conviction in the 8 delusional beliefs over 19 sessions. A = baseline phase, B = treatment phase. F1, F3, F6 = Follow up at 1m, 3m, 6m.





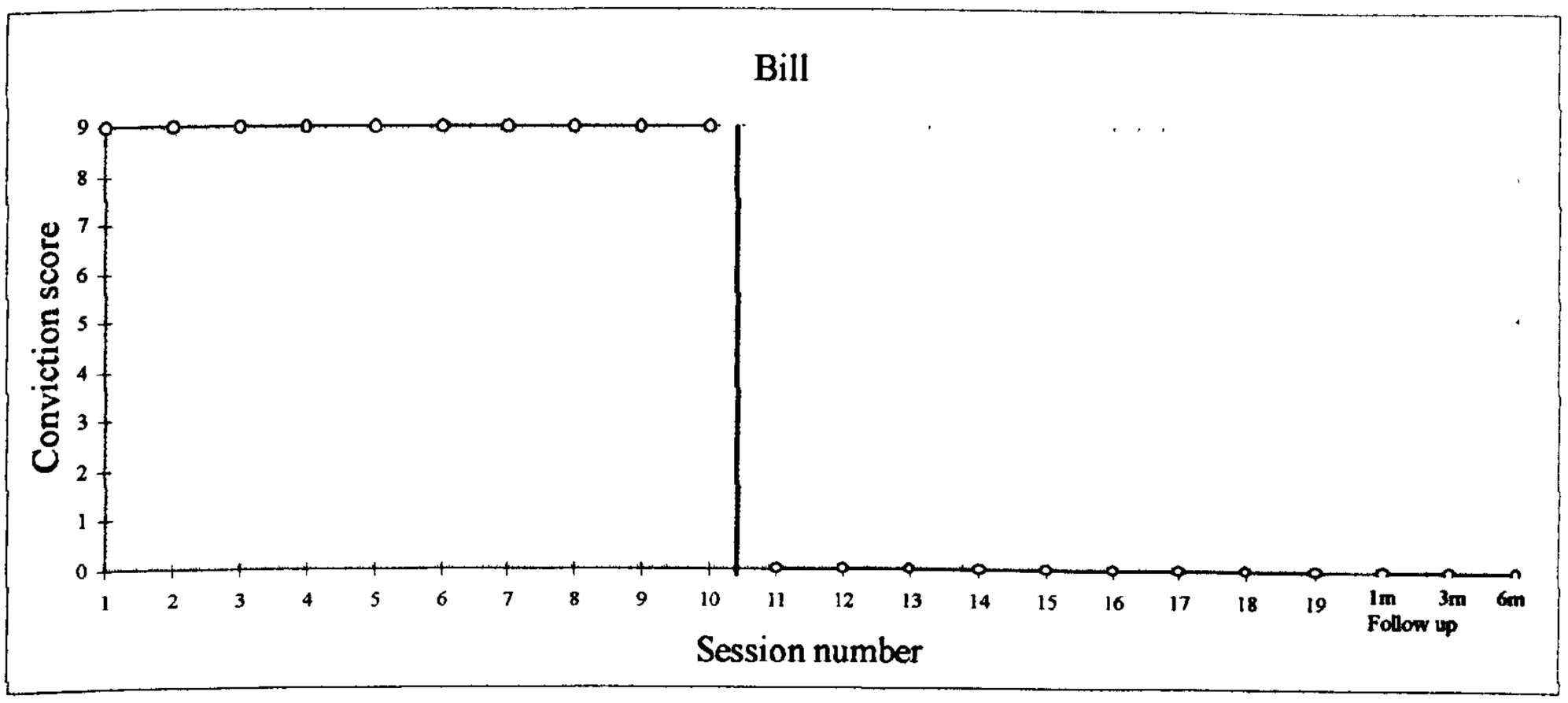


Figure 4.1a

Strength of conviction in delusional beliefs over sessions 1 to 19 and at follow up of 1m, 3m and 6m. Conviction score is from modified PQRST. Vertical lines separate baseline from treatment sessions.

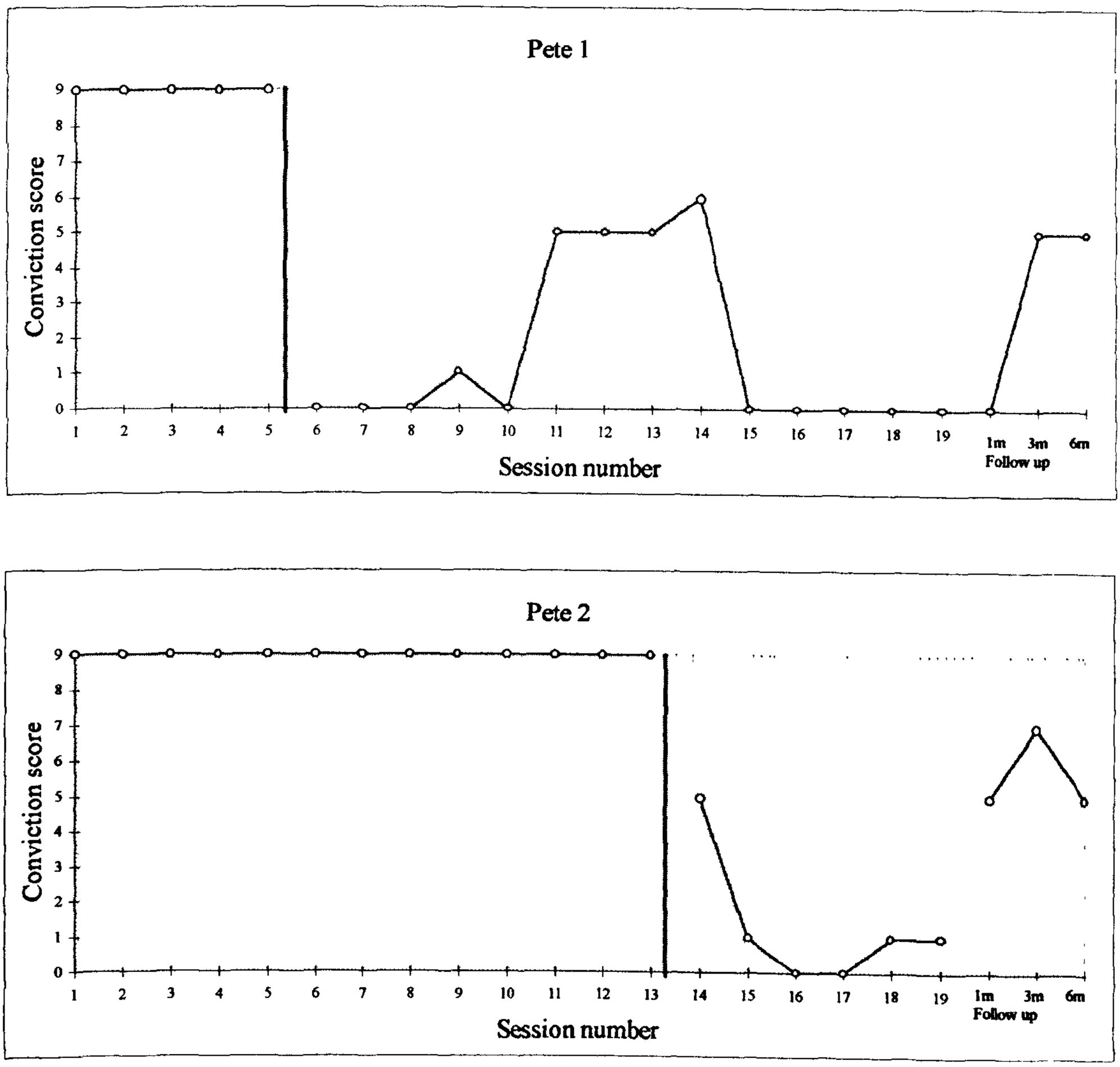
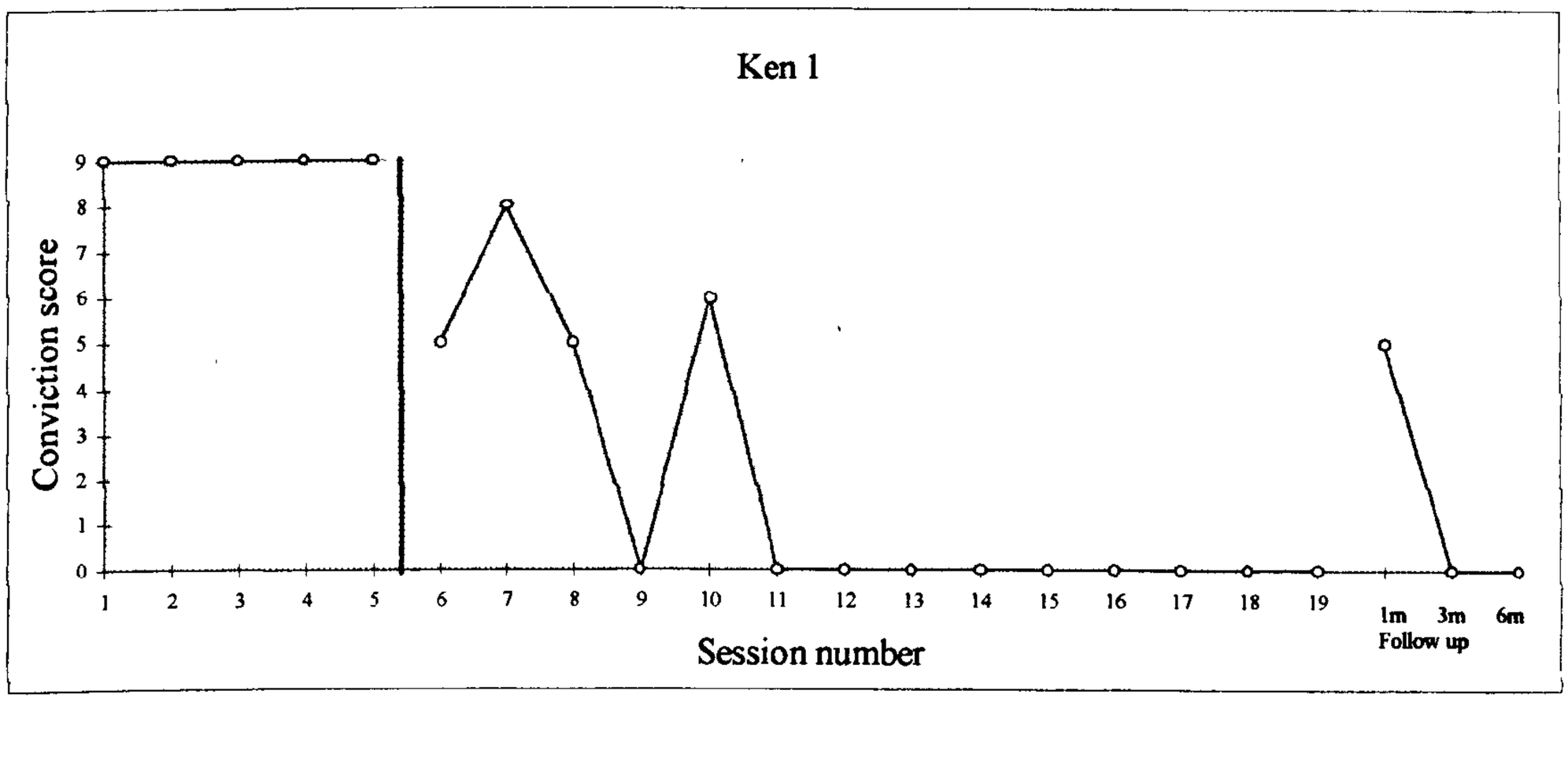
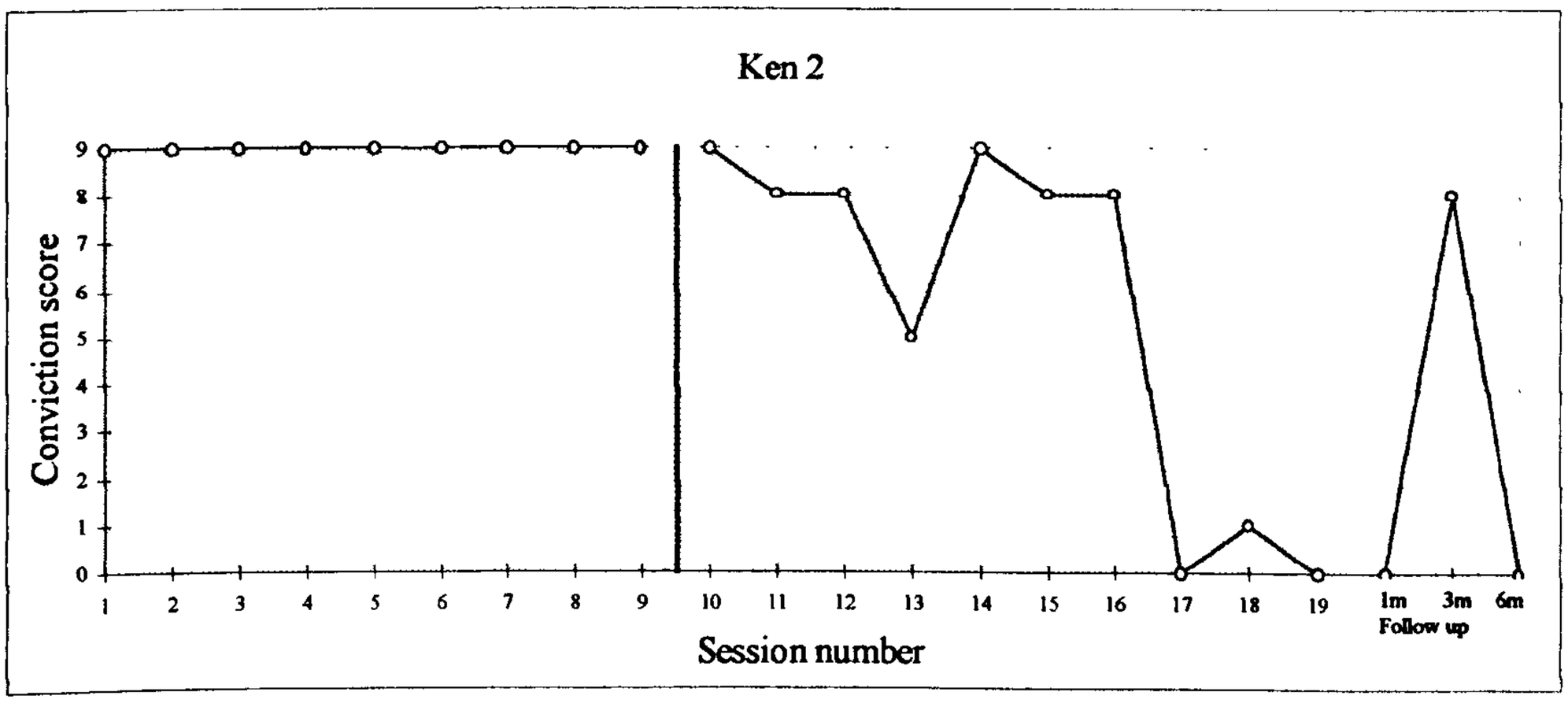


Figure 4.1b

Strength of conviction in delusional beliefs over sessions 1 to 19 and at follow up of 1m, 3m and 6m. Conviction score is from modified PQRST. Vertical lines separate baseline from treatment sessions.





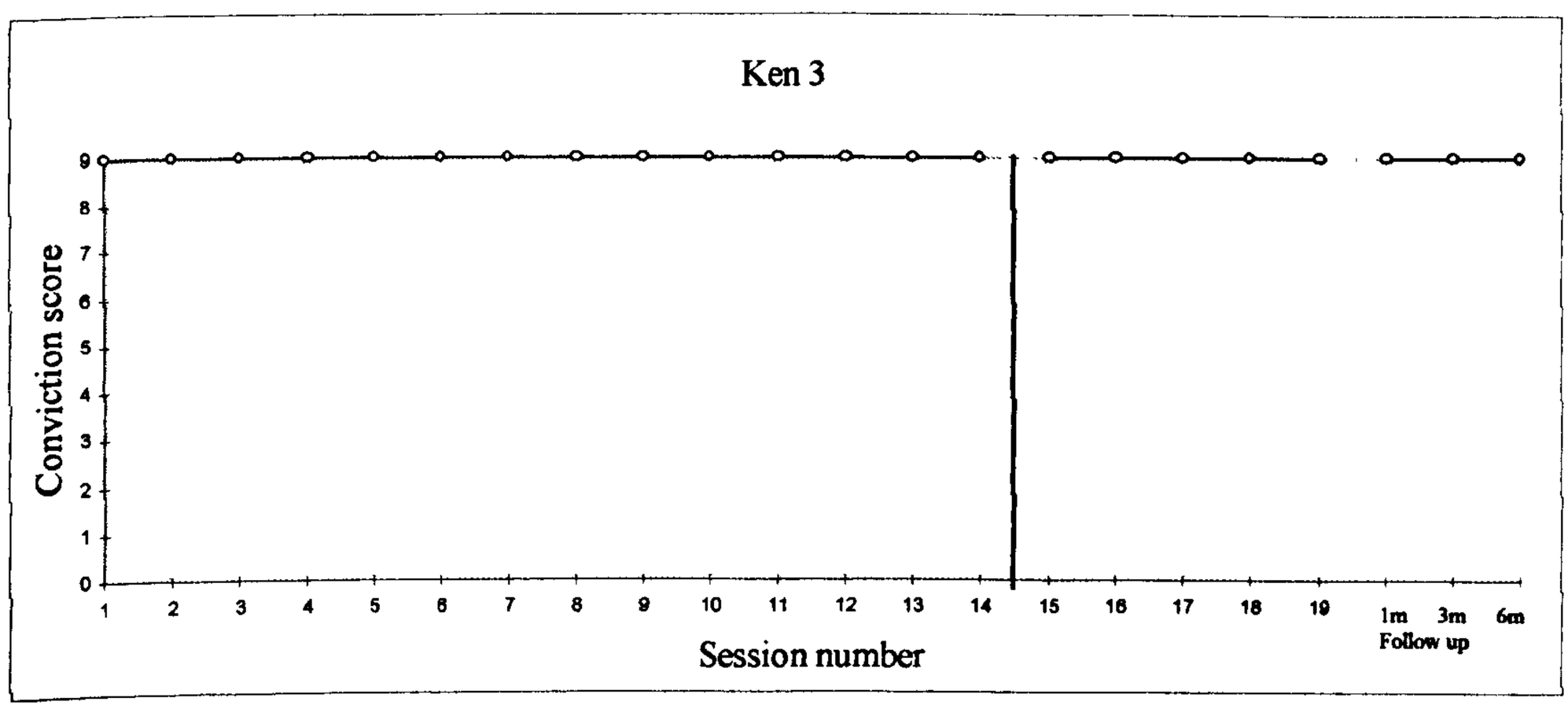
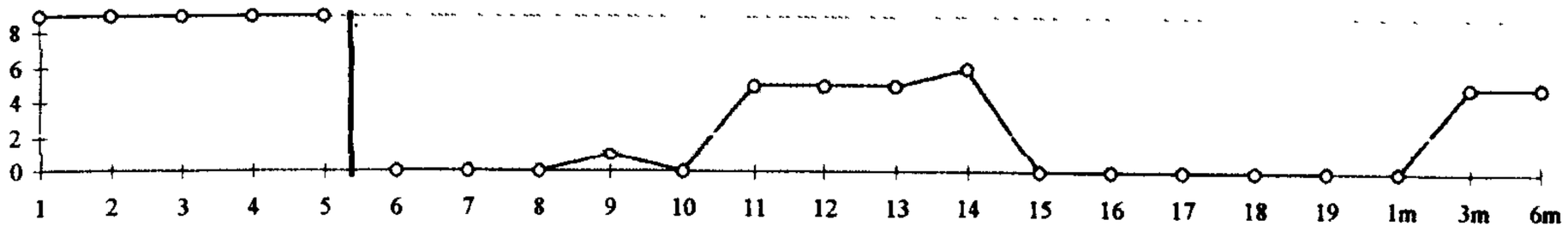


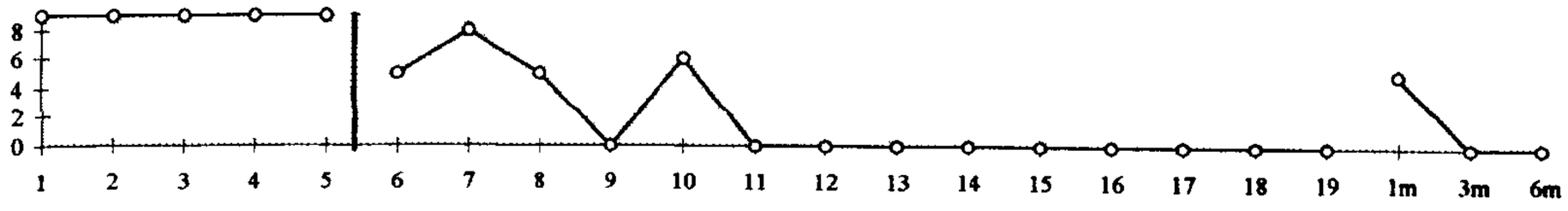
Figure 4.1c

Strength of conviction in delusional beliefs over sessions 1 to 19 and at follow up of 1m, 3m and 6m. Conviction score is from modified PQRST. Vertical lines separate baseline from treatment sessions.

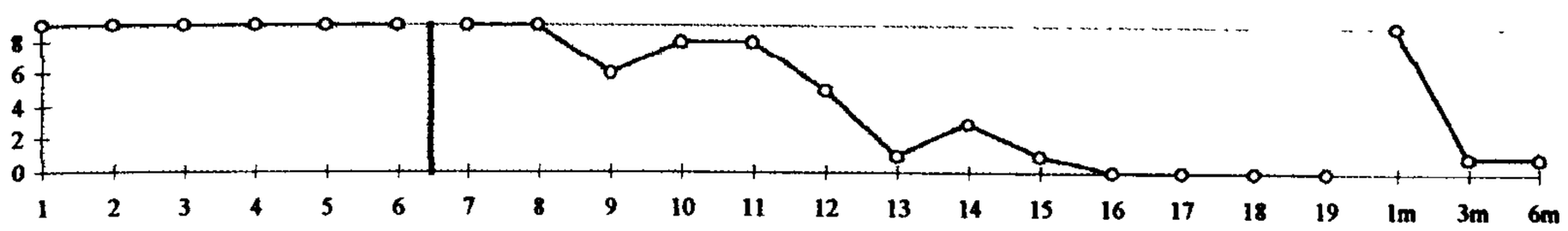
Pete 1



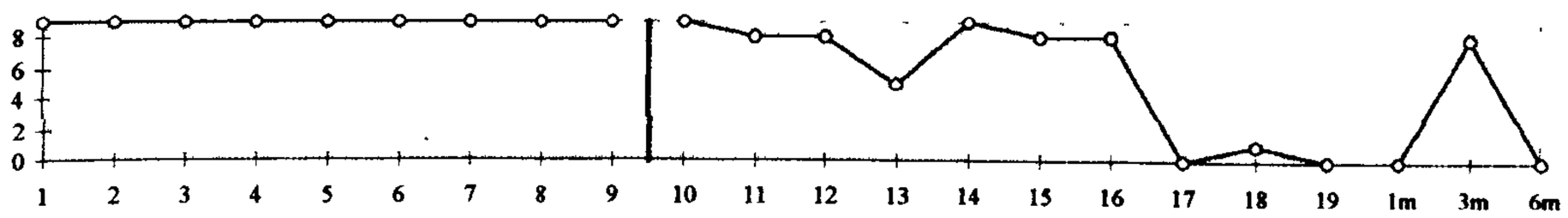
Ken 1



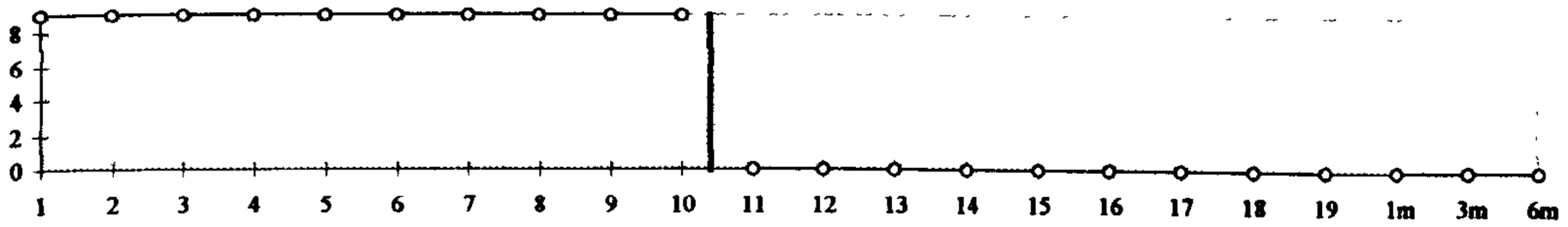
Tom 1



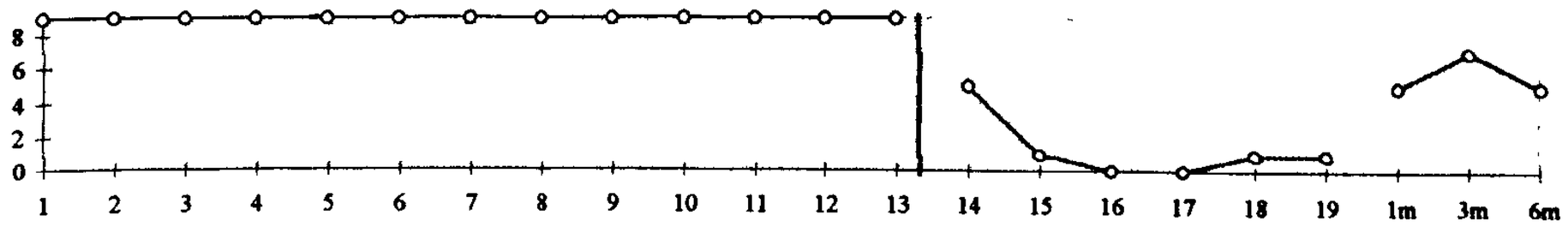
Ken 2



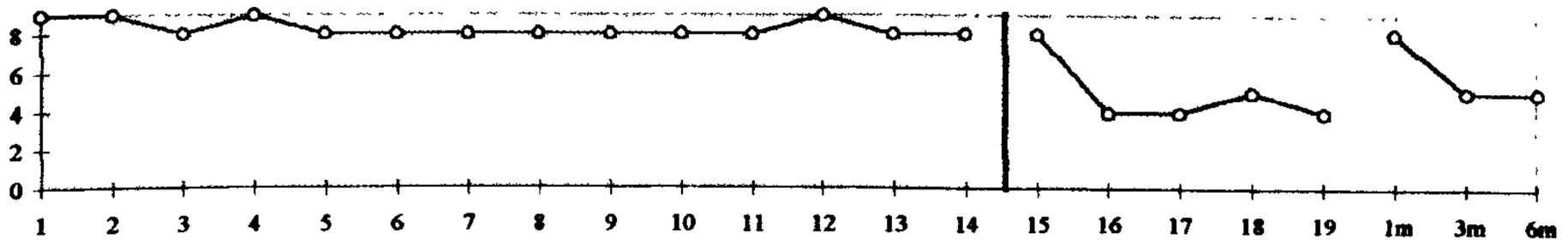
Bill



Pete 2



Tom 2



Ken 3

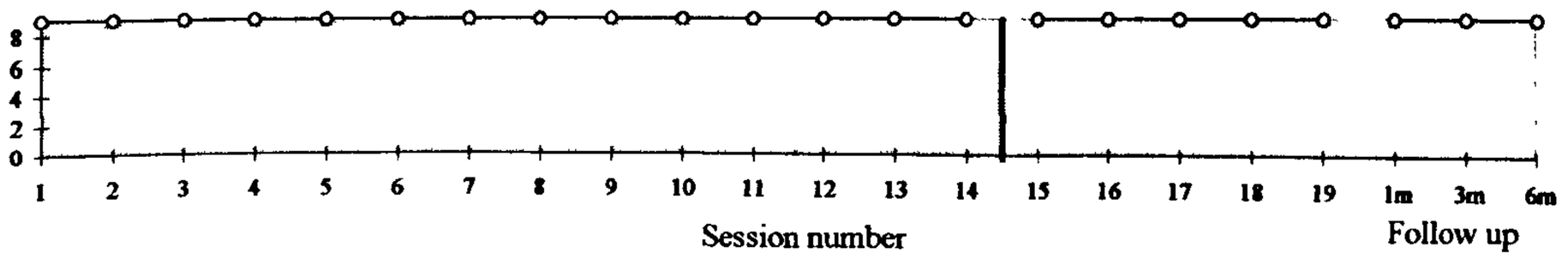


Figure 4.2 Data from Figures 4.1a, 4.1b and 4.1c arranged in order of intervention point.

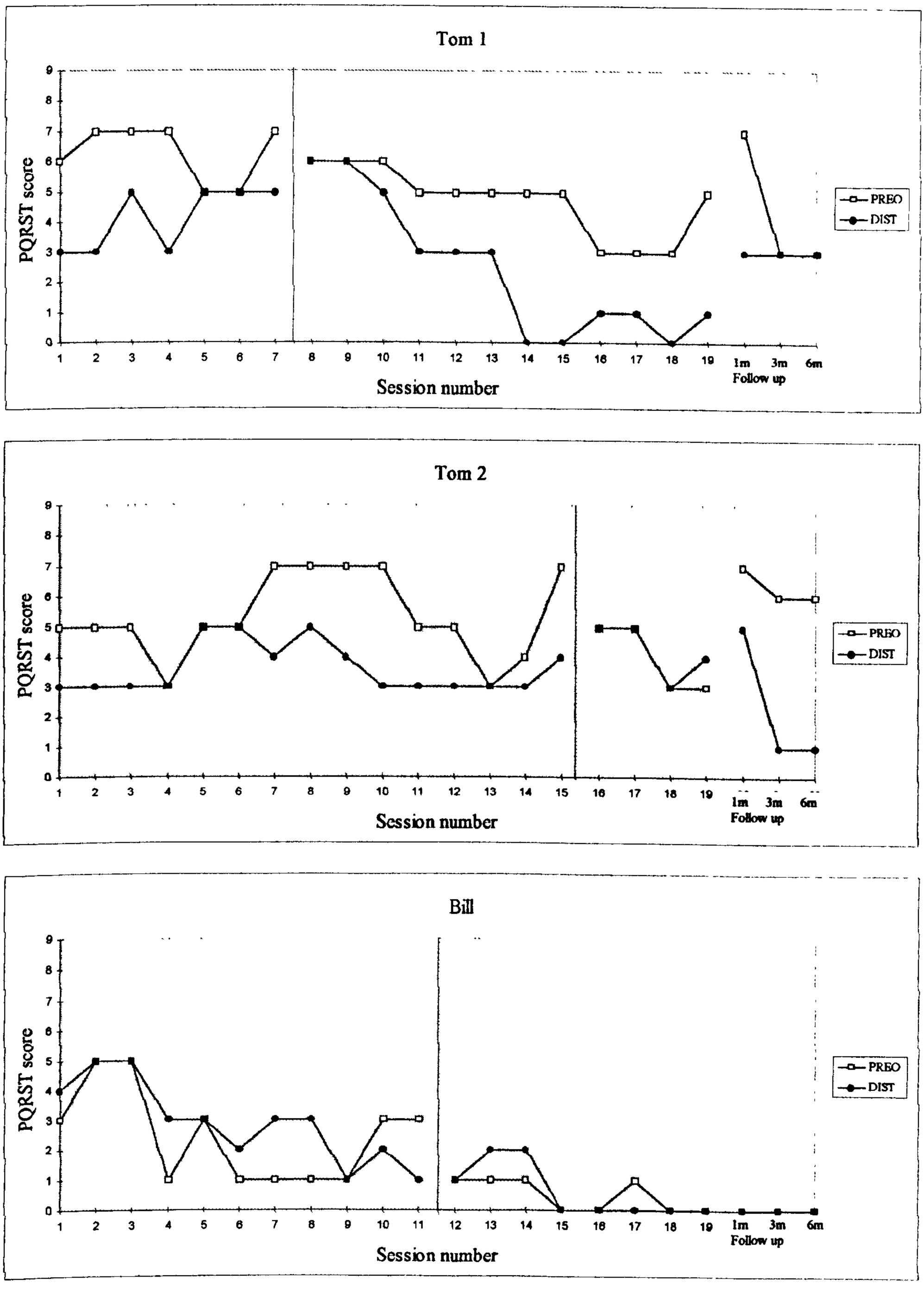


Figure 4.3.a

PQRST scores for preoccupation (PREO) and distress (DIST) in the time since the preceding session. Because the measures are retrospective, the vertical lines marking the introduction of the treatment intervention appear 1 session later than in Figure 4.2.

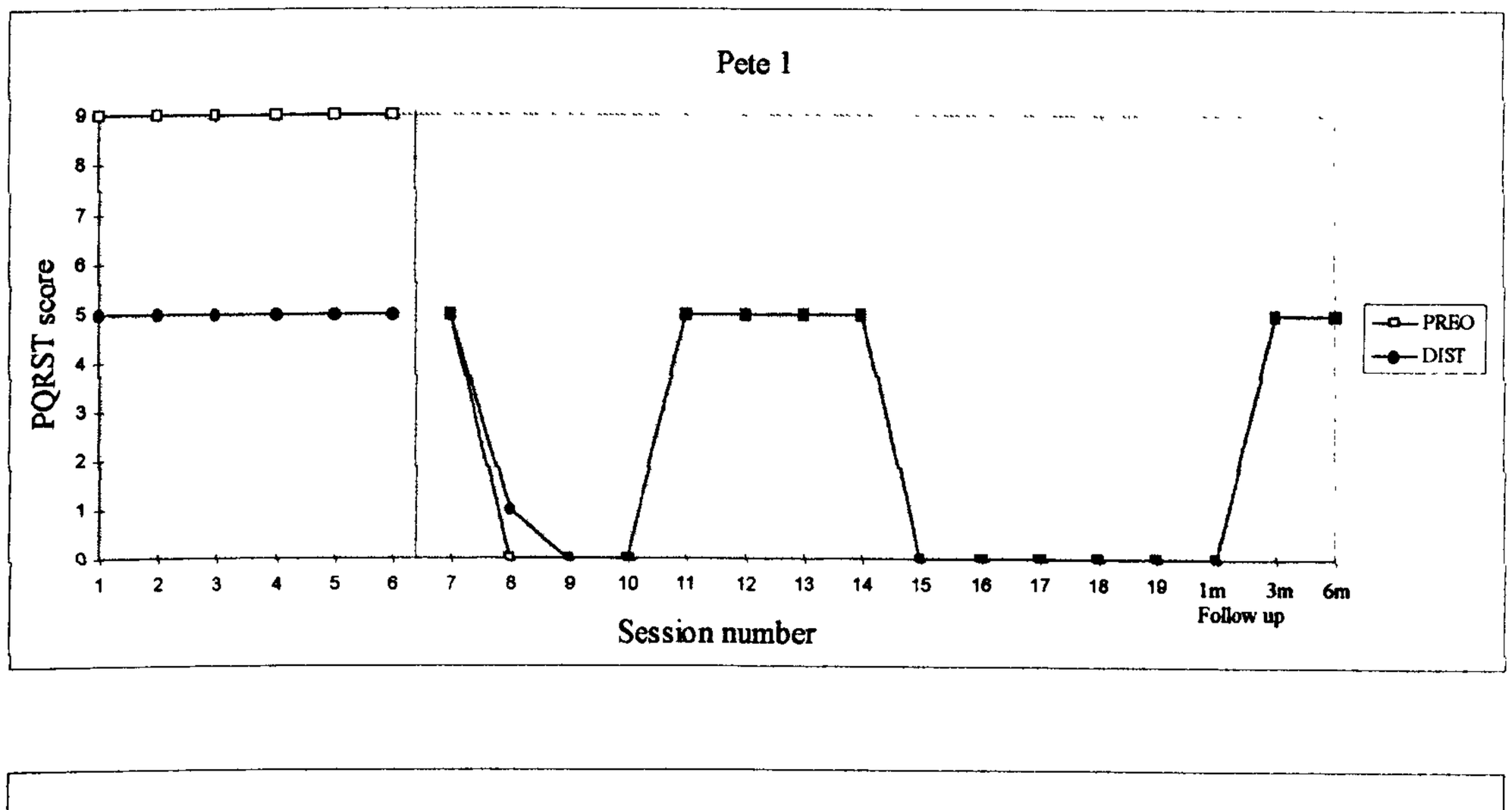


Figure 4.3.b.

PQRST scores for preoccupation (PREO) and distress (DIST) in the time since the preceding session. Because the measures are retrospective, the vertical lines marking the introduction of the treatment intervention appear 1 session later than in Figure 4.2.

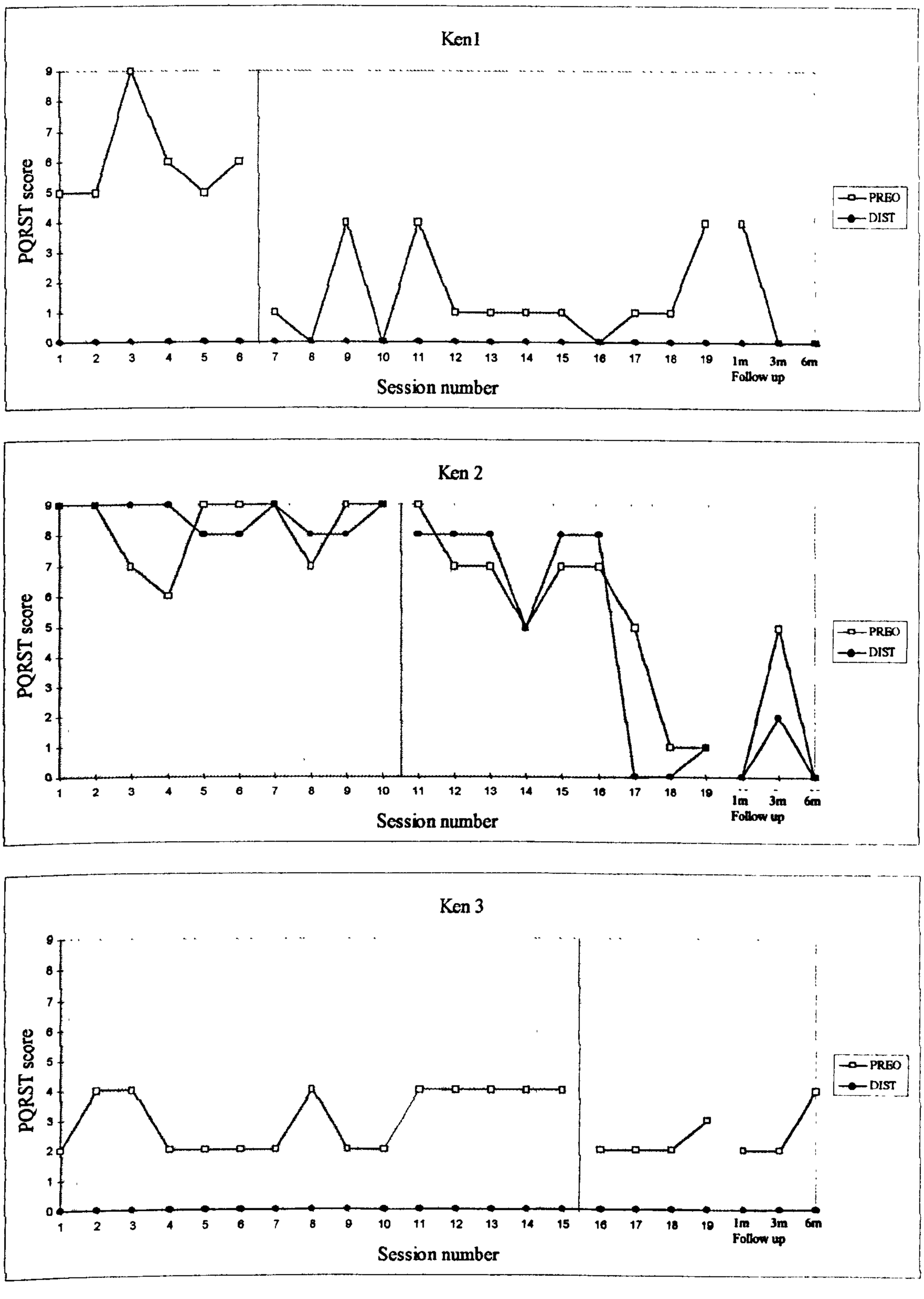
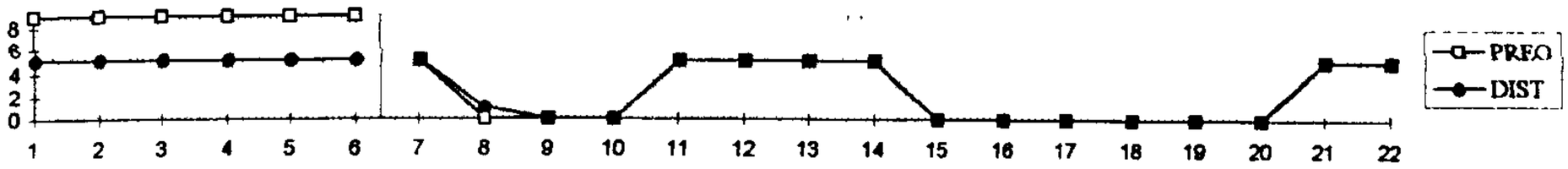


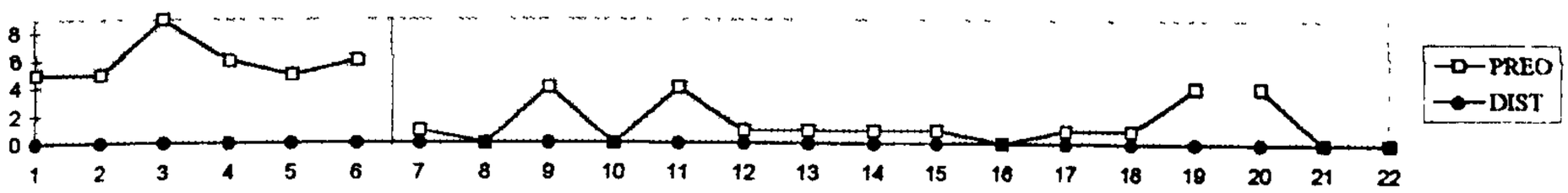
Figure 4.3.c.

PQRST scores for preoccupation (PREO) and distress (DIST) in the time since the preceding session. Because the measures are retrospective, the vertical lines marking the introduction of the treatment intervention appear 1 session later than in Figure 4.2.

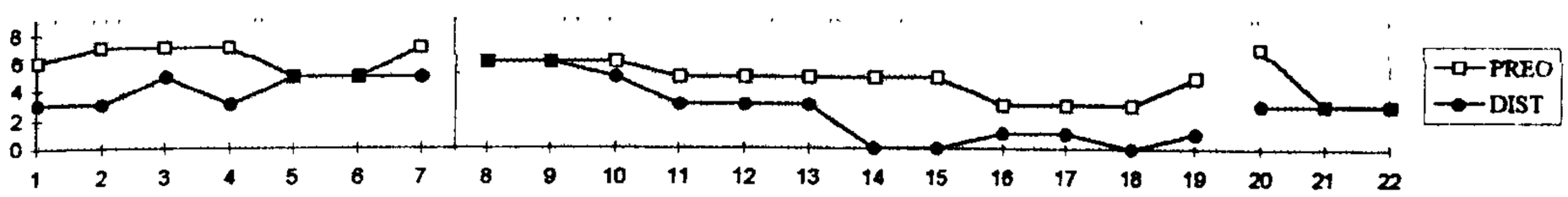
Pete 1



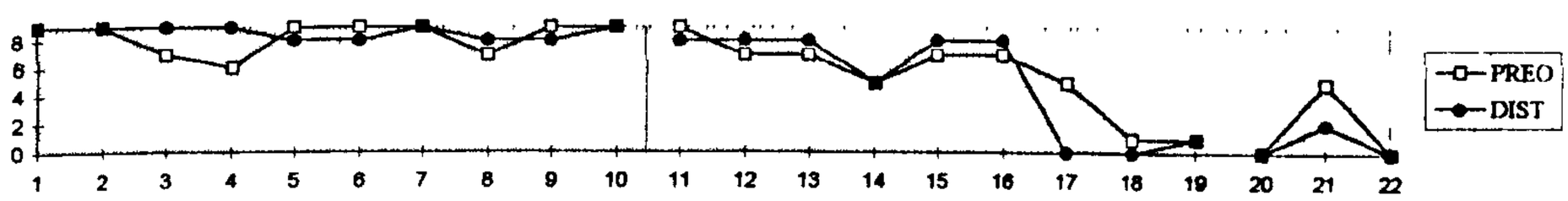
Ken1



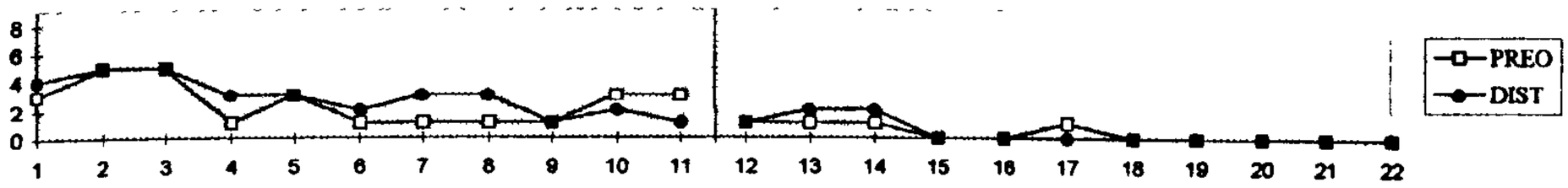
Tom 1



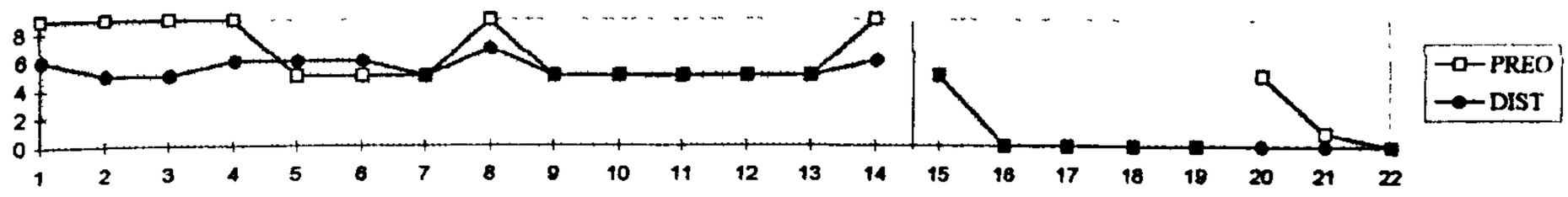
Ken 2



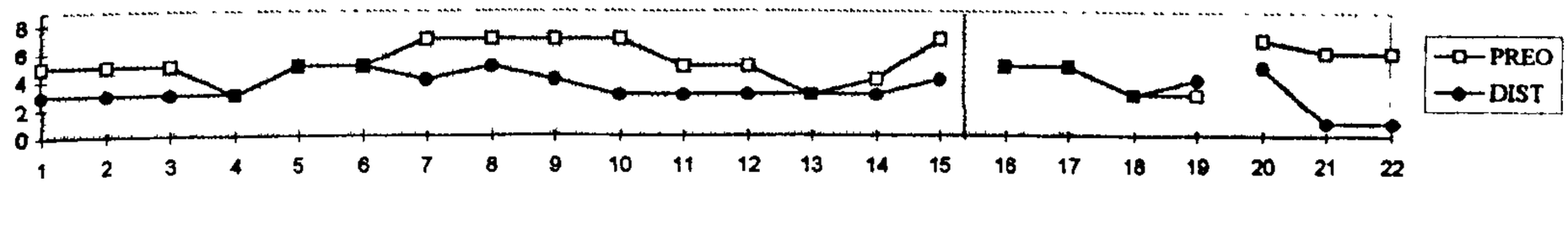
Bill



Pete 2



Tom 2



Ken 3

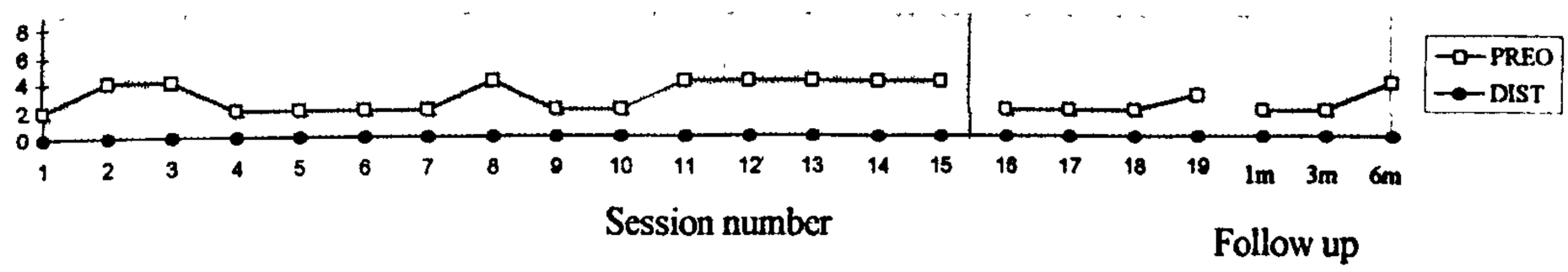


Figure 4.4.

Data from Figures 4.3.a, 4.3.b. and 4.3.c. arranged in order of intervention point.

"Tom"'s first belief was that trains that he observed were carrying gold bullion for some sinister purpose connected with subversive groups attempting to undermine Great Britain. No attempt was made to challenge this belief during the baseline phase, during which strength of conviction was measured and rapport established. Treatment began in session 7, using cognitive behavioural techniques described in the contemporary literature. The belief was not directly challenged, but he was asked to reflect on why he held the belief, what evidence there was to support it, and its likelihood in light of other knowledge he had about the world. There was no shift from maximum conviction for the first two treatment sessions, followed by a fluctuating strength of conviction reducing to zero. At one month follow-up, conviction had risen to maximum once more but fell to a score of 1 at later follow-ups. He now tended to believe that goods trains were transporting ordinary commercial goods for commercial reasons rather than bullion for sinister purposes. Preoccupation scores were at middle to high values during baseline, with evidence of a slight fall during treatment. The score rose at first follow-up before settling to medium level. Distress scores declined during the treatment phase before returning to medium value at follow-up.

For the second delusion concerning trains carrying "Soviet" troops, treatment commenced in session 15 using similar techniques. He was asked to look at the trains closely for evidence for or against his belief and was engaged in discussion of world events such as the collapse of the former Soviet Union which might be relevant. Strength of conviction remained at maximum (9) after the first treatment session and then fluctuated around middling levels of conviction. He was able to entertain doubts but not to give up the belief entirely. Conviction returned to near baseline level at 1 month follow-up and then returned to intermediate values. It is noteworthy that although the two beliefs were rather similar, involving trains, change in strength of conviction of the second belief occurred only when it was directly addressed. There was no evidence of a collateral effect on the second belief following treatment of the first. There was no apparent trend in preoccupation or distress scores during the 19 sessions and follow-up,

except that distress had decreased at 3m and 6m follow-up.

"Bill" believed the number "666" was tattooed on the back of his head. He had held this belief for many years. Strength of conviction was maximal throughout the baseline period, but during one session he volunteered that if he could examine the back of his head and see that the number was not there he would cease to believe it. Treatment began in session 11 when this test was performed using two mirrors. When the number was not seen, strength of conviction immediately fell to zero where it remained for all succeeding sessions and follow-up sessions. He was able to discuss other possible reasons to which he could attribute the course of his life. Scores for preoccupation and distress appeared to be on a downward trend during baseline, which continued until both scores reached zero at session 18 and remained there through follow-up.

"Pete" had beliefs about staff contaminating his depot injection and poisoning his food. In discussion he volunteered that if he were able to observe nursing staff unwrapping the injection kit he might be less sure they were contaminating it. Baseline sessions were planned leading up to the next depot injection when the test was performed in session 6, the previously randomly determined intervention point. Strength of conviction immediately fell to zero. In later sessions it fluctuated about middling values before returning to zero. First follow-up showed a return to a medium value and the later follow-ups showed a return to zero. Preoccupation scores were at maximum during baseline and thereafter alternated between medium values and zero. Distress scores were at a medium level during baseline and thereafter closely followed the preoccupation scores.

In connection with the second delusional belief he had similarly agreed that he might be less convinced that his food were being poisoned if he could observe it from the point of delivery from the catering van to the point of consumption. This intervention was performed at session 14, the randomly determined intervention point. Strength of

conviction immediately fell to medium and low values, returning to moderately high values at follow-up. Again it is noteworthy that despite the apparent similarity of the two beliefs there was no evidence of a collateral effect on the second belief following treatment of the first. That is, there was no change in strength of conviction in the second belief following successful intervention with the first. Preoccupation scores were at maximum to middle values during baseline, falling to zero during treatment. There was a return to middle value at 1m follow-up followed by return to zero. Distress scores were at medium levels during baseline but fell to zero at session 18 where they remained.

"Ken"'s first belief was that he was a member of the SAS. Treatment intervention was randomly allocated to session 6. In discussion it emerged that he had met an actor at a fan-club function, who had acted the role of an SAS soldier in films, and had become a member of the fan-club. In discussion he began to tease out reality from delusional fantasy, was encouraged to explore the question whether membership of the fan-club really implied membership of the SAS, and was able to show some uncertainty about the belief. Strength of conviction fluctuated below maximum, tailing off to zero. There was a rise to maximum at first follow-up and then a return to zero. Scores for preoccupation showed a sharp drop following treatment intervention. Despite some fluctuation, there was no overlap with baseline scores. Distress scores for this delusion were at zero throughout.

The second belief was that he was being pursued by the IRA. He thought that this was because his family had service connections. Session 10 was the randomly determined intervention point. There was no shift in strength of conviction following the first treatment session. Thereafter he began to understand that his connection with the military was tenuous and unlikely to be of interest to the IRA. Strength of conviction remained at high levels for several sessions before falling. At the follow-ups it fluctuated from low to high to low. Preoccupation and distress were both relatively high for this delusion, which is understandable because he thought his life was in danger. The pattern

was similar to that for strength of conviction. Scores for preoccupation and distress remained at relatively high levels for several sessions following treatment intervention before falling to low levels by session 18. They rose at 3m follow-up and returned to zero at 6m follow-up.

The third belief was that he had access to a special telephone number in Whitehall through which he could summon assistance. Strength of conviction remained at the maximum level of 9 throughout, and treatment intervention commencing in session 15 had no effect. It had been noted that he had attempted to make telephone calls from the ward to the local Military Police when in an agitated condition but he would not discuss whether this was what he meant by the "special number". The distress score for this delusion was at zero throughout, probably because the delusional idea was reassuring and beneficial to self-esteem. There was no discernible trend in preoccupation scores, which varied about low to medium values throughout.

Statistical analysis of the conviction score data

1. The randomization test

Over all 8 beliefs, the mean score for strength of conviction in the baseline phase was 8.87. The mean score in the treatment phase was 3.03. The experimental effect (not to be confused with the Effect Size which was computed separately, see below), defined as the difference between the mean score in the baseline phase and the treatment phase, summed over the 8 beliefs, was 43.22. In order to compute the statistical significance for the overall experiment, a randomization test was performed. With 10^8 possible ways of allocating intervention points across the 8 beliefs with 10 possible intervention points per belief, the computation time to perform a complete, exhaustive analysis would have been too great. A non-exhaustive randomization test was performed with 10^5 permutations, using a computer program written by this author using the guidance of Edgington

(1995). The program marbus5. bas is presented at appendix (3). The test statistic was the difference between the mean score in the baseline phase and the treatment phase, summed over the 8 beliefs. In 303 of the 10^5 permutations an "effect" as large as or greater than the obtained effect of 43.22 was obtained. The probability was therefore $303/10^5$, or approximately 0.003. Insofar as this probability was less than the predetermined significance level alpha = 0.01, the null hypothesis was rejected.

2. The Cohen Effect Size

Over all 8 beliefs, the mean score for strength of conviction in the baseline phase was 8.87, in the treatment phase 3.03. The difference between the means was (8.87 - 3.03) = 5.84. The standard deviation of the pooled data was 3.86. From Equation 1, the Effect Size was computed as (5.84 / 3.86) = 1.51.

3. The Common Language Effect Size

The CL ES was computed, following McGraw and Wong (1992) and Bjorgvinsson and Kerr (1995) as 0.95. Thus 95 times in 100, a score randomly sampled from the baseline data would exceed a score randomly sampled from the treatment data (see Chapter 1).

Statistical analysis of the preoccupation data

1. The randomization test

Over all 8 beliefs, the mean score for preoccupation in the baseline phase was 5.52. The mean score in the treatment phase was 2.71. The experimental effect (not to be confused with the Effect Size which was computed separately, see below), defined as the difference between the mean score in the baseline phase and the treatment phase,

summed over the 8 beliefs, was 25.82. The non-exhaustive randomization test identical to that employed in analysis of the conviction score data was computed. The test statistic was the difference between the mean score in the baseline phase and the treatment phase, summed over the 8 beliefs. In 156 of the 10^5 permutations an "effect" as large as or greater than the obtained effect of 25.82 was obtained. The probability was therefore $156 / 10^5$, or 0.00156. Insofar as this probability was less than the predetermined significance level alpha = 0.01, the null hypothesis was rejected.

2. The Cohen Effect Size

Over all 8 beliefs, the mean score for preoccupation in the baseline phase was 5.52, in the treatment phase 2.71. The difference between the means was (5.52 - 2.71) = 2.81. The standard deviation of the pooled data was 2.86. From Equation 1, the Effect Size was computed as (2.81 / 2.86) = 0.98.

3. The Common Language Effect Size.

The CL ES was computed, following McGraw and Wong (1992) and Bjorgvinsson and Kerr (1995) as 0.79. Thus 79 times in 100, a score randomly sampled from the baseline data would exceed a score randomly sampled from the treatment data.

Statistical analysis of the distress data.

1. The randomization test

Over all 8 beliefs, the mean score for distress in the baseline phase was 3.67. The mean score in the treatment phase was 1.88. The experimental effect (not to be confused with the Effect Size which was computed separately, see below), defined as the difference between the mean score in the baseline phase and the treatment phase,

summed over the 8 beliefs, was 14.35. The non-exhaustive randomization test identical to that employed in analysis of the conviction and the preoccupation data was performed.. The test statistic was the difference between the mean score in the baseline phase and the treatment phase, summed over the 8 beliefs. In 58,874 of the 10^5 permutations an "effect" as large as or greater than the obtained effect of 14.35 was obtained. The probability was therefore 58,874 / 10^5 , or approximately 0.59. Insofar as this was greater than the predetermined significance level alpha, the null hypothesis was not rejected.

2. The Cohen Effect Size

Over all 8 beliefs, the mean score for distress in the baseline phase was 3.67, in the treatment phase 1.88. The difference between the means was (3.67 - 1.88) = 1.79. The standard deviationof the pooled data was 2.84. From Equation 1, the Effect Size was computed as (1.79 / 2.84) = 0.63.

3. The Common Language Effect Size

The CL ES was computed, following McGraw and Wong (1992) and Bjorgvinsson and Kerr (1995) as 0.68. Thus 68 times in 100, a score randomly sampled from the baseline data would exceed a score randomly sampled from the treatment data.

4.4 DISCUSSION

This experiment represents a successful attempt at the implementation of the randomized multiple baseline experimental design which from a literature search appears not to have been previously applied. A randomization test showed a statistically significant effect of treatment intervention for the main dependent variable, strength of conviction and for the secondary dependent variable, preoccupation, but not for the third

independent variable, distress. In terms of both Cohen's ES and of the Common Language ES, the effect size was greatest for the conviction data, smaller for the preoccupation data and least for the distress data.

Although the ES was greater (1.51) for the conviction data than for the preoccupation data (0.98), the randomization test gave a higher degree of statistical significance for the preoccupation data (0.00156) than for the conviction data (0.00303). This suggests that the effect of treatment intervention, although not as great, was slightly more clearly defined in the case of the preocupation data. This seems to be supported by visual examination and comparison of the data in Figures 4.2 and 4.4 although such analysis is problematical (see Chapter 2). This finding highlights the need to employ an appropriate statistical analysis in conjunction with purely graphical data analysis in small-n research of this type.

This small scale experiment complies with 2 of the main recommendations made by Bouchard et al. (1996) for work on cognitive restructuring in schizophrenia. They suggested that the use of sophisticated belief scales such as the Personal Scaling Technique (Brett-Jones et al., 1987 ; Shapiro, 1961) is an important asset in such studies. Mulhall's (1978) PQRST and its development by Cliffe, Possamai and Mulhall (1995), both of which were used here to measure the independent variables, are scales of this type. Bouchard et al. were also concerned with the adequacy of experimental design in such studies. The randomized multiple baseline design used here is a powerful extension of the conventional multiple baseline design giving even greater control for historical effects and allowing statistical evaluation by randomization tests

The statistical power of the randomized baseline design will be examined in the following Chapter.

CHAPTER 5. THE POWER OF RANDOMIZATION TESTS APPLIED TO RANDOMIZED MULTIPLE BASELINE DESIGNS.

"Why, she doth hang on him, *as if increase of appetite had grown by what it fed on.*"
(Shakespeare on exponential growth - Hamlet. Italics added.)

5.1. INTRODUCTION

The power of conventional parametric statistical tests has received considerable attention (see Chapter 1). Similarly, the power of rank tests relative to their parametric counterparts has been investigated. There has however been until recently little examination of the power of randomization tests for N = 1 or N = few experiments that lack large N counterparts (Edgington, 1995). Thus, Onghena (1994) points out that in none of the handbooks on applied power analysis are tables or formulae provided to calculate the power of randomization tests (see e.g. Cohen, 1988; Kraemer and Thiemann, 1987; Lipsey, 1990), and that on the other hand, none of the handbooks on randomization tests give advice on how to choose the number of observations that are assigned to each treatment on the basis of power considerations (e.g. Edgington, 1987; Manly, 1991; Noreen, 1989).

Onghena (1994) suggests a possible reason for the above. With a randomization test, no assumptions are made about a population distribution, and the test is applied to the data at hand. On the other hand, an a priori power analysis without assumptions about the responses one can expect, is intrinsically impossible. Therefore, in order to compute an a priori power analysis, additional assumptions have to be invoked, and some randomization testers may be reluctant to do this. Onghena argues however that power analysis is the only statistical tool available for deciding on the number of observations when designing experiments, so that the invocation of additional assumptions can be worthwhile, as long as it is recalled that the validity of the power calculations depends

on the validity of the additional assumptions.

Onghena (1994) discussed N = 1 randomization tests and provided detailed information on the power of two types of design, AB designs and alternating treatments designs. His Monte Carlo simulations demonstrated that power was relatively low for these designs. He showed that approximately N = 500 observations are required for an AB randomization test, and that approximately N = 50 observations are required for an alternating treatments design, to reach the conventional power level of 0.80 with alpha = 0.05, even for large effects (d = 0.80). These figures are in marked contrast to the number of observations usually reported in single-case designs. Thus Huitema (1985) found a mean and median number of observations of approximately 7.9 and 5.5 respectively, contained in initial baselines of 881 single-case designs published in the Journal of Applied Behavior Analysis. The modal number of observations in the baseline was 3 to 4, the number of observations contained in the other phases was even smaller, and the number of phases was rarely larger than 5. Similarly, Center, Skiba and Casey (1985) found a mean number of observations to be approximately N = 43 in 105 single-case designs published in Behavior Therapy, Journal of Abnormal Child Psychology, and Psychology in the Schools. Taken together, these two reviews emphasise the prevalence of low statistical power in single-case research. The undesirable effects of this were discussed in Chapter 1.

In the main clinical experiment reported in Chapter 4, use was made of an experimental design suggested by Marascuilo and Busk (1988), modified to encompass multiple baselines across behaviours within the same individual. Although there are no reported data in the literature on the power of such designs, it was considered likely that reasonable power would be available, for the following reason. Where i represents the number of possible intervention points for a given subject (or behaviour), and k represents the number of subjects (or behaviours), there are i^k possible data permutations. The number of possible data permutations therefore grows rapidly and

exponentially as more subjects (or behaviours) are added. Given i = 10, when k = 1 there are 10 possible data permutations. When k = 2 there are 100, with k = 3 there are 1,000 and so on. In the experiment in Chapter 4, with i = 10 and k = 8, the number of possible data permutations was 10^8.

Because no data have been reported on the power of these designs based on replications of the randomized baseline design for treatment intervention, it was necessary to perform Monte Carlo simulations (Manly, 1991; Noreen, 1989; Onghena, 1994) in order to generate at least approximate estimates of power.

It has already been emphasised that randomization tests in the present context are computer intensive. Monte Carlo simulations are themselves computer intensive. The task was therefore one of applying computer intensive simulations to computer intensive tests. Onghena (1994) referred to this situation as "computer-intensive raised to the square because the power estimation is derived through an iteration of an iterative test procedure".

For the purpose of the present dissertation, to keep the task within reasonable bounds it was necessary to be restrictive. There are many possible forms that a randomized baseline experiment could take. One parameter is the restriction on the minimum number of baseline sessions and treatment sessions that are stipulated. In the simulations reported here this number was five in both cases. Thus the assumption in all the simulations is that treatment intervention cannot be randomly allocated to any session number less than six. Similarly, where n is the total number of sessions available, treatment intervention cannot be randomly allocated to any session number less than (n - 4). For example in the experiment of Chapter 4 there were 19 sessions, so treatment intervention could be randomly allocated among sessions 6 to 15. Restrictions such as these affect statistical power (Onghena, 1994), so it is not possible to generalise the obtained results to experimental situations with different parameters. The software

reported here can be readily modified for other parameters to allow the necessary further simulations to be performed.

Section 3.7.2 above considered the Marascuilo and Busk (1988) randomization test for a randomized multiple baseline experimental design. Marascuilo and Busk also proposed in their paper the use of a normal approximation to their randomization test, the use of which would greatly reduce the amount of computation required. Onghena (1994) argued that "although their normal approximation has some value for quick computations, it is imposible to assess the accuracy of the approximation in a particular research situation. Therefore, the application of exact randomization tests using fast algorithms is to be preferred to using approximations."

The position adopted here, contrary to Onghena's above argument, is that in some circumstances it may not be "impossible" to assess the accuracy of the approximation. If it can be shown that the normal approximation of Marascuilo and Busk (1988) is a sufficiently good approximation to the exact randomization test, the greatly reduced amount of computation required for an a priori power analysis might render the project feasible. Data will be presented below which, it is argued, show that the approximation is close enough for the purpose.

5.2. THE COMPUTER SIMULATIONS FOR POWER ANALYSIS

The basic procedure here is the computer simulation of a multiple baseline experiment with given parameters. This is repeated, or iterated, many times (e.g. 1,000). For each iteration, a determination is made on whether the randomization test has yielded a statistically significant result at a given significance level alpha (alphas 0.05, 0.01, 0.001 are reported on here). Power for the experiment with the given parameters is obtained simply by dividing the number of iterations in which statistical significance was found at a given alpha level, by the total number of iterations. For example, a simulated experiment with 10 baselines, an ES of

0.5 and 20 possible intervention points, might give a significant result at alpha = 0.05 in 739 out of the total of 1,000 iterations. The power is then computed as 739/1000, or 0.739.

The simulated experiments depend on the generation of normally distributed random data with a defined mean and standard deviation. In all of the computer programs presented in the Appendix, when such data needed to be called up it was generated by the algorithm contained in the program sim2.bas (Appendix 1). This algorithm involves the consultation of normal curve tables. Appendix 2 shows the normal curve DATA that was READ by each of the programs. The programs were written in QBasic (Perry, 1993) and were executed on an NEC Versa 2000C laptop computer running at 75 MHz. It was necessary to demonstrate that this algorithm generated data to specification, i.e. normally distributed data with a defined mean and standard deviation. Tests showed that the data generated did conform to the specification. For example, Figure 5.1 shows the observed and the expected frequencies of 10,000 numbers generated by the program sim2.bas with a specified mean of 100 and standard deviation of 2.5. The observed and the expected frequencies were the same to a good approximation and the chi-square was not significant.

The principle of the simulations can be demonstrated by reference to a single hypothetical baseline. Firstly, the intervention point is selected at random from the defined range of possible intervention points. For example, with 19 sessions and a stipulated minimum of 5 baseline and 5 treatment sessions, possible intervention points are sessions 6 to 15. Suppose the randomly selected intervention point to be session 10. In that case, sessions 1 to 9 will be baseline sessions and sessions 10 to 19 will be treatment sessions.

Normally distributed, randomly generated data with a defined mean and standard deviation are assigned to each of the 9 baseline sessions. Normally distributed, randomly generated data with a different mean, reflecting the treatment Effect Size, but with the same standard deviation, are assigned to each of the 10 treatment sessions. In the reported simulations here, the mean for the baseline data was defined as 100. The standard deviation for the baseline and the treatment data was defined as 5. The mean for the treatment data was selected to

reflect the Effect Size. For example, if the chosen Effect Size were 2. the mean for the treatment data would be 90, that is, 100 minus 2 standard deviations (from Equation 1). The process is repeated for all of the baselines in the simulated experiment with the given parameters. Once the data have been generated for all of the baselines required for a simulated experiment with the given parameters, statistical significance is computed using either a non-exhaustive randomization test (see Chapter 3) similar to that used in the clinical experiment in Chapter 4, with 100 iterations, (program marsim3.bas, Appendix 4) or the normal approximation of Marascuilo and Busk (program normdis2.bas, Appendix 5).

5.2.1 RANDOMIZATION TEST VERSUS NORMAL APPROXIMATION

It was noted above that computation could be greatly reduced in these simulations if it could be shown that the normal approximation provided a sufficiently close approximation to the results obtained from the non-exhaustive randomization test. Simulations were computed to enable a comparison to be made.

Simulations were computed with number of baselines from 2 to 8 in even numbers, and with number of possible intervention points from 6 to 20 in even numbers. Four sets of simulations were computed, with parameters as follows: (1) alpha = 0.05, ES = 0.5; (2) alpha = 0.05, ES = 0.8; (3) alpha = 0.01, ES = 0.5; and (4) alpha = 0.01, ES = 0.8.

Simulations were computed with only 100 iterations for the non-exhaustive randomization test because of the great amount of computing time that wold be required for 1,000 iterations. Because computation is much quicker for the normal approximation it was possible to compute at 1,000 iterations. Tables 5.1 to 5.4, and Figures 5.2 to 5.5 show the results of these simulations. In each of Figures 5.2 to 5.5, there are 4 pairs of curves. In each pair, the front curve shows data from the non-exhaustive randomization test with 100 iterations, while the rear curve shows data from the normal approximation with 1,000 iterations.

Two points arise from Figures 5.2 to 5.5. Firstly, the front curve of each pair shows greater variability than the rear curve. The rear curves tend to be smoother. It is probable that this is due to the smaller number of iterations (100) computed for the randomization test than for the normal approximation (1,000). Secondly, allowing for the greater variability, the front and the rear curves are congruent to a rough approximation. This was considered to be reasonable evidence to justify the use of the normal approximation in the further, more detailed computer power analysis simulations. There is a trade-off between the ability to compute more iterations (1,000) with the normal approximation, and the "approximateness" of the approximation. The more detailed simulations below were therefore computed using the normal approximation method, by the program normdis2.bas (Appendix 5). It should be emphasised that the validity of these simulations depends on the validity of the assumption of relatively close approximation of the two procedures.

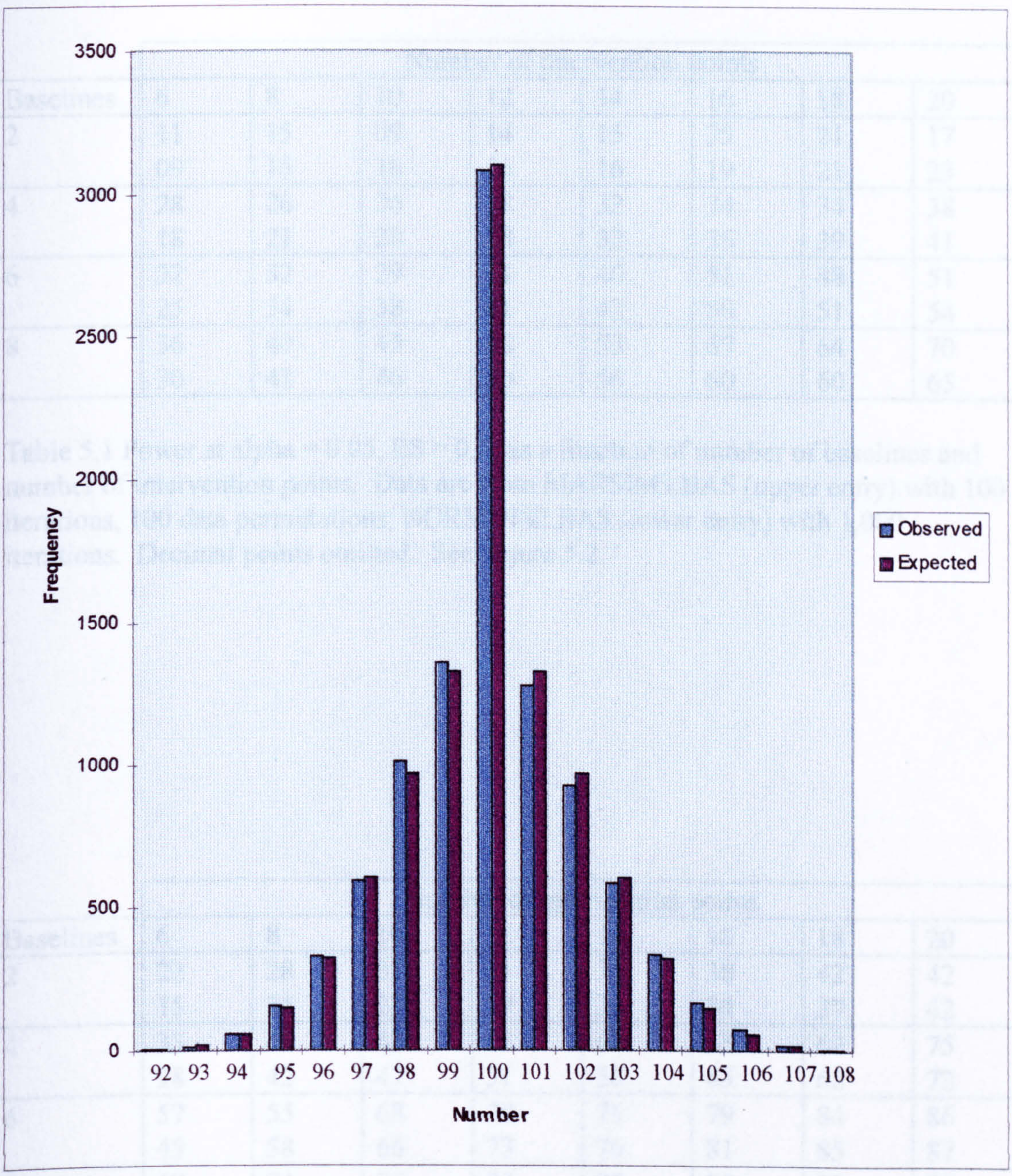
5.2.2. RESULTS OF THE POWER ANALYSIS SIMULATIONS

Power analysis simulations were computed using the program normdis2.bas as described above. Simulated experiments were computed with the number of baselines ranging from 2 to 10, and number of possible intervention points ranging from 6 to 20 in even numbers. Effect sizes ranged from 0.2 to 2.0 in increments of 0.3. Three levels of the significance level alpha were examined: alpha = 0.05; 0.01; and 0.001.

Tables 5.5 to 5.11, and Figures 5.6 to 5.14 show the results of the simulations. The tables and figures allow the approximate determination of the power of a projected multiple baseline study which is to be analysed by a randomization test of the type used here and described by Marascuilo and Busk(1988). For example, consider a projected experiment similar to that reported in Chapter 4, with 8 baselines and 10 possible intervention points in each baseline. Suppose there were grounds to expect an effect size of 0.8, based for example on a meta-analysis of previously reported studies in the field. Figure 5.12 shows the expected power of the experiment to be approximately 0.8 for alpha = 0.05. This would be considered a

reasonable power in Cohen's terms (see Chapter 1). There is an approximately 80% chance of detecting an effect of this size with this projected experiment. Projected power can be read from the tables, with interpolation where necessary, for number of baselines ranging from 2 to 10, and number of possible intervention points from 6 to 20. Should the projected power of a planned experiment be regarded as too low, the tables can be consulted to determine the number of baselines and the number of intervention points required for adequate power for a given effect size.

The type of experimental design used here, a combination of randomized multiple baseline design and statistical analysis by a randomization test, appears to be an efficient design in terms of the number of subjects required. In the hypothetical experiment referred to in the paragraph above, assume that the 8 baselines represented 8 subjects each of whom was measured on a single behaviour. How many subjects would be needed in a group design to detect a similar effect (ES = 0.80) at alpha = 0.05 between two independent sample means, with a power of 0.80? Cohen (1992) shows the number to be 26 in each group. Thus the experimental design and statistical analysis used in the present study achieves approximately the same power with 8 baselines, as a group study employing 52 subjects. This is an example of the recommended (Section 1.6) use of repeated measures designs to increase power.



	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108
O	9	15	54	153	328	593	1014	1361	3084	1278	923	589	332	169	72	18	8
E	5	19	56	146	320	603	968	1327	3108	1327	968	603	320	146	56	19	5

Figure 5.1. Observed (O) vs. expected (E) frequencies of 10,000 numbers generated by SIM2.BAS. Mean = 100, s.d. = 2.5. $\chi^2 = 23.13$, d.f. = 16, NS

	Number of intervention points							
Baselines	6	8	10	12	14	16	18	20
2	11 09	15 15	09 16	14 16	15 16	25 19	21 21	17 23
4	28 18	26 21	36 28	38 28	32 32	34 36	34 39	38 41
6	32 25	32 34	29 38	41 41	46 42	51 50	48 51	51 54
8	36 30	40 41	45 46	48 50	53 56	67 60	64 60	70 65

Table 5.1 Power at alpha = 0.05, ES = 0.5, as a function of number of baselines and number of intervention points. Data are from MARSIM3.BAS (upper entry) with 100 iterations, 100 data permutations; NORMDIS2.BAS (lower entry) with 1,000 iterations. Decimal points omitted. See Figure 5.2.

	Number of intervention points							
Baselines	6	8	10	12	14	16	18	20
2	22 15	28 21	31 23	31 27	34 32	38 34	42 37	42 42
4	33 28	49 42	42 47	63 51	62 58	65 65	62 68	75 72
6	57 45	55 58	68 66	78 73	75 76	79 81	84 85	86 87
8	67 59	74 70	75 79	83 84	85 88	88 91	82 93	90 95

Table 5.2 Power at alpha = 0.05, ES = 0.8, as a function of number of baselines and number of intervention points. Data are from MARSIM3.BAS (upper entry) with 100 iterations, 100 data permutations; NORMDIS2.BAS (lower entry) with 1,000 iterations. Decimal points omitted. See Figure 5.3.

	Number of intervention points							
Baselines	6	8	10	12	14	16	18	20
2	01 00	01 02	02 02	02 02	04 04	04 03	04 05	05 05
4	06 02	09 05	16 06	14 07	14 09	14 11	13 12	16 15
6	09 04	17 10	07 12	16 13	19 14	21 20	20 22	26 26
8	08 08	15 14	16 17	18 21	24 22	28 28	32 31	36 34

Table 5.3 Power at alpha = 0.01, ES = 0.5, as a function of number of baselines and number of intervention points. Data are from MARSIM3.BAS (upper entry) with 100 iterations, 100 data permutations; NORMDIS2.BAS (lower entry) with 1,000 iterations. Decimal points omitted. See Figure 5.4.

	Number of intervention points							
Baselines	6	8	10	12	14	16	18	20
2	01 01	04 03	09 04	11 05	07 07	06 07	10 08	19 11
4	08 05	17 11	11 13	28 18	33 24	32 29	30 31	39 33
6	22 12	21 24	34 30	50 37	42 44	46 51	54 54	55 58
8	36 22	32 37	53 47	52 55	55 61	70 69	71 73	70 76

Table 5.4 Power at alpha = 0.01, ES = 0.8, as a function of number of baselines and number of intervention points. Data are from MARSIM3.BAS (upper entry) with 100 iterations, 100 data permutations; NORMDIS2. BAS (lower entry) with 1,000 iterations. Decimal points omitted. See Figure 5.5.

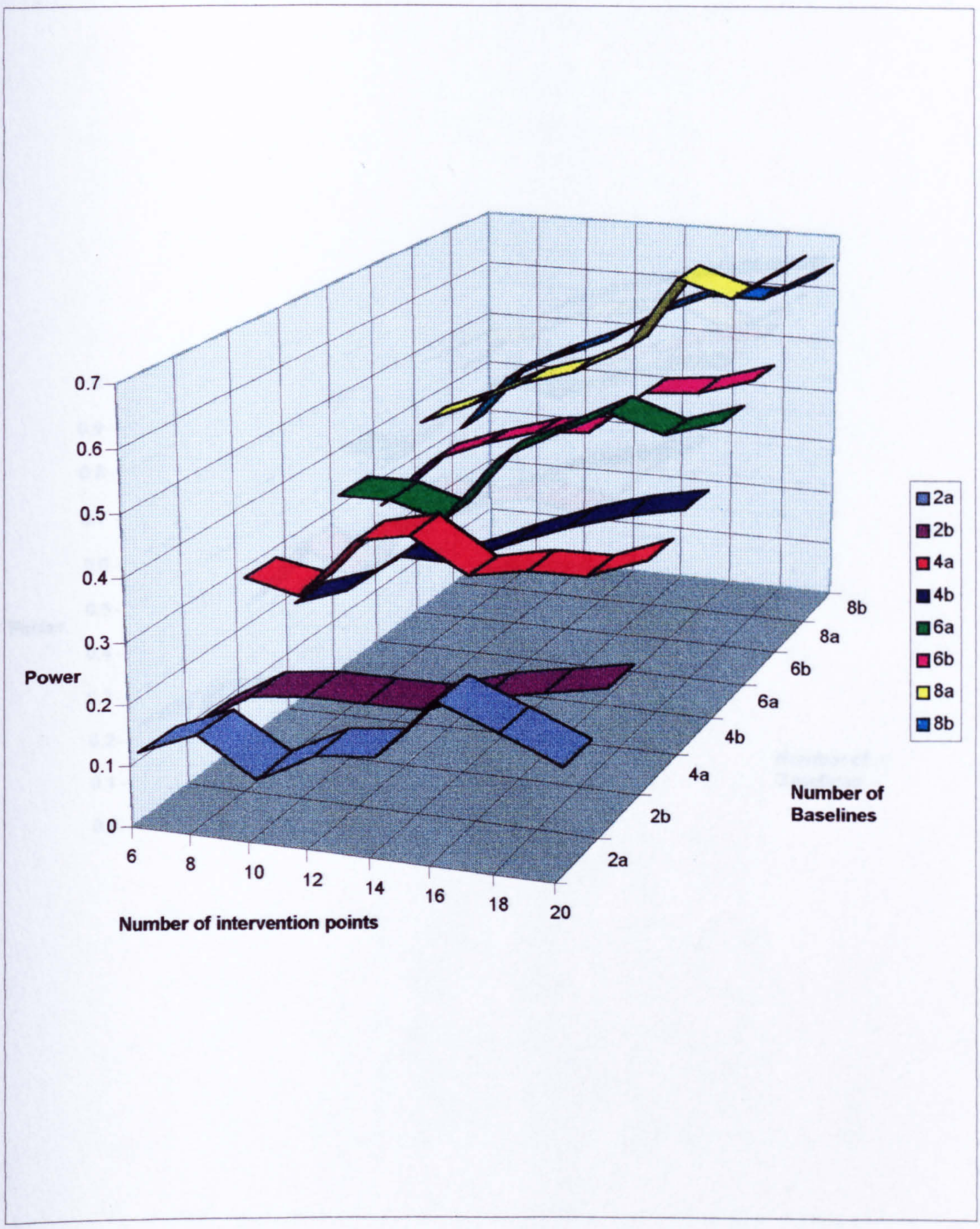


Figure 5.2 . Power at alpha = 0.05, ES = 0.5, as a function of number of baselines and number of intervention points. Data for baselines numbered 2a - 8a are from MARSIM3.BAS with 100 iterations. Data for baselines numbered 2b - 8b are from NORMDIS2.BAS with 1,000 iterations. See Table 5.1.

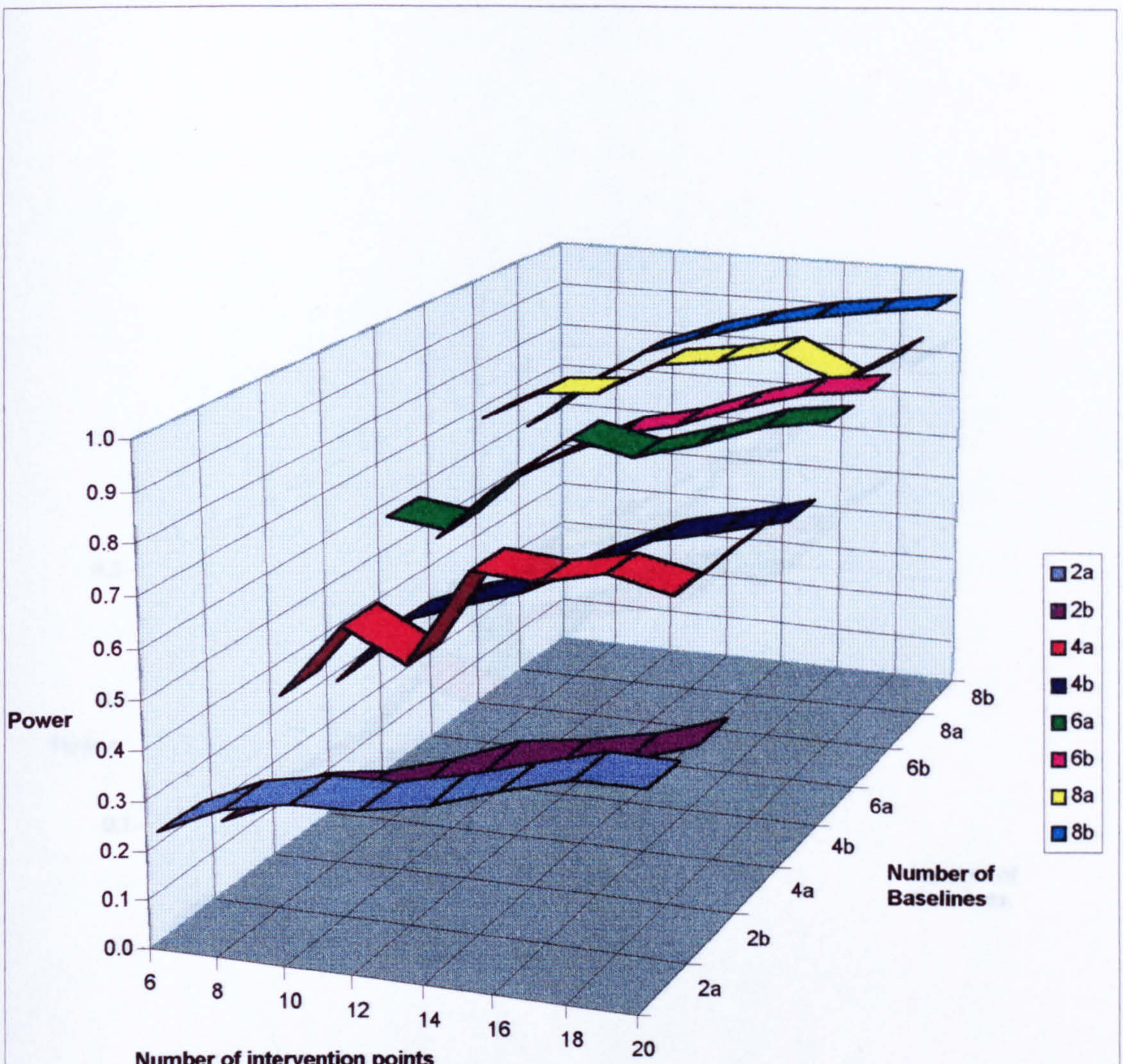


Figure 5.3. Power at alpha = 0.05, ES = 0.8, as a function of number of baselines and number of intervention points. Data for baselines numbered 2a - 8a are from MARSIM3.BAS with 100 iterations. Data for baselines numbered 2b - 8b are from NORMDIS2.BAS with 1,000 iterations. See Table 5.2.

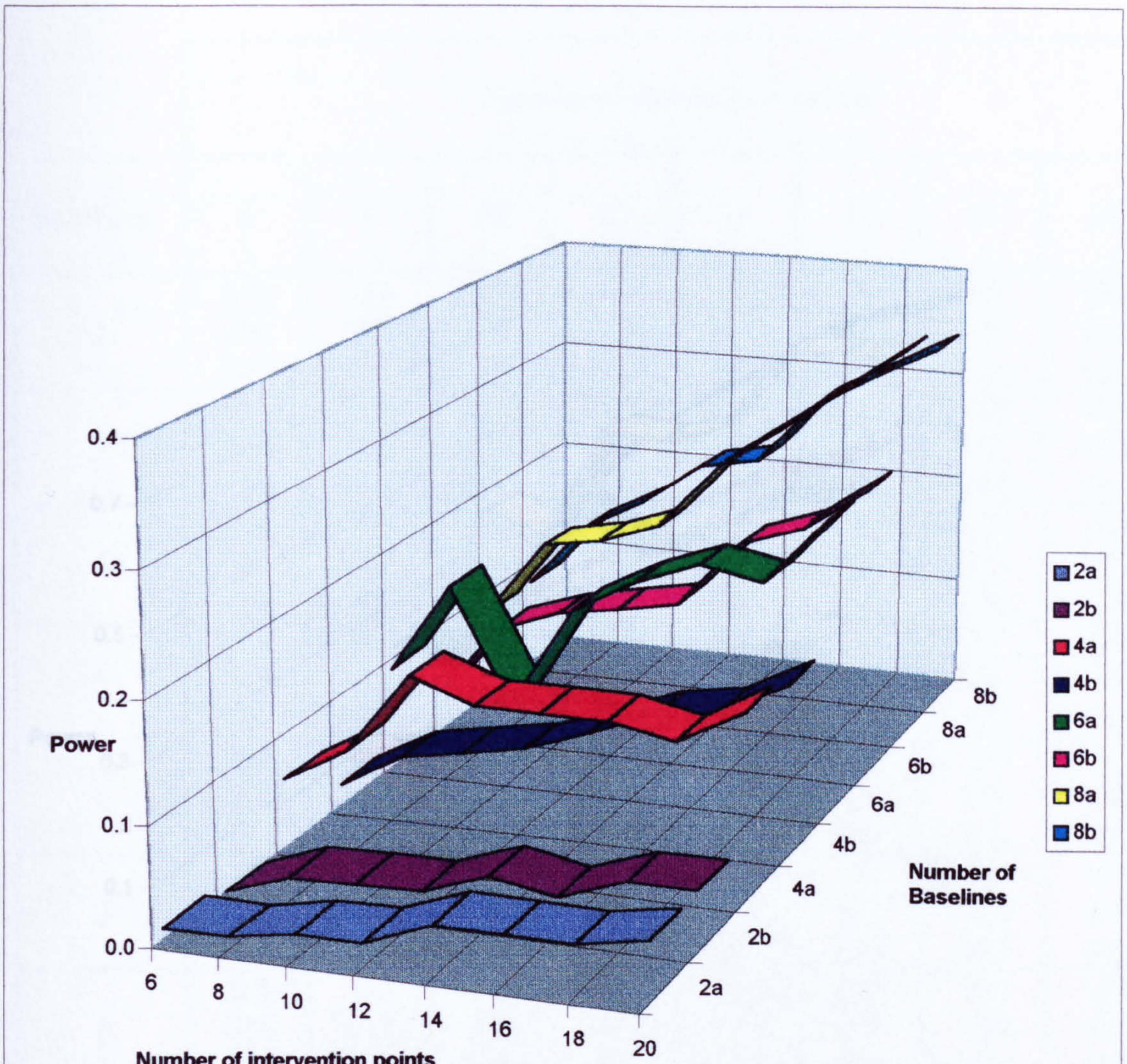


Figure 5.4. Power at alpha = 0.01, ES = 0.5, as a function of number of baselines and number of intervention points. Data for baselines numbered 2a - 8a are from MARSIM3.BAS with 100 iterations. Data for baselines numbered 2b - 8b are from NORMDIS2.BAS with 1,000 iterations. See Table 5.3.

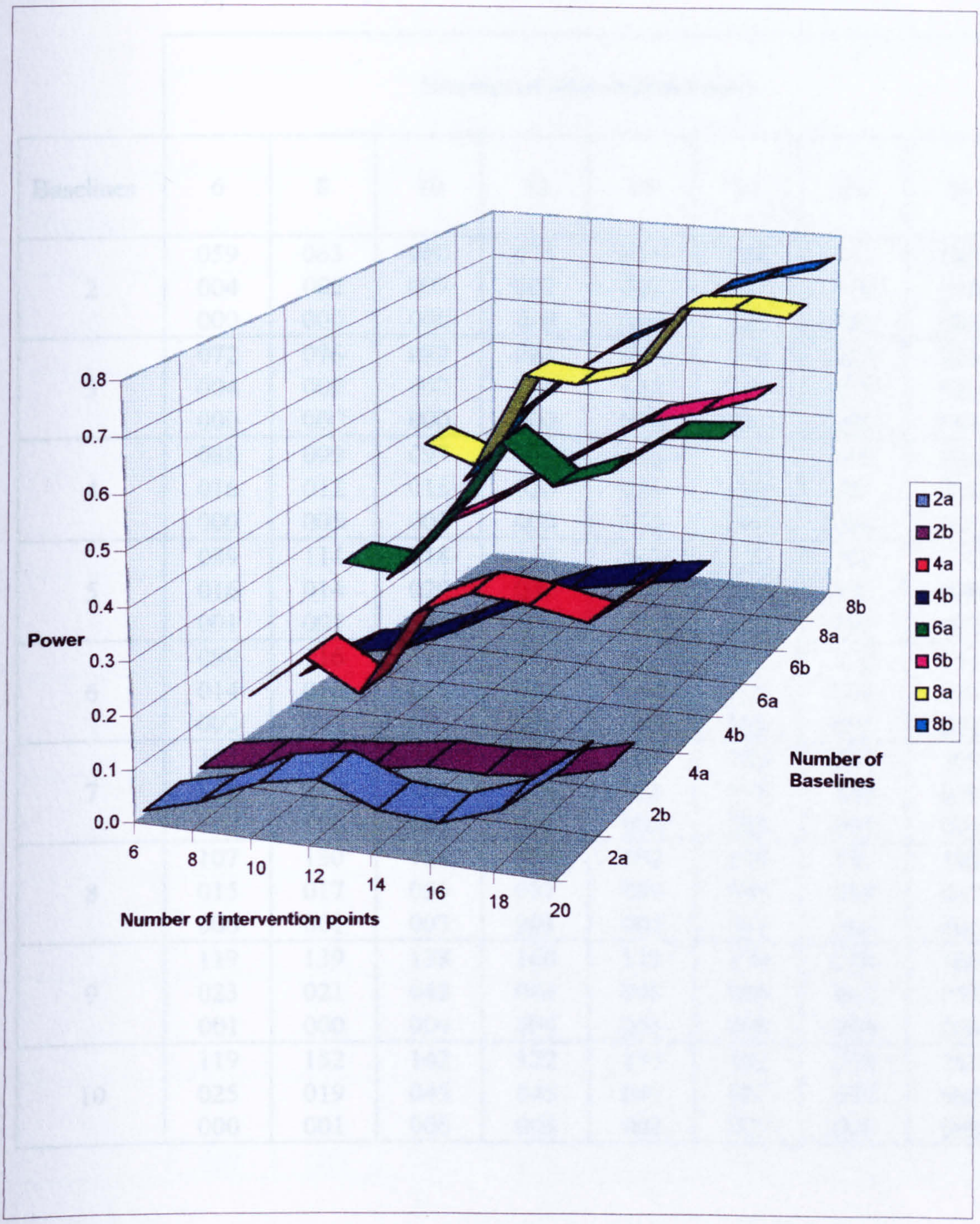


Figure 5.5. Power at alpha = 0.01, ES = 0.8, as a function of number of baselines and number of intervention points. Data for baselines numbered 2a - 8a are from MARSIM3.BAS with 100 iterations. Data for baselines numbered 2b - 8b are from NORMDIS2.BAS with 1,000 iterations. See Table 5.4.

	Number of intervention points							
Baselines	6	8	10	12	14	16	18	20
2	059 004 000	063 002 000	080 007 000	075 007 000	084 007 000	074 012 001	103 018 001	085 016 001
3	072 004 000	076 007 000	087 007 000	083 014 000	097 022 000	105 018 001	129 025 000	103 016 001
4	088 016 000	099 012 000	097 015 001	104 020 000	113 024 000	112 022 000	143 031 002	101 022 000
5	089 016 001	114 014 001	118 020 000	112 020 000	145 030 002	127 030 002	155 033 003	122 029 002
6	099 014 002	116 016 001	114 023 001	133 026 002	143 040 001	149 037 003	158 040 005	151 033 001
7	114 015 000	126 015 001	119 034 001	149 034 002	148 038 002	168 038 007	174 048 007	169 039 004
8	107 015 000	130 017 001	124 036 003	144 037 001	162 039 005	176 045 011	182 050 006	194 047 006
9	119 023 001	139 021 000	138 042 004	160 044 004	173 048 003	174 056 008	206 064 004	205 057 008
10	119 025 000	152 019 001	142 045 006	172 045 003	185 049 002	192 057 009	218 072 008	207 062 008

Table 5.5. Power with ES = 0.2, at alpha = 0.05 (upper entry), 0.01 (middle entry), 0.001 (lower entry) as a function of number of baselines and number of intervention points. Data are from NORMDIS2.BAS with 1,000 iterations. Decimal points omitted.

	Number of intervention points							
Baselines	6	8	10	12	14	16	18	20
2	088 000 000	149 016 000	159 019 000	163 022 001	163 036 001	189 034 003	206 050 002	232 052 004
3	135 010 000	186 034 001	225 045 005	231 048 001	245 049 001	283 055 006	278 096 006	326 087 010
4	176 021 001	210 049 001	278 059 006	281 066 007	318 089 006	363 112 014	386 124 021	408 148 017
5	200 030 002	269 068 002	330 084 014	346 089 008	385 116 012	425 149 024	433 185 024	487 185 031
6	253 043 004	337 096 004	382 120 013	415 128 017	422 145 024	502 196 029	510 221 033	542 260 036
7	283 055 004	376 116 008	425 137 020	459 174 027	497 174 033	560 238 040	564 254 051	591 301 065
8	304 079 006	414 140 011	458 175 032	499 207 032	564 221 045	605 283 054	599 306 080	647 342 096
9	345 093 007	436 159 020	502 197 041	563 228 043	603 267 059	652 332 071	652 367 106	709 395 116
10	374 120 009	479 186 030	546 241 049	607 289 056	637 310 070	699 388 092	706 413 128	739 453 147

Table 5.6. Power with ES = 0.5, at alpha = 0.05 (upper entry), 0.01 (middle entry), 0.001 (lower entry) as a function of number of baselines and number of intervention points. Data are from NORMDIS2.BAS with 1,000 iterations. Decimal points omitted.

	Number of intervention points							
Baselines	6	8	10	12	14	16	18	20
2	149 006 000	214 029 000	226 045 000	275 046 004	322 069 008	339 069 006	375 084 005	420 112 008
3	227 027 000	321 064 003	354 077 005	410 108 008	460 153 017	510 174 016	563 175 016	563 219 024
4	278 049 001	417 107 007	469 133 013	514 180 021	582 240 035	646 286 041	680 315 044	721 333 058
5	370 084 005	498 162 014	578 219 023	633 277 038	681 320 068	756 389 079	775 438 086	793 461 124
6	448 122 008	578 237 027	660 301 044	727 371 065	764 442 106	810 506 140	847 543 167	870 582 207
7	528 160 010	654 304 047	732 384 072	781 445 115	834 530 159	872 611 201	890 655 259	908 676 283
8	589 223 023	701 375 070	789 473 108	837 550 173	876 612 236	911 688 273	930 732 350	953 759 375
9	635 264 036	756 432 099	833 526 159	891 621 226	909 676 294	945 765 361	953 804 454	968 832 471
10	681 329 049	799 477 137	872 594 208	914 675 293	932 733 373	954 817 453	974 852 511	984 880 577

Table 5.7. Power with ES = 0.8, at alpha = 0.05 (upper entry), 0.01 (middle entry), 0.001 (lower entry) as a function of number of baselines and number of intervention points. Data are from NORMDIS2.BAS with 1,000 iterations. Decimal points omitted.

	Number of intervention points							
Baselines	6	8	10	12	14	16	18	20
2	247 010 000	317 036 000	382 065 001	420 097 005	499 136 011	531 144 012	555 142 009	590 191 032
3	408 066 000	492 118 001	584 167 011	626 222 021	682 315 040	714 318 036	777 348 056	808 409 086
4	518 124 001	634 219 014	716 312 034	778 405 047	816 479 101	854 492 111	895 553 145	915 599 180
5	596 226 007	736 341 031	808 432 071	875 545 120	890 622 182	919 677 216	949 729 265	961 761 337
6	698 301 028	811 449 066	889 567 135	927 671 229	945 737 292	967 800 338	976 839 407	982 973 499
7	774 379 060	888 568 132	924 665 211	957 784 341	967 830 424	985 875 488	986 920 561	997 928 644
8	831 485 094	917 668 191	952 775 306	983 854 442	984 910 558	988 943 624	992 954 704	998 962 765
9	879 570 150	941 750 287	968 841 416	987 913 559	992 930 665	995 953 734	999 980 803	999 983 846
10	917 647 203	966 794 388	982 890 528	993 942 642	997 959 968	999 976 814	999 987 872	1000 993 898

Table 5.8. Power with ES = 1.1, at alpha = 0.05 (upper entry), 0.01 (middle entry), 0.001 (lower entry) as a function of number of baselines and number of intervention points. Data are from NORMDIS2.BAS with 1,000 iterations. Decimal points omitted.

	Number of intervention points							
Baselines	6	8	10	12	14	16	18	20
2	318 016 000	422 065 001	501 096 001	592 138 010	658 199 013	695 221 023	695 267 032	751 288 028
3	530 089 000	658 190 007	729 262 017	806 378 036	860 457 061	890 498 081	897 521 118	936 597 123
4	684 208 009	810 380 034	862 483 060	930 606 106	970 688 193	969 730 221	975 772 264	977 818 310
5	800 335 021	902 570 098	935 668 161	975 783 249	985 856 361	992 875 418	997 914 472	991 937 532
6	858 499 067	941 716 191	974 800 317	993 891 444	994 935 561	999 946 618	999 962 689	998 975 734
7	911 608 128	977 807 337	992 891 482	999 950 618	997 977 731	1000 978 789	1000 982 836	999 991 868
8	957 712 224	987 885 445	993 938 598	1000 982 758	1000 992 844	1000 996 890	999 996 917	1000 997 942
9	965 792 335	993 933 563	998 967 745	1000 992 853	1000 993 918	1000 997 944	1000 999 961	1000 1000 972
10	980 845 420	997 960 678	998 985 826	1000 995 929	1000 998 960	1000 998 976	1000 1000 981	1000 1000 988

Table 5.9. Power with ES = 1.4, at alpha = 0.05 (upper entry), 0.01 (middle entry), 0.001 (lower entry) as a function of number of baselines and number of intervention points. Data are from NORMDIS2.BAS with 1,000 iterations. Decimal points omitted.

	Number of intervention points							
Baselines	6	8	10	12	14	16	18	20
2	437 017 000	590 095 000	664 143 004	698 223 013	751 239 018	814 319 034	837 313 039	878 402 049
3	690 145 000	812 284 008	882 404 027	913 506 065	949 555 102	964 675 142	973 684 151	981 770 195
4	832 336 010	932 550 056	967 687 111	982 763 203	992 823 290	991 880 348	997 911 386	999 947 477
5	926 535 040	979 766 170	992 851 295	994 923 423	997 935 507	999 962 610	1000 972 666	1000 991 756
6	952 711 134	990 894 316	999 933 505	998 968 660	1000 985 749	1000 992 826	1000 995 858	1000 999 928
7	981 836 284	996 957 526	999 983 698	1000 993 823	1000 999 886	1000 998 933	1000 999 944	1000 1000 980
8	987 904 411	1000 979 705	1000 996 836	1000 997 913	1000 1000 950	1000 1000 972	1000 999 988	1000 1000 993
9	994 948 570	1000 991 832	1000 999 836	1000 1000 968	1000 1000 985	1000 1000 997	1000 1000 992	1000 1000 998
10	1000 972 695	1000 996 905	1000 1000 972	1000 1000 990	1000 1000 998	1000 1000 999	1000 1000 998	1000 1000 999

Table 5.10 Power with ES = 1.7, at alpha = 0.05 (upper entry), 0.01 (middle entry), 0.001 (lower entry) as a function of number of baselines and number of intervention points. Data are from NORMDIS2.BAS with 1,000 iterations. Decimal points omitted.

	Number of intervention points							
Baselines	6	8	10	12	14	16	18	20
2	501 027 000	652 101 001	756 178 002	833 233 006	853 322 017	902 366 027	927 382 034	930 458 054
3	796 188 000	894 393 010	948 533 030	970 646 080	982 696 132	990 763 168	995 801 190	995 856 258
4	916 443 014	981 695 084	991 810 160	997 877 288	999 924 391	999 964 484	999 968 513	999 978 592
5	974 691 081	994 870 282	999 936 411	1000 971 580	1000 991 692	1000 995 782	1000 994 813	1000 998 870
6	993 841 224	997 955 507	1000 978 693	999 992 820	1000 1000 885	1000 1000 929	1000 999 947	1000 1000 965
7	997 935 425	1000 987 706	1000 997 853	1000 999 927	1000 1000 966	1000 1000 985	1000 1000 983	1000 1000 995
8	999 975 620	1000 995 865	1000 998 936	1000 1000 981	1000 1000 996	1000 1000 999	1000 1000 999	1000 1000 1000
9	1000 983 759	1000 999 943	1000 1000 975	1000 1000 995	1000 1000 1000	1000 1000 1000	1000 1000 1000	1000 1000 1000
10	1000 993 875	1000 1000 975	1000 1000 993	1000 1000 1000	1000 1000 1000	1000 1000 1000	1000 1000 1000	1000 1000 1000

Table 5.11. Power with ES = 2.0, at alpha = 0.05 (upper entry), 0.01 (middle entry), 0.001 (lower entry) as a function of number of baselines and number of intervention points. Data are from NORMDIS2.BAS with 1,000 iterations. Decimal points omitted.

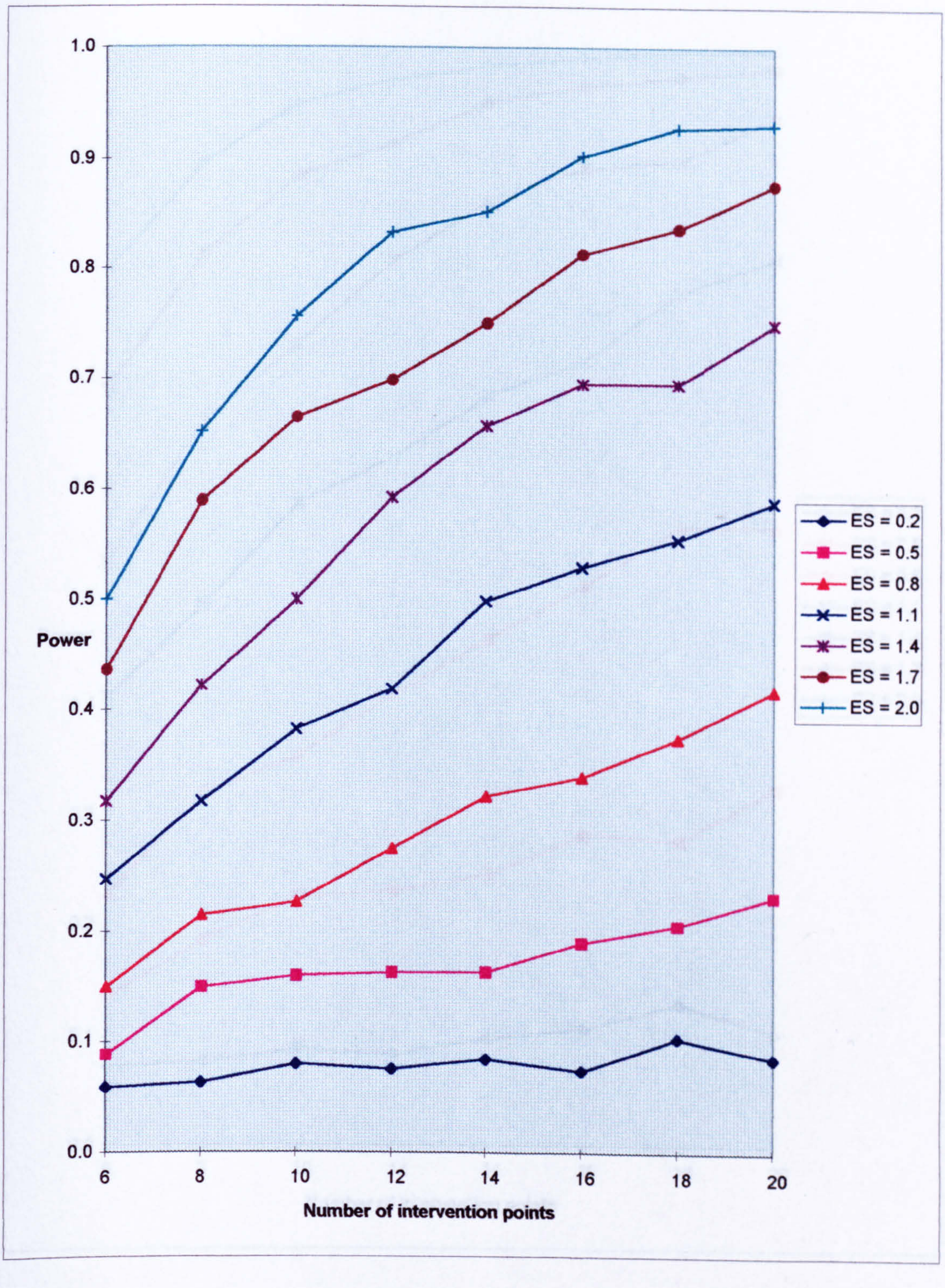


Figure 5.6. Power with 2 baselines at alpha = 0.05, as a function of effect size (ES) and number of intervention points. Data are from NORMDIS2.BAS with 1,000 iterations.

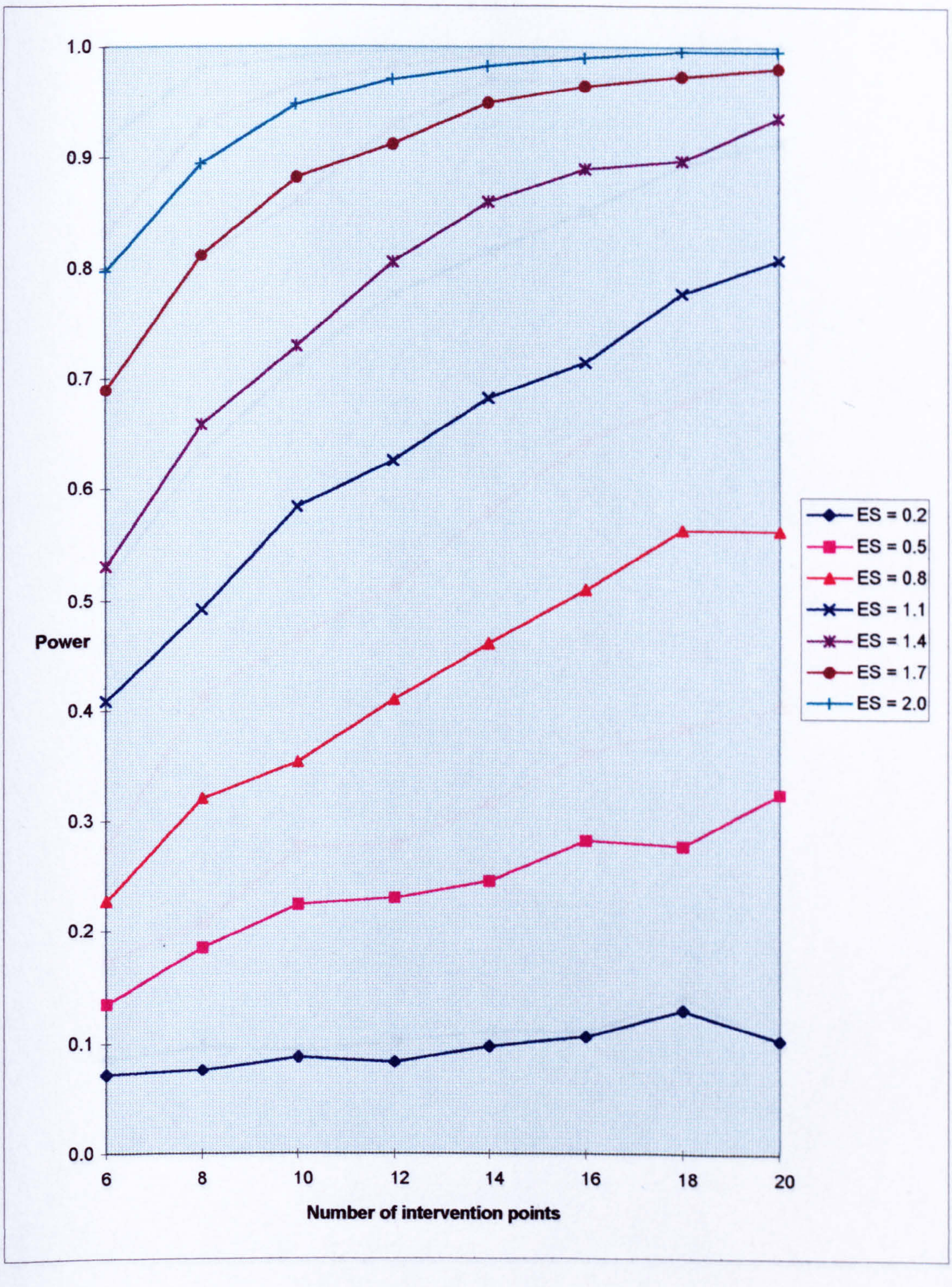


Figure 5.7. Power with 3 baselines at alpha = 0.05, as a function of effect size (ES) and number of intervention points. Data are from NORMDIS2.BAS with 1,000 iterations.

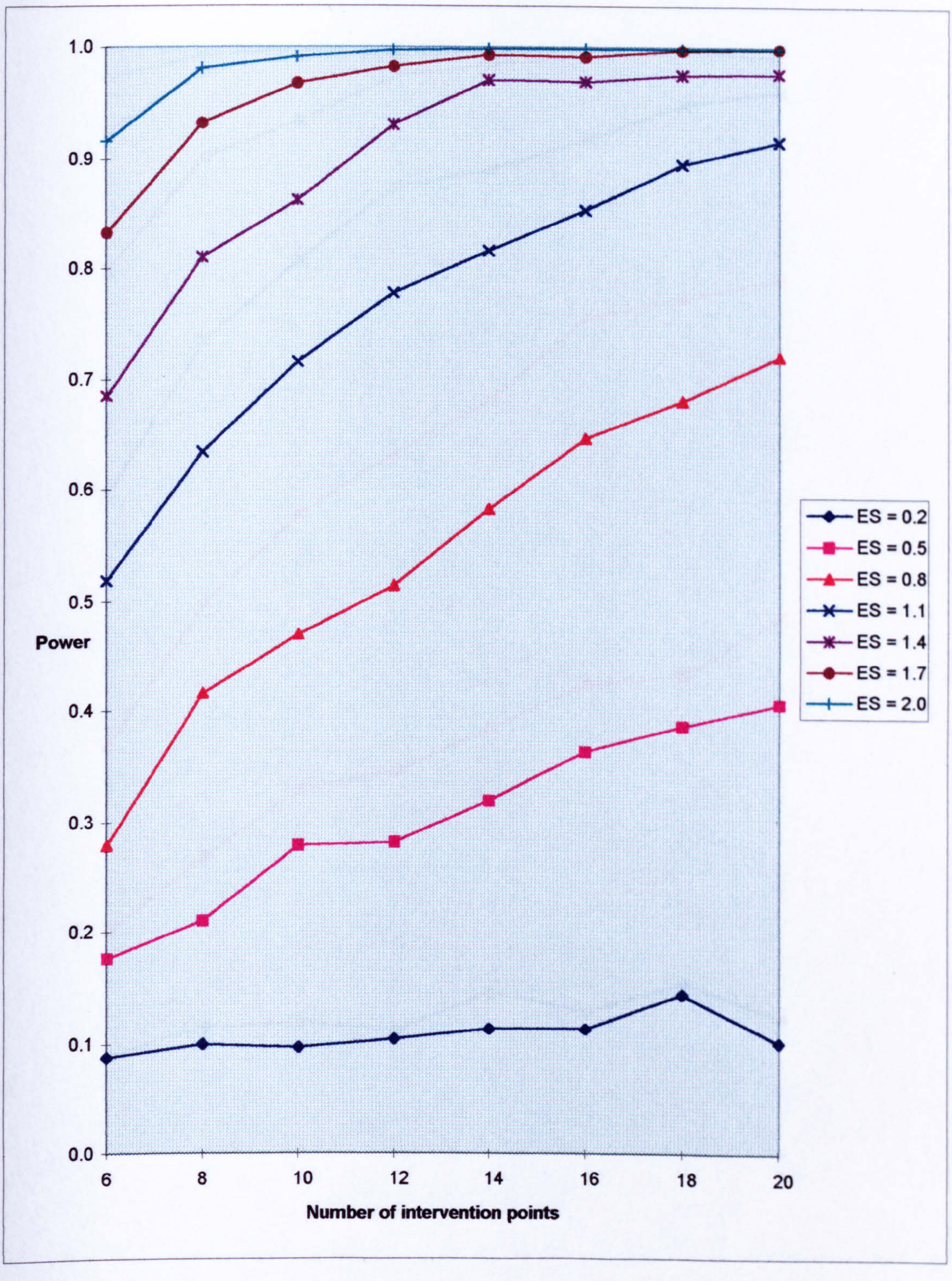


Figure 5.8. Power with 4 baselines at alpha = 0.05, as a function of effect size (ES) and number of intervention points. Data are from NORMDIS2.BAS with 1,000 iterations.

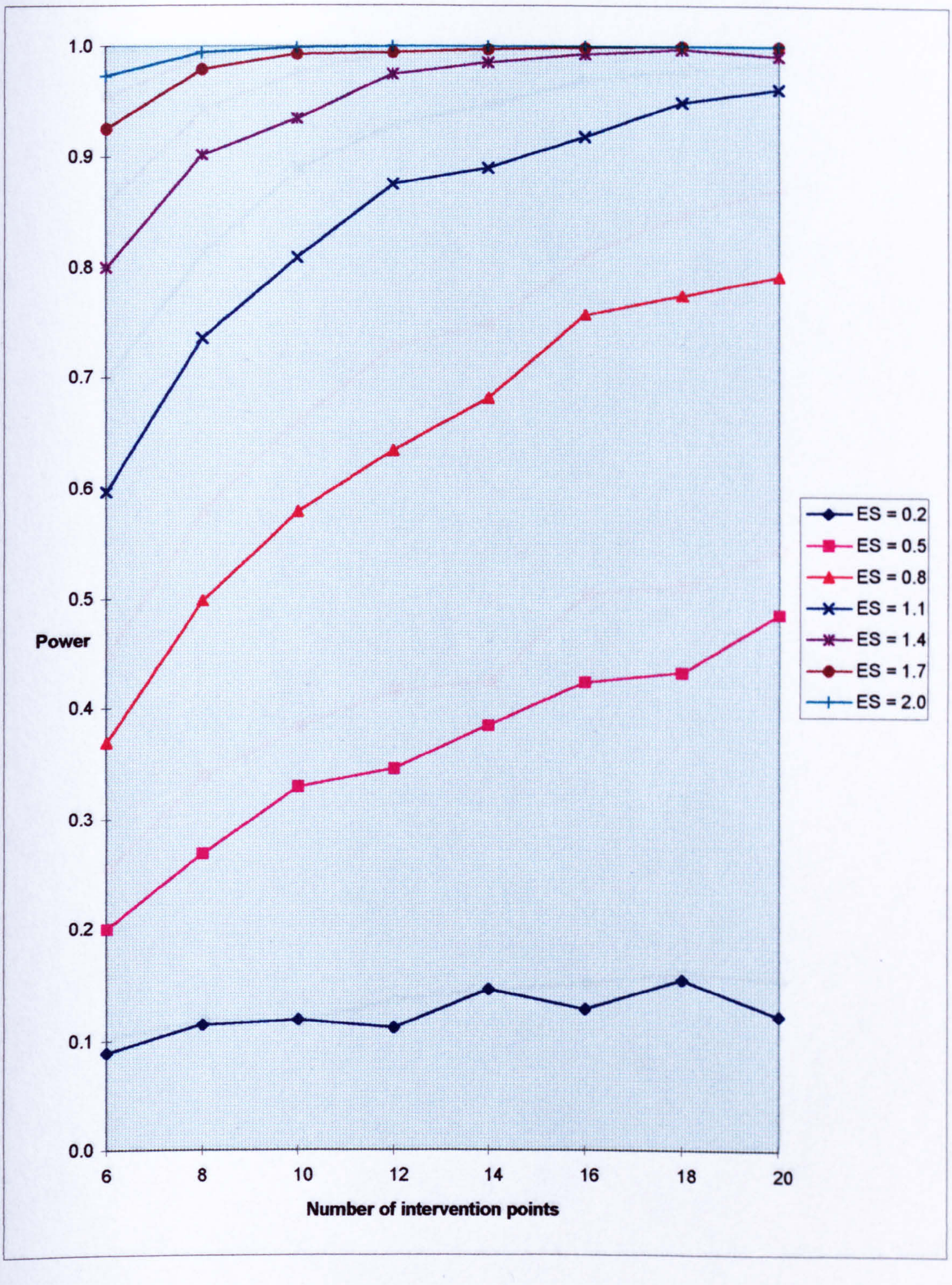


Figure 5.9. Power with 5 baselines at alpha = 0.05, as a function of effect size (ES) and number of intervention points. Data are from NORMDIS2.BAS with 1,000 iterations.

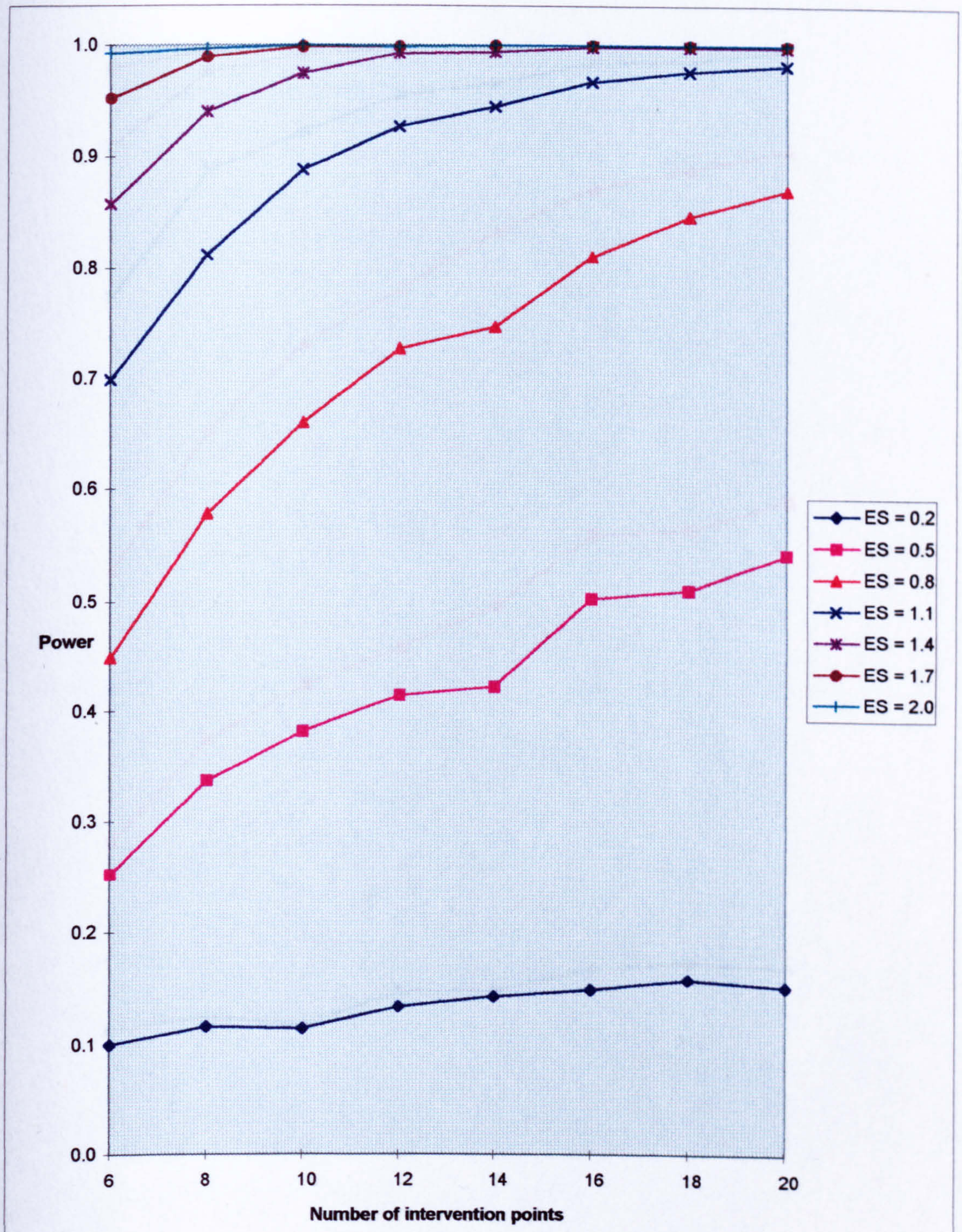


Figure 5.10. Power with 6 baselines at alpha = 0.05, as a function of effect size (ES) and number of intervention points. Data are from NORMDIS2.BAS with 1,000 iterations.

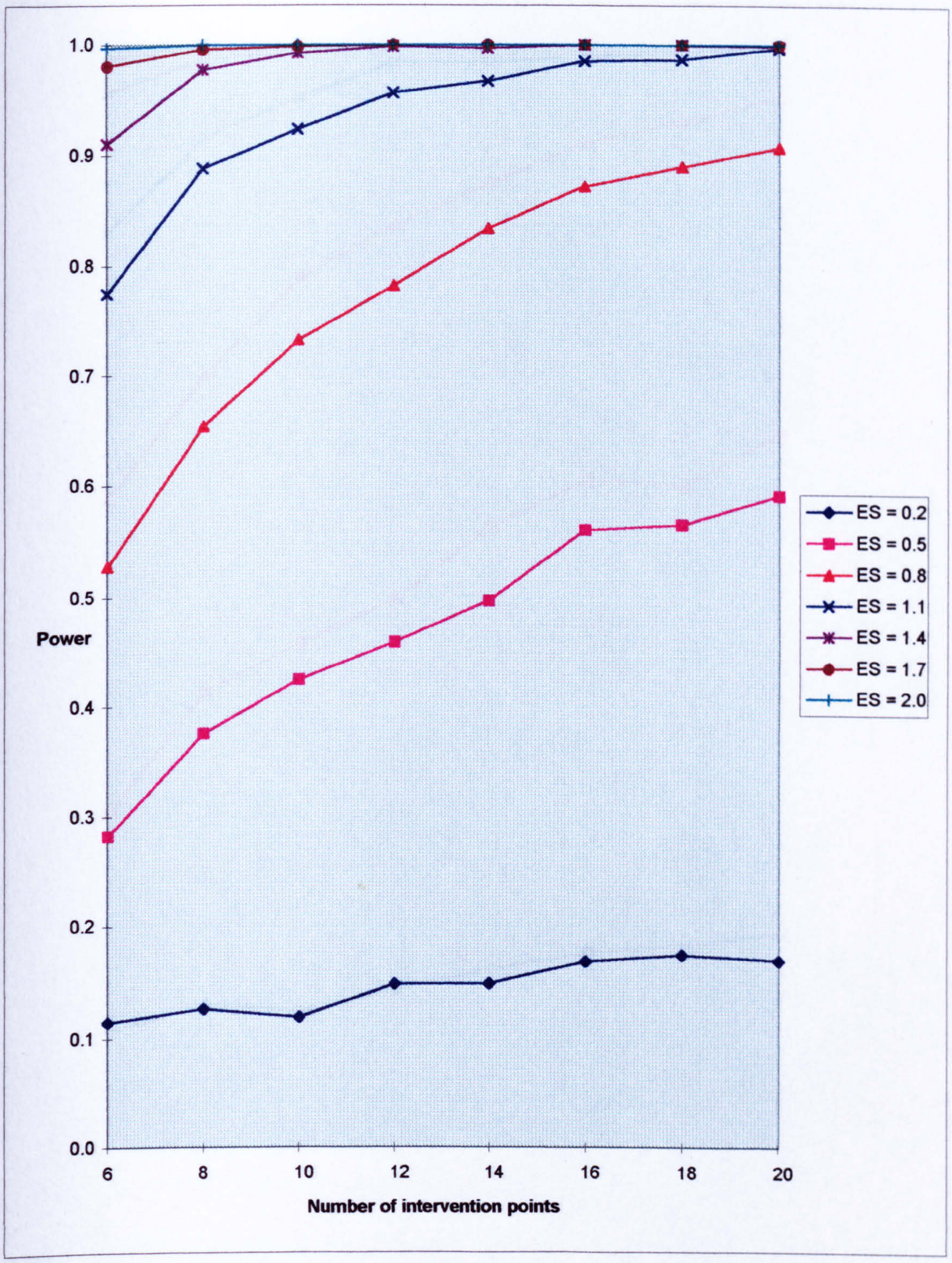


Figure 5.11. Power with 7 baselines at alpha = 0.05, as a function of effect size (ES) and number of intervention points. Data are from NORMDIS2.BAS with 1,000 iterations.

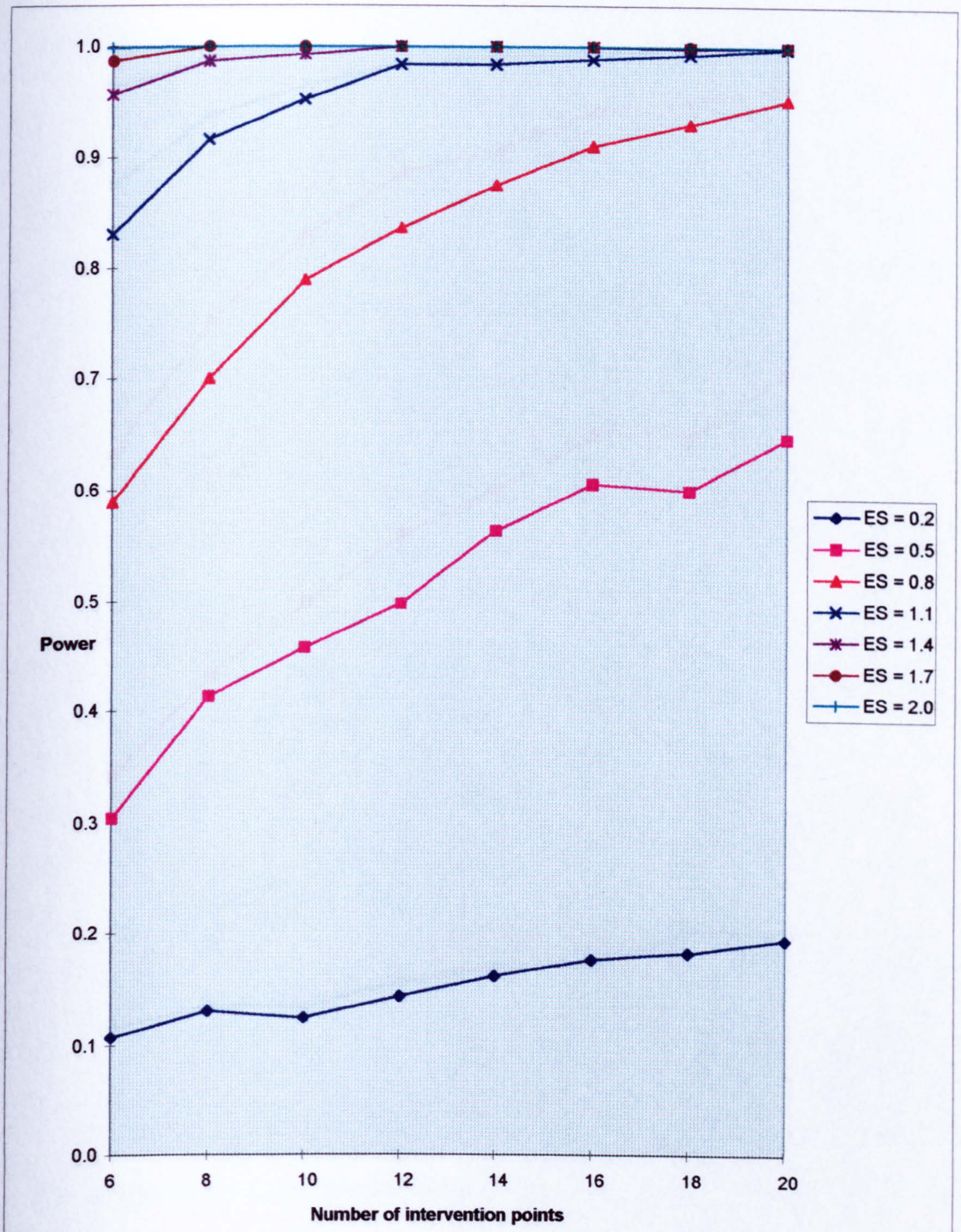


Figure 5.12. Power with 8 baselines at alpha = 0.05, as a function of effect size (ES) and number of intervention points. Data are from NORMDIS2.BAS with 1,000 iterations.

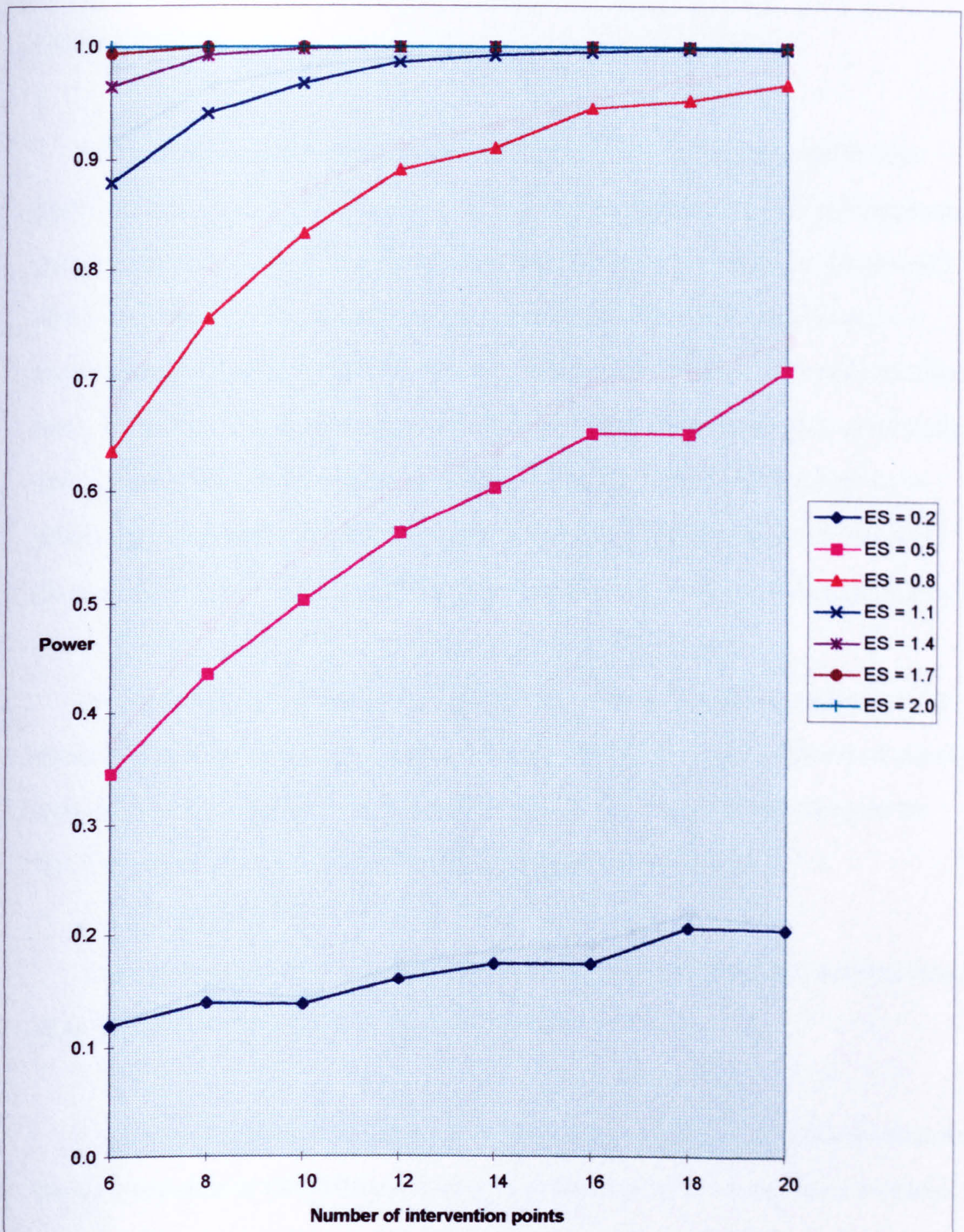


Figure 5.13. Power with 9 baselines at alpha = 0.05, as a function of effect size (ES) and number of intervention points. Data are from NORMDIS2.BAS with 1,000 iterations.

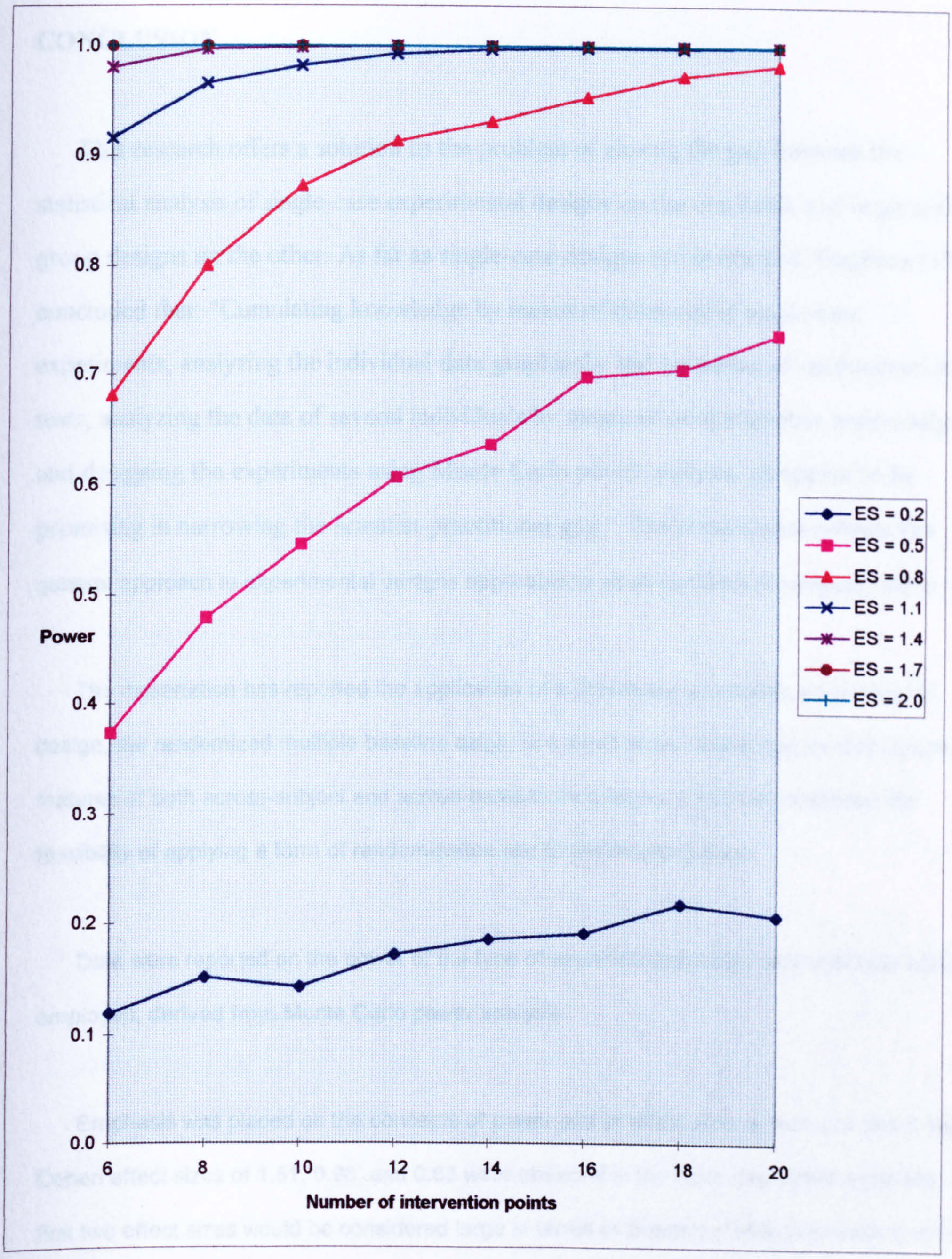


Figure 5.14. Power with 10 baselines at alpha = 0.05, as a function of effect size (ES) and number of intervention points. Data are from NORMDIS2.BAS with 1,000 iterations.

CONCLUSION

This research offers a solution to the problem of closing the gap between the statistical analysis of single-case experimental designs on the one hand, and large scale group designs on the other. As far as single-case designs are concerned, Onghena (1994) concluded that: "Cumulating knowledge by means of consecutive single-case experiments, analyzing the individual data graphically and by means of randomization tests, analyzing the data of several individuals by means of nonparametric meta-analysis, and designing the experiments using Monte Carlo power analysis, all appear to be promising in narrowing the scientist-practitioner gap." The present work extends this general approach to experimental designs applicable to small numbers of subjects (with n > 1).

The dissertation has reported the application of a previously unreported experimental design, the randomized multiple baseline dsign, in a small scale clinical experiment combining features of both across-subject and across-behaviours designs. It has demonstrated the feasibility of applying a form of randomization test to the resulting data.

Data were reported on the power of the type of experimental design and statistical analysis employed, derived from Monte Carlo power analysis.

Emphasis was placed on the concepts of power and of effect size. In terms of effect size, Cohen effect sizes of 1.51, 0.98, and 0.63 were obtained in the three dependent variables. The first two effect sizes would be considered large in terms of Cohen's (1988) proposals in which he suggests that an effect size of 0.80 be defined as "large". In the wider context however, Matyas and Greenwood (1990) found in their survey of 182 baselines published in JABA that the median effect size obtained from 100 AB panels with n >= 10 was 9.2, the 25th percentile was 4.9 and the 75th was 17.1. Thus an effect size of 1 was well below the 25th percentile. The great discrepancy between the magnitude of effect sizes found in such work and by contrast in work in areas such as that reported on in the present clinical experiment, may help to explain

the enthusiasm of operant workers for reliance on graphical data analysis.

The randomized multiple baseline experimental design, with analysis based on a combination of graphical data analysis and randomization statistics, together with the design of experiments using Monte Carlo power analysis, is commended as a useful addition to the armoury of experimental designs for small scale applied clinical psychological research.

REFERENCES

Alford, B.A. (1986). Behavioural treatment of schizophrenic delusions: a single case experimental analysis. Behaviour Therapy, 17, 637-644.

Alford, B.A. and Beck, A.T. (1994). Cognitive therapy for delusions. Behaviour Research and Therapy, 32, 369-380.

Alford, G.S., Fleece, L. & Rothblum, E. (1982). Hallucinatory-delusional verbalizations: modification in a chronic schizophrenic by self-control and cognitive restructuring. Behavior Modification, 6, 421-435.

American Psychiatric Association (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC: Author.

Bailey, D.B. (1984). Effects of lines of progress and semilogarithmic charts on ratings of charted data. Journal of Applied Behavior Ananlysis, 17, 359-365.

Barlow, D.H., Hayes, S.C. & Nelson, R.O. (1984). The scientist practitioner: research and accountability in clinical and educational settings. New York: Pergamon.

Barlow, D.H. & Hersen, M. (1984). Single case experimental designs: strategies for studying behavior change (2nd ed.). New York: Pergamon.

Beck, A.T. (1952). Successful outpatient psychotherapy of a chronic schizophrenic with a delusion based on borrowed guilt. Psychiatry, 15, 305-312.

Bellack, A.S. & Hersen, M. (1988). Behavioral assessment: a practical handbook (3rd ed.). New York: Pergamon Press.

Bentall, R.P. (Ed.) (1990). Reconstructing schizophrenia. London and New York: Routledge.

Bentall, R.P., Haddock, G. & Slade, P.D. (1994). Cognitive behaviour therapy for persistent auditory hallucinations: from theory to therapy. Behaviour Therapy, 25, 51-66.

Bentall, R.P., Jackson, H.F. & Pilgrim, D. (1988). Abandoning the concept of schizophrenia: some implications of validity arguments for psychological research into psychotic phenomena. British Journal of Clinical Psychology, 27, 303-324.

Berger, M. (1996). Outcomes and effectiveness in clinical psychology practice. The British Psychological Society. Division of Clinical Psychology Occasional Paper No. 1.

Birchwood, M. & Tarrier, N. (1994). Psychological management of schizophrenia. Chichester: Wiley.

Bjorgvinsson, T. & Kerr, P. (1995) (letter). Use of a common language effect size statistic. American Journal of Psychiatry, 152, 151.

Bouchard, S., Vallieres, A., Roy, M. & Maziade, M. (1996). Cognitive restructuring in the treatment of psychotic symptoms in schizophrenia: a critical analysis. Behavior Therapy, 27, 257-277.

Boyle, M. (1990). Schizophrenia - a scientific delusion? London: Routledge.

Busk, P.L. & Marascuilo, L.A. (1988). Autocorrelation in single-subject research: a counterargument to the myth of no autocorrelation. Behavioral Assessment, 10, 229-

242.

Busk, P.L. & Marascuilo, L.A. (1992). Statistical analysis in single-case research: issues, procedures, and recommendations, with applications to multiple behaviors. In: Kratochwill & Levin, op. cit..

Campbell, D.T. & Stanley, J.C. (1966). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.

Center, B.A., Skiba, R.J. & Casey, A. (1985). A methodology for the quantitative synthesis of intra-subject design research. The Journal of Special Education, 19, 387-400.

Chadwick, P.D.J. & Birchwood, M. (1994). The omnipotence of voices: a cognitive approach to auditory hallucinations. British Journal of Psychiatry, 164, 190-201.

Chadwick, P., Birchwood, M. & Trower, P. (1996). Cognitive therapy for delusions, voices and paranoia. Chichester: Wiley.

Chadwick, P.D.J. & Lowe, C.F. (1990). Measurement and modification of delusional beliefs. Journal of Consulting and Clinical Psychology, 58, 225-232.

Chadwick, P.D.J., Lowe, C.F., Horne, P.J. & Higson, P. (1994). Modifying delusions: the role of empirical testing. Behaviour Therapy, 25, 35-49.

Chadwick, P. & Trower, P. (1996). Cognitive therapy for punishment paranoia: a single case experiment. Behaviour Research & Therapy, 34, 351-356.

Chassan, J.B. (1979). Research design in clinical psychology and psychiatry (2nd ed.).

New York: Irvington.

Clements, J.C. & Hand, D.J. (1985). Permutation statistics in single case designs. Behavioural Psychotherapy, 13, 288-299.

Cliffe, M.J., Possamai, A. & Mulhall, D. (1995). Modified Personal Questionnaire Rapid Scaling Technique for measuring delusional beliefs. British Journal of Clinical Psychology, 34, 251-253.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. Journal of Abnormal and Social Psychology, 65, 145-153.

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. San Diego, CA: Academic Press.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Cook, T.D. & Campbell, D.T. (1979). Quasi-experimentation: design and analysis issues for field settings. Chicago, IL: Rand McNally.

Cowles, M. (1989). Statistics in psychology: an historical perspective. Hillsdale, NJ: Erlbaum.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. Journal of Consulting and Clinical Psychology, 61, 966 - 974.

Davidson, P.O. & Costello, C.G. (Eds.) (1969). N = 1: Experimental studies of single cases. New York: Van Nostrand/Reinhold Co.

De Prospero, A. & Cohen, S. (1979). Inconsistent visual analysis of intra-subject data. Journal of Applied Behavior Analysis, 12, 573-579.

Dukes, W.F. (1965). N = 1. Psychological Bulletin, 64, 74-79.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. Annals of Mathematical Statistics, 28, 181-187.

Edgington, E.S. (1964). Randomization tests. Journal of Psychology, 57, 445-449.

Edgington, E.S. (1966). Statistical inference and nonrandom samples. Psychological Bulletin, 66, 485-487.

Edgington, E.S. (1967). Statistical inference from N = 1 experiments. The Journal of Psychology, 65, 195-199.

Edgington, E.S. (1969a). Approximate randomization tests. Journal of Psychology, 72, 143-179.

Edgington, E.S. (1969b). Statistical inference: the distribution-free approach. New York: McGraw-Hill.

Edgington, E. S.(1970). Hypothesis testing without fixed levels of significance. Journal of Psychology, 76, 109-115.

Edgington, E.S. (1972a). A normal curve method for combining probability values from independent experiments. The Journal of Psychology, 82, 85-89.

Edgington, E.S. (1972b). An additive method for combining probability values from independent experiments. The Journal of Psychology, 80, 351-363.

Edgington, E.S. (1973). The random-sampling assumption in "Comment on component-randomization test". Psychological Bulletin, 80, 84-85.

Edgington, E.S. (1975a). Randomization tests for one-subject operant experiments. Journal of Psychology, 90, 57-68.

Edgington, E.S. (1975b). Randomization tests for predicted trends. Canadian Psychological Review, 16, 49-53.

Edgington, E.S. (1980a). Overcoming obstacles to single-subject experimentation. Journal of Educational Statistics, 5, 261-267.

Edgington, E.S. (1980b). Randomization Tests. New York: Marcel Dekker.

Edgington, E.S. (1980c). Validity of randomization tests for one-subject experiments. Journal of Educational Statistics, 5, 235-251.

Edgington, E.S. (1984). Statistics and single case analysis. In: Hersen, M., Eisler, R.M. & Miller, P. (eds.) Progress in behavior modification, 16. New York: Academic Press.

Edgington, E.S. (1986). Randomization tests. In S. Kotz & N.L. Johnson (eds.), Encyclopedia of statistical sciences, Vol. 7. New York: Wiley.

Edgington, E.S. (1987). Randomization tests (2nd ed.). New York: Marcel Dekker.

Edgington, E.S. (1992). Nonparametric tests for single-case experiments. In:

Kratochwill & Levin, op. cit..

Edgington, E.S.(1995). Randomization tests (3rd ed.). New York: Marcel Dekker.

Edgington, E.S. (1996). Randomized single-subject experimental designs. Behaviour Research & Therapy, 34, 567-574.

Edgington, E.S. & Haller, O. (1984). Combining probabilities from discrete probability distributions. Educational and Psychological Measurement, 44, 265-274.

Ewart, C.K., Burnett, K.F. & Taylor, C.B. (1983). Communication behaviors that affect blood pressure. Behavior Modification, 7, 331-344.

Ferron, J. & Ware, W. (1994). Using randomization tests with responsive single-case designs. Behavior Research and Therapy, 32, 787-791.

Fisher, R.A. (1955). Statistical methods and scientific induction. Journal of the Royal Statistical Society, Series B, 17, 69-78.

Fisher, R.A. (1966). The design of experiments (8th ed.). Edinburgh: Oliver & Boyd. (Oriiginal work published in 1935).

Fisher, R.A. (1956). Statistical methods and scientific inference (2nd ed.). Edinburgh: Oliver & Boyd.

Fowler, D., Garety, P. & Kuipers, E. (1995). Cognitive behaviour therapy for psychosis. Chichester: Wiley.

Fowler, D., Garety, P. & Kuipers, E. (1998). Cognitive therapy for psychosis:

formulation, treatment, effects and service implications. Journal of Mental Health, 7, 2, 123-133.

Fowler, D. & Morley, S. (1989). The cognitive-behavioural treatment of hallucinations and delusions: a preliminary study. Behavioural Psychotherapy, 17, 267-282.

Furlong, M.J. & Wampold, B.E. (1981). Visual analysis of single-subject studies by school psychologists. Psychology in the Schools, 18, 80-86.

Furlong, M.J. & Wampold, B.E. (1982). Intervention effects and relative variation as dimensions in experts' use of visual inspection. Journal of Applied Behavior Analysis, 15, 415-421.

Gabriel, K.R. & Hall, W.J. (1983). Re-randomization inference on regression and shift effects: computationally feasible methods. Journal of the American Statistical Association, 78, 827-836.

Gabriel K.R. & Hsu, C.F. (1983). Power studies of re-randomization tests, with application to weather modification experiments. Journal of tha American Statistical Association, 78, 766-775.

Garety, P. (1985). Delusions: problems in definition and measurement. British Journal of Medical Psychology, 58, 25-34.

Garety, P. (1992). Making sense of delusions. Psychiatry, 55, 282-291.

Garety, P., Kuipers, L., Fowler, D., Chamberlain, F. & Dunn, G. (1994). Cognitive behavioural therapy for drug-resistant psychosis. British Journal of Medical Psychology, 67, 259-271.

Gentile, J.R., Roden, A.H. & Klein, R.D. (1972). An analysis of variance model for the intrasubject replication design. Journal of Applied Behavior Analysis, 5, 193-198.

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In: L. Kruger, G. Gigerenzer & M. Morgan (eds.) The probabilistic revolution. Vol. 2. Cambridge, MA: The MIT Press.

Gigerenzer, G. & Murray, D.J. (1987). Cognition as intuitive statistics. Hillsdale, NJ: LEA.

Goldstein, R. (1989). Power and sample size via MS/PC-DOS computers. American Statistician, 43, 253-260.

Gorsuch, R.L.(1983). Three methods for analysing limited time-series (N of 1) data. Behavioral Assessment, 5, 141-154.

Greenwood, V.B. (1983). Cognitive therapy with couples and groups. In A. Freeman (Ed.), Cognitive therapy with the young adult chronic patient. New York: Plenum Press.

Guyatt, G.H., Heyting, A., Jaeschke, R., Keller, J., Adachi, J.D. & Roberts, R.S. (1990a). N of 1 randomized trials for investigating new drugs. Controlled Clinical Trials, 11, 88-100.

Guyatt, G.H., Keller, J., Jaeschke, R., Rosenbloom, D., Adachi, J.D. & Newhouse, M.T. (1990b). The n-of-1 randomized controlled trial: clinical usefulness. Annals of Internal Medicine, 112, 293-299.

Haddock, G. & Slade, P.D. (Eds.) (1996). Cognitive-behavioural interventions with

psychotic disorders. London & New York: Routledge.

Hallahan, M. & Rosenthal, R. (1966). Statistical power: concepts, procedures and applications. Behaviour Research & Therapy, 34, 489-499.

Hand, D.J. (1982). Statistical tests in experimental psychiatric research. Psychological Medicine, 12, 415-421.

Hartman, L.M. & Cashman, F.E. (1983). Cognitive behavioural and psychopharmacological treatment of delusional symptoms: a preliminary report. Behavioural Psychotherapy, 11, 50-61.

Hayes, S.C. (1981). Single case experimental design and empirical clinical practice. Journal of Consulting and Clinical Psychology, 49, 193-211.

Hersen, M. & Barlow, D.H. (1976). Single-case experimental designs: strategies for studying behavior change. New York: Pergamon.

Himadi, B. & Kaiser, A.J. (1992). The modification of delusional beliefs: a single-subject evaluation. Behavioral Residential Treatment, 7, 1-14.

Holtzman, W.H. (1963). Statistical methods for studying change in the single case. In C.W. Harris (Ed.), Problems in measuring change. Madison: University of Wisconsin Press.

Hole, R.W., Rush, A.J. & Beck, A.T. (1979). A cognitive investigation of schizophrenic delusions. Psychiatry, 42, 312-319.

Honig, W.K. (Ed.) (1966). Operant behavior: areas of research and application. New

York: Appleton-Century Crofts.

Hope, A.C.A. (1968). A simplified Monte Carlo significance test procedure. Journal of the Royal Statistical Society Series B, 30, 582-598.

Huitema, B.E. (1985). Autocorrelation in applied behavior analysis: a myth. Behavioral Assessment, 7, 107-118.

Ischi, N. (1980). Analyse des fondements technologiques de la modification des contingences sociales en classe. Revue Suisse de Psychologie, 39, 113-182.

Johnson, J.M. & Pennypacker, H.S. (1980). Strategies and tactics of human behavioral research. Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson, W.G., Ross, J.M. & Mastria, M.A. (1977). Delusional behaviour: an attributional analysis. Journal of Abnormal Psychology, 86, 421-426.

Jones, H.G. (1971). In search of as idiographic psychology. Bulletin of the British Psychological Society, 24, 279-290.

Jones, R.R., Weinrott, M. & Vaught, R.S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. Journal of Applied Behavior Analysis, 11, 277-283.

Kazdin, A.E. (1976). Statistical analyses for single-case experimental designs. In: Hersen & Barlow, op. cit.

Kazdin, A. E. (1980). Behavior modification in applied settings. Homewood, Ill: Dorsey Press.

Kazdin, A.E. (1981). Drawing valid inferences from case studies. Journal of Consulting and Clinical Psychology, 49, 183-192.

Kazdin, A.E. (1982). Single-case research designs. New York & Oxford: Oxford University Press.

Kazdin, A.E. & Bass, D. (1989). Power to detect differences between alternative treatments in comparative psychotherapy outcome research. Journal of Consulting and Clinical Psychology, 57, 138-147.

Kempthorne, O. (1952). The design and analysis of experiments. New York: Wiley.

Kempthorne, O. (1955). The randomization theory of experimental inference. Journal of the American Statistical Association, 50, 964-967.

Kenny, D.A. (1994). Interpersonal perception. New York: Guilford Press.

Kingdon, D. (1997). The Wellcome study of cognitive therapy for treatment resistant schizophrenia. Paper presented at 2nd international conference on psychological treatments for schizophrenia. Oxford, UK, June 1997.

Kingdon, D.G. & Turkington, D. (1994). Cognitive-behavioural therapy of schizophrenia. Hove: Lawrence Erlbaum.

Kingdon, D.G., Turkington, D. & John, C. (1994). Cognitive behaviour therapy of schizophrenia: the amenability of delusions and hallucinations to reasoning. British Journal of Psychiatry, 164, 581-587.

Knapp, T.J. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. Behavioral Assessment, 5, 155-164.

Kraemer, H.C. & Thiemann, S. (1987). How many subjects? Statistical power analysis in research. Newbury Park, CA: Sage.

Kratochwill, T.R. (Ed.) (1978). Single subject research: strategies for evaluating change. New York: Academic Press.

Kratochwill, T.R. (1992). Single-case research design and analysis: an overview. In: Kratochwill & Levin, op. cit.

Kratochwill, T., Alden, K., Demuth, D., Dawson, D., Panicucci, C., Arntson, P., McMurray, N., Hempstead, J. & Levin, J. (1974). A further consideration in the application of an analysis-of-variance model for the intrasubject replication design. Journal of Applied Behavior Analysis, 7, 629-633.

Kratochwill, T.R., & Brody, G.H. (1978). Single subject designs: a perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. Behavior Modification, 2, 291-307.

Kratochwill, T.R. & Levin, J.R. (Eds.) (1992). Single-case research design and analysis. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kuipers, E., Garety, P., Fowler, D., Dunn, G., Bebbington, P., Freeman, D. & Hadley, C. (1997). The London-East Anglia randomized controlled trial of cognitive-behavioural therapy for psychosis 1: effects of the treatment phase. British Journal of Psychiatry, 171, 319-227.

Lehmann, E.L. (1986). Testing statistical hypotheses (2nd ed.). New York: Wiley.

Levin, J.R., Marascuilo, L.A. & Hubert, L.J. (1978). N = nonparametric randomization tests. In T.R. Kratochwill (Ed.), op. cit.

Lipsey, M.W. (1990). Design sensitivity: statistical power for experimental research. Newbury Park, CA: Sage.

Loftus, G.R., (1991). On the tyranny of hypothesis testing in the social sciences. Contemporary Psychology, 36, 102-105.

Loftus, G.R. (1993). One picture is worth a thousand p values: on the irrelevance of hypothesis testing in the microcomputer age. Behavior Research Methods, Instruments and Computers, 25, 250-256.

Long, C.G. & Hollin, C.R. (1995). Single case design: a critique of methodology and analysis of recent trends. Clinical Psychology and Psychotherapy, 2 , 177-191.

Lowe, C.F. & Chadwick, P.D.J. (1990). Verbal control of delusions. Behaviour Research and Therapy, 21, 461-479.

Ludbrook, J. (1994). Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology. Clinical and Experimental Pharmacology and Physiology, 21, 673-686.

Manly, B.F.J. (1991). Randomization and Monte Carlo methods in biology. London: Chapman and Hall.

Marascuilo, L.A. & Busk, P.L. (1988). Combining statistics for multiple-baseline AB

and replicated ABAB designs across subjects. Behavioral Assessment, 10, 1-28.

Matyas, T.A. & Greenwood, K.M. (1990). Visual analysis of single-case time series: effects of variability, serial dependence, and magnitude of intervention effects. Journal of Applied Behavior Analysis, 23, 341-351.

McGraw, K.O. & Wong, S.P. (1992). A comon language effect size statistic. Psychological Bulletin, 111, 361-365.

McNally, R.J. (1994). Introduction to the special series: innovations in cognitive-behavioural approaches to schizophrenia. Behaviour Therapy, 25, 1-4.

Milton. F., Patwa, V.K. & Hafner, R.J. (1978). Confrontation vs belief modification in persistently deluded patients. British Journal of Medical Psychology, 51, 127-130.

Morley, S. (1989). Single-case research. In: G. Parry & F.N. Watts (Eds.). Behavioural and mental health research: a handbook of skills and methods. Hove, Sussex: Lawrence Erlbaum Associates.

Morley, S. & Adams, M.(1989). Some simple statistical tests for exploring single-case time series data. British Journal of Clinical Psychology, 28, 1-18.

Morley, S. & Adams, M. (1991). Graphical analysis of single-case time series data. British Journal of Clinical Psychology, 30, 97-115.

Mulhall, D.J. (1976). Systematic self-assessment by PQRST. Psychological Medicine, 6, 591-597.

Mulhall, D.J. (1978). Manual and booklet for the Personal Questionnaire Rapid Scaling

Technique. Windsor, Berks.: NFER/Nelson.

Neyman, J. & Pearson, E.S. (1933). The testing of statistical hypotheses in relation to probabilities a priori. Proceedings of the Cambridge Philosophical Society, 29, 492-510.

Noreen, E.W. (1989). Computer-intensive methods for testing hypotheses: an introduction. New York: Wiley.

Onghena, P. (1992). Randomization tests for extensions and variations of ABAB single-case experimental designs: a rejoinder. Behavioral Assessment, 14, 153-171.

Onghena, P. (1993). A theoretical and empirical comparison of mainframe, microcomputer, and pocket calculator pseudorandom number generators. Behavior Research Methods, Instruments & Computers, 25, 384-395.

Onghena, P. (1994). The power of randomization tests for single-case designs. Ph.D. Thesis, University of Leuven.

Onghena, P. & Delbeke, L. (1992). Power analysis of randomization tests for single-case designs. International Journal of Psycholgy, 27, 379.

Onghena, P. & Edgington, E.S. (1994). Randomization tests for restricted alternating treatments designs. Behavior Research and Therapy, 32, 783-786.

Onghena, P. & May, R.B. (1995). Pitfalls in computing and interpreting randomization test p values: a commentary on Chen and Dunlap. Behavior Research Methods, Instruments & Computers, 27, 408-411.

Onghena, P. & Van Damme, G. (1994). SCRT 1.1: Single case randomization tests.

Behavior Research Methods, Instruments & Computers, 26, 369.

Parsonson, B.S. & Baer, D.M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In: Kratochwill & Levin, op. cit.

Perris, C., Ingelsson, U. & Jonsson, P.A. (1993). Cognitive therapy as a general framework in the treatment of psychotic patients. In K.T Kuehlwein & H. Rosen (Eds), Cognitive therapies in action. Evolving innovative practice. San Francisco: Jossey Bass.

Perry, G. (1993). QBasic by example. Indianapolis, IN: QUE.

Phillips, J.P.N. (1977). Generalised personal questionnaire techniques. In: P. Slater (Ed.). Dimensions of interpersonal space, vol. 2. Chichester: Wiley.

Phillips, J.P.N. (1983). Serially correlated errors in some single-subject designs. British Journal of Mathematical and Statistical Psychology, 36, 269-280.

Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any population. Journal of the Royal Statistical Society, B, 4, 119-130.

Revusky, S.H. (1967). Some statistical treatments compatible with individual organism methodolgy. Journal of the Experimental Analysis of Behavior, 10, 319-330.

Rosenthal, R. (1978). Combining results of independent studies. Psychological Bulletin, 85, 185-193.

Rosenthal, R. (1991). Meta-analytic procedures for social research (rev. edn.). Beverly Hill, CA: Sage.

Rosenthal, R. & Rubin, D.B. (1982). A simple, general purpose display of magnitude of experimental effect. Journal of Educational Psychology, 74, 166-169.

Rosenthal, R. & Rubin, D.B. (1994). The counternull value of an effect size : a new statistic. Psychological Science, 5, 329-334.

Rossi, J.S. (1990). Statistical power of psychological research: what have we gained in 20 years? Journal of Consulting and Clinical Psychology, 58, 646-656.

Sacks, S.B.Z. (1987). Peer-mediated social skills training: enhancing the social competence of visually handicapped children in a mainstreamed school setting. Ph.D. thesis, University of California, Berkeley with San Francisco State University. (Dissertation Abstracts International, 48 no. 10, April 1988, 2601-A.

Scheffe, H. (1959). The Analysis of Variance. New York: Wiley.

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309-316.

Shapiro, M.B. & Ravenette, A.T. (1959). A preliminary experiment on paranoid delusions. Journal of Mental Science, 105, 296-312.

Shapiro, M.B. (1961). A method of measuring psychological changes specific to the individual patient. British Journal of Medical Psychology, 34, 151-155.

Sharpley, C.F. (1988). Single-subject research. In J.P. Keeves, Educational research, methodology and measurement: an international handbook (pp. 580-586). Oxford: Pergamon Press.

Shapiro, M.B. (1961). A method of measuring psychological changes specific to the individual patient. British Journal of Medical Psychology, 34, 151-155.

Shapiro, M.B. & Ravenette, A.T. (1959). A preliminary experiment in paranoid delusions. Journal of Mental Science, 105, 295-312.

Sidman, M. (1960). Tactics of scientific research. New York: Basic Books.

Siegel, S. & Castellan, N.J. (1988). Nonparametric statistics for the behavioral sciences (2nd ed.). New York: McGraw-Hill.

Silvey, S.D. (1975). Statistical inference. London: Chapman & Hall.

Snedecor, G.W. (1937). Statistical Methods. Ames, IA: Collegiate Press.

Snedecor, G.W. & Cochran, W.G. (1989). Statistical methods (8th ed.). Ames, IA: Iowa State University Press.

Strauss, J.S. (1969). Hallucinations and delusions as points on continua function. Archives of General Psychiatry, 21, 581-586.

"Student". (1908). The Lanarkshire milk experiment. Biometrika, 23, 398-406.

Suen, H.K. & Ary, D. (1987). Autocorrelation in applied behavior analysis: myth or reality? Behavioral Assessment, 9, 125-130.

Tarrier, N. (1997). Coping and problem solving in the treatment of persistent psychotic symptoms. Paper presented at 2nd international conference on psychological treatments for schizophrenia. Oxford, UK, June 1997.

Thoresen, C.E. & Elsahoff, J.D. (1974). "An analysis-of-variance model for intra-subject replication design": some additional comments. Journal of Applied Behavior Analysis, 7, 639-641.

Toothaker, L.E., Banz, M., Noble, C., Camp, J. & Davis, D. (1983). N = 1 designs: the failure of ANOVA-based tests. Journal of Educational Statistics, 8, 289-309.

Wampold, B.E. & Furlong, M.J. (1981a). Randomization tests in single-subject designs: illustrative examples. Journal of Behavioral Assessment, 3, 329-341.

Wampold, B.E. & Furlong, M.J. (1981b). The heuristics of visual inspection. Behavioural Assessment, 3, 79-92.

Wampold, B.E. & Worsham, N.L. (1986). Randomization tests for multiple-baseline designs. Behavioral Assessment, 8, 135-143.

Watson, P.J. & Workman, E.A. (1981). The non-concurrent multiple baseline across-individuals design: an extension of the traditional multiple baseline design. Journal of Behavior Therapy and Experimental Psychiatry, 12, 257-259.

Watts, F.N., Powell, G.E. & Austin, S.V. (1973). The modification of abnormal beliefs. British Journal of Medical Psychology, 46, 359-363.

Welch, B.L. (1937). On the z-test in randomized blocks and Latin squares. Biometrika, 29, 21-52.

Wilson, B. (1987). Rehabilitation of memory. New York: Guilford Press.

Wilson, B. (1991). Behaviour therapy in the treatment of neurologically impaired adults. In P.R. Martin (Ed.), Handbook of behavior therapy and psychological science: an integrative approach. New York: Pergamon Press.

Wolery, M., & Billingsley, F.F. (1982). The application of Revusky's Rn test to slope and level of change. Behavioral Assessment, 4, 93-103.

```
'APPENDIX 1


'sim2.bas
'generates random normally distributed data


CLS : DIM ctr(200): DIM norm(310): DIM zedd(310)

FOR i = 1 TO 308
 READ norm(i)
NEXT i

FOR i = 1 TO 308
 READ zedd(i)
NEXT i

FOR i = 1 TO 200
 ctr(i) = 0
NEXT i

RANDOMIZE TIMER
PRINT "input mean": INPUT m
PRINT "input standard deviation": INPUT s

FOR i = 1 TO 10000

 x = RND: IF x > .5 THEN x = 1 - x

 FOR enn = 1 TO 308
  IF norm(enn) >= x THEN z = zedd(enn): EXIT FOR
 NEXT enn

 sign = RND
 sc = INT(z * s)
 IF sign > .5 THEN y = m + sc ELSE y = m - sc

 FOR j = 0 TO 200
  IF y = j THEN ctr(j) = ctr(j) + 1
 NEXT j

NEXT i

FOR j = 90 TO 110
 LPRINT j; ctr(j)
NEXT j

'normal curve DATA are READ as in Appendix 2
```

```
'APPENDIX 2

'Percentage of scores under the Normal Curve from 0 to z

'Percentage of scores

DATA .0040,.0080,.0120,.0160,.0199,.0239,.0279,.0319,.0359
DATA .0398,.0438,.0478,.0517,.0557,.0596,.0636,.0675,.0714,.0754
DATA .0793,.0832,.0871,.0910,.0948,.0987,.1026,.1064,.1103,.1141
DATA .1179,.1217,.1255,.1293,.1331,.1368,.1406,.1443,.148,.1517
DATA .1554,.1591,.1628,.1664,.17,.1736,.1772,.1808,.1844,.1879
DATA .1915,.195,.1985,.2019,.2054,.2088,.2123,.2157,.219,.2224
DATA .2258,.2291,.2324,.2357,.2389,.2422,.2454,.2486,.2518,.2549
DATA .258,.2612,.2642,.2673,.2704,.2734,.2764,.2794,.2823,.2852
DATA .2881,.291,.2939,.2967,.2996,.3023,.3051,.3078,.3106,.3133
DATA .3159,.3180,.3212,.3238,.3264,.3289,.3315,.334,.3365,.3389
DATA .3413,.3438,.3461,.3485,.3508,.3531,.3554,.3577,.3599,.3621
DATA .3643,.3665,.3686,.3708,.3729,.3749,.3770,.379,.381,.383
DATA .3849,.3869,.3888,.3907,.3925,.3944,.3962,.398,.3997,.4015
DATA .4032,.4049,.4066,.4082,.4099,.4115,.4131,.4147,.4162,.4177
DATA .4192,.4207,.4222,.4236,.4251,.4265,.4279,.4292,.4306,.4319
DATA .4332,.4345,.4357,.437,.4382,.4394,.4406,.4418,.4429,.4441
DATA .4452,.4463,.4474,.4484,.4495,.4505,.4515,.4525,.4535,.4545
DATA .4554,.4564,.4573,.4582,.4591,.4599,.4608,.4616,.4625,.4633
DATA .4641,.4649,.4656,.4664,.4671,.4678,.4686,.4693,.4699,.4706
DATA .4713,.4719,.4726,.4732,.4738,.4744,.475,.4756,.4761,.4767
DATA .4772,.4778,.4783,.4788,.4793,.4798,.4803,.4808,.4812,.4817
DATA .4821,.4826,.483,.4834,.4838,.4842,.4846,.485,.4854,.4857
DATA .4861,.4864,.4868,.4871,.4875,.4878,.4881,.4884,.4887,.489
DATA .4893,.4896,.4898,.4901,.4904,.4906,.4909,.4911,.4913,.4916
DATA .4918,.492,.4922,.4925,.4927,.4929,.4931,.4932,.4934,.4936
DATA .4938,.494,.4941,.4943,.4945,.4946,.4948,.4949,.4951,.4952
DATA .4953,.4955,.4956,.4957,.4959,.496,.4961,.4962,.4963,.4964
DATA .4965,.4966,.4967,.4968,.4969,.4970,.4971,.4972,.4973,.4974
DATA .4975,.4976,.4977,.4978,.4979,.498,.4981,.4982,.4983,.4984
DATA .4985,.4986,.49865,.4988,.4989,.499,.49903,.4991,.4992,.4993
DATA .49931, .4994, 4995,.4996,.4997,.4998,.49984,.49993,.5

'z score

DATA .01, .02, .03, .04, .05, .06, .07, .08, .09
DATA .1, .11, .12, .13, .14, .15, .16, .17, .18, .19
DATA .2, .21, .22, .23, .24, .25, .26, .27, .28, .29
DATA .3, .31, .32, .33, .34, .35, .36, .37, .38, .39
DATA .4, .41, .42, .43, .44, .45, .46, .47, .48, .49
DATA .5, .51, .52, .53, .54, .55, .56, .57, .58, .59
DATA .6, .61, .62, .63, .64, .65, .66, .67, .68, .69
DATA .7, .71, .72, .73, .74, .75, .76, .77, .78, .79
DATA .8, .81, .82, .83, .84, .85, .86, .87, .88, .89
DATA .9, .91, .92, .93, .94, .95, .96, .97, .98, .99
DATA 1, 1.01, 1.02, 1.03, 1.04, 1.05, 1.06, 1.07, 1.08, 1.09
DATA 1.1, 1.11, 1.12, 1.13, 1.14, 1.15, 1.16, 1.17, 1.18, 1.19
DATA 1.2, 1.21, 1.22, 1.23, 1.24, 1.25, 1.26, 1.27, 1.28, 1.29
DATA 1.3, 1.31, 1.32, 1.33, 1.34, 1.35, 1.36, 1.37, 1.38, 1.39
DATA 1.4, 1.41, 1.42, 1.43, 1.44, 1.45, 1.46, 1.47, 1.48, 1.49
DATA 1.5, 1.51, 1.52, 1.53, 1.54, 1.55, 1.56, 1.57, 1.58, 1.59
DATA 1.6, 1.61, 1.62, 1.63, 1.64, 1.65, 1.66, 1.67, 1.68, 1.69
```

```
DATA 1.7, 1.71, 1.72, 1.73, 1.74, 1.75, 1.76, 1.77, 1.78, 1.79
DATA 1.8, 1.81, 1.82, 1.83, 1.84, 1.85, 1.86, 1.87, 1.88, 1.89
DATA 1.9, 1.91, 1.92, 1.93, 1.94, 1.95, 1.96, 1.97, 1.98, 1.99
DATA 2, 2.01, 2.02, 2.03, 2.04, 2.05, 2.06, 2.07, 2.08, 2.09
DATA 2.1, 2.11, 2.12, 2.13, 2.14, 2.15, 2.16, 2.17, 2.18, 2.19
DATA 2.2, 2.21, 2.22, 2.23, 2.24, 2.25, 2.26, 2.27, 2.28, 2.29
DATA 2.3, 2.31, 2.32, 2.33, 2.34, 2.35, 2.36, 2.37, 2.38, 2.39
DATA 2.4, 2.41, 2.42, 2.43, 2.44, 2.45, 2.46, 2.47, 2.48, 2.49
DATA 2.5, 2.51, 2.52, 2.53, 2.54, 2.55, 2.56, 2.57, 2.58, 2.59
DATA 2.6, 2.61, 2.62, 2.63, 2.64, 2.65, 2.66, 2.67, 2.68, 2.69
DATA 2.7, 2.71, 2.72, 2.73, 2.74, 2.75, 2.76, 2.77, 2.78, 2.79
DATA 2.8, 2.82, 2.83, 2.85, 2.86, 2.88, 2.89, 2.91, 2.93, 2.94
DATA 2.96, 2.98, 3, 3.03, 3.05, 3.08, 3.1, 3.11, 3.14, 3.18
DATA 3.2, 3.22, 3.3, 3.33, 3.4, 3.5, 3.6, 3.8, 4
```

```
' APPENDIX 3

'filename marbus5.bas
'Marascuilo & Busk for s subjects
'random data permutation

PRINT "How many baselines?": INPUT s
PRINT "How many data points per row?": INPUT number
PRINT "How many random samples?": INPUT nsample
PRINT "How many possible intervention points?": INPUT ip
PRINT "Minimum data points per phase?": INPUT q: min = q + 1

CLS : nge = 0: nse = 0: effect = 0
DIM array(s, number): RANDOMIZE TIMER

FOR i = 1 TO s
 sb(i) = 0: nb(i) = 0: st(i) = 0: nt(i) = 0
NEXT i

FOR i = 1 TO s
 FOR j = 1 TO number
  READ array(i, j)
 NEXT j
NEXT i

' input actual intervention points
FOR i = 1 TO s
 PRINT "Intervention point "; i; "?": INPUT v(i)
NEXT i

'compute obtained effect

FOR i = 1 TO s
 FOR j = 1 TO number
  x = array(i, j)
  IF j < v(i) THEN sb(i) = sb(i) + x: nb(i) = nb(i) + 1
  IF j >= v(i) THEN st(i) = st(i) + x: nt(i) = nt(i) + 1
 NEXT j
NEXT i

FOR i = 1 TO s
 bmean(i) = sb(i) / nb(i): tmean(i) = st(i) / nt(i)
 d(i) = bmean(i) - tmean(i): effect = effect + d(i)
NEXT i

PRINT "effect = "; effect

FOR i = 1 TO s
 sb(i) = 0: st(i) = 0: nb(i) = 0: nt(i) = 0
NEXT i

nge = 1
```

```
'randomly permute data (nsample - 1) times

FOR enn = 1 TO (nsample - 1)

 RANDOMIZE TIMER: diff = 0

 FOR i = 1 TO s
  v(i) = (INT(RND * ip) + min)
 NEXT i

FOR i = 1 TO s
 FOR j = 1 TO number
  x = array(i, j)
  IF j < v(i) THEN sb(i) = sb(i) + x: nb(i) = nb(i) + 1
  IF j >= v(i) THEN st(i) = st(i) + x: nt(i) = nt(i) + 1
 NEXT j
NEXT i

FOR i = 1 TO s
 bmean(i) = sb(i) / nb(i): tmean(i) = st(i) / nt(i)
 d(i) = bmean(i) - tmean(i): diff = diff + d(i)
NEXT i

IF diff >= effect THEN nge = nge + 1 ELSE nse = nse + 1

FOR i = 1 TO s
 sb(i) = 0: st(i) = 0: nb(i) = 0: nt(i) = 0
 NEXT i

 PRINT enn

NEXT enn

'print probability

PRINT "nge = "; nge; " nse = "; nse; " p = "; nge / nsample

DATA 9,9,9,9,9,0,0,0,1,0,5,5,5,6,0,0,0,0,0

DATA 9,9,9,9,9,5,8,5,0,6,0,0,0,0,0,0,0,0,0

DATA 9,9,9,9,9,9,9,9,6,8,8,5,1,3,1,0,0,0,0

DATA 9,9,9,9,9,9,9,9,9,9,8,8,5,9,8,8,0,1,0

DATA 9,9,9,9,9,9,9,9,9,9,0,0,0,0,0,0,0,0,0

DATA 9,9,9,9,9,9,9,9,9,9,9,9,9,5,1,0,0,1,1

DATA 9,9,8,9,8,8,8,8,8,8,8,9,8,8,8,4,4,5,4

DATA 9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9
```

```
'APPENDIX 4

'MARSIM3.BAS
'simulates Marascuilo & Busk for up to 12 baselines
'minimum of 5 baseline points, 5 treatment points

CLS : DIM norm(310), zedd(310)
DIM sb(16), st(16), nb(16), nt(16)
nge = 0: nse = 0: effect = 0
s05 = 0: ns05 = 0: s01 = 0: ns01 = 0

'read normal distribution
FOR i = 1 TO 308: READ norm(i): NEXT i
FOR i = 1 TO 308: READ zedd(i): NEXT i

m = 100: s = 5: size = .8
DIM array(12, 40): DIM v(16)

FOR baselines = 2 TO 8


 RANDOMIZE TIMER
 FOR i = 1 TO baselines
  sb(i) = 0: nb(i) = 0: st(i) = 0: nt(i) = 0
 NEXT i

 FOR intpoints = 6 TO 2 STEP 2

number = (5 + intpoints + 4): last = (number - 4)

FOR iteration = 1 TO 100: effect = 0

 'generate random intervention point for all baselines

 FOR i = 1 TO baselines
  v(i) = (INT(RND * (last - 5) + 1) + 5)
 NEXT i

 'generate random data

 mb = m: mt = m - (size * s)
 FOR i = 1 TO baselines
  FOR j = 1 TO number
   IF j < v(i) THEN av = mb ELSE av = mt
   x = RND: IF x > .5 THEN x = 1 - x
   FOR enn = 1 TO 308
    IF norm(enn) >= x THEN z = zedd(enn): EXIT FOR
   NEXT enn
   sign = RND
   sc = INT(z * s)
   IF sign > .5 THEN y = av + sc ELSE y = av - sc
   array(i, j) = y
  NEXT j
NEXT i
```

```
'compute effect = basemean - treatmean

FOR i = 1 TO baselines
 FOR j = 1 TO number
  x = array(i, j)
  IF j < v(i) THEN sb(i) = sb(i) + x: nb(i) = nb(i) + 1
  IF j >= v(i) THEN st(i) = st(i) + x: nt(i) = nt(i) + 1
 NEXT j
NEXT i

FOR i = 1 TO baselines
 bmean(i) = sb(i) / nb(i): tmean(i) = st(i) / nt(i)
 d(i) = bmean(i) - tmean(i): effect = effect + d(i)
NEXT i

FOR i = 1 TO baselines
 sb(i) = 0: nb(i) = 0: st(i) = 0: nt(i) = 0
NEXT i

nge = 1

FOR perm = 1 TO 100
 diff = 0

 'generate new random intervention points

 FOR i = 1 TO baselines
  v(i) = (INT(RND * (last - 5) + 1) + 5)
 NEXT i

 'compute diff for each data permutation

 FOR i = 1 TO baselines
  FOR j = 1 TO number
   x = array(i, j)
   IF j < v(i) THEN sb(i) = sb(i) + x: nb(i) = nb(i) + 1
   IF j >= v(i) THEN st(i) = st(i) + x: nt(i) = nt(i) + 1
  NEXT j
 NEXT i

 FOR i = 1 TO baselines
  bmean(i) = sb(i) / nb(i): tmean(i) = st(i) / nt(i)
  d(i) = bmean(i) - tmean(i): diff = diff + d(i)
 NEXT i

 FOR i = 1 TO baselines
  sb(i) = 0: st(i) = 0: nb(i) = 0: nt(i) = 0
 NEXT i

 IF diff >= effect THEN nge = nge + 1 ELSE nse = nse + 1

NEXT perm
```

```
 'print results

 p = nge / (nge + nse)
 PRINT "nge = "; nge; " nse = "; nse
 PRINT "p = "; p
 nge = 0: nse = 0
 IF p < .05 THEN s05 = s05 + 1 ELSE ns05 = ns05 + 1
 IF p < .01 THEN s01 = s01 + 1 ELSE ns01 = ns01 + 1
 PRINT "size = "; size; " iteration = "; iteration
 PRINT

NEXT iteration
   LPRINT "marsim3.bas effect size = "; size
   LPRINT "baselines = "; baselines; "intpoints = "; intpoints
   LPRINT "power05 = "; s05 / (s05 + ns05)
   LPRINT "power01 = "; s01 / (s01 + ns01)
   s05 = 0: ns05 = 0: s01 = 0: ns01 = 0
   LPRINT

 NEXT intpoints

NEXT baselines

'normal curve DATA are READ as in Appendix 2
```

```
'APPENDIX 5


'normdis2.bas
'Marascuilo & Busk normal distribution method
'for up to 10 "baselines"

CLS : DIM a(40): DIM norm(310): DIM zedd(310): DIM df(40)
DIM d(10): DIM ed(10): DIM vard(10): DIM statz(10)
DIM t(10): DIM et(10): DIM vart(10)
sb = 0: st = 0: nb = 0: nt = 0

FOR i = 1 TO 308
 READ norm(i)
NEXT i

FOR i = 1 TO 308
 READ zedd(i)
NEXT i

'm = mean, s = standard deviation, size = effect size
m = 100: s = 5

FOR intpoints = 6 TO 20 STEP 2

 RANDOMIZE TIMER
 number = (5 + intpoints + 4)
first = 6: last = (number - 4)

FOR size = .2 TO 2 STEP .3

 FOR i = 1 TO 10
  s05(i) = 0: s01(i) = 0: s001(i) = 0
 NEXT i

 FOR iteration = 1 TO 1000

  PRINT "size = "; size; " iteration = "; iteration
  PRINT "intpoints = "; intpoints
  PRINT
  d = 0: ed = 0: vard = 0

  FOR baselines = 1 TO 10

'compute random intervention point

inta = (INT(RND * (last - 5) + 1) + 5)

'generate simulated data

ma = m: mb = m - (s * size)

FOR j = 1 TO number
 IF j < inta THEN av = ma ELSE av = mb
 x = RND: IF x > .5 THEN x = 1 - x
```

```
 FOR enn = 1 TO 308
  IF norm(enn) >= x THEN z = zedd(enn): EXIT FOR
 NEXT enn
 sign = RND
 sc = INT(z * s)
 IF sign >= .5 THEN y = av + sc ELSE y = av - sc
 a(j) = y
NEXT j

'compute effect = basemean - treatmean

sb = 0: st = 0: nb = 0: nt = 0
FOR j = 1 TO number
 IF j < inta THEN sb = sb + a(j): nb = nb + 1
 IF j >= inta THEN st = st + a(j): nt = nt + 1
NEXT j
bmean = sb / nb: tmean = st / nt
effect = bmean - tmean
d(baselines) = effect
sb = 0: st = 0: nb = 0: nt = 0: bmean = 0: tmean = 0

'compute diff for all data permutations

FOR e = first TO last
 FOR j = 1 TO number
  IF j < e THEN sb = sb + a(j): nb = nb + 1
  IF j >= e THEN st = st + a(j): nt = nt + 1
 NEXT j
 bmean = sb / nb: tmean = st / nt
 diff = bmean - tmean: df(e) = diff
 sb = 0: st = 0: nb = 0: nt = 0
NEXT e

'compute mean of differences

x = 0
FOR e = first TO last
 x = x + df(e)
NEXT e
ed(baselines) = (x / intpoints)

'compute variance of differences

 g = 0
 FOR e = first TO last
  w = df(e) - (x / intpoints): r = w ^ 2
  g = g + r
 NEXT e
 h = intpoints - 1
 x = (g / h)
 vard(baselines) = x

NEXT baselines

t = d(1): et = ed(1): vart = vard(1)
```

```
FOR i = 2 TO 10
 t = t + d(i)
 et = et + ed(i)
 vart = vart + vard(i)
 statz(i) = (t - et) / (SQR(vart))
NEXT i

FOR i = 1 TO 10
  IF statz(i) > 1.645 THEN s05(i) = s05(i) + 1
  IF statz(i) > 2.325 THEN s01(i) = s01(i) + 1
  IF statz(i) > 3.08 THEN s001(i) = s001(i) + 1
 NEXT i

 NEXT iteration

 FOR i = 1 TO 10
  LPRINT "intpoints = "; intpoints; "size = "; size
  LPRINT "baselines = "; i
  LPRINT "power at p < .05 = "; s05(i)
  LPRINT "power at p < .01 = "; s01(i)
  LPRINT "power at p < .001 = "; s001(i)
  LPRINT
 NEXT i

NEXT size

 NEXT intpoints

'normal curve DATA are READ as in Appendix 2
```

```
'Appendix 6

'filename marbus3.bas
'Marascuilo & Busk for 4 subjects
'systematic data permutation
'this program contains DATA from Wampold & Worsham (1986)

CLS : nge = 0: nse = 0: effect = 0: s = 4: DIM v(10)

PRINT "How many data points per row?": INPUT number
PRINT "First possible intervention point?": INPUT first
PRINT "Last possible intervention point": INPUT last

FOR i = 1 TO s
 PRINT "Actual intervention point "; i: INPUT v(i)
NEXT i

DIM array(s, number)

FOR i = 1 TO s
 FOR j = 1 TO number
  READ array(i, j)
 NEXT j
NEXT i

FOR i = 1 TO s
 sb(i) = 0: nb(i) = 0: st(i) = 0: nt(i) = 0
NEXT i

'compute effect for actual intervention points

FOR i = 1 TO s
 FOR j = 1 TO number
  x = array(i, j)
  IF j < v(i) THEN sb(i) = sb(i) + x: nb(i) = nb(i) + 1
  IF j >= v(i) THEN st(i) = st(i) + x: nt(i) = nt(i) + 1
 NEXT j
NEXT i

FOR i = 1 TO s
 bmean(i) = sb(i) / nb(i): tmean(i) = st(i) / nt(i)
 d(i) = bmean(i) - tmean(i): effect = effect + d(i)
NEXT i

PRINT "effect = "; effect

FOR i = 1 TO s
 sb(i) = 0: st(i) = 0: nb(i) = 0: nt(i) = 0
NEXT i

'systematically permute the data
```

```
FOR a = first TO last
 FOR b = first TO last
  FOR c = first TO last
   FOR d = first TO last
v(1) = a: v(2) = b: v(3) = c: v(4) = d: diff = 0

FOR i = 1 TO s
 FOR j = 1 TO number
  x = array(i, j)
  IF j < v(i) THEN sb(i) = sb(i) + x: nb(i) = nb(i) + 1
  IF j >= v(i) THEN st(i) = st(i) + x: nt(i) = nt(i) + 1
 NEXT j
NEXT i

FOR i = 1 TO s
 bmean(i) = sb(i) / nb(i): tmean(i) = st(i) / nt(i)
 d(i) = bmean(i) - tmean(i): diff = diff + d(i)
NEXT i

IF diff >= effect THEN nge = nge + 1 ELSE nse = nse + 1

    FOR i = 1 TO s
     sb(i) = 0: nb(i) = 0: st(i) = 0: nt(i) = 0
    NEXT i

   NEXT d
  NEXT c
 NEXT b
NEXT a

PRINT "nge = "; nge
PRINT "nse= "; nse
PRINT "Probability = nge/(nge+nse) = "; nge / (nge + nse)


DATA 8,7,6,7,4,5,6,5,4,4,5,2,4,3,4,5,4,3,2,2
DATA 6,7,8,7,5,7,6,8,6,5,4,4,4,3,2,5,3,4,3,6
DATA 5,5,4,6,4,5,6,7,4,5,6,5,2,3,2,4,1,0,2,3
DATA 8,6,7,7,8,5,7,8,7,6,7,8,5,6,8,8,6,4,4,5
```