ED 361 347 TM 020 397

AUTHOR Ferron, John

TITLE Suggested Solutions to Problems Facing the Use of Randomization Tests with Single-Case Designs.

PUB DATE Apr 93

NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Case Studies; *Data Analysis; *Research Design; Research Methodology; Test Construction; *Test Use

IDENTIFIERS Power (Statistics); *Randomization; Randomized Experiments; *Responsive Design; Single Subject Designs

ABSTRACT

Randomization tests have been suggested as a method for analyzing the data from single-case designs. This paper explores three concerns which have arisen: the nonresponsive nature of randomization tests, the appropriateness of specific test statistics and the power of these tests. An example is given to illustrate how partial control can be given to the researcher so that a design can be responsive and incorporate random assignment. A general test statistic is given that can be used when the researcher is unsure of the specific nature of the treatment effect and thus uncomfortable with the specific test statistics typically employed. Finally a method of combining types of randomization is given to provide a way of increasing the number of assignments and the power of the randomization test. (Author)

Manuscript for Poster Presentation

AERA April, 1993. Atlanta, GA.

Suggested Solutions to Problems Facing the Use of Randomization Tests with Single-Case Designs

John Ferron

Abstract

Randomization tests have been suggested as a method for analyzing the data from single-case designs. This paper explores three concerns which have arisen: the nonresponsive nature of randomization tests, the appropriateness of specific test statistics and the power of these tests. An example is given to illustrate how partial control can be given to the researcher so that a design can be responsive and incorporate random assignment. A general test statistic is given which can be used when the researcher is unsure of the nature of the treatment effect and thus uncomfortable with the specific test statistics typically employed. Finally a method of combining types of randomization is given to provide a way of increasing the number of assignments and the power of the randomization test.

Suggested Solutions to Problems Facing the Use of Randomization Tests with Single-Case Designs

The analysis of the data resulting from single-case designs presents a problematic issue which has been debated over the last two decades. Randomization tests are one method of analysis which has been proposed and discussed. The statistical validity of using randomization tests for single-case designs has been argued convincingly by Edgington (1980b). Other concerns, however, have arisen. These concerns include: (a) the nonresponsive nature of randomization tests, (b) the appropriateness of specific test statistics and (c) the power associated with randomization tests. After a brief review of the logic of randomization tests, each of these issues will be discussed. The relevant suggestions from the literature will be noted and new suggestions aimed at overcoming these obstacles will be provided.

A researcher conducting a randomization test will go through the following procedure. The study is designed in a way which incorporates random assignment. A test statistic is chosen based on the nature of the anticipated effect. The study is then conducted, and the test statistic is calculated based on the obtained assignment. A randomization distribution is formed by calculating the test statistic for each assignment which logically could have resulted from the randomization scheme. Statistical significance is determined by finding where the obtained test statistic falls within the randomization distribution.

There are two general types of randomization which underlie the numerous examples and models which have been discussed in the literature. One type of randomization involves the random assignment of treatments to times, while the other involves the random assignment of intervention points within the sequence. The assignment of treatments to times can be illustrated by considering a researcher who wishes to determine if one treatment, say A, is more beneficial than another treatment, say B. The researcher could set up an experiment where treatment A was randomly assigned to 4 of 8 points in time. Treatment B would be assigned to the other 4 times. For this experiment, there are 135 (8!/4!4!) possible treatment sequences that could result. The probability of the obtained test statistic being the largest of the 135 possibilities is .007 (1/135). This logic has been demonstrated in Fisher's classic example of the lady tasting tea (1951), as well as in other examples given by Edgington (1967, 1980a, 1980b, 1987, 1992).

By using the same logic, but restricting the random assignment to a subset of the time points other variations can be obtained. Edgington has considered the random assignment of treatments to successive pairs of times (1980a, 1980b) and the random assignment of treatments within halves of the data (1987).

Levin, Marascuilo and Hubert (1978) described a design with the restriction that treatments are randomly assigned to blocks of time. The researcher envisioning the common ABAB type design has two treatments and four blocks of time to incorporate into the randomization scheme. Treatment A could be randomly assigned to two of the four blocks of time, resulting in six possible sequences and a minimum p-value of .17. The majority of the possible assignments, however, lead to sequences which are not consistent with logic of the ABAB design (Onghena, 1992).

The random assignment of treatments to blocks of time for multiple baseline AB designs has been considered by Wampold and Worsham (1986). With 5 subjects, there are 32 (2^5) possible assignments corresponding to a minimum p-value of.03. Marascuilo and Busk (1988) extended this logic to multiple baseline ABAB designs. With 5 subjects this logic would result in 7776 (6^5) possible assignments and a minimum p-value of .0001.

The second way of accomplishing randomization within a single-case design is to randomly assign the point at which an intervention is made. For example, in a series of 20 planned measurements, the intervention could be randomly chosen to start with any of the 6th through 15th measurements. This would result in 10 possible assignments and a minimum p-value of .10. Examples involving this type of randomization have been given by Edgington (1975, 1980a, 1980b & 1987) and Wampold and Furlong (1981). This logic can also be extended for use with multiple baseline AB designs (Marascuilo and Busk, 1988). A design incorporating 10 possible intervention points per subject and three subjects, will have 1000 (10^3) possible assignments.

Random assignment of intervention points has also been suggested and illustrated for ABA designs by Edgington (1975, 1980a, 1980b & 1987). In this design the researcher randomly selects a pair of points, one for intervention and one for withdraw. Onghena (1992) further generalized the random assignment of intervention points to designs involving any number of phases. In an ABAB design, for example, the researcher randomly selects a triplet of time points from the set of possible triplets. The set of possible triplets can be constrained to ensure a minimum number of points per phase.

The Nonresponsive Nature of Randomization Tests

The nonresponsive nature of randomization tests has been noted as an obstacle by Kazdin (1980), Matyas & Greenwood (1991) and Onghena (1992). Kazdin (1980) pointed out that the design of a single-case experiment should depend on the pattern of data emerging within a particular phase. This implies the researcher should choose the times of intervention as the experiment unfolds, not based on some initial randomization scheme. The logic behind this argument is illustrated by Parsonson and Baer (1986) in their discussion of a fined grained graphic analysis of a hypothetical experiment. At several points they noted the need to extend a phase to ensure that an apparent trend is truly stable.

As pointed out previously, the valid use of a randomization test requires some sort of randomization in the design; therefore, one can not validly turn complete control over to the researcher. Edgington (1980c), however, gave an example of a randomization scheme which gives partial control to the researcher. In this example, a random assignment of treatments to times is carried out under the constraint that if the data reaches a predetermined critical level then the other treatment must be given. Although this example illustrates partial control it does not address the question of within phase stability.

The following discussion will demonstrate another way of turning partial control over to the researcher. This method will allow for the extension of phases in an ABAB type experiment based on the pattern of data which emerges during the experiment. First, a nonresponsive example will be given to illustrate an alternative method for randomly assigning intervention points in an ABAB design. This method will then be modified to take into account responsive designs.

Imagine a study where the researcher wishes to carry out an ABAB design. The researcher randomly assigns three points of change within confined regions in a sequence of 32 observations. The first point of change is randomly assigned so the B phase starts with either the 6th, 7th, 8th, 9th or 10th observation. This defines a 5 point randomization region which falls between observation 6 and 10. The second point of change is randomly assigned to fall in the region from 15 to 19, and the third point of change in the region from 24 to 28. In this design there is a minimum of 5 points in any of the phases and each of the assignments is made independently. The number of possible assignments for this ABAB design would be 125 (5x5x5), and the smallest possible p-value would be .008.

Let us further assume the researcher wishes to restrict the design such that points of change only occur after the within phase data shows an "acceptable" pattern. The researcher would initially decide on the length of the three intervention regions, say 5, and an objective criteria for pattern acceptability, say a minimum of five points in which the last three points do not form a monotonic trend. Figure 1 shows a sequence of observations. Although this sequence is not known to the researcher prior to the study, it is effectively fixed, if the null hypothesis of no effect is true.

The study begins and the observations in Figure 1 begin to emerge. The fifth point is the third in a monotonically increasing trend, therefore, the pattern of the first phase is not deemed acceptable. The six point breaks this pattern, making the phase acceptable. The first randomization region is then assigned to fall from observation 7 to 11. The point of change is randomly assigned within this region, but the actual value need not concern us now. Assuming the treatment has no effect the points will continue to emerge as depicted in Figure 1.

We know the treatment has started by observation 11, so this is the first observation which must fall in phase B. Beginning with this observation, we again wait for an acceptable within phase pattern. After observation 15 we have met the requirement of a minimum of 5 observations without the final three forming a monotonic trend. The second intervention region is then assigned to fall from observation 16 to 20 and a random assignment is made.

Again we need not worry about the actual value of the random assignment.

Knowing that observation 20 must fall in the second A phase, we can again begin evaluating within phase acceptability. A unacceptable trend is present after point 24 and continues through point 26. The pattern becomes acceptable after observation 27. The third randomization region is assigned t‿ fall from 28 to 32, and the final point of change is randomly assigned within this region. Knowing the final phase must have started by observation 32, we can again start judging acceptability. Observation 36 gives us the final acceptable phase and ends the study.

The establishment of the total study length and the placement of the intervention regions is independent of the values of the actual random assignments as long as the null hypothesis is true and phase acceptability is based on the points which must fall in the given phase. In the study described above, there is 125 (5X5X5) possible random assignments. The randomization distribution can be formed from the test statistics of the 125 possible divisions of the data, and significance can be established by finding where the obtained test statistic falls within this distribution.

The use of randomization regions, as illustrated above, allows the researcher to exercise some control to ensure stability, but uses random assignment so valid p-values can be determined. The number of intervention regions as well as the length of each region contribute to minimal size of the possible p-values.

The Appropriateness of Specific Test Statistics

Prior to gathering data, a test statistic is chosen based on the expected pattern of the intervention effect. Implicit in this process is the necessity of the researcher to choose a test statistic that will describe the intervention effect. The usefulness of the randomization test, therefore, is limited by the ability of the researcher to choose an appropriate test statistic (Matyas & Greenwood, 1991). Test statistics have been proposed for a wide variety of possible effects.

When an immediate, permanent change in level is expected, the difference between phase means, is a useful statistic (Wampold & Worsham 1986, Wampold & Furlong 1981, Marascuilo & Busk 1988, Edgington 1975, 1980a, 1980b, 1987). $\bar{X}_B - \bar{X}_A$, can be used with a one tail hypothesis, and $/\bar{X}_B - \bar{X}_A/$ can be used with a two tail hypothesis. The difference between phase means gives the same results as using a t-statistic, but has the advantage of being easier to calculate (Wampold & Worsham 1986). When an immediate, permanent change in level exists, but a trend exists throughout the data, the difference between phase means may be an insensitive test statistic (Edgington 1987). When this situation is anticipated, an ANCOVA F can be used where the sequence number of the data point can be used as a covariate (Wampold & Worsham 1986, Edgington 1980a, 1987).

Another immediate, permanent effect that may occur is a change in slope. In this case, the difference between slopes of the regression lines associated with each phase, $b_{1A} - b_{1B}$, may be a useful test statistic (Wampold & Worsham 1986, Wampold & Furlong 1981). A change in variance is another possible effect due to an intervention. This effect can be tested for by using the ratio of the two phase variances, σ_A^2/σ_B^2 , as the test statistic (Edgington 1975, 1980b).

The test statistics discussed thus far have been limited in use to situations where immediate and permanent effects are expected as a result of the intervention. A delayed or transient effect may also result. When a delayed effect is anticipated, one of the above test statistics may be adapted to account for the expected delay. This adaption involves the omission of a predetermined number of points from the beginning of phase 2 (Wampold & Worsham 1986). When a transient effect is expected, the difference between the first point in phase 2 and the last point in phase 1 has been suggested as a possible test statistic (Edgington, 1980b).

The above test statistics cover a variety of expected outcomes; however, others come to mind. For example a change in both level and slope may occur or a change in trend, say from liner to curvilinear. As different expected outcomes arise different test statistics can be derived. As Edgington stated (1980b, p.247) "The experimenter has the freedom to choose whatever conventional or unconventional test statistic he desires. Any statistical test, no matter how complex, can be employed validly, when significance is determined by the randomization test procedure."

The probability of committing a type I error is not affected by the choice of test statistic; therefore, the choice of the test statistic can be made solely on the sensitivity of the test statistic, its ability to detect differences when they exist. The test statistics discussed previously are sensitive to specific types of intervention effects. The researcher, therefore, must correctly anticipate the nature of the intervention effect to pick the most sensitive test statistic. No problem occurs as long as the intervention effect is easily anticipated; however, as the nature of the effect becomes less clear, the choice of the test statistic becomes more problematic. In short, the use of the above test statistics confine the researcher to test for specific changes. The need of the researcher to test a more general hypothesis about an intervention effect has been discussed by McCain and McCleary (1979) in their discussion of time series analysis. With this in mind it seem reasonable to pursue test statistics which are sensitive to multiple types of intervention effects One such test statistic is discussed below.

Regression lines could be fitted to each phase of the data sequence and an r^2 value could be calculated. The r^2 value indicates the proportion of the variance explained and how well the regression line fits the data. Given that a permanent change in trend coinciding with the intervention occurs, one would expect the sum of the r^2s to be a maximum when the data are split at the point of the intervention. Splitting the data at any of the other possible intervention points would lead to some of the data from one trend being grouped with the data from the other trend. This would tend to decrease the value of the r^2 in this "contaminated" phase, thus decreasing the sum of the r^2s. The sum of the r^2s appears to be a test statistic which may be sensitive to a variety of immediate, permanent effects; namely, changes in level, slope or changes in both. Extending this notion to nonlinear regression would enable the detection of changes when the trend may be nonlinear in either phase or both.

One may question the use of a statistic based on r^2 values when the data is autocorrelated and/or the number of points per phase is relatively small, say under 30. The logic of the randomization test ensures the validity of the p-values assuming the treatment had no effect. The power of the test, however, is suspect. The general question raised when a more general test statistic is considered, is whether the more general statistic is as sensitive as the appropriate specific test statistic for a given effect. Would we expect to pick up "true" differences in levels more often employing $\bar{X}_B - \bar{X}_A$ than employing the sum of the r^2s? If so, how much more often, what price is being paid for the luxury of using a more general test statistic? The answers to these specific questions await future research.

The Low Power of Randomization Tests

Several of the designs discussed above yield minimum p-values which are greater than the typical alpha of .05. When this situation arises there is no chance of statistically detecting true intervention effects regardless of their size. One occasion where this occurs is when the random assignment of treatments to blocks of time is made in an ABAB design. Recall this design has 6 possible sequences and a minimum p-value of .17. Marascuilo and Busk (1988) showed how the number of possible assignments could be increased by extending this logic to multiple baseline ABAB designs.

A second solution is evident from the work of Onghena (1992). Prior to this work the examples and discussions involving designs with more than one treatment phase, such as the ABAB design, had focused on randomly assigning treatments to phases. Onghena's method is based on randomly assigning intervention points within the sequence. This latter type of randomization has the potential of dramatically increasing the number of possible assignments.

The number of possible assignments could be further increased by combining types of randomization. The examples given above and in the literature to this point use one method of randomization to the exclusion of the other. Either treatments are randomly assigned to times, or intervention points are randomly assigned within the sequence. There exists, however, situations when either form of randomization may be reasonable and hence the possibility of combining the two types of randomization exists. For example, the experimenter may wish to incorporate 4 phases in the experiment, two with treatment A and two with treatment B. The researcher could randomly choose between the sequence ABAB and BABA, in addition to randomly assigning the triplet of intervention points. In this case, assuming a one-tail test, the random choice between treatment sequences doubles the size of the randomization distribution, and thus halves the minimum p-value.

The above discussion of power has primarily revolved around designing the experiment to increase the number of possible assignments. Although power tends to be related to the number of possible data division, it is also affected by a host of other factors (Edgington, 1987). It is clear that the randomization test has no power when the minimum possible p-value is greater than the chosen alpha; however, the power of randomization tests when the minimum p-value is less than the chosen alpha is unknown. Some single-case researchers would prefer relatively low power tests so that only large effects are detected (eg. Baer 1977, Parsonson and Baer 1986), while others maintain sensitivity is desirable in at least some situations (Gottman & Glass, 1978; Kazdin, 1976). In either case, knowledge of the power of randomization tests is important. Simulation research directed toward this issue would be helpful, since it would aid researchers in designing experiments to the desired degree of sensitivity.

REFERENCES

Baer, D. M. (1977). Perhaps it would be better not to know everything. Journal of Applied Behavioral Analysis, 10, 167-172.

Edgington, E. S. (1992). Nonparametric tests for single-case experiments. In T. R. Kratochwill, & J. R. Levin (Eds.), Single-case research design and analysis: New directions for psychology and education. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Edgington, E. S. (1987). Randomization Tests. New York: Marcel Dekker, Inc.

Edgington, E. S. (1980a). Random assignment and statistical tests for one-subject experiments. Behavioral Assessment, 2, 19-28.

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

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

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

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

Fisher, R. A. (1951). The design of experiments (6th ed.). London: Hafner.

Gottman, J. M., & Glass, G. V. (1978). Analysis of interrupted time-series experiments. In T. R. Kratochwill (Ed.), Single subject research: Strategies for evaluating change. New York: Academic Press.

Kazdin, A. E. (1980). Obstacles in using randomization tests in single-case experimentation. Journal of Educational Statistics, 5, 253-260.

Kazdin, A. E. (1976). Statistical analysis for single-case experimental designs. In M. Hersen & D. Barlow, Single-case experimental designs: Strategies for studying behavior change. New York: Pergamon Press.

Levin, J. R., Marascuilo, L. A., & Hubert, L. J. (1978). N = nonparametric randomization tests. In T. E. Kratochwill (Ed.), Single subject research: Strategies for evaluating change. New York: Academic Press.

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. (1991). Problems in the estimation of autocorrelation in brief time series and some implications for behavioral data. Behavioral Assessment. 13, 137-157.

McCain, L. J., & McCleary, R. (1979). The statistical analysis of the simple interrupted time-series quasi-experiment. In T. D. Cook, & D. T. Campbell, Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton Mifflin Company.

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

Parsonson, B. S., & Baer, D. M. (1986). The graphic analysis of data. In A. Poling, & W. R. Fuqua (Eds.), Research methods in applied behavior analysis: Issues and advances. New York: Plenum Press.

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

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

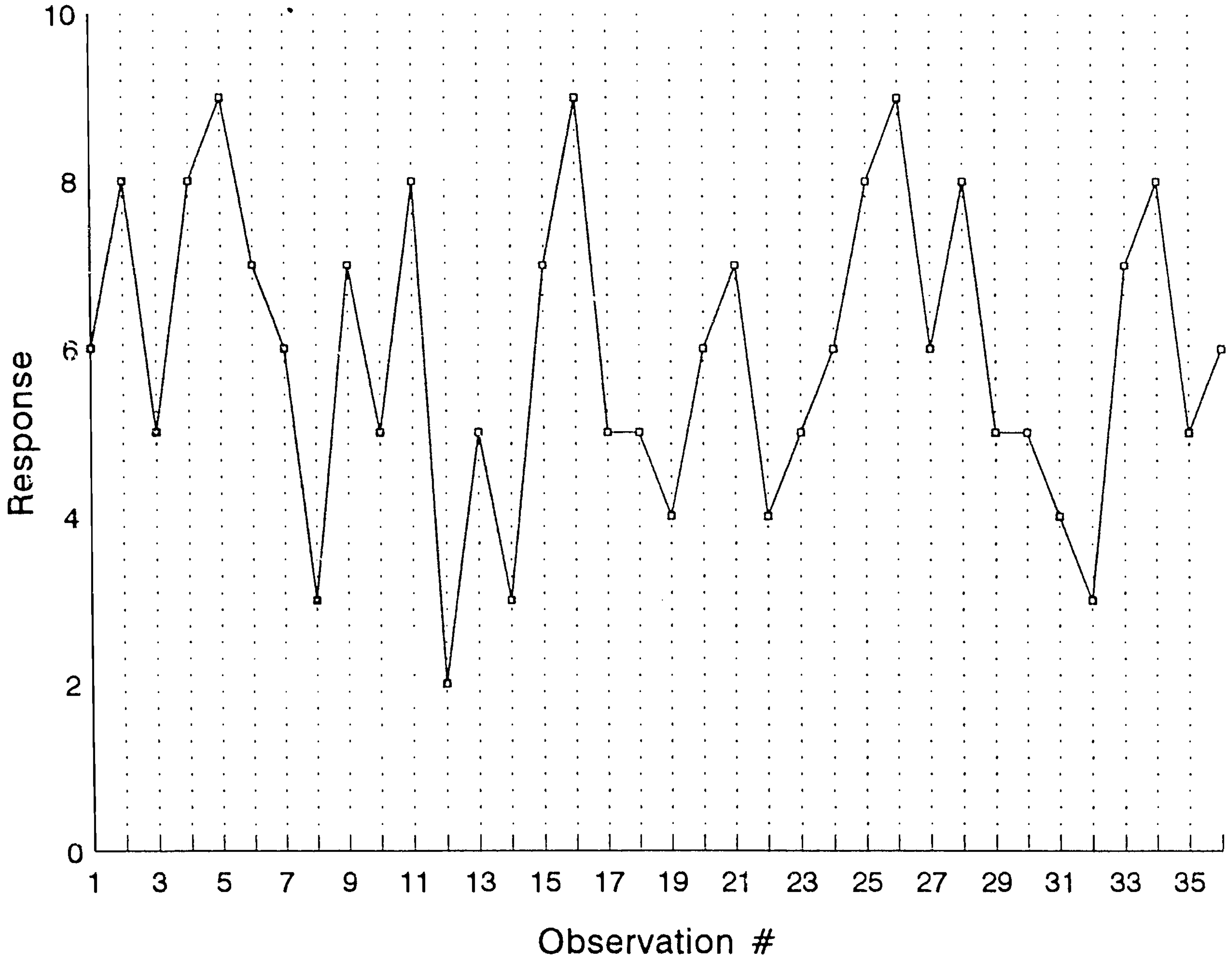

Figure 1: Graphic Representation of Example Data for a Responsive ABAB design