

# Spinning Off The Multiple Baseline Design In Its Clinical Application

Jan Bisset

G. Albyn Davis

Memphis State University  
Memphis, Tennessee

Throughout the past year and a half at the Memphis Speech and Hearing Center we have attempted to refine a research method for addressing fundamental questions in our clinical work. The title of our presentation, "Spinning Off the Multiple Baseline Design in its Clinical Application" implies that the multiple baseline has been our starting point, but that we're developing a modified version of the design. The uniqueness of our work with aphasic individuals presents a dilemma when trying to use a basic research design appropriate for other fields of behavior analysis. We want to address this dilemma by discussing several of the assumptions underlying the multiple baseline design. Then we will present three cases to illustrate our research spinning off the design. The practical problems we've encountered have been numerous. We hope that by sharing the growth of our work, solutions to some problems and the current challenges we're faced with we can encourage you to use such a design in your clinical research and at the same time ward off some frustration for you.

Let us briefly review the multiple baseline design. In this design a number of responses are identified and measured over time to provide baselines against which changes can be evaluated (Baer, Wolf & Risley, 1968). Baseline and subsequent treatment intervention for each targeted behavior can be conceptualized as separate A-B designs with the A phase further extended for each of the succeeding behaviors until the treatment variable is finally applied (Hersen & Barlow, 1976). It is predicted that the behavior treated will be changed, while the untreated behavior will remain unchanged until treated in turn (Leitenberg, 1973). What we hope to establish by employing such a research design is a systematic framework for addressing such questions as: (1) Is aphasia therapy effective in promoting behavior change? and (2) What is the nature of response generalization across language modalities?

## Basic Assumptions of the Multiple Baseline Design

The multiple-baseline design has had its widest application within an operant model of behavior analysis. Some assumptions met by an operant condition paradigm do not directly apply to language therapy. For example, Leitenberg (1973) notes that one underlying assumption of the multiple-baseline design is that the targeted behaviors under evaluation are relatively independent from one another. The relative independence of language behaviors treated in aphasia therapy is not well known. Aphasia assumes by definition that language behavior are not completely independent. Hersen & Barlow (1976) are talking to us when they state that experimenters involved in new areas of study where no precedents apply are placed in an interesting position; assumptions of independence of behavior cannot be made without empirical tests. So in fact, what this implies is that the work we're engaged in when employing this research design is preliminary. We feel that we can use a modified multiple baseline design to help us answer the question of independence among language modalities and thereby

ultimately determine the appropriateness of the multiple baseline design or one similar to it for aphasia research.

Another major reason why we've had to consider our work a spin-off of the multiple baseline design is that we're in clear violation of one other basic principle. That principle assumes that the same treatment procedure is applied to each behavior treated. In our clinical work, however, the treatment procedure differs from modality to modality. Again, we can see how this principle is met within an operant model of behavior analysis, but not in aphasia therapy.

#### Selection of Behaviors for Which Treatment Was Delayed

We have chosen three cases to illustrate the growth of our work with the modified multiple baseline design. In each case the treated behaviors from the onset of therapy were auditory comprehension and expression. Therapy in these modalities was maintained throughout the duration of the study. The problem was in selecting the behavior for which treatment was delayed. Our experience suggests that reading and writing are the two modalities most suitable for analysis within the design for two reasons: 1) Delaying treatment of reading and writing does not interrupt the normal course of therapy as we most typically delay treatment on these modalities in order to concentrate our time on auditory comprehension and expression. 2) We feel that reading and writing most nearly meet the inherent assumption of relative independence of behaviors treated.

#### A Case in Which Reading Was the Behavior For Which Therapy Was Delayed

In our first case, the behavior for which treatment was delayed was reading (Figure 1). This example, taken from our initial research design work, dates back to Spring of 1977, at which time we looked at five cases. This case illustrates the basic procedure used in our modification of the multiple baseline design. A series of three observations constituted the initial baseline for verbal repetition, the behavior selected to represent the treated language modalities. Four observations were initially taken on reading comprehension, the behavior for which treatment was delayed. Note that in an operant model of behavior analysis the initial baseline reflects the frequency of the natural occurrence of a behavior, whereas in the work presented here, the initial baseline reflects pretreatment performance on an imposed task. For example, the task for measuring comprehension was matching two-word phrases to pictures.

Treatment of verbal repetition was begun following baseline observations, while treatment of reading comprehension was delayed for five weeks. Baseline observations were continued for reading comprehension while weekly performance measures were taken of verbal repetition. Pretreatment performance level of reading comprehension ranged between 10 and 40% correct. Coinciding with the onset of therapy on 3-17, the weekly measure of the target behavior, reading, reflects an increase in performance throughout the following three weeks of the study.

In this initial work with the research design, verbal repetition of two-word phrases was selected as a criterion measure of treated behaviors. We had selected verbal repetition as a criterion measure because we were working on it in therapy. However, we soon became dissatisfied with choosing such a behavior as a criterion measure because we really wanted to see if our therapy efforts were serving to improve some other more functional behavior. Therefore, we concluded that the criterion measure of the treated behavior should be something other than that which was being treated in therapy. In addition, we

decided that a weekly measure of auditory comprehension was needed, along with a weekly measure of expression. We felt that a more comprehensive picture of behavior change could be provided if performance in both expression and auditory comprehension was measured. We felt that in order to address the question of generalization of treatment effects across modalities both of the treated behavior modalities needed to be measured.

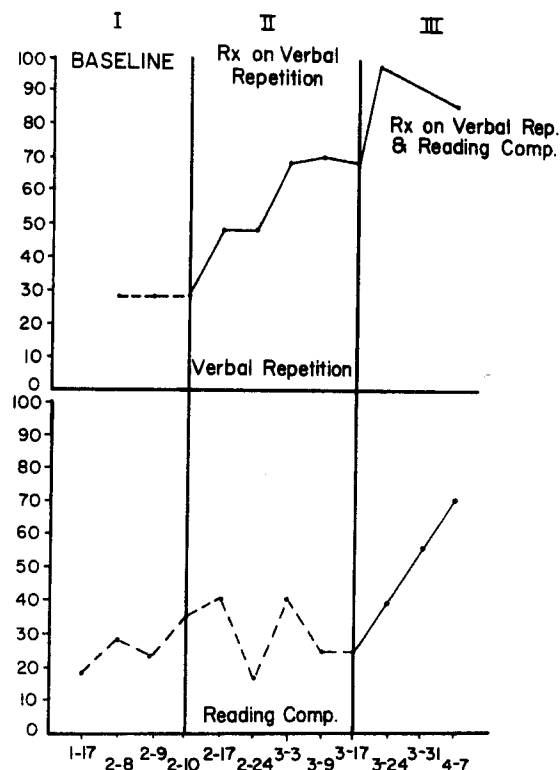


Figure 1. Multiple Baseline Design In Which The Delayed Treatment Behavior Was Reading.

## Practical Problems Which Emerged in Carrying Out The Design

From the case just discussed and the other four cases that we focused on that Spring, several additional practical problems were brought to our attention. One problem concerned what point of the week to take the weekly measure. Originally, we started collecting data at the end of the week. The problem we ran into was that when the patient missed therapy we ended up postponing taking the measure for three days because of the weekend, thus increasing time between measures. We would recommend that weekly measures be scheduled early in the week, in the event that a follow-up day must be arranged. In this way, time between measures can be kept as consistent as possible. Another problem that we encountered was running out of time for the treatment period, as opposed to the pretreatment period. Several times we would have liked to extend treatment over a longer time period, as in the case just discussed. Careful scrutiny of the semester calendar, if you have such time constraints, will avoid this problem. Those of you who don't work in a university setting won't have to be concerned about this particular problem. We were also alerted to the need for a more sensitive scoring scale so that small changes in behavior could be noted. This issue will be addressed again shortly.

A legitimate concern you might be having is the amount of time involved when doing such clinical research. Since this research entails weekly measures the amount of time required by the patient is a major consideration. From our experience the total amount of time invested by each patient in such a study was approximately 25 minutes per week.

## Most Recent Work With Single Subject Research Designs

Our most recent work with the research design has solved some, but not all of the problems. The data which follow are rather unimpressive. Either the patient didn't improve in anything or patient improvement was not detected because our method to measure performance was poor. Daily progress notes report improvement in auditory comprehension and expression. But that's not good enough. What we need are data in a graph that will reflect this improvement. The problem is twofold: (1) We need to create a more diverse and subtle measure of auditory comprehension. (2) We need to develop a scoring system that will reflect small changes in behavior.

Following are two studies in which the behavior for which treatment was delayed was writing, while auditory comprehension and expression were worked on daily four days a week. Baseline responses from which the target behavior stimuli were selected were obtained by presenting a set of pictures three times within a two-day period. (This represents a modification of the usual sense in which baseline data is used). The pictures represented specific levels of frequency on the Thorndike & Lorge Word Frequency List (1944). We've learned that the weekly performance measure is most informative if it represents a range of difficulty of the target behavior. In our most recent studies, the word lists used to test weekly performance on writing ability were made up of nouns representing two levels of frequency of occurrence (a la Thorndike and Lorge) plus verbs. We saw the need to develop a series of word lists in order to control for possible learning effects. We feel that by alternating between two word lists in obtaining the weekly measure of writing performance and holding a third word list until the end of the study we can control for learning effects. We've also learned the importance of carefully recording the correct word next to the patient's response following a set of stimuli. There is often little

correspondence between the patient's written response and the target word.

This brings us to the issue of scoring weekly written responses. As already mentioned, the problem of developing a scoring system sensitive to small changes in behavior is a critical one. We can report both success and failure in the solution of this problem. We have developed an eleven-point scale for evaluating written responses (see Appendix A). We feel that this scale reflects a hierarchy of performance and can be used across patients. This eleven-point scale was used to derive percent correct performance on the weekly measure displayed in figures 2 and 3. Figure 2 shows that there was improvement with the initiation of treatment prior to treatment as well as additional improvement with the initiation of treatment. The top graph in both Figure 2 and Figure 3 reflects weekly measure of auditory comprehension. The middle graph in both figures reflects expression based on picture description. Both measures were based on a 5 point rating scale. The rating scale for measuring auditory comprehension was: 4 = accurate, 3 = delayed, 2 = self correct, 1 = with repetition, 0 = inaccurate. The rating scale for measuring expression was: 4 = use of phrases in picture description, 3 = single word description, 2 = extended English jargon, 1 = extended English jargon with neologistic jargon, 0 = neologistic jargon.

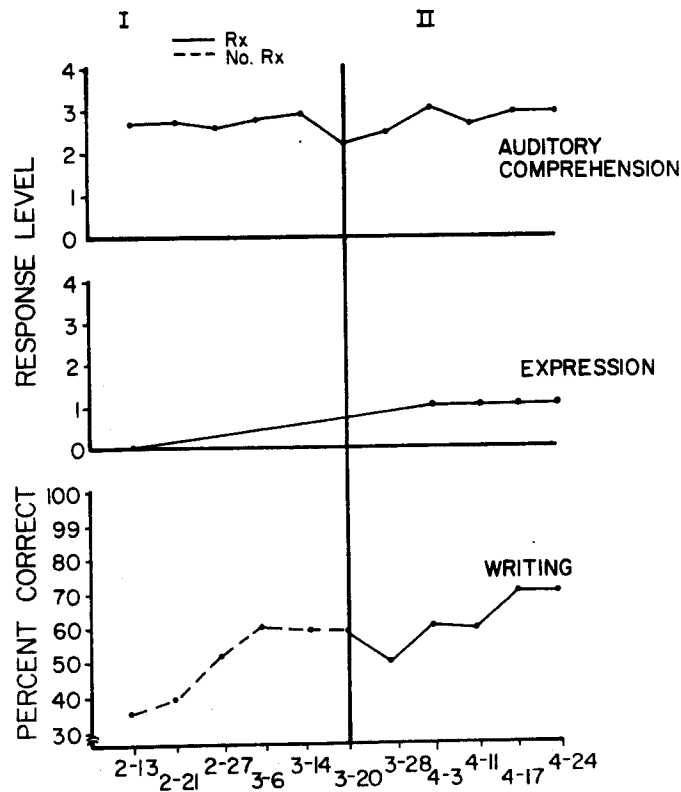


Figure 2. Multiple Baseline Design in Which The Delayed Treatment Behavior Was Writing.

The more traditional auditory comprehension battery of yes/no questions, digit memory, following directions and phrase comprehension used in this study does not reflect a wide enough range of performance on auditory comprehension tasks nor does the 5 point scale reflect a wide enough range of difficulty within each task to sensitively measure behavior change. We feel that a performance measure of auditory comprehension would be most valuable if it reflected functional behavior such as determining what to do in hypothetical life situations presented in role playing activities. Such a battery of activities is currently being developed at our Center by Jeanne Wilcox.

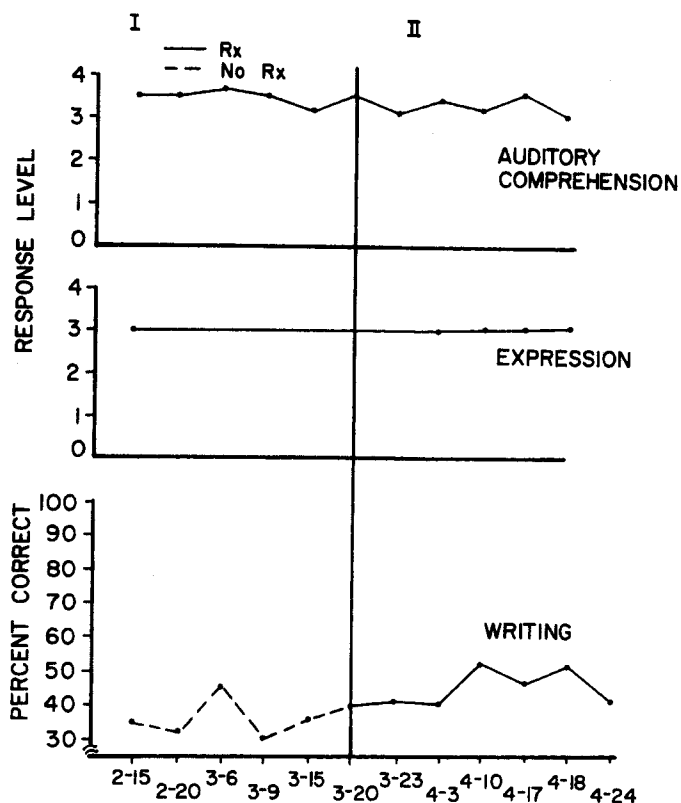


Figure 3. Multiple Baseline Design in Which the Delayed Treatment Behavior Was Writing.

Figure 3 shows that again a small amount of improvement in writing was noted both before and after therapy on writing was begun. While the conclusions are highly speculative due to our current inadequacies in measuring behavior changes in auditory comprehension and expression, both cases seem to indicate that performance in a third modality, such as writing, improves as a function of work in other language modalities.

I would like to note that one additional benefit of systematically collecting samples of writing performance within the construct of the multiple baseline design is that it provides a permanent record of behavior that may lend itself to further linguistic analysis, such as aphasic vs. apraxic types of errors in writing. Initially one patient's writing behavior was like a flood of graphemic jargon much like her expressive speech. But with time, control of written responses was evident as responses became closer to the target as noted, for example, in responding with the correct number of syllables for the intended word.

In summary, there are many problems yet to be worked out in our clinical application of single subject research designs. The major problem deals with violating basic assumptions of a design appropriate for other fields of behavior analysis and consequently, the need for modification of the design. Selecting the measuring device and developing a sensitive scoring system to assess behavior change are crucial considerations for successful research design work.

Despite the many frustrations we've encountered we're still encouraged by what we might learn about therapy effectiveness and the nature of language behaviors through systematic measurement of aphasic persons' performance. The number of questions this sort of clinical research has generated for us about the underlying processes of language behavior and hierarchies of performance within behaviors have been numerous and exciting. Perhaps with further refinement of our measurement system and a more thorough understanding of language modalities we can convincingly talk about the efficacy of aphasia therapy.

#### Appendix A

##### 11 Point Rating Scale For Assessing Writing Responses

- 10 Accurate
- 9 Self Correct
- 8 Recognized as Target but Misspelled
- 7 Related Noun or Related Process for Verb
- 6 Due to one Letter Omission, Substitution or Letter Reversal Word is Inaccurate and Unrecognizable
- 5 At Least Two Letters at the Beginning or End of the Word are Correct
- 4 Related Noun with Other Graphemic Jargon or Noun Description for Verb or Adjective for Noun
- 3 Unrelated Word
- 2 Nonsense Word, but Different Graphemic String, or Attempt is so Incomplete as to be Unrecognizable
- 1 Perseverative Grapheme String
- 0 No Attempt at the Task

## References

- Baer, D.M., Wolf, M.M., and Risley, T.R. Some Current Dimensions of Applied Behavior Analysis. Journal of Applied Behavior Analysis, 1968, 1, 91-97.
- Hersen, M., and Barlow, D. Single Case Experimental Designs: Strategies for Studying Behavior Change. New York: Pergamon Press, 1976.
- LaPointe, L.L. Base-10 Programmed Stimulation: Task Specification, Scoring, and Plotting Performance in Aphasia Therapy. Journal of Speech and Hearing Disorders, 1977, 42, 90-105.
- Leitenberg, H. The Use of Single-Case Methodolgy. Journal of Abnormal Psychology, 1973, 82, 87-101.
- Thorndike, E.L. and Lorge, I. The Teacher's Word Book of 30,000 Words. New York: Teachers College Press, 1944.

## Acknowledgement

This work was supported by Veterans Administration Education and Training Grant 2 C (75) III-43 0295-02.