*A PROBABILISTIC MODEL OF INTENSIVE DESIGNS*TERRY F. PECHACEK¹

UNIVERSITY OF MINNESOTA

Without internal validity, experimental data are uninterpretable. With intensive designs, most methods presented to quantify a design's internal validity have been subject to criticism. A probabilistic model of intensive designs is presented that demonstrates the high degree of internal validity of these designs without relying on adaptations from traditional inferential statistics. Where the experimenter is able to conform to the restrictions of the model, the equations provide an estimation of internal validity for either reversal or multiple-baseline designs. More importantly, the model provides mathematical bases for some of the common recommendations and design considerations in intensive research (such as the desirability of within-subject replications and of four or more multiple baselines).

DESCRIPTORS: experimental design, intensive design, probabilistic model of significance, methodology, multiple baselines, guidelines, single organism research

Demonstration of internal validity of inferences about the effects of an experimental (independent) variable on dependent variables is central to experimental investigation (Campbell and Stanley 1966). While traditional inferential statistics provide numeric estimations of internal validity, it can be meaningfully argued that they can obscure the true relationship between independent and dependent variables by averaging within-subject variability of response to treatment (Hersen and Barlow, 1976; Sidman, 1960). Hersen and Barlow (1976) outlined the advantages of single-subject designs in this area, but the internal validity of these designs bears further elucidation.

Internal validity concerns the ability of the design to discount or eliminate plausible rival

hypotheses not based on independent variables, especially those based on extraneous variables such as history (psychosocial events concurrent with but unrelated to experimental influences) (Campbell and Stanley, 1966, Kazdin and Koppel, 1975).

Kazdin (1976) reviewed both the appropriateness and controversy regarding efforts to adapt group-comparison statistics for single-subject designs. The following is an attempt to demonstrate by means of a probabilistic model that intensive designs do have a high degree of internal validity without relying on adaptations from traditional group statistics. In formulations of a more general nature than those presented by Revusky (1967), the internal validity of both the reversal and multiple-baseline designs can be demonstrated. Also, these formulations offer mathematical rationales for some of the common recommendations for using these designs.

In this model, the issue of what magnitude of observed effects or change is significant will not be addressed;² rather, consistent with the basic

¹Portions of this paper were presented at the annual meeting of the Association for Advancement of Behavior Therapy, Atlanta, December 1977. The basic work on this manuscript was completed while the author was a clinical psychology intern at the Palo Alto Veterans Administration Hospital, Palo Alto, California. The author would like to express his gratitude to Alan E. Kazdin for his editorial assistance on this manuscript. Reprints may be obtained from Terry F. Pechacek, Laboratory of Physiological Hygiene, Stadium Gate 27, University of Minnesota, Minneapolis, Minnesota 55455.

²The reader is referred to the discussion of clinical *versus* statistical significance in Hersen and Barlow (1976) and to the chapter by Kazdin (1976) on the

philosophy of single-subject research, "change" will be defined as a departure from baseline of a magnitude that would be judged of *clinical* significance by the research community at large; "no change" is defined as all other occurrences. Therefore, these formulations are not dependent on traditional statistical analyses, nor do they preclude their use.

In the development of this model, a basic assumption must be made. Consistent with inferential statistics, plausible rival hypotheses for observed treatment effects are considered as a single null hypothesis and expressed as the potential for random change (that is, not systematically related to the experimental variables). This basic assumption states that *the observed behavior has an equal probability of not changing or changing (increasing or decreasing) at any time when measured in the natural state.*

This assumption does not imply that proximate observation or measurement are *not* correlated. Rather, this assumption states that the effects of the extraneous factors can become manifest at any time and, *a priori*, there is no more than a random chance to suspect that they will coincide with the experimental variables. If this assumption is accepted, the model computes an estimate of the likelihood that rival hypotheses based on extraneous, random factors (such as history) could account for the observed results in reversal or multiple-baseline intensive designs.³

REVERSAL DESIGNS

In reversal designs, the experimenter defines the direction of change that is expected in each

appropriations of various statistical analysis in single-case designs.

³While this assumption is constructed to be as consistent as possible with the phenomenon of intensive research, the application of the hypothesis testing approach, which is standard to most psychological research, does not always suit the endeavors of the applied researcher whose goal often is the discovery, rather than testing of potent independent variables.

phase. A statistic can be formulated that estimates the likelihood that treatment and reversal effects could be due to extraneous, random factors. This statistic is based on the number of possible patterns of results that could occur if the outcomes were due only to the random factors. In the most general cases, random factors could produce three outcomes at each phase. (no change, an increase, or a decrease). Since the experimenter is *a priori* defining a single, expected outcome for each phase, there is a probability of 0.33 that this definition is correct if the outcome is due to random factors. In a design with N experimental, postbaseline⁴ phases, the probability that the outcome in all phases would be correctly defined is equal to $0.33 * N$.⁵

In some experimental analyses, the observed behavior may not have the freedom either to increase or decrease. Such a case would occur if the frequency was at or reduced almost to a zero rate or near the physical or biological maximum rate of occurrence. In such cases, the probability of correct definition would be increased to 0.50. Table 1 presents the probabilities for increasing N s in both the unrestricted and common restricted types of cases.

The statistic can be simply and directly computed for any design. Moreover, if each phase is continued for a sufficient period of observations, the intervention or reversal phases can be subdivided. For example, the standard A-B-A design of two experimental phases can be expanded

⁴The initial baseline cannot be considered as an experimental phase in this model since the requirements for obtaining a stable baseline preclude acceptance of the basic assumption. If such extraneous factors were to produce significant variability, the logic of the design requires the baseline to be continued until it restabilized.

⁵In reality, the probability (P_β) of the common patterns of data may not be $0.33 * N$ but would actually be a function of the true probability (P_{a1}) of the behavior changing in the desired direction in each of the N phases. Hence, the true probability (P_β) = $(P_{a_1})(P_{a_2}) \dots (P_{a_N})$, where $0 < P_{a1} \leq 1.0$. In many situations, such as in the first experimental phase after the initial baseline, P_{a1} may be quite small, making $P_{a1} = 0.33$ a generally conservative estimate.

Table 1

Probabilities for reversal designs with N experimental phases.

N	Type of Design	Number of Patterns ^a	Probability ^b
DESIGNS WITH UNRESTRICTED RANGE			
2	A-B-A	9	0.111
3	A-B-A-B	27	0.037
4	A-B-A-B-A	81	0.012
5	A-B-A-B-A-B	243	0.004
6	A-B-A-B-A-B-A	729	0.001
DESIGNS WITH INITIAL BASELINE HAVING RESTRICTED RANGE			
2	A-B-A	6	0.167
3	A-B-A-B	12	0.083
4	A-B-A-B-A	36	0.028
5	A-B-A-B-A-B	72	0.014
6	A-B-A-B-A-B-A	210	0.005
DESIGNS WITH RESTRICTED RANGE AFTER INTERVENTION			
2	A-B-A	6	0.167
3	A-B-A-B	18	0.056
4	A-B-A-B-A	36	0.028
5	A-B-A-B-A-B	108	0.009
6	A-B-A-B-A-B-A	216	0.005

^aNumber of Patterns = number of permutations of change that could occur due to random factors.

^bProbability that the expected pattern occurred due to random factors ($p = (1 \div \text{number of patterns})$).

to a A-B₁-B₂-A design of three experimental phases. However, such subdivision accentuates within-subject variability, which may obscure ability to define whether the predicted outcome occurred or not. Within the model, it is primary that results be distinguishable as either significant change from baseline level or reversal back toward baseline level.

On the basis of this model, the internal validity of reversal designs can be demonstrated. While the model concludes that the traditional A-B-A design produces a probability estimate greater than the generally accepted criterion of $p < 0.05$, this estimate may be conservative (see Footnote 4) and does not take into consideration that the data from many A-B-A experiments could have been considered as A-B₁-B₂-A or A-B-A₁-A₂ experiments. Nevertheless, in line with recommendations for direct replications (A-B-A-B-A designs) and for termination of

the experiment during the intervention phase (A-B-A-B design) (Barlow and Hersen, 1973; Hersen and Barlow, 1976), the experimenter may want to include more than the two experimental phases. This is especially true where the dependent measure is at or near either a zero rate or the physical or biological maximum rate during baseline or after intervention. Likewise, the model suggests that more complex designs (see Barlow and Hersen, 1973, or Hersen and Barlow, 1976) should have good internal validity.

MULTIPLE-BASELINE DESIGNS

Within multiple-baseline experiments, several baselines are formed and then the independent variables (intervention) are sequentially introduced into the environments of each baseline. Treatment effects are demonstrated when each baseline changes in the desired direction *only* after the intervention. Kazdin and Kopel (1975) stated that the internal validity of the multiple-baseline experiment is a function of a number of factors, including the rapidity with which the behaviors change in the desired direction, the variability in the data, and the number of baselines used. Again, a probabilistic model can quantify some of these factors, demonstrating the high validity of a well-constructed, multiple-baseline experiment and providing specific recommendations regarding the important issues of independence of baselines and the number of baselines required. The model is derived from the basic process of the multiple-baseline design. When an experimenter selects the order of introducing the independent variables into baselines, an *a priori*, expected order of independent, sequential changes in the baselines has been defined.⁶ If

⁶This requirement for *a priori* definition can run counter to the logic of some intensive experiments, since it restricts the flexibility of the experimenter to a predefined order of treatment. Hence, the model applies only to those designs wherein the order of treatment can be flexible and not affect the intent of the experiment.

Table 2
Probabilities for Multiple-Baseline Designs

		Number of Baselines				
		3	4	5	6	7
SIMPLE CASE ^a						
Number of Patterns ^b	6	24	120	720	5040	
Probability	0.1670	0.0417	0.0083	0.0014	0.0002	
PROBABILITIES WITH POTENTIAL FOR SIMULTANEOUS CHANGE ^d						
Number of Patterns	15	64	325	1956	1.37 × 10 ⁴	
Probability	0.0667	0.0156	0.0031	0.0005	7.3 × 10 ⁻⁵	
PROBABILITIES WITH POTENTIAL FOR SIMULTANEOUS AND/OR NO CHANGE ^e						
Number of Patterns	35	208	1233	2.10 × 10 ⁴	1.00 × 10 ⁵	
Probability	0.0286	0.0048	0.0008	4.8 × 10 ⁻⁵	1.0 × 10 ⁻⁵	

^aDerived with formula $f_1(n) = n!$.
^bNumber of permutations of change that could occur due to random factors.
^cProbability of predicted pattern occurring due to random factors.
^dDerived with Equation 1.
^eDerived with Equation 2.

the observed effects are inferred to be due to extraneous, random factors, then the observed order of change could have been any one of several. If the basic assumption of this model, namely: *that each observed behavior has an equal probability of not changing, changing in the undesired direction, or changing in the desired direction at any time when measured in the natural state,*⁷ is accepted, then the number of possible orders of changes can be computed by the formula, $f_1(n) = n!$, where n is the number of baselines in the design. When the experimenter can *a priori* define a single pattern of expected results (see Footnote 6), the likelihood that this would occur due to random factors would be $1 \div n!$. Table 2 provides those probabilities for increasing numbers (ns) of baselines.

⁷While each behavior may not have an equal probability of changing or not changing, the likelihood that baselines will change in the desired direction is of critical importance. Therefore, the experimenter may need to select randomly the order of introducing the intervention into baselines or, where that is not possible for therapeutic or experimental reasons, begin with the baseline with the least probability of change and conclude with the baseline of highest probability for desired change. The model is specifically inappropriate where the order of treatment is selected to maximize the likelihood of producing the desired results at each stage, since this approach maximizes the interdependence of the baselines.

Unfortunately, the multiple-baseline experiment is somewhat more complex. First, the requirement of independence of change across baselines is difficult to achieve (Kazdin and Kopel, 1975). Often, the baselines are not truly independent, and change in one increases the probability of change in others. Moreover, the simple formula expressed above only considers sequential, *nonsimultaneous* changes in baselines. The reality of interdependence creates the potential for *simultaneous* changes (defined as the case where two or more baselines display significant changes in the desired direction the intervention is introduced in only one of the baselines). When this aspect of the design is incorporated into the model, the formula becomes more complex. Equation 1 will compute the number of possible orders of change under these conditions.

$$f_2(n) = \sum_{t=1}^n \left(\frac{n!}{(n-t)!} \right).$$

(1)

Table 2 provides probability estimates for these more complex cases.
In the simple formula, $f_1 = n!$, and in Equation 1, it is assumed that all baselines will be observed to change during the experimental analysis. In reality, one or more of these baselines may in fact show no changes judged as

significant (see Footnote 2). Equation 2 computes the number of possible patterns of results that could be observed under conditions that allow for one or more baselines to change simultaneously and/or one or more baselines not to change significantly.

$$f_3(n) = n + \sum_{t=2}^n \left\{ t + \sum_{r=2}^{t-1} \left[r \cdot (t-r) \cdot \left(\frac{t!}{(t-r)!} \right) \right] + t! \right\} \quad (2)$$

Table 2 provides the resulting probabilities for increasing numbers of baselines under these conditions.

All probabilities in Table 2 are based on the optimum outcome in a well-designed, multiple-baseline experiment, namely, an observed pattern of independent, sequential, and significant changes across baselines in the exact order in which the independent variables were introduced. When this does occur, even under conditions where the baselines are somewhat interdependent, if the order of introducing the intervention into the baselines was not biased toward the desired results (see Footnote 7), the model would predict that possibly with three and more definitely with four baselines, plausible rival hypotheses for the results could be rejected with a $p < 0.05$. Therefore, the model

provides a mathematical rationale supporting the most common recommendations regarding the need to use four or more baselines (Jeffrey, 1975; Kazdin and Kopel, 1975). In fact, the model demonstrates the rapidly increasing power of the design when more than four baselines are used (see Table 2).

The model also enables the experimenter to evaluate the less desirable pattern of results when interdependence of baselines produces nonsequential or simultaneous change in two or more baselines. Kazdin and Kopel (1975) discussed the problems in this area and the difficulty in predicting the sought interdependence in baselines. However, the probabilistic model can quantify the situation of two baselines changing simultaneously and displaying interdependence and offers some estimates about the relevance of the results. Equation 3 computes the number of ways that this type of pattern could occur with n baselines plus the single, desired pattern of independent change across all baselines.

$$f_4(n) = 1 + \left(\frac{(n-2)!}{n!} \right) \quad (3)$$

The results of Equation 3 define the subsample of patterns that, when compared to the results from Equations 1 and 2, provide ratio estimates of the number of baselines that will be needed in order to reject the influences of ex-

Table 3
Ratios with Designs Having Potential for Interdependent Baselines

	Number of Baselines				
	3	4	5	6	7
RATIOS BASED ON EQUATION 1					
Possible experimental patterns ^a	6	12	20	30	42
Total patterns ^b	15	64	325	1956	1.37×10^4
Ratio	0.4000	0.1875	0.0615	0.0153	0.0031
RATIOS BASED ON EQUATION 2					
Possible experimental patterns	6	12	20	30	42
Total patterns ^c	35	208	1233	2.10×10^4	1.00×10^5
Ratio	0.1714	0.0577	0.0162	0.0014	0.0004

^aDerived with Equation 3.

^bDerived with Equation 1.

^cDerived with Equation 2.

traneous factors with an acceptable degree of certainty should some pair of baselines display a degree of interdependence. Table 3 shows these ratios for increasing numbers of baselines. In the interpretation of these results, a decision must be made regarding the appropriateness of the ratios based on either Equation 1 or Equation 2. The ratios based on Equation 2 require that all baselines display unequivocal change in the desired direction. Ratios based on Equation 1 are valid as long as all baselines display a pattern of change distinctly related to the intervention, except for the dependent member of the pair. Therefore, the model provides a mathematical rationale for using five, six, or more baselines depending on the experimenter's confidence in the expected results.

Overall, this probabilistic model demonstrates the high degree of internal validity that well-constructed, multiple-baseline experiments can display. In fact the model predicts that as few as three baselines are sufficient for an experiment with independent baselines displaying sequential changes distinctly related to the intervention. Moreover, the model permits the evaluation of less obvious results, such as those that occur when a pair of baselines display interdependent change; however, ensuring against this occurrence may require the planned use of at least six baselines.

In summary, where the experimenter is able to conform to the *a priori* definition process outlined in the model, and when the obtained results are judged sufficiently definitive and unequivocal, the probabilistic model provides an estimation of internal validity for either a reversal or multiple-baseline type intensive experiment. In a process similar to inferential statistics, plausible, rival hypotheses are combined into a single null hypothesis that can be rejected with a specific limit of confidence. In this process, the model provides specific evidence

of the strength of the obtained results. However, the model does not offer a solution to the controversy regarding the best method in which equivocal results should be analyzed. Rather, the model is based on the assumption that interventions will produce clinically significant results, with the significance defined by the experimenter or experiment as appropriate. Hopefully, the application of the model where appropriate will help to support the legitimacy of single-subject designs and encourage their application to even more diverse experimental questions.

REFERENCES

- Barlow, D. H. and Hersen, M. Single-case experimental designs: Uses in applied clinical research. *Archives of General Psychiatry*, 1973, **29**, 319-325.
- Campbell, D. T. and Stanley, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally, 1966.
- Hersen, M. and Barlow, D. H. *Single case experimental designs: strategies for studying behavior change*. New York: Pergamon Press, 1976.
- Jeffrey, D. B. Treatment evaluation issues in research on addictive behaviors. *Addictive Behaviors*, 1975, **1**, 23-26.
- Kazdin, A. E. Methodological and assessment considerations in evaluating reinforcement programs in applied settings. *Journal of Applied Behavior Analysis*, 1973, **6**, 517-531.
- Kazdin, A. E. Statistical analyses for single-case experimental designs. In M. Hersen and D. H. Barlow, (Eds), *Single case experimental designs: strategies for studying behavior change*. New York: Pergamon Press, 1976.
- Kazdin, A. E. and Kopel, S. A. On resolving ambiguities of the multiple-baseline design: Problems and recommendations. *Behavior Therapy*, 1975, **6**, 601-608.
- Revusky, S. H. Some statistical treatments compatible with individual organism methodology. *Journal of Experimental Analysis of Behavior*, 1967, **10**, 319-330.
- Sidman, M. *Tactics of scientific research*. New York: Basic, 1960.

Received 18 November 1975.

(Final Acceptance 3 February 1978.)