A Sharper Analytical Tool for the Multiple-Baseline Design

Matthew J. Koehler and Joel R. Levin

University of Wisconsin – Madison

ABSTRACT

Despite its widespread applicability, the multiple-baseline design remains an underused approach in the educational researcher's methodological bag of tricks. The new analytical tool described here is sharper than previously proposed multiple-baseline nonparametric statistical approaches, in at least two different respects. First, it is sharper *conceptually* and *methodologically* than the Marascuilo and Busk (1988) approach, insofar as it maintains the basic integrity of the multiple-baseline design, namely the systematically staggered introduction of the intervention across replicates. Second, the present analytical tool is sharper *statistically* than either of the previously proposed multiple-baseline randomization procedures (Revusky, 1967; Wampold & Worsham, 1986), in that: (*a*) it is statistically practicable with fewer replicates ($N < 4$); and (*b*) it can be shown to be more sensitive to patterns that reflect desired, as well as other typically observed, effects of an educational or clinical intervention. Moreover, the "number of permutations" formula related to the specific research application described here (KL_2) is shown to be a special case of an all-encompassing formula (KL_1), which provides the researcher with a flexible analytic tool that can take into account the methodological and statistical tradeoffs alluded to above.

Paper presented at the annual meeting of the American Educational Research Association, New York, April 1996.

A Sharper Analytical Tool for the Multiple-Baseline Design

Matthew J. Koehler and Joel R. Levin

University of Wisconsin – Madison

Background

Despite its widespread applicability, the multiple-baseline design remains an underused approach in the educational researcher's methodological bag of tricks. This is puzzling insofar as the design satisfies critical empirical validity criteria in a variety of research contexts – specifically, internal validity, discriminant validity, and, to some extent, external validity (see, for example, Kazdin, 1992; and Levin, 1992a). Two particularly fertile areas of application include single-case interventions (which originate from clinical "behavior analysis" studies) and classroom- or other group-based educational interventions.

The multiple-baseline design may be diagrammed in adapted Campbell and Stanley (1966) notation for four experimental "units" as follows:

	T_1	T_2	T_3	T_4	T_5
U_1	O	I O	I O	I O	I O
U_2	O	O	I O	I O	I O
U_3	O	O	O	I O	I O
U_4	O	O	O	O	I O

where: the *T*s represent time periods; I represents the intervention; O represents a measured outcome; and the *U*s represent the units or replicates to which the intervention is administered (usually an individual or group).

Thus, the experimental intervention begins prior to the first measured outcome at Time 1 with the first randomly determined individual or group, while the remaining units serve as nonintervention controls. (Alternatively, any Os prior to the introduction of the intervention belong to that unit's "baseline" phase, whereas those following the intervention belong to the

unit's "intervention" phase.) The intervention is maintained for the first unit for the remainder of the time periods. The intervention for the second unit commences just prior to Time 2 (while Units 3 and 4 serve as controls) and continues throughout the duration of the study. Observations or measures are taken within each time period and are used to assess between- and/or within-unit changes in performance.

Relative to competing single-case designs, the multiple-baseline framework is noteworthy for its qualities of internal validity (concerning plausible rival hypotheses that could account for intervention effects), replication/generalization (across units) and selectivity/discrimination (in producing the desired effects of the intervention). That is, if it can be demonstrated on logical or statistical grounds that a replicable effect is selectively produced during the targeted intervention phases while other potentially contributing variables are controlled, then one's confidence in the intervention's efficacy is enhanced (Levin, 1992b, pp. 216-217). The same degree of confidence is not as easily inspired by alternative single-case designs – incuding the replicated AB design, to be mentioned shortly.

Previously Proposed Statistical Procedures

A variety of statistical procedures have been proposed to analyze the data from multiple-baseline designs. These procedures fall primarily into two general classes, those based on time-series models and those incorporating a permutation (or nonparametric randomization) rationale. In this paper, we consider only the latter of these two general classes. Earliest of the nonparametric procedures was the straightforward method of Revusky (1967), which entails determining the joint probability of independent between-unit outcomes. Next came the approach of Wampold and Worsham (1986), which arguably improved both the appropriateness and precision of the analysis by incorporating a within-unit comparison component. In these initial randomization-based statistical approaches, the "phased in" intervention is assumed to occur at certain constant times (e.g., immediately, after 3 weeks,

after 6 weeks, after 9 weeks). What is randomized is the order in which the units receive the phased-in intervention (i.e., units are randomly assigned to the points at which the intervention is first introduced). Adopting a fundamentally different randomization notion for single-case experiments, Edgington (1975) had earlier proposed an ingenious design-and-analysis procedure for the basic unreplicated AB design (where A and B are baseline and intervention phases, respectively) – or, consistent with the above notation, O O O .... I O I O I O... The novelty of the Edgington approach consists of the kind of randomization demanded of the researcher, specifically that the particular time period (*T*) for introducing the intervention must be determined *randomly*, in advance of the study. Thus, in a study containing 12 time periods, rather than the researcher deciding to provide six baseline periods followed by six intervention periods, randomization according to the Edgington model might produce 10 baseline periods followed by two intervention periods. The Edgington model *does* allow a researcher to specify the minimum number of within-phase observations that are required, which is taken into account in the analysis. Moreover, Onghena (1994) has derived the general numerical form of Edgington's restricted randomization procedure, which can be fruitfully applied to randomization analyses of other single-case and small-sample designs. Indeed, the restricted randomization notion was adapted, though in a different sense, to the multiple-baseline model that is presented here.

More recently, Marascuilo and Busk (1988) extended the Edgington (1975) approach to incorporate baseline (A) vs. intervention (B) comparison data from more than one unit by computing a joint probability. They do this in a perfectly appropriate manner for the "replicated AB" design and their analysis is a powerful one. However, there is the sense (conveyed by the authors themselves both in the title of their article and the discussion and examples contained therein) that the same analysis can be routinely applied to the multiple-baseline design. There

is a conceptual shortcoming with that argument, however, which is summarized in the following paragraph.

The beauty and logic of the multiple-baseline design lie in the credibility and discriminant validity associated with its temporal contiguity and sequencing of the units. That is: (1) *At the same point in time* that one or more units is receiving the intervention, other units are still in the baseline phase; and (2) The intervention is phased in sequentially to the randomly designated units. Because of these temporal features, various threats to the design's internal validity can be easily dismissed. The same cannot be said of replicated AB designs. At one extreme, it is not even necessary that the replications take place concurrently; the different units' data could be collected at entirely different points in time, even in different settings or sites. At the other extreme, with the Edgington model as applied by Marascuilo and Busk, depending on the "luck of the draw" it would be possible for all units to receive the intervention close to, or exactly at, the same points in time. For example, in a four-unit, 12-period design, it could happen that the four units were randomly selected to commence their interventions just prior to Times 4, 5, 4, and 6, respectively. Such an unfortunate coincidence would serve to eliminate the desired discriminant validity provided by the multiple-baseline's staggered and balanced introduction of the intervention. Moreover, were the units to consist of classrooms within a school, simultaneous scheduling of an instructional intervention might well be an unwanted consequence, if not an impossibility. Thus, although well-suited to the generalized or replicated AB design, the Marascuilo and Busk (1988) solution is one that does not fit well with either the conceptual basis, the esthetic character, or the practical implementation of the multiple-baseline design.

A Sharper Analytical Tool (KL_2)

The sharper analytical tool discussed in this section can be thought of most easily as a modified version of the Wampold and Worsham (1986) approach. It is one that retains the

basic integrity of the multiple-baseline design, while at the same time capitalizing on two randomization schemes for the analysis: (1) the random assignment of replicates to the different points at which the intervention is to be phased in, as is required for both the Revusky (1967) and Wampold-Worsham analyses; and (2) the determination of a specific intervention "start point" for each unit, based on a random selection from a designated interval of acceptable start points within the unit's assigned phase-in stage – representing a novel adaptation of Edgington's (1975) notion of a "minimum phase length." The subsequent nonparametric statistical analysis takes advantage of these two randomization components, thereby improving its sensitivity relative to the earlier nonparametric randomization alternatives. With each of these procedures, the analysis consists of determining the likelihood of the obtained outcome – that associated with the difference between intervention and control units and/or intervention and baseline phases – and those outcomes more extreme, relative to all outcomes that could have been produced (i.e., given all possible randomizations of the data).

To provide a hypothetical example of this dual randomization scheme vis-à-vis the single one that is inherent in the previous Revusky (1967) and Wampold and Worsham (1986) procedures, consider the multiple-baseline intervention study outlined in Table 1, which contains $N = 3$ classrooms and $T' = 6$ outcome-assessment time periods (excluding T_1, which includes the initial baseline assessment). According to both of the previous statistical procedures, the associated randomization distributions consider a total of $N!$ possible rank-ordered outcomes (Revusky) or intervention vs. baseline mean differences (Wampold & Worsham), which for this example is equal to $3! = 6$. Let us now add the component of randomly selecting either of two ($k = 2$) designated potential staggered multiple-baseline start points for each classroom, namely: prior to either T_2 or T_3 for the first randomly assigned classroom, between that and either T_4 or T_5 for the second classroom, and between that and either T_6 or T_7 for the third classroom. The randomization distribution associated with the

present statistical procedure now considers a total of $N! \times k^N$ possible intervention vs. baseline mean differences, which for this example is equal to $3! \times 2^3 = 48$.

The effect of increasing the number of randomization outcomes possible is to increase the sensitivity of the present approach, relative to its competitors, with respect to detecting intervention effects. A unique advantage of this can be seen in the present example: Neither the Revusky nor the Wampold-Worsham procedure is capable of detecting an intervention effect based on $\alpha \leq .05$. The minimum sample size required for either of those procedures is $N = 4$ units; for this example, the best that one could do with $N = 3$ is $\alpha = 1/3! = 1/6 = .167$. In contrast, with the present approach, one could detect a statistically significant ($\alpha \leq .05$) intervention effect with $N = 3$ units and $k = 2$ potential start points, in that a study that yielded either of the *two* most extreme differences in the expected direction would be associated with $\alpha = 2 / 3! \, 2^3 = 2/48 = .0417$. For $N = 3$ units and $k = 3$ potential start points, the *eight* most extreme differences could be included in the rejection region, for $\alpha = .049$. Indeed, the present procedure is capable of detecting an intervention effect based on $\alpha \leq .05$ with only $N = 2$ units, as long as there are at least $k = 4$ potential intervention start points for each unit. Consistent with our previous discussion, however, with as many potential start points as in the latter two cases, one could begin to lose control of the important "temporal contiguity" character of the multiple-baseline design. For a summary comparison of the three nonparametric procedures discussed here, see Table 2.

Although the dual randomization scheme is the one that initially motivated the present work, the associated "number of possible outcomes" formula will henceforth be referred to as KL_2. As will now be shown, that is because KL_2 is subsumed by a more general statistical formula for the multiple-baseline design (dubbed KL_1), which is able to encompass all of the previously proposed randomization approaches.

A General Statistical Model for the Multiple-Baseline Design (KL_1)

The specific dual randomization multiple-baseline model just discussed can be subsumed by KL_1, a more general statistical model.[1] In addition to the previously given N = the number of units and T' = the number of outcome assessment periods (excluding the initial baseline assessment), for this model we need to specify P = the desired number of partitionings of the T' assessments, k_i = the number of potential start points per partition, and n_i = the number of units per partition. With these specifications, the number of possible outcomes associated with each of the previously discussed multiple-baseline schemes can be determined from KL_1 as:

$$N! \prod_{i=1}^{P} \binom{k_i}{n_i}$$

with $\sum_{i=1}^{P} k_i = T'$ and $\sum_{i=1}^{P} n_i = N$.

To illustrate, we reconsider our example for which $N = 3$ units are to used in a study with $T' = 6$ post-baseline assessment outcomes (T_2 - T_7). In addition, we specify that these 6 outcomes are to be partitioned into $P = 3$ groups, with $k_1 = k_2 = k_3 = 2$ intervention start points per partition and $n_1 = n_2 = n_3 = 1$ unit randomly assigned to each start point. Accordingly, there are:

$$3! \binom{2}{1}\binom{2}{1}\binom{2}{1} = 6(2)(2)(2) = 48$$

randomization outcomes (intervention-baseline mean differences), as was determined through the more specialized KL_2. Note that this general formula can be readily adapted to situations in which T cannot be equally divided into the P partitions. In the present example, suppose that only $T' = 5$ assessment outcomes are possible, rather than 6. In addition, the researcher decides that $k = 2$ potential start points are to be associated with the first partition, $k = 1$ with the second partition, and $k = 2$ with the third partition. All other specifications are the same. With these changes, the total number of possible intervention-mean differences would be reduced by a factor of two (i.e., halved), as can be seen in the following KL_2 calculation:

$$3!\binom{2}{1}\binom{1}{1}\binom{2}{1} = 6(2)(1)(2) = 24$$

What is more, it can be seen from Table 3 that, with some terminological modifications, the "number of possible outcome" calculations associated with all of the previously proposed nonparametric multiple-baseline approaches can be reproduced by the present general formula (KL_1). As is summarized in Table 4, the present KL approach (restricted start point randomization for each unit) represents a methodological compromise between that of Marascuilo and Busk (1988), with its complete start point randomization for each unit (i.e., no randomization restrictions), and that of both Revusky (1967) and Wampold and Worsham (1986), with its absence of start point randomization for each unit (i.e., no randomization). As a result, the present approach would be expected to lead to statistical improvements over the latter two.[2] In addition, the present general approach allows for the possibility of incorporating more than one unit at each specified partition (i.e., per-partition replications, or $n_i > 1$).

At the same time, it should be noted that with the increased flexibility afforded by this approach comes the potential for ethical abuses. For example, following a "close, but not statistically significant outcome," a researcher might be tempted to reconduct the analysis based on some number of within-partition potential start points even when start-point randomization was not incorporated into the study as conducted (i.e., when the traditional multiple-baseline approach was employed). Such potential for researcher misconduct needs to be taken into consideration and balanced against the previously indicated strengths of the present approach.

Actual Research Application

We now describe an empirical research application of the present procedure, in a study conducted by the first author of college students' ability to identify the strategies used by young children when solving mathematics story problems. The study compared the utility of computer-based hypermedia training materials with traditional text-based instruction. The

lesson outlined research findings about young children's solution strategies for addition and subtraction word problems. Students in the baseline phase received a training manual that contained only text. During the intervention phase, students received a hypermedia version of the training materials (for purposes of this example, the design is based on $N = 4$ students). The materials in this phase contained essentially the same content as the baseline phase text materials, but also included video examples of children solving word problems to augment the text examples. After each of the periods of study, students were administered a computer task that required them to freely arrange 12 text examples of children's solution strategies within a 3 x 4 grid. Students' sorts were scored on a 0-14 point scale, with higher scores reflecting an organization more consonant with the conceptual model presented in the training materials. This ideal organization uses one of the dimensions (the three-level dimension) to group strategies of like developmental level, and uses the other dimension to group strategies used to solve identical problem types. This grid used in the assessment is not presented in any form in either the baseline or the treatment materials. To assess the effectiveness of the two approaches, difference scores between measurement points are used in the statistical analysis so that rates of learning can be investigated.

Thus, for the present example, the data consist of $T' = 8$ difference scores. In addition, prior to the study, it was specified that the hypermedia phase would begin at T_2 for one of the four students ($k_1 = 1$), at either T_3 or T_4 for one of the students ($k_2 = 2$), at either T_5 or T_6 for one of the students ($k_3 = 2$), and at either T_7 or T_8 for one of the students ($k_4 = 2$). The startpoints randomly selected from those specified were T_2, T_4, T_6, and T_7, and the four students were randomly assigned to these. The data from the study are presented in Table 5, which yield an observed statistic (an across-student average of the intervention minus baseline phase means) of 1.11. According to KL_1, with these specifications there are:

$$4!\binom{1}{1}\binom{2}{1}\binom{2}{1}\binom{2}{1} = 24(1)(2)(2)(2) = 192$$

possible permutations of the various mean difference outcomes. The value of 1.11 observed here turns out to be the third most extreme in the predicted direction, which is therefore associated with a one-tailed probability of 3/192 = .016.[3] Accordingly, with a one-tailed test based on $\alpha = .05$, one could conclude that students gained more during the hypermedia lessons than they did during the standard computer lessons.

Summary

In summary, the new analytical tool described here is sharper than previously proposed multiple-baseline approaches in at least two different respects. First, it is sharper conceptually and methodologically than the Marascuilo and Busk (1988) approach, insofar as it maintains the basic integrity of the multiple-baseline design, namely the systematically staggered introduction of the intervention across replicates. Second, the present analytical tool is sharper statistically than either of the previously proposed multiple-baseline randomization procedures (Revusky, 1967; Wampold & Worsham, 1986), in that: (*a*) it is statistically practicable with fewer experimental units ($N < 4$); and (*b*) it can be shown to be more sensitive to patterns that reflect desired, as well as other typically observed, effects of an educational or clinical intervention.

References

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

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

Kazdin, A. E. (1992). *Research design in clinical psychology* (2nd ed.). Boston: Allyn and Bacon.

Levin, J. R. (1992a). On research in classrooms. *Mid-Western Educational Researcher*, *5*, 2-6, 16.

Levin, J. R. (1992b). Single-case research design and analysis: Comments and concerns. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New developments for psychology and education*. Hillsdale, NJ: Erlbaum.

Marascuilo, L. A., & Busk, P. L. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behavioral Assessment*, *10*, 1-28.

Onghena, P. (1994). *The power of randomization tests for single-case designs*. Faculteit der psychologie en pedagogische wetenschappen, Katholieke Universiteit Leuven, Leuven, Belgium.

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

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

Footnotes

1. We are most grateful to Carol Blumberg, who provided the impetus that prompted us to uncover this generalization.
2. Such potential improvements are currently being investigated by the present authors. In particular, we are examining the competing procedures' abilities to detect multiple-baseline effects of various magnitudes (e.g., strong vs. weak) and types (e.g., immediate vs. delayed), as a function of the number of outcome assessments, the number of randomized units, and the width of the within-unit randomized start-point interval.
3. Macintosh-based software has been developed to incorporate the design and analysis, and when in final form, will be made available to requesters.

Table 1. Illustration of the Koehler-Levin Restricted Randomization Scheme for the Multiple-Baseline Design ($N = 3$, $T' = T\text{-}1 = 6$, and $k = 2$)

	T_1		T_2		T_3		T_4		T_5		T_6		T_7
C_1	O	I[a]	O	I[a]	O	I	O	I	O	I	O	I	O
C_2	O		O		O	I[a]	O	I[a]	O	I	O	I	O
C_3	O		O		O		O		O	I[a]	O	I[a]	O

[a] Within each classroom, one of these two designated potential start points is randomly selected as the actual start point for the intervention.

Table 2. Minimum Number of Units (*N*), Outcome Assessments (*T*), and Potential Intervention Start Points for Each Unit (*k*) In Order To Detect an Intervention Effect Based on $\alpha \leq .05$: Comparison of Three Nonparametric Multiple-Baseline Statistical Procedures

Procedure	Basis of Comparison	*N*	*T*	*k*
Revusky (1967)	Between Units[a]	4	3	1
Wampold-Worsham (1986)	Between and Within Units	4	5	1
Koehler-Levin (1996)	Between and Within Units	3	7	2

[a] With the Revusky procedure, no pre-intervention outcome assessment period is necessary. Moreover, for each successive unit, the intervention need not be continued beyond the first outcome assessment following its introduction. These are two practical advantages of the Revusky procedure that should be considered.

Table 3. Application of The General Formula (KL_1) to Various Multiple-Baseline Nonparametric Procedures

Procedure	Sample Specifications	No. of Possible Outcomes According to: Original Formula	KL_1
Marascuilo-Busk (1988)[a]	$N = 3$, $T' = 6$	$6^3 = 216$	$0!\binom{6}{1}\binom{6}{1}\binom{6}{1} = 216$
Wampold-Worsham (1986)[b]	$N = 3$ $P = 1$, $k_1 = 1$, $n_1 = 1$	$3! = 6$	$3!\binom{1}{1}\binom{1}{1}\binom{1}{1} = 6$
Revusky (1967)[b]	$N = 3$ $P = 1$, $k_1 = 1$, $n_1 = 1$	$3! = 6$	$3!\binom{1}{1}\binom{1}{1}\binom{1}{1} = 6$

Note: N = the number of units (or randomized units in KL_1); T' = the number of outcome assessment periods excluding the initial baseline assessment; P = the desired number of partitionings of the T' assessments; k_i = the number of specified start points per partition; and n_i = the number of units per partition

[a] In the general formula, $N = 0$ because this model does not incorporate between-unit randomization.

[b] In the general formula, k_1 and n_1 each equals 1 because within-unit randomization is not incorporated (and, therefore, $P = 1$ fixed partition is used for each unit).

Table 4. Comparison of Multiple-Baseline Nonparametric Procedures

Procedure	Overlap Constraints	Internal Validity	No. of Permutations
Marascuilo-Busk (1988)	None: Potential for much phase overlap among units	Low	T'^{N}
Koehler-Levin 1 (1996)	Mild: Small range of phase overlap among units	Medium	$N! \prod_{i=1}^{P} \binom{k_i}{n_i}$
Koehler-Levin 2 (1996)	Moderate: No phase overlap among units	Medium to High[a]	$N! \times k^N$
Wampold-Worsham (1986)	Complete: No phase overlap among units	Very High	$N!$
Revusky (1967)	Complete: No phase overlap among units	Very High	$N!$

Note: T' is one less than the number of outcome assessments (i.e., excluding the initial baseline assessment). In the present case, we also assume that the number of required baseline and intervention assessments is each set at the minimum possible, namely 1.

[a] With all else held constant, internal validity increases as k decreases.

Table 5. Data (Difference Scores) from the Hypermedia Study, Excluding Initial Baseline Phase (T_1)

Student	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	*A Mean*	*B Mean*	$M_B - M_A$
1	3	-2	1	-3	1 ▌	3	0	-1	.000	.667	.667
2	3	0	-3 ▌	1	0	-1	3	-4	.000	-.200	-.200
3	-3 ▌	-1	3	0	-4	4	-2	1	-3.000	.143	3.143
4	3	-3	3	-2	1	-1 ▌	1	1	.167	1.000	.833
				Across-Student Means					-.708	.402	1.11

Note: ▌ indicates the beginning of the hypermedia phase