Psychological Methods
1998, Vol. 3, No. 2, 206–217

1082-989X/98/$3.00

Regulated Randomization: A Potentially Sharper Analytical Tool for the Multiple-Baseline Design

Matthew J. Koehler and Joel R. Levin
University of Wisconsin—Madison

A dual-randomization procedure, *regulated randomization,* is proposed for behavioral and educational interventions that incorporate the logic of single-case multiple-baseline designs. The new approach is sharper conceptually and methodologically than previously developed approaches in that regulated randomization maintains the basic integrity of the multiple-baseline design (namely, the systematically staggered introduction of the intervention across the experimental units) while being statistically practicable with fewer units ($N < 4$). Moreover, previously suggested nonparametric analyses of multiple-baseline data can be subsumed by a general regulated randomization formula. The regulated randomization approach provides researchers with a flexible analytic tool that can take into account specific substantive, methodological, and statistical trade-offs.

Despite its widespread applicability, the single-case multiple-baseline design remains an underused approach in researchers' bag of tricks for behavioral and educational interventions. 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, e.g., Kazdin, 1992, pp. 168–172; Levin, 1992a). Two particularly fertile areas of application of the multiple-baseline design include true *single-subject* interventions (which originate from clinical behavior analysis studies) and classroom- or other group-based instructional interventions.

The multiple-baseline design is diagrammed in adapted Campbell and Stanley (1966) notation for four experimental units in Table 1, in which the *T*s represent time periods *I* represents the intervention, *O*s represent measured outcomes and *U*s represent the experimental units to which the intervention is administered (usually individuals, small groups, or classrooms). Thus, for the first randomly designated individual or group, the experimental intervention begins following the first measured outcome at Time 1 (or following a series of preintervention outcomes), while the remaining units serve as nonintervention controls. The one or more *O*s prior to the intervention belong to that unit's *baseline* (or A) phase, and those *O*s following the introduction of intervention belong to the unit's *intervention* (or B) phase. The intervention is maintained for Unit 1 for the remainder of the time periods. The intervention for the Unit 2 commences following a later measured outcome, say, following Time 3 (while Units 3 and 4 serve as controls) and continues throughout the duration of the study. Units 3 and 4 are subsequently phased into the intervention in a similar systematically staggered fashion (i.e., following Times 5 and 7, respectively). Observations or measures are taken at each time period and are used to assess between- and within-unit changes in pre- to postintervention performance.

Compared with 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 and generalization (across units),

Matthew J. Koehler and Joel R. Levin, Department of Educational Psychology, University of Wisconsin—Madison.

We are grateful to Carol Blumberg for her helpful comments on earlier versions of this article.

Correspondence concerning this article should be addressed to Joel R. Levin, Department of Educational Psychology, 1025 West Johnson Street, University of Wisconsin, Madison, Wisconsin 53706. Electronic mail may be sent to levin@macc.wisc.edu.

Table 1
Basic Multiple-Baseline Design for N = 4 Units

	Time period																
Unit	T_1		T_2		T_3		T_4		T_5		T_6		T_7		T_8		T_9
U_1	O	I	O	I	O	I	O	I	O	I	O	I	O	I	O	I	O
U_2	O		O		O	I	O	I	O	I	O	I	O	I	O	I	O
U_3	O		O		O		O		O	I	O	I	O	I	O	I	O
U_4	O		O		O		O		O		O		O	I	O	I	O

Note. T = time period; U = experimental unit (usually individuals or groups); I = intervention; O = measured outcome.

and selectivity and 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—including the replicated AB design, mentioned shortly.

But how does a researcher interpret the data produced by multiple-baseline designs? A historically prevalent (and still prevailing) school of thought among behavior analysts is essentially to let the data from single-case intervention studies speak for themselves through graphical plots and visual comparisons of the preintervention and postintervention phases' constituent observation points (see, e.g., Hersen & Barlow, 1976; Kazdin, 1992, pp. 340–348; Parsonson & Baer, 1992). For example, a multiple-baseline intervention effect might be inferred if, for each sequentially phased-in unit, a noticeable difference is apparent between that unit's preintervention and postintervention observation series (i.e., between phases A and B) with respect to either their respective levels (e.g., means) or slopes (i.e., steepnesses). Another school of thought is that more formal, and assumedly more objective data-analytic procedures should be enlisted (e.g., Jones, Weinrott, & Vaught, 1978): inferential statistical procedures in particular. We do not intend to either resolve or extend the single-case analysis debate with the new method we propose here. Neither do we claim that the method we propose is the single best or "the" correct way to analyze data from all multiple-baseline experiments. In fact, we later specifically acknowledge some of its likely limitations. Rather, proposed here is what we believe represents a novel analytical tool that (a) lends itself conceptually and methodologically to many multiple-baseline investigations and (b) has the potential to be statistically superior to contemporary competitors of the same genre. In short, the present methodological–statistical approach is intended for application specifically by multiple-baseline researchers who are receptive (if not already accustomed) to conducting formal inferential statistical analyses of their data.

Previously Proposed Statistical Procedures for the Multiple-Baseline Design

As implied, a number of statistical procedures currently are being applied to the analysis of multiple-baseline data. These procedures fall primarily into two general classes: those based on time-series models and those incorporating a permutation-based (or nonparametric randomization) rationale. Although each class of procedures has its own specific advantages and disadvantages (cf. Kratochwill, 1978), in this article we consider only the latter of these two general classes. In a nutshell, nonparametric analysis of single-case data considers *all possible* intervention-versus-baseline outcomes (derived from mathematical combinations or permutations) given the study's specific design and unit-assignment features and according to the null hypothesis assumption of no intervention effect. On the basis of all such possible outcomes, one constructs a complete randomization distribution, and the actual outcome's location within that distribution is noted along with its statistical probability.

Previous Nonparametric Analyses of Multiple-Baseline Data

Earliest among the nonparametric procedures was the straightforward method of Revusky (1967), which entails determining the joint probability of independent between-units 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 following the initial assessment (e.g., immediately, after 3 weeks, after 6 weeks, after 9 weeks).[1] What is randomized is the order in which the units receive the phased-in intervention (i.e., the order in which units are randomly assigned to the different predetermined start points of the intervention).

Edgington's Randomization Model

Adopting a fundamentally different randomization notion for single-case experiments, Edgington (1975) proposed an ingenious design-and-analysis procedure for the basic unreplicated AB design. Consistent with the above notation, this can be diagrammed as

O O O . . . I O I O I O . . .

or, more simply, by phase as

AAAA . . . BBBB . . . .

The novelty of the Edgington approach is 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 (please refer again to Footnote 1). Thus, in a study containing 12 time periods, rather than the researcher deciding to provide 6 baseline periods followed by 6 intervention periods, randomization according to the Edgington model might produce an intervention start point between the 10th and 11th observations, thereby yielding 10 baseline periods and 2 intervention periods. The Edgington model does allow a researcher to specify the minimum number of within-phase observations that are required, which is subsequently taken into account in the statistical analysis. Moreover, Onghena (1994) has derived the general numerical form of Edgington's restricted randomization procedure, which can be applied fruitfully to randomization analyses of other single-case and small-sample designs. Indeed, the restricted randomization notion has been adapted, albeit in a different sense, to the multiple-baseline model that is presented here.

Marascuilo and Busk's Extension

More recently, Marascuilo and Busk (1988) extended Edgington's (1975) approach to incorporate intervention versus baseline phase comparisons from more than one unit by computing a joint probability. They did 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 in the discussion and the 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 paragraphs.

Problem

The beauty and logic of the multiple-baseline design lie in the credibility and discriminant validity associated with its temporal contiguity and planned systematic sequencing of the units (see our previous discussion in this article and Kazdin, 1992, pp. 168–172). That is to say (a) *at the same point in time,* one or more units are receiving the intervention while other units are still in the baseline phase and (b) the intervention is phased in *systematically and sequentially* to the randomly designated units. By virtue of these features, various threats to the design's internal validity can be dismissed easily. However, the same cannot be said of replicated AB designs. At one extreme, it is not 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 at different sites. At the other extreme, with Edgington's (1975) model as applied by Marascuilo and Busk (1988), dependent on the luck of the draw, it is possible for all units to receive the intervention close to or exactly at the same points in time. That circumstance could cloud a key desired property of multiple-baseline designs—namely, clear temporal separability of the replicated intervention effects:

[1] This fixed and predetermined number of baseline observations for each unit is a stipulation that is not universally accepted by multiple-baseline researchers out of the behavior-analytic tradition. Many such researchers argue that the length of the baseline phase for each unit should be individually determined in the context of the investigation itself and should be based on how long it takes for the particular unit to achieve a "stable" set of baseline observations (Kazdin, 1992, p. 169). We later return to this issue, as it is one that has direct implications for both the statistical analyses and the conclusions that follow from them.

> The *multiple-baseline design* demonstrates the effect of an intervention by showing that behavior change accompanies introduction of the intervention at different points in time. . . . A causal relation between the intervention and behavior is clearly demonstrated if each response changes only when the intervention is introduced and not before. (Kazdin, 1992, pp. 168–169)

For example, in a four-unit, 12-period design, with the Marascuilo and Busk (1988) approach, it might happen that the four units are randomly selected to commence their interventions just prior to Times 4, 5, 4, and 6, respectively. Such an unfortunate start-point determination would serve both to reduce the internal validity and to eliminate the desired discriminant validity provided by the multiple baseline's systematic, staggered introduction of the intervention. In addition, if the units are classrooms within a school, simultaneous scheduling of an instructional intervention might well be an impractical logistical consequence if not an impossibility. Thus, although well suited to the generalized or replicated AB design, the Marascuilo–Busk solution is one that does not fit well with the conceptual basis, the aesthetic character, or the practical implementation of the multiple-baseline design.

The Marascuilo–Busk (1988) approach essentially represents a sampling (of intervention start points) *with* replacement randomization scheme across units. Yet the same concerns as just expressed also apply to a sampling *without* replacement framework.[2] Suppose, for the above example, that Times 4, 5, 3, and 6 result from sampling 4 of 12 potential start points without replacement. Alternatively, suppose that Times 2, 3, 4, and 11 happen to be selected for the same example. Although some would not view outcomes of this kind as worrisome consequences of random selection, others (including the present authors) would be bothered by either the near simultaneous intervention start points for all of the replicates (in the first case) or the unsystematic staggering of the intervention introduction (in the second case) in a multiple-baseline design. Overlap and stagger issues of the kind just mentioned remain an interpretive concern.[3] With the decision to stagger the intervention start points systematically and at clearly separable intervals, one is essentially adapting a blocked random assignment strategy to the present context. Such a strategy is incorporated into traditional between-groups experiments for purposes of (a) similarly enhancing the study's internal validity (e.g., by controlling for potential temporal factors associated with treatment administration), (b) increasing the precision of the statistical analysis, or (c) both. Blocked random assignment is the basis of the multiple-baseline strategy proposed here.

Regulated Randomization: A Potentially Sharper Analytical Tool

The new procedure developed 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 capitalizing on two randomization schemes for the analysis: (a) the random assignment of units 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 (b) 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 particular unit's assigned phase-in stage, an 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 logic and sensitivity relative to the earlier nonparametric multiple-baseline alternatives. With this dual-component *regulated randomization* procedure, the analysis consists of determining the likelihood of the obtained outcome—that associated with the difference between intervention and baseline phases—and those outcomes as extreme as or more extreme than that obtained relative to all outcomes that could have been produced (i.e., given all possible permissible randomizations of the data).

To provide a concrete example of the dual regulated-randomization scheme as opposed to the single scheme characteristic of the previous Revusky (1967) and Wampold and Worsham (1986) procedures, we consider the multiple-baseline intervention study outlined in Table 2. This example contains $N = 3$ classrooms as experimental units observed across 10 outcome-assessment time periods (T). According to both of the earlier statistical procedures, the associated randomization distributions would consider a total of $N!$ possible rank-ordered outcomes (Revusky, 1967) or

[2] This point was made by a reviewer of an earlier version of this article.

[3] Nonetheless, we give additional consideration to this suggested sampling-without-replacement approach in a following section.

Table 2

Regulated Randomization Scheme for the Multiple-Baseline Design (N = 3 and k = 2)

	Time period																		
Unit	T_1		T_2		T_3		T_4		T_5		T_6		T_7		T_8		T_9		T_{10}
C_1	O	I[a]	O	I[a]	O	I	O	I	O	I	O	I	O	I	O	I	O	I	O
C_2	O		O		O		O	I[a]	O	I[a]	O	I	O	I	O	I	O	I	O
C_3	O		O		O		O		O		O		O	I[a]	O	I[a]	O	I	O

Note. T = time period; C_1, C_2, and C_3 represent N = 3 different classrooms; I = intervention; O = measured outcome.

[a] With the present procedure, one of the two designated potential start points is randomly selected as the actual start point for the intervention within each classroom.

intervention-versus-baseline mean differences (Wampold & Worsham, 1986), which, for this example, is equal to 3! = 6. Our method adds the component of randomly selecting either of two (k = 2) designated potential staggered multiple-baseline start points for each classroom: prior to either T_2 or T_3 for the first randomly assigned classroom (C_1), between that and either T_5 or T_6 for the second classroom (C_2), and between that and either T_8 or T_9 for the third classroom (C_3). The distribution associated with the present regulated randomization procedure considers a total of $N! \times k^N$ possible intervention-versus-baseline mean differences, which, for this example, is equal to $3! \times 2^3 = 48$ (i.e., eight times as many randomization outcomes as there are in the two earlier approaches).

A hypothetical data set based on these specifications and randomization test calculations, are presented in Table 3 and Table 4, respectively. We further assume that the actual intervention start points, randomly selected, are just prior to T_3 for Classroom 1, T_6 for Classroom 2, and T_8 for Classroom 3. In Table 4 (test calculations), over the data-set columns are specified the eight permissible start-point combinations given the preceding regulations. The unit-order column specifies the six possible permutations of three classrooms. Included in Table 4 are baseline phase (A) and intervention phase (B) means, listed by classroom permutation and corresponding to all permissible intervention start-point combinations. Also included are the mean B − A differences along with the ranks of those differences (R), with largest R = 1 and smallest R = 48. We illustrate the calculations for the first set of summary statistics in Table 3, which correspond to potential intervention start points just prior to T_2, T_5, and T_8 for Classrooms 1, 2, and 3, respectively. If Classroom 1's intervention actually had begun just prior to T_2, then the A mean for that classroom would be the single outcome of 4, or 4/1 = 4.000. Similarly, intervention start points just prior to T_5 and T_8 for Classrooms 2 and 3 would yield respective A means of (6 + 7 + 5 + 6)/4 = 6.000 and (9 + 9 + 7 + 10 + 10 + 8 + 9)/7 = 8.857. Across the three classrooms, the combined A mean is therefore given by (4.000 + 6.000 + 8.857)/3 = 6.286. With the same potential intervention start points, the respective B means are calculated to be (3 + 5 + 7 + 6 + 8 + 7 + 8 + 9 + 7)/9 = 6.667, (6 + 7 + 10 + 9 + 10 + 10)/6 = 8.667, and (12 + 11 + 14)/3 = 12.333, for a combined B mean of (6.667 + 8.667 + 12.333)/3 = 9.222. The difference (B − A) in these means is, therefore, equal to 9.222 − 6.286 = 2.936 ≈ 2.94, which turns out to be the 7th largest difference in the set of 48. Likewise, with the three intervention start points just prior to T_3 for Classroom 1, T_6 for Classroom 2, and T_8 for Classroom 3 (i.e., $O_3O_6O_8$ combined with $C_1C_2C_3$), the mean B − A difference would be equal to 3.43, which represents the largest of all 48 possible mean differences. Because this difference is the one associated with the intervention start points that were actually randomly selected for the three classrooms, one can conclude, with p = 1/48 = .021 (one-tailed), that there is evidence for higher performance after the introduction of the intervention than before it (i.e., a positive intervention effect).

Table 3

Hypothetical Data Associated With the Design in Table 2

Unit	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}
C_1	4	3[a]	5[b]	7	6	8	7	8	9	7
C_2	6	7	5	6	6[a]	7[b]	10	9	10	10
C_3	9	9	7	10	10	8	9	12[b]	11[a]	14

Note. $C_{1...3}$ represent N = 3 different classrooms; $O_{1...10}$ represent measured outcomes across 10 different observation periods ($T_{1...10}$).

[a] First observation associated with unselected intervention start point.

[b] First observation associated with selected intervention start point.

Table 4
The 48 Possible Randomization Outcome Sets Associated With the Design in Table 2

Unit order	M_A	M_B	M_B–M_A	R	M_A	M_B	M_B–M_A	R	M_A	M_B	M_B–M_A	R	M_A	M_B	M_B–M_A	R
	$O_2O_5O_8$				$O_2O_5O_9$				$O_2O_6O_8$				$O_2O_6O_9$			
$C_1C_2C_3$	6.29	9.22	2.94	7	6.42	9.28	2.86	8	6.29	9.40	3.11	5	6.42	9.46	3.04	6
$C_1C_3C_2$	6.49	9.00	2.51	20	6.58	9.11	2.53	19	6.57	9.04	2.47	23	6.67	9.16	2.49	21
$C_2C_1C_3$	6.54	9.20	2.67	14	6.67	9.26	2.59	16	6.62	9.30	2.68	13	6.75	9.36	2.61	15
$C_2C_3C_1$	6.82	9.06	2.23	32	6.92	9.17	2.25	31	6.90	9.16	2.25	30	7.00	9.27	2.27	29
$C_3C_1C_2$	6.82	8.81	1.99	38	6.92	8.81	1.90	42	6.90	8.86	1.95	40	7.00	8.86	1.86	44
$C_3C_2C_1$	6.90	8.89	1.98	39	7.00	8.89	1.89	43	6.90	9.07	2.16	34	7.00	9.07	2.07	36
	$O_3O_5O_8$				$O_3O_5O_9$				$O_3O_6O_8$				$O_3O_6O_9$			
$C_1C_2C_3$	6.12	9.38	3.26	3	6.25	9.43	3.18	4	6.12	9.55	3.43	1[a]	6.25	9.61	3.36	2
$C_1C_3C_2$	6.32	9.15	2.83	10	6.42	9.26	2.85	9	6.40	9.20	2.79	12	6.50	9.31	2.81	11
$C_2C_1C_3$	6.70	9.24	2.54	18	6.83	9.29	2.46	24	6.79	9.34	2.55	17	6.92	9.39	2.48	22
$C_2C_3C_1$	6.82	9.10	2.28	28	6.92	9.21	2.29	27	6.90	9.20	2.29	26	7.00	9.31	2.31	25
$C_3C_1C_2$	6.99	8.85	1.86	45	7.08	8.85	1.76	47	7.07	8.89	1.82	46	7.17	8.89	1.73	48
$C_3C_2C_1$	6.90	8.93	2.03	37	7.00	8.93	1.93	41	6.90	9.11	2.20	33	7.00	9.11	2.11	35

Note. $C_{1\ldots3}$ represent $N = 3$ different classrooms; $O_{2\ldots9}$ represent observation periods selected and then randomized to produce the eight possible start-point order indexes, which in turn are cross-tabulated with the six classroom order indexes to produce the above 48 data sets; A = baseline phase; B = intervention phase; R = ranked mean difference.
[a] Ranked mean difference corresponding to the selected intervention start point.

The consequence of increasing the number of possible randomization outcomes in the present regulated-randomization approach is its increased capacity to detect intervention effects relative to those capacities of its competitors. A unique advantage of this can be seen in the present example, in which neither the Revusky (1967) nor the Wampold and Worsham (1986) procedure is capable of detecting an intervention effect based on a Type I error probability, α, of .05 or less. The minimum sample size required for either of those procedures is $N = 4$ units; for this example, the best that one could do with $N = 3$ units is $\alpha = 1/3! = 1/6 = .167$. In contrast, with the present approach, one can detect a statistically significant ($\alpha \leq .05$) intervention effect with $N = 3$ units and $k = 2$ potential start points per unit (as was just demonstrated); a study that yields either of the *two* most extreme differences in the expected direction can be included in a rejection region with $\alpha = 2/3!\,2^3 = 2/48 = .0417$ (one-tailed). For $N = 3$ units and $k = 3$ potential start points per unit, the *eight* most extreme differences can be included in the rejection region, for $\alpha = 8/3!\,3^3 = .049$. Indeed, the regulated randomization 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. In light of our preceding "internal validity" 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 randomization procedures discussed here, see Table 5.

The regulated randomization example just discussed is the one that initially motivated the present work. As shown in the following discussion, however, the associated number-of-possible-outcomes formula is a special case of a more general formula for the multiple-baseline design, which is able to encompass all of the previously proposed randomization approaches.

A General Randomization Model for the Multiple-Baseline Design

The specific regulated randomization multiple-baseline model under consideration generalizes to a two-component randomization model for which N = the number of units and k_i = the number of potential start points associated with the ith unit's partition. With these dual specifications, the number of null hypothesis–compatible possible outcomes associated with each of the previously discussed multiple-baseline schemes can be determined simply as

$$N! \prod_{i=1}^{N} k_i. \qquad (1)$$

Table 5

Comparison of Three Nonparametric Multiple-Baseline Statistical Procedures

Procedure	Basis of comparison	N	T	k
Revusky (1967)	Between units	4	3–5[a]	1
Wampold and Worsham (1986)	Between and within units	4	5	1
Regulated randomization[b]	Between and within units	3	7	2

Note. Minimum number of units (N), outcome assessment periods (T), and potential intervention start points for each unit (k) in order to detect an intervention effect are based on $\alpha \leq .05$.

[a] With the Revusky procedure, if only raw postintervention outcomes (rather than standardized or regression-adjusted outcomes) are analyzed, no preintervention outcome assessment period is necessary. Moreover, if only raw, standardized, or adjusted postintervention outcomes are analyzed (rather than within-unit, preintervention vs. postintervention differences), 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.

[b] As discussed in text, the minimum requisites here can be further reduced with a more general regulated randomization procedure.

To illustrate, we reconsider our example for which $N = 3$ units are included in a study with $T = 10$ outcome assessment periods. In addition, we again specify that these 10 periods are to be separated into three start-point partitions (T_2–T_3, T_5–T_6, and T_8–T_9) so that each partition contains $k_1 = k_2 = k_3 = 2$ potential intervention start points. Accordingly, there are $3!(2)(2)(2) = 3! \times 2^3 = 48$ randomization outcomes (intervention-baseline mean differences), as was determined through the special-case regulated randomization procedure. Note that the general regulated randomization Formula 1 can be readily adapted to situations in which T (here, number of assessment periods) cannot be equally divided into the specified number of partitions. In the present example, suppose that only $T = 6$ assessment periods are possible rather than $T = 10$. In addition, suppose that the researcher decides that $k_1 = 2$ potential start points are to be associated with the first partition (following an initial baseline assessment), $k_2 = 1$ with the second partition, and $k_3 = 2$ 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 calculation based on Formula 1: $3!(2)(1)(2) = 24$. Note that in terms of the regulated randomization test minima specified in Table 5, application of the present unequal start-point specifications (namely, $k_1 = 2$, $k_2 = 1$, and $k_3 = 2$) would reduce the minimum number of required outcome assessments from $T = 7$ to $T = 6$.

What is more, it can be seen from Table 6 that, with a few terminological modifications, the number-of-possible-outcomes calculations associated with all of the previously proposed nonparametric multiple-baseline approaches can be reproduced with Formula 1. As is summarized in Table 7, the present approach (with regulated 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 those of Revusky (1967) and Wampold and Worsham (1986), with their absence of start-point randomization for each unit

Table 6

Application of the General Regulated-Randomization Formula to Various Multiple-Baseline Nonparametric Procedures

	Sample specifications				
Procedure	N	k_1	k_2	k_3	No. of possible outcomes according to Formula 1
Marascuilo and Busk (1988)[a]	3	6	6	6	0! (6)(6)(6) = 216
Sampling-without-replacement analog to Marascuilo–Busk[a]	3	6	5	4	0! (6)(5)(4) = 120
Wampold and Worsham (1986)[b]	3	1	1	1	3! (1)(1)(1) = 6
Revusky (1967)[b]	3	1	1	1	3! (1)(1)(1) = 6

Note. N = the number of units (or randomized units in the regulated-randomization approach); k_i = the number of specified start points associated with each partition.

[a] In the general formula, $N = 0$ because this model does not incorporate between-units randomization. [b] In the general formula, all $k_i = 1$ because within-unit randomization is not incorporated (i.e., $k = 1$ fixed partition is used for each unit).

Table 7
Comparison of Multiple-Baseline Nonparametric Procedures

Procedure	Overlap constraints	Internal validity	No. of permutations
Marascuilo and Busk (1988)	None: potential for much phase overlap among units	Low	T'^N
Regulated randomization (general formula)	Mild: small range of phase over lap among units	Medium	$N! \prod_{i=1}^{N} k_i$
Regulated randomization (special case)	Moderate: no phase overlap among units	Medium to high[a]	$N! \times k^N$
Wampold and 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 numbers of mandatory baseline and intervention assessments are each designated pre-experimentally to be the minimum possible, namely 1.

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

(i.e., no within-unit randomization). As a result, the present regulated randomization approach might be expected to lead to statistical improvements over the latter two. Conceptually, the present regulated randomization Formula 1 combines a between-units randomization component ($N!$), as is incorporated into the Revusky and Wampold–Worsham procedures, with a within-unit randomization component,

$$\prod_{i=1}^{N} k_i ,$$

as is incorporated into the Marascuilo-Busk extension of Edgington's (1975) procedure (as well as the sampling-without-replacement analog).

An even more general regulated randomization formula, which allows for nonoverlapping within-partition replications, can be given as

$$N! \prod_{i=1}^{N} \binom{k_i}{n_i}, \qquad (2)$$

where n_i represents the number of nonoverlapping units associated with the ith partition. Although this replication adaptation has some statistical and implementation advantages associated with it (including the fact that the number of outcome assessments does not need to be increased for a fixed number of units), its application is not illustrated here because it diminishes certain of the methodological niceties of the multiple-baseline design (as was discussed previously). Note that for the nonreplicated version of the design, where $n_i = 1$ for all partitions, the second (combinatorial) piece of Formula 2 reduces to the corresponding piece of Formula 1, in that $\binom{k}{1} = k$ for each partition.

Research Application

We now describe a research application of the present regulated randomization procedure derived from a study of college students' ability to identify the strategies used by young children when solving mathematics story problems (Koehler & Lehrer, in press). The actual study, conducted with 6 college students, compared the utility of computer-based hypermedia training materials with traditional text-based instruction. (For illustrative purposes, the present example is based on $N = 4$ students, and modified data are used.) The lessons outlined research findings about young children's solution strategies for addition and subtraction word problems (Carpenter, Fennema, Peterson, Chiang, & Loef, 1989). A total of $T = 9$ individual sessions was held with each student. Students in the baseline sessions received a training manual that contained only text. During the intervention sessions, students received a hypermedia version of the training materials. The materials in these sessions featured essentially the same content as the baseline-session 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 arrange freely 12 text examples of children's solution strategies within a 3 × 4 grid.

Students' sorts were scored on a 15-point-scale (0–14), with higher scores reflecting an organization

Table 8
Session-to-Session Gains in the Hypermedia Study

Student	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	M_A	M_B	$M_B - M_A$
1	3	−2	1	−3	1	▌ 3	0	−1	0.00	0.67	0.67
2	3	0	−3	▌ 1	0	−1	3	−4	0.00	−0.20	−0.20
3	−3	▌ −1	3	0	−4	4	−2	1	−3.00	0.14	3.14
4	3	−3	3	−2	1	−1	▌ 1	1	0.17	1.00	0.83
	Across students								−0.71	0.40	1.11

Note. T = training assessment period (T_1 is excluded); A = baseline phase; B = hypermedia (treatment) phase; ▌ indicates the beginning of the hypermedia phase.

more consonant with the conceptual model presented in the training materials. The ideal organization uses one of the grid's dimensions (the three-level dimension) to group strategies of like developmental level and uses the other (four-level) 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, we incorporate into the statistical analysis difference scores (gains) between measurement points so that rates of learning can be investigated.[4]

Thus, for the present example, the data consist of the $T - 1 = 8$ difference scores. In addition, suppose that prior to the study, it was specified that the hypermedia sessions would begin at T_2 for 1 of the 4 students ($k_1 = 1$), at either T_3 or T_4 for another of the students ($k_2 = 2$), at either T_5 or T_6 for yet another of the students ($k_3 = 2$), and at either T_7 or T_8 for the remaining student ($k_4 = 2$). The start points randomly selected from those specified were T_2, T_4, T_6, and T_7, and the 4 students were randomly assigned to these. The adapted outcome data are presented in Table 8, and they yield an observed statistic (an across-students average of the intervention-minus-baseline session means) of 1.11. According to Formula 1, with these specifications, there are $4!(1)(2)(2)(2) = 192$ 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 $p = 3/192 = .016$.[5] Accordingly, with a one-tailed test based on $\alpha = .05$, one could conclude that students gained statistically more during the hypermedia lessons than during the standard text lessons.

Discussion

As noted previously, the present regulated randomization design-and-analysis approach affords greater flexibility and coherence relative to previously suggested multiple-baseline alternatives (e.g., Marascuilo & Busk, 1988; Revusky, 1967; Wampold & Worsham, 1986), but does it have potential to exhibit greater precision under a variety of patterns of change over time? The term *potential* is carefully chosen in that we are currently examining both the present and competing procedures' abilities to detect multiple-baseline effects of various magnitudes (e.g., strong vs. weak) and types (e.g., immediate vs. delayed), each as a function of such factors as the number of randomized units, the width of the within-unit randomized start-point interval, the number and type of outcome assessments, and the stability of the baseline series. The last two of these factors are briefly elaborated on in turn.

Sensitive Summary Measures

One yet-to-be-resolved issue concerns the use of baseline and intervention means in this and other single-case randomization analyses, which might be argued are not appropriately sensitive to the effects of an intervention, especially when the number of pre- and postintervention observations is large. Regardless

[4] A difference score approach is particularly appropriate for time-series situations in which a stable preintervention baseline is difficult or impossible to achieve: as, in the present example, when improvement would be expected to occur from one preintervention assessment period to the next. The present regulated randomization procedure can readily accommodate difference scores as well as other measures (e.g., within-classroom slopes, covariate-adjusted means).

[5] Macintosh-based microcomputer software has been developed by the authors to accommodate the design and analysis options associated with Formula 2 and, when in final form, will be made available on request.

of the general merit of that argument, randomization analyses can readily be adapted to address such concerns. For example, suppose that (a) either the units consist of relatively large aggregates (such as classrooms) or each observation is a relatively stable one (such as a reliable multiple-item test score) and (b) immediate preintervention to postintervention changes are anticipated. In such cases, one could base the present randomization analysis on, say, just the one or few observations immediately preceding and following the intervention rather than on the pre- and postintervention complete series means as was illustrated here. In reference to Table 8, for example, if only the single differences on each side of the actual intervention start point are examined, the obtained effect would consist of $3 - 1 = +2$, $1 - (-3) = +4$, $-1 - (-3) = +2$, and $1 - (-1) = +2$ for Students 1–4, respectively, for an across-student mean effect of $10/4 = +2.5$. This would then be referred to the distribution of all 192 possible randomization outcomes based on just the two observations that immediately precede and follow the intervention, which, in this case, turns out to be one of the 10 most extreme positive outcomes, for a one-tailed significance probability of $p = .052$ (compared with the $p = .016$ reported when the test was based on mean differences). Although we are not inclined to advocate the single-observation of few-observations approach in general (because of the likely lowered reliability associated with it), some might wish to consider that approach when the two conditions specified above are met.[6] However, from a practical standpoint, with this *selected observation* strategy, T observations still need to be collected for all *N* units even though not all of those observations are incorporated into the analysis, which might be regarded as wasteful (of either money or data).

Baseline Phase Considerations

Let us now return to the important distinction between allowing the length of each unit's baseline phase to be flexibly determined within the context of the actual investigation (which many behavior analysts would advocate) versus establishing it on a priori basis either through specification or randomization (an assumption underlying all of the statistical techniques discussed here). This philosophical difference, in fact, boils down to whether or not one subscribes to the notion that formal statistical analyses should be conducted on single-case (here, multiple-baseline) data rather than, say, the more informal visual analyses that were mentioned previously. Often, the philosophical difference translates into defensible tradeoffs between statistical concerns (typically for the staunch data analyst) and substantive-scientific concerns (typically for the stauch behavior analyst). Even though all of the statistical analyses discussed here would be compromised to varying degrees by data-based baseline-phase determinations, modifications of these procedures can be made to accommodate the behavior analyst's desire to establish a stable baseline for each unit prior to introducing the intervention or the unit.

One such modification that readily comes to mind would be for the researcher to decide not to begin the randomization process (i.e., not to determine each unit's intervention start point) until *all N* units have achieved a stable baseline. In a 4-unit design, if it takes nine observation periods before all 4 units demonstrate a stable baseline, with the present approach, one could wait until after the ninth observation period and then randomly determine for each unit at which of *k* potential start points beyond that period the intervention should commence.[7] Alternatively (and as we have suggested in Footnote 3), in learning situations in which improvements within the baseline phase would be expected (and thus achieving a stable baseline would not be expected), a researcher can circumvent the stable-baseline concern by analyzing differences between adjacent observations—or even between slopes of the complete intervention and baseline phases—rather than between intervention and baseline means based on all of the associated within-phase observations. The former (adjacent difference) measures reflect differential amounts or degrees of improvement in the context of continual improvement

[6] Although not elaborated on here, another recently offered suggestion in a related single-case context is that all within-series differences be based on the same constant number of both pre and post observations—in contrast to the standard Edgington (1975) approach of taking complete-series mean differences. With the latter approach, different means based on differing numbers of observations are necessarily associated with different stabilities, which can be shown to have unwanted consequences on the randomization distribution and its test statistic (Levin & Wampold, 1997).

[7] If achieving a stable baseline for each unit is deemed important, the Marascuilo and Busk (1988) approach can be similarly modified to allow for the *sequential* random selection of one of T_k intervention start points for each unit as that unit's baseline series stabilizes.

and thus are conceptually appropriate in such contexts.

Multiple-Baseline Design Specifications

Another important statistical-versus-substantive trade-off needs to be mentioned before concluding. We reiterate that the regulated randomization procedure introduced and described in this article is not the only one that can be applied to the analysis of multiple-baseline-like data. It is not necessarily even the best in the class of nonparametric competitors. From a purely statistical perspective, for example, both the Marascuilo and Busk (1988) sampling of start points with replacement and its sampling without replacement analog can be shown to have a greater number of possible randomization outcomes than the present procedure and thus might be claimed to be more sensitive to effects of an intervention. We counter that statistical criteria are not the only factors that count here. Indeed, the present dual-randomization procedure itself represents an attempted balance of dual concerns: (a) responsiveness to the intervention demonstration-and-interpretation standards of both the behavior analyst and the methodologist and (b) improving the statistical performance of the associated analysis relative to currently available nonparametric alternatives (i.e., those of Revusky, 1967, and Wampold & Worsham, 1986). Procedures that open the door to overlapping, near-overlapping, or unsystematic intervention stagger (i.e., those of Marascuilo & Busk and the sampling-without-replacement variation discussed earlier) are not likely to be universally embraced by multiple-baseline behavior analysts and methodologists. However, if potential start-point adjacencies or irregularities are either not a critical concern or necessitated by a reduced number of outcome observation periods, then a sampling-without-replacement randomization approach represents a reasonable methodological–statistical compromise between the Marascuilo–Busk procedure and the one proposed here. It should also be pointed out that the microcomputer program alluded to in Footnote 5 can handle any of the design specifications mentioned here (including *replicated* regulated randomization, discussed in conjunction with Formula 2) and can perform the associated randomization analyses. The only challenge, therefore, is for the researcher to decide which of these approaches is most applicable, acceptable, or appropriate for his or her own particular single-case situation.

Cautionary Comment

Finally, it should be noted that with the increased flexibility afforded by regulated randomization 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 on the basis of 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 opportunities for researcher misconduct need to be taken into consideration and weighed against the indicated strengths of the present approach.

Conclusion

In summary, the new regulated randomization 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, by the systematically staggered introduction of the intervention across experimental units. Second, the tool is sharper analytically than either of the previously proposed multiple-baseline nonparametric procedures (Revusky, 1967; Wampold & Worsham, 1986) in that it is statistically practicable with fewer units ($N < 4$). Whether it turns out to be sharper in a third respect—in terms of its sensitivity to patterns that reflect desired as well as other typically observed effects of an educational or a behavioral intervention—is a critical, yet-to-be-answered, question.

References

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

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal, 26,* 499–531.

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

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

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

Kazdin, A. E. (1992). *Research design in clinical psychology* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Koehler, M. J., & Lehrer, R. (in press). Designing a hypermedia tool for learning about children's mathematical cognition. *Journal of Educational Computing Research.*

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

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* (pp. 213–224). Hillsdale, NJ: Erlbaum.

Levin, J. R., & Wampold, B. E. (1997, July). *Single-case randomization tests for a variety of situations.* Paper presented at the 10th European meeting of the Psychometric Society, Santiago de Compostela, Spain.

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.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New developments for psychology and education* (pp. 15–40). Hillsdale, NJ: Erlbaum.

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.

Received February 7, 1997
Revision received June 16, 1997
Accepted November 12, 1997 ■