Abstract Title Page

Title:
On internal validity in multiple baseline designs

Authors and Affiliations:
James E. Pustejovsky
The Unversity of Texas at Austin
Educational Psychology Department

Abstract Body.

Background / Context:

Single-case designs are a class of research designs for evaluating intervention effects on individual cases. The designs are widely applied in certain fields, including special education, school psychology, clinical psychology, social work, and applied behavior analysis. The multiple baseline design (MBD) is the most frequently used single-case design (Shadish & Sullivan, 2011; Smith, 2012), and is often described as having desirable internal validity characteristics. The MBD involves periodically measuring an outcome variable on several cases in the absence of intervention, then introducing the intervention to each case in turn while continuing to measure the outcome. Two key features of the design thought to bear on its internal validity. The first feature is that the intervention is introduced deliberately by the researcher, rather than as a result of naturally occurring events (Horner et al., 2005; Kratochwill et al., 2012). In practice, researchers seldom use formal randomization of intervention starting points, though it is recognized that doing so strengthens internal validity (Kratochwill & Levin, 2010). The second key feature is that each case begins treatment at a different time, thus making it easier to rule out history-related threats (Kazdin, 2011; Kratochwill & Levin, 2010).

In many of the disciplines that use MBDs, visual inspection of graphed outcome data is the primary method of analysis (Gast & Spriggs, 2010; Kazdin, 2011; Smith, 2012). Systematic procedures for creation of graphic displays and for visual analysis have been proposed (e.g., Horner, Swaminathan, Sugai, & Smolkowski, 2012) and incorporated in the What Works Clearinghouse procedures for reviewing single-case studies (Kratochwill et al., 2012). However, recent years have also seen renewed interest in statistical analysis of MBDs, particularly as a means for estimating summary effect sizes. Much of the statistical work has focused on the use of piece-wise regression models, either applied separately to each case (e.g., Center, Skiba, & Casey, 1985; Crosbie, 1993; Maggin et al., 2011) or formulated as a hierarchical linear model (e.g., Hedges, Pustejovsky, & Shadish, 2012, 2013; Van den Noortgate & Onghena, 2003, 2008).

Purpose / Objective / Research Question / Focus of Study:

This paper examines how the use of certain hierarchical linear models that have been proposed for the MBD affect internal validity, focusing specifically on the two key features described above. First, I consider a deliberate but non-random method of treatment assignment. I demonstrate that the treatment effect estimate from a conventional multi-level model can be biased under this mechanism, then provide an expression for the magnitude of the bias. Second, I argue that the main analytic models proposed for the MBD fail to capture the benefits of staggered treatment introduction. I propose an alternative model that does account for this key feature and, based on the model, I define an index measuring the strength of control offered by the MBD.

Significance / Novelty of study:

It is widely recognized that parameter estimates from hierarchical linear models can be biased if the covariates are correlated with the higher-level error terms (e.g., Wooldridge, 2002). One contribution of the present investigation is to describe a specific model that produces such correlation and to characterize the size of the resulting bias. The other contribution is to introduce an analytic model that more closely captures the key features of the MBD. While similar models have been used in other disciplines, I believe that this one warrants greater attention for analysis of single-case studies.

Statistical, Measurement, or Econometric Model:

The data from a MBD with *m* cases and *N* measurement occasions can be described as follows. Let Y_{ij} denote the outcome measurement from the j^{th} occasion for case *i*. Case *i* is in the baseline phase for the first B_i outcome measurements, after which it begins the treatment phase for the remaining $N - B_i$ measurements. Let X_{ij} indicate the treatment status of case *i* on occasion *j*, so $X_{ij} = 0$ for $j \leq B_i$ and $X_{ij} = 1$ for $j > B_i$.

A simple but commonly considered model for these data (see, e.g., Hedges et al., 2013) assumes that cases vary in their baseline levels and that the treatment leads to a shift in the mean level of the outcome that is constant across cases:

$$Y_{ij} = \beta + \delta X_{ij} + \nu_i + \epsilon_{ij}, \tag{1}$$

where the within-case errors $\epsilon_{i1},...,\epsilon_{iN}$ have mean 0 and variance σ^2, and are assumed for sake of simplicity to be independent. The terms $\nu_1,...,\nu_m$ represent case-specific variation in the mean level of the outcome during the baseline phase, and are assumed to be independent and normally distributed with mean 0 and variance τ^2. Note that the covariate is not case-centered so as to preserve the interpretation of τ^2 as between-case variation in baseline outcomes.

Selection mechanisms. In the hierarchical modeling framework, the parameters of Model (1) would typically be estimated using restricted maximum likelihood (for σ^2 and τ^2) and weighted least squares (WLS) for β and the treatment effect δ. However, the WLS estimator of δ may be biased if treatment assignment times are selected using a deliberate but non-random method. One plausible method would be to "triage" cases according to the severity of their baseline outcome levels. Suppose that the experimenter plans to introduce the treatment at a fixed set of times $b_1,...,b_m$, where $1 < b_1 < b_2 < \cdots < b_m < T$. Suppose further that the experimenter has knowledge of $\nu_1,...,\nu_m$ based on experience or outside diagnostic information. The cases are assigned to treatment in order of their baseline severity: for a treatment intended to raise the level of the outcome, the case with the lowest expected baseline level enters the treatment phase first and the case with the highest expected baseline level enters the treatment phase last. It follows that $B_i = b_{r_i}$, where $(r_1,...,r_m)$ are the index ranks of $\nu_1,...,\nu_m$. This selection mechanism induces dependence between the covariate and the case-specific errors because $(X_{i1},..., X_{iN})$ is a function of B_i, which depends on v_i. I term this mechanism "triage on known baseline ranks," in order to emphasize that selection is based on the known ranks of $\nu_1,...,\nu_m$.

Denote the WLS estimator based on Model (1) as $\hat{\delta}_{(1)}$. Given estimates of the variance parameters σ^2 and τ^2, it can be shown that the standardized bias of $\hat{\delta}_{(1)}$ under triage on known baseline ranks is

$$\frac{E\left(\hat{\delta}_{(1)} - \delta\right)}{\sqrt{\tau^2 + \sigma^2}} = -\frac{\sqrt{\rho}\left(1-\rho\right)}{\left[\rho\bar{w} + \left(1-\rho\right)\bar{b}\left(1 - \bar{b}/N\right)\right]} \times \frac{1}{m}\sum_{i=1}^{m}\left(b_i - \bar{b}\right)\mu_{(i)}, \tag{2}$$

where $\mu_{(1)},...,\mu_{(m)}$ are the mean order statistics from *m* independent, unit-normal random variates, $\bar{b} = \frac{1}{m}\sum_{i=1}^{m} b_i$, $\bar{w} = \frac{1}{m}\sum_{i=1}^{m} b_i\left(N - b_i\right)$, and $\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$. Figure 1 plots the magnitude of the bias as a function of the intra-class correlation ρ, for varying values of *m* and *N*. The bias is

always negative, is never larger in magnitude than -0.1, and is largest when the between-case variation is small relative to the within-case variation.

Triage on known baseline ranks might be considered unrealistic because it assumes knowledge of the true case-level effects (or at minimum, their ranks). A more realistic mechanism would involve assigning treatment times to cases based on the baseline outcomes actually *observed* as the experiment progresses. Such a selection mechanism induces dependence between the covariate and both the case-level errors and the within-case errors, making it difficult to derive analytic expressions for the magnitude of the resulting bias. In on-going work, I am using simulation methods to examine the bias under this and other related mechanisms.

Model for staggered treatment introduction. No aspect of Model (1) accounts for the other key feature of the MBD—that treatment introduction is staggered rather than simultaneous. Instead, the model is structurally identical to a model for a collection of replicated AB designs, which are considered to have lower internal validity than the MBD (Kratochwill et al., 2012). The same remark holds true for a variety of other models proposed for use with the MBD, including piece-wise regression models (e.g., Center et al., 1985; Crosbie, 1993; Maggin et al., 2011) and other hierarchical linear models (e.g., Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; Shadish, Kyse, & Rindskopf, 2013; Van den Noortgate & Onghena, 2003, 2008).

Heuristically, staggering the introduction of treatment allows the analyst to control for history threats that are *common across cases* (Shadish, Cook, & Campbell, 2002, Chapter 6). Such threats are called sometimes called "common shocks," meaning influences that impact the outcome equally for all cases. To capture these common shocks, Model (1) can be modified to include fixed effects for each measurement occasion:

$$Y_{ij} = \beta_j + \delta X_{ij} + \nu_i + \epsilon_{ij}, \tag{3}$$

This analytic models is applied commonly in the econometric literature on panel data (e.g., Wooldridge, 2002) and sometimes in the public health literature on "stepped wedge" trials (e.g., Hussey & Hughes, 2007); however, it has been entirely overlooked for analysis of MBDs.

Three inter-related problems arise due to the focus on statistical models that do not capture this essential feature of the MBD. First, the naïve statistical analysis will negate whatever improvements in internal validity that the feature provides. To illustrate, observed that the treatment effect estimator based on Model (1) will be biased in the presence of common shocks (even when m is large). Assuming that treatment times are determined independently of either the known or observed baseline levels, the exact bias is given by

$$E\left(\hat{\delta}_{(1)} - \delta\right) = \frac{\bar{w}\left[1+(N-1)\rho\right]}{\rho\bar{w} + (1-\rho)\bar{b}\left(1-\bar{b}/N\right)} \sum_{i=0}^{m}\left(\frac{i}{m\bar{w}} - \frac{N-\bar{b}}{N\bar{w}}\right)\sum_{j=b_i+1}^{b_{(i+1)}} \beta_j, \tag{4}$$

where $b_0 = 0$ and $b_{m+1} = N$. Clearly, there is no reason to expect the bias to be zero unless each of the occasion-specific common shocks $\beta_1, ..., \beta_N$ is zero.

Second, the naïve statistical analysis is incongruent with best practices for visual analysis of the MBD, where a "vertical analysis" is used to assess whether the introduction of treatment for one case coincides with changes in the pattern of outcomes for other cases (Horner et al., 2012). Including occasion-specific common shocks is the statistical analogue of the vertical analysis technique.

Finally, Model (1) does not allow the analyst to quantify the extent to which cases are subject to common influences, even though this bears directly on internal validity. Intuitively, a design in which six cases are sampled from six separate schools does not provide the same strength of control as a design in which six cases are sampled from the same school. In the

former case, the only influences controlled are those that are common across schools; in the latter case, influences that are unique to the school may also be ruled out. Under Model (3), an index for quantifying the magnitude of common shocks can be defined by comparing the variation between estimated common shocks to the variation in the outcome (both within and across cases) on a given measurement occasion. Let $\hat{\beta}_1,...,\hat{\beta}_N$ be the WLS estimates of the common shocks, with estimated covariance between shocks *j* and *k* denoted $v_{jk} = \text{Var}\left(\hat{\beta}_j, \hat{\beta}_k\right)$. Define the index

$$Q_{shock} = \frac{1}{\tau^2 + \sigma^2} \max\left(0, \quad \sum_{j=1}^{N} \frac{\left(\hat{\beta}_j - \bar{\beta}\right)^2}{N-1} - \sum_{j=1}^{N} v_{jj} + \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{N} v_{jk}\right), \tag{5}$$

where $\bar{\beta} = \frac{1}{N}\sum_{j=1}^{N}\hat{\beta}_j$. The Q_{shock} index is strictly positive, with a value of zero indicating no common shocks and larger values corresponding to greater control.

Usefulness / Applicability of Method:

I estimated Models (1) and (3) on each of four outcomes from a MBD reported by Laski, Charlop, and Shreibman (1988). Table 1 reports standardized treatment effect estimates from each model along with the value of the Q_{shock} index and the p-value from a test of the hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_N$. For only one outcome measure—parent verbalization in free-play settings—is there strong statistical evidence of period-specific common shocks. For this outcome, controlling for common shocks substantially reduces the treatment effect estimate, from 1.57 s.d. to 0.42 s.d. The lack of evidence for common shocks in the other outcomes may suggest that the staggered treatment introduction in this design does not offer much additional control, or it may merely indicate low power to detect such shocks.

Conclusions:

Under triage on known baseline ranks, the bias of the treatment effect estimator can be removed either by treating the case-level intercepts as fixed effects or by group-centering the treatment indicator. Future work on statistical models for single-case designs will need to attend carefully to the possibility of bias induced by correlation between treatment assignment times and case-level characteristics. Further, it would be beneficial if applied single-case researchers described their treatment assignment procedures in greater detail; as the above example shows, "deliberate" assignment does not necessarily mitigate selection threats to internal validity.

I have offered several arguments regarding the benefits of analytic models that control for common history threats. Compared to models without common shocks, these models have distinctly different implications for statistical power, which need to be explored further. While I have focused on perhaps the simplest possible specification, models that include more complex features (such as case-specific time trends) are also possible and warrant further study. Finally, further statistical and empirical work is needed in order to understand the properties of the Q_{shock} index and the range of values likely to be observed in practice.

Appendices

Appendix A. References

Center, B. A., Skiba, R. J., & Casey, A. (1985). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education, 19*(4), 387–400.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology*, *61*(6), 966–74.

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods, 41*(2), 372–84. doi:10.3758/BRM.41.2.372

Gast, D. L., & Spriggs, A. D. (2010). Visual analysis of graphic data. In D. L. Gast (Ed.), *Single Subject Research Methodology in Behavioral Sciences* (pp. 199–233). New York, NY: Routledge.

Hedges, L. V, Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods*, *3*, 224–239. doi:10.1002/jrsm.1052

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods*, (October 2012), n/a–n/a. doi:10.1002/jrsm.1086

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. L., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*(2), 165–179.

Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education and Treatment of Children*, *35*(2), 269–290. doi:10.1353/etc.2012.0011

Hussey, M. A., & Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials, 28*(2), 182–91. doi:10.1016/j.cct.2006.05.007

Kazdin, A. E. (2011). *Single-Case Research Designs: Methods for Clinical and Applied Settings*. New York, NY: Oxford University Press.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2012). Single-case intervention research design standards. *Remedial and Special Education*, *34*(1), 26–38. doi:10.1177/0741932512452794

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, *15*(2), 124–144. doi:10.1037/a0017736

Laski, K. E., Charlop, M. H., & Schreibman, L. (1988). Training parents to use the natural language paradigm to increase their autistic children's speech. *Journal of Applied Behavior Analysis*, *21*(4), 391–400.

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keeffe, B. V, Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology*, *49*(3), 301–21. doi:10.1016/j.jsp.2011.03.004

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton, Mifflin and Company.

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods*, 1–43.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, *43*(4), 971–80. doi:10.3758/s13428-011-0111-y

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, *17*(4), 510–50. doi:10.1037/a0029312

Van den Noortgate, W., & Onghena, P. (2003). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly*, *18*(3), 325–346.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 142–151. doi:10.1080/17489530802505362

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Appendix B. Tables and Figures

Figure 1. Standardized bias of weighted least squares estimator under the triage mechanism, for varying values of the intra-class correlation (ρ), phase length (n), and number of cases (m). It is assumed that treatment assignment times are equally spaced, so that $b_i = i \times n$ and the total number of measurement occasions is $N = n \times (m + 1)$.

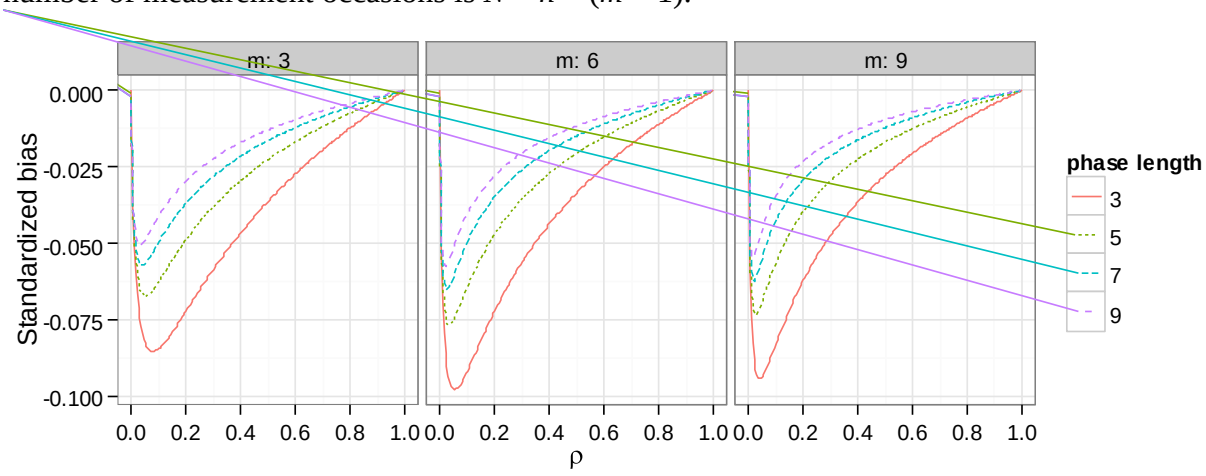

Table 1. Standardized treatment effect estimates, Qshock index values, and p-values from test for no period-specific shocks, as applied to outcome measures from MBD by Laski, Charlop, & Shreibman (1988).

Outcome measure	Standardize treatment effect estimate (standard error)		Q_{shock}	*p*-value for $\beta_1 = \beta_2 = \cdots = \beta_N$
	Model (1)	Model (3)		
Child vocalizations – free play setting	1.47 (0.12)	1.46 (0.25)	0.0	0.909
Parent verbalizations – free play setting	1.76 (0.11)	1.60 (0.24)	0.0	0.994
Child vocalizations – break room setting	1.35 (0.16)	0.96 (0.37)	0.0	0.676
Parent verbalizations – break room setting	1.57 (0.22)	0.42 (0.40)	19.5	0.003