Research in Autism Spectrum Disorders 8 (2014) 933–943

Contents lists available at ScienceDirect

Research in Autism Spectrum Disorders

Journal homepage: http://ees.elsevier.com/RASD/default.asp

False positives with visual analysis for nonconcurrent multiple baseline designs and ABAB designs: Preliminary findings☆

Marissa A. Novotny[a], Katheryne J. Sharp[a], John T. Rapp[b,*], Joel D. Jelinski[a], Elizabeth A. Lood[a], Ayriel K. Steffes[a], Monica Ma[c]

[a] St. Cloud State University, United States
[b] Auburn University, United States
[c] California State University, Sacramento, United States

ARTICLE INFO

Article history:
Received 16 April 2014
Received in revised form 18 April 2014
Accepted 21 April 2014
Available online 21 May 2014

Keywords:
AB designs
ABAB designs
False positives
Nonconcurrent multiple baseline designs
Reversal designs
Single-subject designs

ABSTRACT

This study evaluated the probability of generating false positives with three-tier nonconcurrent multiple baseline (NMBL) designs and ABAB designs. For Experiment 1, we generated four sets of three-tier NMBL design graphs. The first, second, and third sets consisted of fixed A-phase data points for all three tiers at 0%, 25% and 50%, respectively, and randomly generated data points in the B phases. The fourth set consisted of randomly generated data points in the A and B phases for all three tiers. Across all four sets ($N = 1000$), results show that false positives were produced with 7.5% of three-tier NMBL design graphs and were most probable when baseline levels were set at 0% or 25%. For Experiment 2, we generated 3000 ABAB design graphs consisting of three to five data points per phase. Results indicate that no false positives were produced, regardless of the number of data points included in each phase. Results of this study support specific guidelines for the use of NMBL designs and ABAB designs.

1. Introduction

Smith (2012) systematically reviewed treatment studies that utilized single-subject designs and were published between the years 2000 and 2010. In part, Smith found over 400 articles published in various peer-reviewed journals, most of which have applied focus (e.g., applied behavior analysis, rehabilitation, education, sports performance). In a similar endeavor, Shadish and Sullivan (2011) found 113 studies published in the year 2008 across 21 journals, most of which had applied focus. To this end, Dallery, Cassidy, and Raiff (2013) recently asserted that data analysis with single-subject designs prioritizes “clinically significant change over statistically significant change” (p. 8). Bulkeley, Bundy, Roberts, and Einfeld (2013) argued that single-subject designs offer a “contemporary solution” to many problems related to the conduct of clinical research for individuals with Autism Spectrum Disorders. In particular, Bulkeley et al. (2013) noted the potential utility of the multiple baseline across participants design for evaluating various interventions in “real-world contexts.”

☆ Portions of this study were conducted in partial fulfillment of a Master degree in Applied Behavior Analysis at St. Cloud State University, St. Cloud, MN by the first and second authors.

* Corresponding author at: Department of Psychology, Auburn University, 226 Thach, Auburn, AL 36849-5214, United States.
E-mail address: jtr0014@auburn.edu (J.T. Rapp).

http://dx.doi.org/10.1016/j.rasd.2014.04.009
1750-9467/

Consistent with that assertion, Smith (2012) and Shadish and Sullivan (2011) found that over 330 articles in their reviews utilized multiple baseline designs.

Over the past four decades, experimenters and practitioners in applied behavior analysis have used many variations of multiple baseline designs (i.e., across participants, behaviors, and settings) to evaluate training programs, educational programs, and other treatments programs that influence overt behavior (Cooper, Heron, & Heward, 2007; for a brief history, see Matson, Turygin, Beighley, & Matson, 2012). More recently, researchers and practitioners in others areas of psychology (Smith, Erard, & Handler, 2013) and related disciplines such as preventative medicine (Hawkins, Sanson-Fisher, Shakeshaft, D'Este, & Green, 2007) have extolled the virtues of multiple baseline designs. Across various disciplines, there appears to be a general increase in the use of multiple baseline across participants or groups designs. According to Baer, Wolf, and Risley (1968), experimental control is shown in a multiple baseline design only if behavior change is seen when treatment is applied to at least two baselines. For concurrent multiple baseline (CMBL) designs, Cooper et al. (2007) and Kazdin (2011) recommended using at least three baselines or tiers (i.e., a data series for a targeted variable), with more tiers strengthening the demonstration of experimental control by virtue of the replicated treatment effect. Similarly, Smith (2012) noted that the American Psychological Association Division 12 Task Force (Chambless & Ollendick, 2001) recommended using at least three successive demonstrations (i.e., three tiers) within multiple baseline designs in order to make meaningful conclusions about the effects of a given intervention. In addition, general guidelines for single-subject research typically include at least three data points per phase (e.g., Cooper et al., 2007; Kratochwill et al., 2010; Matson et al., 2012).

As an alternative to the CMBL design, Watson and Workman (1981) proposed the nonconcurrent multiple baseline design (NMBL) for researchers and practitioners to use when participants with the same problem behavior or diagnosis were not concurrently available. In general, the number of tiers required for an NMBL design is the same as for a CMBL design. Although the flexibility offered by NMBL designs may be appealing to both researchers and practitioners, the manner in which experimental control is demonstrated with NMBL and CMBL designs is slightly different.

Carr (2005) described the similarities in CMBL and NMBL across participants designs. Carr argued that both designs use baseline logic and repeated measurement, which allow for the production of baseline data paths to predict the course of the dependent event in the absence of treatment. Following the introduction of an intervention, this prediction allows researchers or practitioners to detect differences between the predicted path and the actual path. In addition, both designs involve replicating the effects of the independent variable (IV) by putting multiple participants through baseline and intervention phases. By contrast, Carr suggested that the two designs differ insofar as only the CMBL design allows for the IV effect to be verified via vertical analysis. This verification is made when behavior change occurs only for the participant in the intervention phase, while the other baselines remain unchanged (Carr, 2005; Cooper et al., 2007). According to Carr, the IV effect cannot be verified with a NMBL design using a vertical analysis because the baselines may not be contemporaneous. Carr noted that the IV effect could be verified with a NMBL design if the experimenter or practitioner performs a reversal (i.e., withholds the IV) with one participant after treatment has been implemented for all participants. However, the process described by Carr may not be possible for dependent variables (DVs) that are not reversible.

Others have suggested that the difference between CMBL and NMBL designs is broader. For example, Cooper et al. (2007) contend that the absence of concurrent measurement in the NMBL design eliminates the experimental logic of the multiple baseline design. Interestingly, both Watson and Workman (1981) and Christ (2007) have argued that NMBL designs rule out numerous threats to internal validity for many interventions *because* the participants are not contemporaries and, therefore, extraneous variables are not likely to affect each individual during their participation in a study. Nevertheless, Christ suggested that NMBL designs may be more vulnerable to threats of mortality (i.e., loss of participants), which, in turn, may reduce the number tiers that are subjected to visual inspection.

To illustrate the broad practical utility of NMBL designs, researchers have used variations of the design to evaluate individual interventions for nighttime sleep disturbances (France & Hudson, 1990), parent training (e.g., Lequia, Machalicek, & Lyons, 2014), decreasing skin picking (Twohig & Woods, 2001), increasing math and other academic skills (Lerman, Vorndan, Addison, & Kuhn, 2004; Rapp et al., 2012), decreasing marijuana use (Twohig, Shoenberger, & Hayes, 2007), increasing compliance with medical prostheses (Richling et al., 2011), and increasing appropriate toilet use (LeBlanc, Carr, Crossett, Bennett, & Detweiler, 2005), as well as group interventions in classrooms (e.g., Donaldson, Vollmer, Krous, Downs, & Berard, 2011). In addition to its utility as a research tool, the NMBL design is particularly well suited to a wide range of practical settings for at least two reasons. First, it can be used to experimentally evaluate interventions for multiple clients, none of whom need be contemporaries, without withdrawing the intervention (Harvey, May, & Kennedy, 2004). Second, it can be used to demonstrate the effectiveness of interventions, which are derived from either single-subject or group-design research, for individual clients within clinical settings (Kazdin, 2011). Specifically, the AB components of a NMBL design lend themselves well to ongoing data collection and treatment evaluation. In addition, Harvey et al. (2004) argued that NMBL designs afford the flexibility that is needed to analyze various aspects of educational practices for individuals or groups of individuals.

Several textbooks provide guidelines for using multiple baseline designs and others single-subject designs (Barlow, Nock, & Herson, 2009; Cooper et al., 2007; Johnston & Pennypacker, 2009; Kazdin, 2011; Miltenberger, 2012); however, there is very limited empirical support for most of these guidelines (Krueger, Rapp, Ott, Lood, & Novotny, 2013). To our knowledge, no study has evaluated the production of false positives with either NMBL designs or ABAB designs (but see Kahng et al., 2010 for a recent discussion and study of visual analysis with ABAB designs). Likewise, there are no clear guidelines for determining an acceptable level of false positives for each design; however, Bartlett, Rapp, and Henrickson (2011) suggested