Evaluation and Program Planning 70 (2018) 83–89

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan

Revisiting the utility of retrospective pre-post designs: The need for mixed-method pilot data

G. John Geldhof[a,*], Danielle A. Warner[a], Jennifer K. Finders[a], Asia A. Thogmartin[a], Adam Clark[b], Kelly A. Longway[a]

[a] Human Development and Family Studies, College of Public Health and Human Sciences, Oregon State University, United States
[b] Family Studies and Human Development, University of Arizona, United States

ARTICLE INFO

Keywords:
Retrospective pretest
Response shift bias
Thentest
Parenting education
Parenting skills ladder

ABSTRACT

The retrospective pre-post design affords many benefits to program staff and, accordingly, has piqued renewed interest among applied program evaluators. In particular, the field has witnessed increasing application of a post-program-only data collection strategy in which only posttest and retrospective pretest data are collected. A post-program-only assessment strategy takes considerably less time than is required for collecting pre-program data and presumably has the added benefit of eliminating the impact of response-shift bias. Response-shift bias occurs when the knowledge, skills, or experiences participants gain through program participation leads them to interpret questionnaire items in a qualitatively different manner at pretest versus posttest. In this article, we discuss the strengths and weaknesses associated with administering retrospective pretest assessments and underscore the importance of thoroughly evaluating any application of a retrospective measurement strategy prior to its broader implementation. We provide a practical illustration of this evaluation process using a mixed-method study that assesses one measure of parenting education program effectiveness—the Parenting Skills Ladder.

1. Introduction

Applied programs aim to change participants' real-world behaviors. Accordingly, social and behavioral scientists have long preferred objective assessments of behavioral change over self-reported change (e.g., Howard, Schmeck, & Bray, 1979). Direct behavioral assessments are not readily available for all program-relevant outcomes, however, and even these seemingly objective measures can be fraught with bias. Take, for instance, the behaviors most directly relevant to assessing a parenting education curriculum—parents' daily interactions with their children. Such behaviors likely have strong consequences for children's development (e.g., Burchinal, Peisner-Feinberg, Pianta, & Howes, 2002), yet these behaviors are wholly inaccessible to researchers, program evaluators, and program staff. Observing participants' parenting behaviors in person would be prohibitively expensive (in terms of both time and financial resources), and these observations still will not reliably reflect parents' actual behaviors behind closed doors. Despite researchers' best efforts to create comfortable, realistic settings for observation, participants will inevitably behave differently when being observed than when they are alone (i.e., participant reactivity; see Gravetter & Forzano, 2012, for a discussion).

Given the costs associated with obtaining direct assessments that, in the end, may still be prone to biases, program staff often seek simpler, more cost-efficient measurement strategies. For instance, self-report questionnaires are inexpensive and are flexible enough to assess a wide variety of behaviors, attitudes, knowledge, and beliefs in a time-efficient manner. Thus, self-report pre-post designs are a common approach to assessing program effectiveness. In this design, program participants complete a self-report assessment prior to beginning the program (i.e., a true pretest) and again after program completion (i.e., a posttest). Although self-report measures certainly face their own limitations (e.g., Nisbett & Wilson, 1977; Schwarz, 2007), a simple difference between pretest and posttest scores can be taken as an approximation of participant change when control group data are not available (i.e., an evaluator must assume the difference in observed scores accurately reflects differences in the underlying constructs as caused by program participation).

Self-report pre-post designs present a unique constellation of difficulties, however. The difference between a participant's scores at pretest and posttest may be influenced by a host of potential confounding factors. Observed differences reflect actual gains caused by program participation, but they also reflect how participants'

* Corresponding author at: 410 Waldo Hall, Oregon State University, Corvallis, OR, 97331, United States.
E-mail address: john.geldhof@oregonstate.edu (G.J. Geldhof).

https://doi.org/10.1016/j.evalprogplan.2018.05.002
Received 22 August 2017; Received in revised form 3 April 2018; Accepted 16 May 2018
Available online 09 July 2018
0149-7189/

perceptions of questionnaire items changed over the course of program participation. Taking the example of a parenting education series, parents may simply 'not know what they do not know' before completing the program (Pratt, McGuigan, & Katzev, 2000). Their pretest self-reports may therefore paint an overly rosy picture of their parenting skills.

The new knowledge gained through program completion will often lead participants to realize their actual skill levels at pretest were lower than they originally thought. This change is often part and parcel to a program's overall impact (Nolte, Elsworth, Sinclair, & Osborne, 2012) but also draws the validity of comparing pretest and posttest measurements into question. Such a change in participants' interpretation of survey items, called response-shift bias (Howard & Dailey, 1979), violates a critical assumption of pre-post measurement designs described by Cronbach and Furby (1970)—that pretest and posttest scores are evaluated using a common metric.

In a seminal series of papers, Howard et al. (e.g., Howard & Dailey, 1979; Howard, Ralph et al., 1979; Howard, Schmeck et al., 1979) proposed the retrospective pre-post design as a useful tool for measuring, and potentially overcoming, response-shift bias. Like a standard pre-post design, the retrospective pre-post design requires that participants respond to a set of questions administered both before and after participants complete a program. Participants also complete a third assessment, the retrospective pretest, at the end of the program. The retrospective pretest asks participants to think back about their attitudes, beliefs, and/or behaviors prior to beginning the program and to respond to questions in a way that reflects what they were like 'then.' Retrospective pretest scores can then be compared to true pretest and posttest data to determine the extent of a program's impact as well as the potential magnitude of response shift bias.[1] This approach may be especially useful for survey items that measure attitudes and beliefs because of their highly subjective nature.

Retrospective pretests consistently indicate larger program effects than direct comparisons of true pretest and posttest scores (e.g., Hill & Betz, 2005; Hoogstraten, 1985; Pratt et al., 2000; Sibthorp, Paisley, Gookin, & Ward, 2007). Mixed-methods studies have found that participants describe response-shifts as one underlying cause of observed discrepancies between true pretest and retrospective pretest scores (Howard, Ralph et al., 1979; Sibthorp et al., 2007). Furthermore, retrospective pretests tend to correlate more strongly with more-objective measures of pre-program skills than do actual pretest self-assessments (Howard, Ralph et al., 1979; Howard, Schmeck et al., 1979; Pratt et al., 2000).

Empirical research has helped elucidate some of the specific causes of response-shift bias. For instance, Hoogstraten (1985) found that the presence of an objective measurement may prevent response-shifts. Nolte and colleagues found that the presence of a retrospective pretest assessment can increase posttest scores (Nolte et al., 2012). Sprangers and Hoogstraten (1989) also found that situation-relevant role-playing can reduce response-shift bias, presumably by giving participants a realistic situation against which they can form self-assessments prior engaging in a program. A substantial body of literature therefore supports the retrospective pre-post design as a promising technique and clarifies conditions when it may be more versus less effective.

The empirical support for using retrospective pretests has led to a variation of the original design that has become especially attractive to program administrators. This variant, which we call the single-assessment retrospective pre-post design (SARPPD), omits the pre-program assessment altogether, resulting in an assessment that occurs only after program completion (e.g., Davis, 2002; Dolenc-Nott, Peters, Sektnan, Rennekamp, & Bowman, 2015; Lam & Bengo, 2003; Rockwell & Kohn, 1989). As these sources note, eliminating the true pretest offers many attractive benefits for program administrators. For instance, administering a single post-program evaluation frees up valuable time and money that staff can instead dedicate to administering their programs (Pratt et al., 2000). The time freed by not administering a true pretest also comes during a critical stage in program implementation—omitting the pre-program survey allows staff to build stronger rapport with participants before asking them to respond to potentially sensitive questions. As an example, many parents enrolled in parenting education series are mandated to attend those series by a legal body (e.g., Child Protective Services, the courts). These parents may initially be reluctant to divulge negative parenting behaviors to program staff. In such a situation, pretest data may not only be tainted by parents' desire to look better than they are, but the very act of asking these questions may also lead parents to mistrust course facilitators.

Omitting the pre-program assessment therefore confers advantages beyond the potential for eliminating response-shift bias. The presumed benefits of a SARPPD have accordingly led evaluation specialists to encourage their wider implementation (Lamb, 2005). Relying only on post-program data collection exacerbates the many limitations to retrospective pre-post evaluation, however, which we describe next. These limitations draw the utility of SARPPDs into serious question.

1.1. Problems with retrospective responses

Response-shift bias presents a substantial problem in evaluation research, but the empirical findings that support retrospective designs should not be taken as unequivocal evidence for their validity. Indeed, the research and evaluation literature has long questioned the validity of retrospective reports (e.g, Campbell & Stanley, 1966). In particular, retrospective reports can be biased by both the fallibility of human memory and by participants' competing desires.

Retrospective assessments are prone to recall bias due to the distortion and/or degradation of memory (Hill & Betz, 2005; Schwartz & Sprangers, 2010). For instance, Wilson and Ross (2001) presented a series of studies supporting what they termed Temporal Self-Appraisal Theory. Results from these studies suggested that participants were more likely to report changes that make themselves look better in the present. In other words, participants were especially critical of their past selves when doing so would not negatively impact perceptions of their current selves. Whether this bias was conscious or subconscious was unclear, however. Applied to the example of a parenting education series, participants may therefore retrospectively inflate their inadequacy as parents before completing a parenting education series because doing so makes the participants look like better parents in the present.

A similar critique draws on Ross' (1989) assertion that people tend to apply implicit theories of change when providing retrospective reports (e.g., Hill & Betz, 2005; Schwartz & Rapkin, 2012; Schwartz & Sprangers, 2010). Because human memory is imperfect, people often use implicit theories of change to help fill in missing information. For instance, most individuals hold an implicit theory that their attitudes and beliefs remain consistent over short periods of time (e.g., a few years). People therefore tend to use their current beliefs and attitudes to inform ratings about their prior selves. Other implicit theories anticipate change, however. For instance, it is likely that parenting education program participants will adopt an implicit theory that the program improved their parenting skills regardless of whether actual change occurred. This implicit theory could then lead participants to artificially inflate the difference between posttest and retrospective pretest scores.

Retrospective pretest scores can also be biased by social factors (e.g., Hill & Betz, 2005; Robinson & Doueck, 1994; Schwartz &

[1] Some readers may be more familiar with the revised version of this method that omits pre-program assessments. We discuss this variation later, but for now note that Howard, Schmeck et al. (1979, p. 134) recommend, "that, when self-report measures must be employed, the researcher should collect a retrospective pre-rating *in addition to* the conventional pre-rating or posttest comparisons," (italics added). Howard and Dailey (1979) make a similar recommendation.