Measuring Program Outcomes: Using Retrospective Pretest Methodology

Clara C. Pratt, William M. McGuigan and Aphra R. Katzev

American Journal of Evaluation 2000 21: 341

DOI: 10.1177/109821400002100305

Method Notes

This section includes shorter (e.g., 10–15 double-spaced manuscript pages or less) papers describing methods and techniques that can improve evaluation practice. Method notes may include reports of new evaluation tools, products, or services that are useful for practicing evaluators. Alternatively, they may describe new uses of existing tools. Also appropriate for this section are user-friendly guidelines for the proper use of conventional tools and methods, particularly for those that are commonly misused in practice.

Measuring Program Outcomes: Using Retrospective Pretest Methodology

CLARA C. PRATT, WILLIAM M. MCGUIGAN, AND APHRA R. KATZEV

ABSTRACT

This study used longitudinal data from 307 mothers with firstborn infants participating in a home-visitation, child-abuse prevention program. A self-report measure of specific constructs the program hoped to affect showed that the retrospective pretest methodology produced a more legitimate assessment of program outcomes than did the traditional pretest-posttest methodology. Results showed that when response shift bias was present, traditional pretest-posttest comparisons resulted in an underestimation of program effects that could easily be avoided by the retrospective pretest methodology. With demands for documenting program outcomes increasing, retrospective pretest designs are shown to be a simple, convenient, and expeditious method for assessing program effects in responsive interventions. The limits of retrospective pretests, and methods for strengthening their use, are discussed.

INTRODUCTION

With the passage of the federal Government Performance and Results Act (GPRA) in the early 1990s, education, health, and human service programs have experienced a dramatic

Clara C. Pratt • Oregon State University, Department of Human Development and Family Sciences, Family Policy Program, 204 Bates Hall, Corvallis, OR 97331; Tel.: (541) 737-49992; Fax: (541) 737-1076; E-mail: prattc@orst.edu.

American Journal of Evaluation, Vol. 21, No. 3, 2000, pp. 341–349.
ISSN: 1098-2140

increase in outcome accountability (Harty, 1997). The demands for documenting the outcomes for programs have expanded well beyond large, federally funded initiatives to include smaller state, local, and nonprofit, community-based programs (United Way of America, 1996). Increasingly, funding agencies link continued program support to progress toward key program outcomes. In this accountability context, it is critical that the positive effects of programs not be underestimated.

Despite the importance of assessing program outcomes, the accurate measurement of outcomes presents a variety of practical and methodological challenges. For example, experimental and quasi-experimental designs require measuring outcomes before and after an intervention as well as measuring outcomes for a comparison group. Many programs find it difficult or impossible to locate or maintain adequate comparison groups, and other programs lack the time and resources to conduct such complex evaluations (Brooks & Gersh, 1998).

Thus, to track outcomes, most government and nonprofit programs rely on performance measurement strategies rather than more expensive and complicated quasi-experimental and experimental designs. Essentially, performance measurement strategies seek to answer the question: Did the program accomplish what it set out to accomplish? Performance measurement relies heavily on single group designs, utilization of records, staff observations, and participant self-reports (Harty, 1997; Newcomer, 1997). Typically, measures are collected at the beginning of a program (pretest) and again at the end of the program (posttest) with the idea that program effects are demonstrated by differences in the two measures. Pretest-posttest designs allow clients to serve as their own baseline of comparison. These within-group designs are often used in social sciences because they provide greater statistical power than between-group designs.

However, traditional pretest-posttest designs have several limitations, especially when participant self-report measures are used. For example, when time for intervention is limited, pretest-posttest questionnaires consume time that may be better spent on program delivery (Marshak, deSilva, & Silberstein, 1998). Further, for pretest-posttest comparisons to be meaningful, participants must be present when the program begins and ends, yet difficulties with consistent attendance are well documented, especially among programs serving high-risk groups. Most importantly, even when complete pretest-posttest information is obtained, actual changes in knowledge and behaviors may be masked if the participants overestimate their knowledge and skills on the pretest.

Pretest overestimation is likely if participants lack a clear understanding of the attitude, behavior, or skill the program is attempting to affect. Ironically, it is the participants' inexperience and lack of knowledge and skills that often necessitate the program intervention. Taking part in the program may show participants that they actually knew much less than they originally reported on the pretest. In such cases, pretest-posttest comparisons are misleading because participants use a changed frame of reference to classify themselves after engaging in the program (Howard, Ralph, Gulanick, Maxwell, Nance, & Gerber, 1979).

This change in an individual's frame of reference because of program participation has been called the response shift bias (Howard & Dailey, 1979). This bias can be defined as a program-produced change in the participants' understanding of the construct being measured. When participants rate themselves on traditional pre-posttests, program-produced changes in the participants' standards are potential threats to internal validity (Howard et al., 1979).

To avoid response-shift bias, researchers have suggested collecting both contemporary and retrospective information at the conclusion of the program (Goedhart & Hoogstraten, 1992; Terborg, Howard, & Maxwell, 1980). This means that at the end of the program,

participants first report on their current (contemporary) knowledge, behavior, or attitudes. Then participants complete the same self-report measure a second time with reference to where they perceive themselves to have been when the program began. This second measure forms a retrospective pretest. Response shift bias is avoided because participants are rating themselves with a single frame of reference on both the posttest and retrospective pretest. Some studies suggest that a more accurate assessment of changes in self-reported knowledge and behavior may be produced by retrospective pretest designs than by the traditional pretest-posttest design (Goedhart & Hoogstraten, 1992; Terborg, et al. 1980).

METHODS

Context for the Study

The present article compares pretest-posttest methodology to a retrospective pretest methodology. Data for the present analysis were obtained from mothers served by Oregon Healthy Start (OHS) between November 1997 and July 1999. OHS is a primary prevention program designed to prevent child maltreatment and other poor child outcomes. The OHS program targeted families with firstborns in 13 of Oregon's 36 counties. At the time of the child's birth, families were screened and assessed for child abuse risk. Families identified as being at risk for child maltreatment or other poor child outcomes were offered home visits, parenting education, and extensive family support services by trained paraprofessionals. Home visits were offered weekly or biweekly, depending on the families' needs and the counties' caseload limitations. Participation in OHS was voluntary.

Participants

Complete data were available for 307 mothers ($n = 307$) served by OHS during the study period. These mothers were representative of the total population served by OHS. Most of the 307 mothers were single (76%), unemployed (77%), white non-Hispanics (82%), under age 21 (62%), with less than a high school diploma (50.2%). Most mothers (53%) lived with their spouse or boyfriend; over one-third (38%) lived with parents or other relatives. Approximately one-half (56%) had a monthly family income below $1,000.

Procedures

OHS used a battery of measures to assess program outcomes. These included developmental assessments of infants, agency reports of maltreatment rates, observational scales, and parent self-report. Among the self-report measures is a seven-item self-report index, the Parent Ladder, which assesses maternal knowledge of child development, confidence in parenting, basic resources, social support, stress, and coping skills. The seven items on the Parent Ladder represent specific constructs that the OHS intervention hopes to impact.

On average, mothers were enrolled in OHS within 1 to 7 days after giving birth to their first child. At the point of enrollment, mothers were shown a picture of a ladder leaning against a wall (Fig. 1). Using this visual aid, mothers were asked to "place themselves on the ladder" by circling the number to the right of the ladder that indicates (a) "your knowledge of how children grow and develop"; (b) "your confidence that you know what is right for

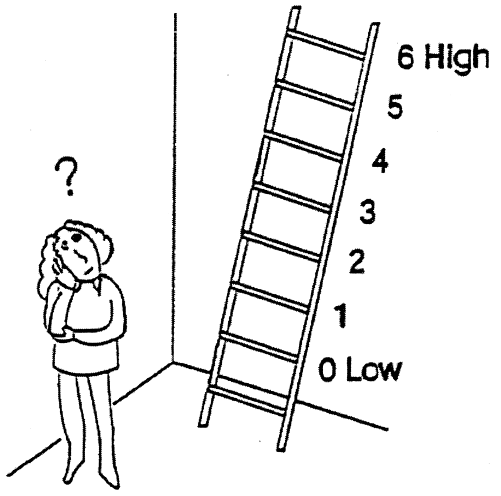

Figure 1. Parent Ladder.

your child"; (c) "your ability to help your child learn"; (d) "your resources, like money, food, and transportation"; (e) "the amount of helpful advice or moral support you get from other people"; (f) "your ability to cope with the stress in your life now"; and (g) "the amount of stress in your life right now." Items were rated on a 7-point scale ranging from low on the ladder (0) to high on the ladder (6). Higher scores indicated greater self-perceived knowledge of child development, confidence in parenting, basic resources, social support, and coping skills. Scores were reversed for the statement "the amount of stress in your life right now" so higher scores indicated less stress. Individual item scores were summed and divided by 7 to produce a total index score ranging from low (0) to high (6).

When their child was 6 months of age, mothers completed the Parent Ladder a second time, describing their current level of functioning on the seven index items. In addition, the 6-month assessment included the retrospective pretest component. Mothers first rated themselves on each of the seven items in terms of "how are you doing now." Then in the next section (retrospective), they were asked to rate themselves on the same seven items again, this time, "thinking back to when your baby was born." This essentially asked mothers to think back to when they completed the traditional pretest Parent Ladder, because that pretest was administered within 1 to 7 days of the child's birth.

Statistical Analysis

As suggested by previous researchers (Bray, Maxwell, & Howard, 1984; Howard et al., 1979), program effects were tested by comparing the difference in mean scores between time 1 and time 2 as measured by the traditional pretest-posttest design. Paired sample *t* tests yield

TABLE 1.
Parent Ladder pretest (Pre), Retrospective Pretest (RPT), and Post-test (Post) Item Means and *t*-values[a]

	Means			*Paired Samples* t-*test*[a]		
Item	*Pre*	*RPT*	*Post*	*Traditional Pre-Post*	*RPT-Post*	*Pre-RPT*
Total Parent Ladder score (mean score of 7 items)	4.04	3.58	4.29	5.84**	15.00***	9.43***
Confidence you know what's right for your child	4.81	3.95	5.17	5.29**	15.66***	9.88***
Knowledge of how children grow and develop	3.84	3.24	4.55	9.98***	17.53***	6.91***
Ability of help your child learn	4.86	4.17	5.13	4.26*	13.77***	8.14***
Resources like money, food, and transportation	3.47	3.57	3.83	3.53*	3.01*	.96
Amount of helpful advance/moral support from others	4.83	4.47	4.79	.44	4.45*	4.39*
The amount of stress in your life right now	2.36	1.93	2.41	.45	4.06*	3.95*
Your ability to cope with the stress in your life	4.17	3.73	4.17	.08	5.18**	4.52*

Note: *$p < .05$, **$p < .01$, ***$p < .001$ (*significance after strict Bonferroni adjustment*).
[a]*For each t-test* df = 306.
[b]*All t-tests are two-tailed.*

results identical to analysis of variance with repeated measures with only one variable; thus, the simpler procedure was used.

Paired sample *t* tests were used to compare mean scores on pretest items assessed at intake with mean scores on the corresponding posttest items assessed after 6 months of program participation. Paired sample *t* tests were also used to compare retrospective pretest scores with the corresponding posttest scores, both assessed after 6 months of program participation. Finally, to examine the existence of response shift bias, paired sample *t* tests were used to compare mean item scores on traditional pretest items with mean item scores on retrospective pretest items. All *t* tests were two-tailed. Because 24 pairwise *t* tests were conducted on the same data set, a strict Bonferroni adjustment was applied to control for inflation of Type I error.

RESULTS

After 6 months in the OHS program, traditional pretest-posttest comparisons (see Table 1) indicated improvements in the mothers' perceived knowledge and skills. Within-subject analyses indicated significant improvement in the total Parent Ladder scores as well as significant improvement on four of the seven Parent Ladder items. Mothers reported higher confidence in knowing what is right for their child, increased knowledge of how children grow and develop, greater ability to help their child learn, and improved resources, such as money, food, and transportation. However, on traditional pretest-posttest comparisons moth-

ers showed no significant improvement in the amount of helpful advice or moral support they received from other people, the amount of stress in their life, or their ability to cope with stress.

In contrast, in the comparison of retrospective pretest scores with posttest scores, mothers showed a significant improvement on all seven Parent Ladder items. To examine response shift bias we compared mean pretest item scores with mean retrospective pretest item scores (Howard, 1982). Consistent with past studies (Conway & Ross, 1984; Goedhart & Hoogstraten, 1992; Howard et al., 1979), Healthy Start mothers rated themselves *lower* on each item of the retrospective pretest than they did on the traditional pretest. Statistically significant differences between the mean scores on pretest items and the mean scores on retrospective pretest items suggested the presence of a response shift bias (Goedhart & Hoogstraten, 1992).

Further examination of individual items indicated that a program-produced change in understanding, or response shift, occurred on the six Parent Ladder items in which respondents were asked to report on some personal characteristic (e.g., knowledge, skill). Only the item concerned with the perception of material goods ("resources like money, food, transportation") was free of response shift. Most importantly, there was evidence of response shift bias on the three Parent Ladder items that failed to show significant change on the traditional prepost tests (see Table 1).

Although response shift is perhaps the most plausible reason for the observed differences in the pretest and retrospective pretest scores, alternative mechanisms may operate under some conditions. For instance, demand characteristics, such as wanting to please the program providers, may affect the change in scores. Implicit theories of change, like thinking that change *should* have occurred (Conway & Ross, 1984) might also have an effect. Finally, memory-related biases such as hindsight bias could also be operating (Hawkins & Hastie, 1990). (For a comprehensive review of the validity of the retrospective pretest, see Schwarz & Sudman, 1993.)

To supply further evidence of a program-produced response shift and to address the possibility of other causes, the study sample was divided into two groups based on the number of home visits received. A program-produced response shift hypothesis would suggest differential effects for these two groups, whereas the alternative hypotheses (e.g., demand characteristics, implicit theories of change, and memory-related biases) would not. Among the 57 (19%) mothers who received approximately nine home visits ($m = 8.52$, $SD = 1.7$) during the 6-month study period, there was evidence of a program-produced change, or response shift bias, on only three Parent Ladder items: confidence in knowing what is right for their child, knowledge of how children grow and develop, and ability to help their child learn. Among the 250 (81%) mothers who averaged 19 ($m = 18.8$, $SD = 5.4$) home visits, there was evidence of a response shift on all of the Parent Ladder items except for the item concerned with the perception of material goods. Mothers for whom the program was implemented with more intensity shifted their frame of reference about their initial knowledge and skill levels more so than mothers who received a smaller "dose" of the program. These results support the hypothesis that the program produced the response shift.

As a final examination of the validity of the retrospective methodology we compared one of the self-report Parent Ladder items ("knowledge of how children grow and develop") with two similar, but more objective, measures. During the first month of home visits, family support workers used a 5-point scale to rate mothers on the frequency of: (a) "demonstrating knowledge of the babies needs and interests" and (b) "accurately interpreting the babies'

signals and cues; understanding the baby's behavior." Both of these items were significantly correlated with the mothers' retrospective pretest score on "knowledge of how children grow and develop" ($r = 0.27$, $p < .01$ and $r = 0.22$, $p < .05$, respectively), but were not significantly correlated with her corresponding traditional pretest score ($r = 0.10$, *ns* and $r = 0.04$, *ns*, respectively). As in past research (Howard, et al., 1981), the retrospective pretest scores were more highly correlated with the more objective measures than were the standard pretest scores. Although these correlations were modest, they provide further evidence of the validity of the retrospective pretest methodology.

DISCUSSION

This study confirms that when response shift bias is present, a retrospective pretest methodology produces a more legitimate assessment of program outcomes than does traditional pretest-posttest methodology. After participating in the OHS program for 6 months, mothers shifted their frame of reference about their initial knowledge and skill levels. On six of the seven index items, mothers' retrospections about their initial abilities were lower than their original pretest ratings. In addition, relative to their retrospective pretests, mothers reported significant improvements in every area. Traditional pretest-posttest comparisons failed to detect some of these improvements, resulting in an underestimation of program effects. This illustrates the primary reason for using the retrospective pretest methodology: to avoid the bias resulting from response shift.

In addition, retrospective designs are a simple, convenient, and expeditious method of assessing changes in self-reported knowledge and skills. The retrospective pretest has an added advantage in that it is only administered a single time. Collecting outcome information only at the end of a program conserves valuable instruction time and requires less complicated data management than traditional pretest-posttest evaluation designs.

The retrospective pretest method is also extremely flexible because questions can be designed to reflect actual program content as it evolves over the time of an intervention. This is especially important with educational and supportive interventions, which attempt to respond to evolving, dynamic needs of participants. Such responsive interventions are increasingly recognized as a "best practice" in family and other supportive programs (Family Resource Coalition of America, 1998). Furthermore, the retrospective approach allows researchers to gather information that would be impossible to gather in a prospective fashion, such as information before an unforeseen traumatic event (Toedter, Lasker, & Campbell, 1990).

In spite of the advantages of the retrospective methodology, its limitations must be acknowledged. Evaluators considering retrospective pretests must consider demand characteristics and memory-related problems that influence the recall process. Demand characteristics may be especially problematic in programs where clients have a subjective motivation to make the program look good (e.g., clients gain from a positive appraisal of the program). Among the most salient memory-related biases are the length and specificity of the time period that is being recalled. Clarifying a defined period, such as "since you began this program," may facilitate recall. In addition, it is prudent to formulate questions in a manner that enhances the recall of events. Behaviors that are more specific are easier to recall and assess than are behaviors that are more global. For example, it is likely that parents can better

assess their current and past ability to "help their child learn" or to "cope with stress" than to assess "how good a parent they are."

Finally, the level of recall accuracy available from any self-report must be considered. Despite the fact that response-shift bias is reduced by retrospective pretests, self-reports remain a form of estimation. Sprangers and Hoogstraten (1991) contend that even the best self-report methodology has the potential for subject bias when subjects voluntarily try to improve their skills. The retrospective methodology also is not free from other possible biases. For example, in some programs, change over time may be influenced by regression to the mean, secular drift, interfering events, maturational effects, or other well-known threats to validity (Rossi & Freeman, 1993).

Although retrospective pretest scores have been shown to be more highly correlated with objective measures than standard pretest scores (Howard et al., 1981), programs may be wise to include both objective measures and self-report measures when substantial precision is required. Stratifying clients by exposure to treatment, as well as retrospectively measuring items for which response shift is not likely to occur, are two techniques that could be used to enhance the retrospective pretest design. Where feasible, including both traditional and retrospective pretests with objective measures would provide evidence about the conditions under which the retrospective pretest design is preferable to a traditional pretest-posttest design.

Despite these caveats, the retrospective pretest methodology offers an effective, manageable strategy to reduce the underestimation of program effects when participant self-report measures are used. In the current contexts of performance measurement and results accountability, reducing underestimation of program effects may be particularly critical.

ACKNOWLEDGMENTS

Data for this study were provided by the Oregon Healthy Start Evaluation (97-59), awarded to Oregon State University Family Policy Program by the Oregon Commission on Children and Families.

REFERENCES

Bray, J. W., Maxwell, S. E., & Howard, G. S. (1984). Methods of analysis with response shift-bias. *Educational and Psychological Measurement, 44*, 781–804.

Brooks, L., & Gersh, T. L. (1998). Assessing the impact of diversity initiatives using the retrospective pretest design. *Journal of College Student Development, 34*, 383–385.

Conway, M., & Ross, M. (1984). Getting what you want by revising what you had. *Journal of Personality and Social Psychology, 47*, 738–748.

Family Resource Coalition of America. (1998). *How are we doing?: A program self-assessment toolkit for the family support field*. Chicago, IL: author.

Goedhart, H., & Hoogstraten, J. (1992). The retrospective pretest and the role of pretest information in evaluation studies. *Psychological Reports, 70*, 699–704.

Hatry, H. (1997). Where the rubber meets the road: Performance measurement for state and local public agencies. *New Directions for Evaluation, 75*, 31–44.

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgements of past events after the outcomes are known. *Psychological Bulletin, 107*, 311–327.

Howard, G. S. (1982). Improving methodology via research on research methods. *Journal of Counseling Psychology, 29*(3), 318–326.

Howard, G. S., & Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology, 64*, 144–150.

Howard, G. S., Millham, J., Slaten, S., & O'Donnel, L. (1981). Influence of subject response style effects on retrospective measures. *Applied Psychological Measurement, 5*, 89–100.

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D., & Gerber, S. L. (1979). Internal invalidity in pretest-posttest self-report evaluations and the re-evaluation of retrospective pretests. *Applied Psychological Measurement, 3*, 1–23.

Marshak, S. H., deSilva, P., & Silberstein, J. (1998). Evaluation of a peer-taught nutrition education program for low-income parents. *Journal of Extension, 27*, 19–21.

Newcomer, K. (1997). Using performance measurement to improve public and non-profit programs. *New Directions for Evaluation, 75*, 5–14.

Rossi, P. H., & Freeman, H. E. (1993). *Evaluation: A systematic approach* (5th ed.). Newbury Park, CA: Sage.

Schwarz, N., & Sudman, S. (Eds.). (1993). *Autobiographical memory and the validity of retrospective reports*. New York: Springer-Verlag.

Sprangers, M., & Hoogstraten, J. (1991). Subject bias in three self-report measures of change. *Methodika, 5*, 1–13.

Terborg, J. R., Howard, G. S., & Maxwell, S. E. (1980). Evaluating planned organizational change: A method for assessing alpha, beta and gamma change. *Academy of Management Review, 5*, 109–121.

Toedter, L. J., Lasker, J. N., & Campbell, D. T. (1990). The comparison group problem in bereavement studies and the retrospective pretest. *Evaluation Review, 14*, 75–90.

United Way of America. (1996). *Measuring program outcomes: A practical approach*. Alexandria, VA: Author