Synthesis of Literature Relative to the Retrospective Pretest Design

John Klatt and Ellen Taylor-Powell
University of Wisconsin-Extension

I Introduction

The retrospective pretest grew out of work by Donald Campbell and Julian Stanley (1963). Campbell and Stanley (1963) outline nine threats to internal validity, the extent to which an evaluator can determine a cause – effect relationship by adequately ruling out alternative explanations. They argue that random assignment of participants to treatment and control conditions is the best method of controlling each of the nine threats. In 1979, George Howard (Howard & Daily, 1979; Howard, Ralph, Gulanick, Maxwell, Nance, & Gerber, 1979) proposed a threat to internal validity that Campbell and Stanley had not considered and that random assignment of participants to control and treatment conditions could not attenuate. This threat is based on instrument effects (Howard, 1980; Howard & Dailey, 1979; Howard, Ralph, et al. 1979), or results produced by the measurement instrument rather than the treatment. Howard called this threat to internal validity "response shift bias" and described it as a change in the participant's metric for answering questions from pretest to posttest due to a new understanding of the concept being investigated (Howard, 1980; Howard, Ralph, et al. 1979). Response shift occurs when participants, rating themselves on self-report measures, use a different internal standard between ratings. To reduce response shift bias, Howard proposed using the retrospective pretest (RPT). Although Campbell and Stanley (1967) had discussed the possible contributions of a retrospective pretest to experimental designs, the notion of response shift was new and created fresh interest in the retrospective design.

With the increasing demand for accountability and measurement of change, the retrospective pretest design has gained prominence as a convenient, valid method for measuring self-reported change. It has been shown to reduce response-shift bias providing more accurate assessments of actual effect, is convenient to implement, provides comparison data in the absence of "pre" data, and may be more appropriate in given situations. This review was undertaken to examine the published literature relative to the retrospective pretest design to see what we know about its use, relevance and validity. It covers 49 articles representing sources from educational measurement, psychology, sociology, health, agricultural education, evaluation, extension, management, training, and social work.

Panel presentation for 2005 Joint CES/AEA Conference**,** Toronto, October 29, 2005
Session Title: More on Retrospective Pre-Test: Developing a Taxonomy of Best Practice Uses

The reviewed articles cluster into those that:

1. report empirical research conducted on testing the design
2. report the use of the design in program evaluation (some do and some don't report results related to validity of the design)
3. describe or advocate the design and are considered "instructional materials"

The RPT goes by a number of names: may be referred to as retrospective posttest; post-then-pre; post-then; then tests; retrospective pretest posttest; after retrospective, and RPT ratings. For brevity, we will refer to it as RPT in the following review.

II Research Review - Overview of articles (see accompanying chart)

1. Empirical research testing the design

Much of the early research on response shift bias and the RPT was conducted by Howard and his colleagues. Their research initially focused on questions related to the utility of the RPT. Howard and colleagues (Howard, & Dailey, 1979; Howard, Ralph, et al. 1979; Howard, Schmeck, & Bray, 1979) asked (a) if response shift actually exists, (b) if the RPT provides different information than the traditional pre–post design, and (c) if it does, which method is more valid. Howard, Ralph, et al. (1979) report a series of studies that address these three questions.

(a) Existence of response shift. The first study reported in Howard, Ralph, et al. (1979) clearly demonstrate response shift bias exists and that the traditional pre–post and retrospective pretest obtain different results. This study was an assessment of an educational program to reduce dogmatism among non-commissioned officers on an Air Force base. The program evaluation used a self-report dogmatism scale with the traditional pre–post method. The results indicated that 62% of the participants became more dogmatic from pretest to posttest and a t-test indicated the group as a whole was significantly more dogmatic at the time of the posttest than at the time of the pretest. The self-assessment scores were quite surprising because they not only indicated the program did not work, but that it actually increased dogmatism. Furthermore, these data were not congruent with the reactions from program facilitators and the participants' evaluations of the program. Follow-up interviews with participants suggested that the program altered the participants' understanding of dogmatism and changed the way they completed the dogmatism scale. Response shift, a change in the participants' understanding of dogmatism from pretest to posttest, appears to be responsible for the misleading findings.

Additional research verifies the existence of response shift in self-report measures (Cantrell, 2003; Ingram, Staten, Cohen, Stewart, & G. deZapien, 2004; Lamb & Tschillard, n.d.; Mann, 1997; Pratt, Mcguigan, & Katzev, 2000). However, the existence of response shift is not always clearcut. It may vary even within the same program. Pratt et al. (2000) found response shift on some items and not on all. In that study, the existence of response

shift appeared to relate to the outcome being measured. Response shift was found on items that reported personal characteristics (knowledge, skill) but not for an item concerned with one's perception of material goods (resources like money, food, transportation). Likewise, Manthei (1997) found that within a counselor training program some participants displayed a response-shift and others did not. In addition, there was a training effect for both. Thus one cannot assume that response-shift, will negate a training effect. Koele and Hoogstraten (1988) suggest researchers and evaluators verify response shift is present before using the RPT design and describe a method to test for response shift bias. Evaluators may also need to ask individuals to explain or describe their responses.

(**b) Difference between pretest and RPT results**. In study two by Howard, Ralph, et al. (1979), the authors compared the traditional pre-post with the RPT-posttest. Workshops on dogmatism were conducted with non-commissioned officers. The participants were randomly assigned to complete either the traditional pre–post or the RPT-posttest. Results indicated that dogmatism decreased for significantly more participants in the RPT-posttest group than for participants completing the traditional pre-post. In addition, the RPT-posttest participants rated themselves as being more dogmatic on their retrospective pretest than the pre–post group rated themselves on their traditional pretest. The authors conclude that the traditional pre-post and the RPT-posttest produce different results and the pattern of results is consistent with the notion of response shift bias.

These results found by Howard and colleagues have been substantiated by others (e.g. Cantrell, 2003; Pratt, McGuigan, & Katzev, 2000). Program effects based on pre-posttest self-reports are masked because people either overestimate or underestimate their preprogram knowledge, skill, etc. In most instances, greater program effects are found with the RPT-posttest than with the traditional pre-post as participants tend to overestimate their preprogram performance on the pretest, thus displaying little or negative effect at post-test. However, there are cases of respondents underestimating their preprogram status (Mann, 1997) which artificially inflates the effect of the program.

One of the issues raised by the notion of response shift, is how much different the results are between the traditional pre-post and the RPT-posttest. There is some evidence that the difference may not be very large. For example, Cantrell (2003) measured two types of self-efficacy in teachers. For one measure, the reported effect size of the RPT-posttest was substantially larger than the effect size of the traditional pre-post. For the other measure however, the traditional pre-post design showed a larger effect size than the RPT-posttest. Difference in results due to response shift and magnitude of effect depends upon what is being measured.

In a more recent study, Schmidt, Nubling, Steffanowshi, Lichtenberg, Wittman (2005) examined the relationship between three different methods of measuring change: the traditional pre-post, the retrospective pre-test, and the "direct measure of change" defined as when the person rates his or her own change as a comparison (e.g., better, unchanged, worse). Using four psychosomatic clinics for inpatient rehabilitation, a total of 858

patients participated in the study with data collection at three points in time: admission, discharge and one year follow-up. The retrospective pre-test demonstrated higher effect sizes and a better estimation of outcome quality in the longer term. Overall, the study found no real difference in the three methods – the three measure of change methods produced the same result for most patients relative to improved/not improved status.

c. Which is more valid. A number of studies have investigated the validity of the RPT and the performance of the RPT in relation to the traditional pre–post design. The retrospective pretest has been shown to be more consistent with objective measures, observations from program judges, and performance measures. Study three in Howard, Ralph, et al. (1979) demonstrated the RPT is more consistent with objective measures than the traditional pre–post design. In an androgyny training course, woman completed a traditional pre–posttest, a retrospective pretest, and an objective measure of femininity. The results of the retrospective pretest were more consistent with the objective measure than the traditional pre–post results. The authors conclude the RPT is therefore a more valid measurement. Howard and Bray (1979) also found evidence the retrospective pretest is more valid than the traditional pretest by correlating results of each type of test to ratings from program judges. Correlations between RPT-posttest scores and judges ratings of participants, were much higher than the correlations between the traditional pre-post results and judges ratings of participants.

Additional studies provide further evidence of the validity of the RPT. Pohl (1982) evaluated a three-week probability unit with business students in order to test response shift in a typical classroom setting. He administered three self report ratings (pre, post and retrospective) and two objective performance measures (pretest and posttest). Pohl found a significantly higher correlation between the retrospective rating and the objective pretest than between the pre rating and objective pretest, indicating that the retrospective rating is a more accurate estimate of actual pretest performance. Pohl, also compared behavioral change to self-reported change. He found greater correlation between the behavioral change and the retrospective self-reported change. Howard, Schmeck, and Bray (1979) conducted a similar study with educational psychology students and found comparable results.

A number of additional questions have been raised about the RPT as reported in the literature:

(d) Effect of preprogram information on response bias. Some researchers have asked if it is possible to reduce or eliminate response shift by providing program participants with a short introduction to the topic prior to the program. Howard, Dailey, and Gulanick (1979) tested this idea and found some evidence of response shift even when a pre-training information session was provided before a traditional pretest. Goedhart and Hoogstraten (1992) also found evidence that preprogram information does not reduce response shift. On the other hand, Sprangers and Hoogstraten (1989) found evidence that behavioral pretesting can reduce or eliminate response shift. In a well-controlled study the authors found administering a behavioral test of an outcome prior to the program can reduce or

eliminate response shift. It should be noted, however, that the behavioral pretest is different than preprogram information. The behavioral pretest is a direct measure of how a person acts whereas the preprogram information familiarizes participants with program concepts.

In other research, Manthei (1997) suggests that the differences he found within the same training program, where some participants displayed response-shift and others did not, might be linked to how the individual actually receives and understands the information as well as the individual's own self-appraisal competence. Pretest self-reports are influenced by individual's perception of what the training (program) will cover. In interviews after a program, Mann (1997) found that participants actually underestimated their ability on the pretest because the training had been pitched at a high level in the preprogram information. She concludes that response-shift can be reduced by ensuring participants receive and understand adequate preprogram information.

(e) Other biases that may result from or be intensified by RPT. Howard and colleagues also tested the retrospective method for **response style effects** (Howard, Millham, Slaten, & O'Donnell, 1981). They wanted to rule out the possibility that the RPT-posttest has outperformed the traditional pre–post due to **social desirability** or complying with **implicit task demands**. In a study of assertiveness, participants were assigned to a training group and a control group. Participants in both groups completed a traditional pre–posttest, a retrospective pretest, a measure of social desirability, and a measure of compliance with implicit task demand. Correlations were computed between the two types of pretests and the measures of social desirability and compliance with implicit task demand. The correlations involving the RPT were lower or equal to the correlations involving the traditional pretest. They conclude that the RPT is no more susceptible to **social desirability** or **compliance with implicit task** demands than the traditional pretest.

2. Use of the RPT design in program evaluation

The RPT is being used in a variety of programming settings and contexts to measure change. Besides the published articles reviewed here, it is likely that practitioners and professionals are using the design even more widely. For example, since the 1989 Journal of Extension article by Rockwell and Kohn, the RPT-posttest is often included in outcome evaluations of Cooperative Extension programs, either as the central research design, as the design of one question within a series of questions, or as a part of a mixed method design. The popularity of RPT in program evaluation is largely due to the following factors:

- Acceptance and widespread use of self-reports for measuring change
- Lack of effect from traditional pre-post test design
- Nature of programs results in existence of response shift bias
- Lack of pre-test data or inappropriate to collect
- Lack of same participants at posttest time
- Resource constraints

- Attrition or change in participants over course of program

In building a taxonomy of best practices, six aspects appear to stand out:

(1) Program Settings. There is great variability and flexibility in the use of the RPT design in program evaluation. From the reviewed literature, RPT is most frequently used in evaluating professional development training courses in such fields as health (Farel, Umble, & Polhamus, 2001; Steckler, Farel, Breny Bontempi, Umble, Polhamus, & Trester, 2001; Upshaw, Umble, Orton, Matthews, 2000), management (Mann, 1997), teacher training (Cantrell, 2003; Lan & Bengo, 2003), occupational therapy (Lee, Paterson, & Chan, 1994). The RPT is also popular in higher education evaluation in such settings as counselor education programs (Manthei, 1997), graduate psychiatry training (Myers, 2004), and MCH degree program (Umble, Shay, & Sollecito, 2003). It is also being used in the evaluation of a variety of community prevention and intervention programs: nutrition program (Rockwell & Kohn, 1989), healthcare workers training (Ingram et al, 2004), HIV attitude programs (Riley & Greene, 1993).

The RPT appears particularly appropriate when the collection of pretest data is infeasible such as in trauma, emergency events, or bereavement studies such as the pregnancy loss study reported by Toedter, Lasker, & Campbell (1990).

(2) Audience. Target populations (respondents) in the programs using the RPT designs vary by age, educational level, gender, socio-economic status, ethnicity, position and nature of the program participation (voluntary vs. mandatory). There appear to be few studies that consider effects introduced by different audience characteristics. Le Rouzic and Cusick (1998) make mention of the influence of cultural diversity in the World Bank training programs and Mann (1997) speculates on the influence that cultural differences might have had in the lack of response-shift bias among an American group of trainees versus a European group. It may well be that the differences were due more to previous experience and abilities. Mann goes on to speculate if the difference could be due to age since the American group was older and older people might have a better self-awareness of their own abilities. Le Rouzic, Ouchi, and Zhou (1999) found participants who had completed a Masters Degree and had less than five years work experience were more likely to self-assess accurately.

(3) What is measured. Likewise, there is great variation in what is measured. Program outcomes of interest include changes in knowledge, skills, attitudes, motivations, self-efficacy, and behaviors. Often, one retrospective instrument may ask responds to self-report on several of these changes and be the reason for reported differences in response shift within instruments. Le Rouzic and Cusick (1998) among others argue that self-assessments are more appropriate for measuring self-efficacy, one's confidence in his or her ability to use what is taught, than as a measure of knowledge or learning. Self-efficacy can be used as an important predictor of motivation and skill performance (Applebaum, 1996). They point to Dixon's (1990) research that shows no relationship between trainee perceptions of amount learned and actual learning. In another article, Le Rouzic et al.

(1999) did find significant relationships between what people felt they learned (knowledge) and what they actually learned but the relationships were weak. They caution against replacing assessments of actual knowledge with self-reports of perceived knowledge. However, there is a secondary benefit of self-reports in that the process of self-reporting promotes reflection and learning, often objectives of an adult learning program (Fuque, Newman, Scott, & Gade, 1986).

Manthei (1997) recommends collecting qualitative data as part of RPT to better understand response-shift and the different ways participants may view their initial skill levels and impact of the training. The patterns of scores and reasons given suggest that effects of training on individual students are more varied and complex than can be explained by a single concept, such as response-shift.

Another aspect that influences self-reported change depends upon the way in which the change is reported (or the question is worded). Lam and Bengo (2003) found that estimating actual frequency of practice has an effect on measuring change. Participants were more liberal in reporting change when they did not have to indicate the actual frequency of their practice; specifically, teachers found it easier to report change, and they tended to report larger amounts of increase in desirable practices if they did not have to first determine their levels of instructional practice. The authors attribute this to the difficulty of the task required by the question that leads to differential satisficing and socially desirable responses. Citing Krosnick (1991), Lam and Bengo discuss three determinants of satisficing: task difficulty, respondent ability, and respondent motivation.

(4) Methods. The reviewed evaluations employ a variety of ways of using the RPT, including mixed methods and various quasi-experimental designs. Evaluations that also are testing the validity of the RPT commonly include a pretest, plus the RPT and posttest, often with a control or comparison group (e.g., Zweibel, 1987; Umble, Shay & Sollecito, 2003; Rhodes & Jason, 1987). Several studies have employed the RPT as part of a follow-up survey (e.g., Farel et al., 2001; Steckler et al., 2001). The community-based program evaluations appear to be using the RPT + posttest only (Davis, 2003; Gamon, Harold, & Creswell, 1994). Numerous authors recommend the use of multiple methods, not relying solely on RPT-posttest (Manthei, 1997; Pratt et al., 2000; Umble et al., 2000) When precision is required, behavioral and self-report measures are recommended (Pratt et al., 2000).

(5) Memory recall period. In the reviewed evaluations, the memory recall period varied widely dependent upon the timeframe of the program, from relatively short time frames to over 3 years. Pratt et al. (2000) discuss several salient memory-related biases in RPT: length and specificity of the period that is recalled. They recommend clarifying the time period that is to be considered for the participant, using such clues as, "since you began the program" and formulating questions in a way that facilitates recall. For example, specific behaviors are easier to recall. Participant recall of past events is affected by methods used to elicit information. It is also dependent upon cognitive abilities, for example, stages of cognitive growth and social development their influence childrens'

responses (Borgers, de Leeuw, & Hox, 2000). Toedter, Lasker, & Campbell (1990) speculate that emotion may effect recall in the bereavement study.

Related research is relevant to here. Hawkins and Hastie (1990) argue that hindsight exerts an effect on memory. When people know the outcome of events, they remember and recall information consistent with the way events actually occurred. However, when people do not know the outcome of events, their recall is much broader, including information consistent with and contrary to what eventually occurred. Knowing how events unfold seems to organize and restrict recall to make it consistent with actual events.

(6) Formats. Various ways the RPT-posttest is included and formatted: as single questionnaire or as two separate questionnaires; as a single question within a series of questions or as the total questionnaire; in vertical or horizontal layouts and various page designs; different font and style characteristics (Umble, Orton & Matthews, 2000; Rockwell, 1995; Klatt & Taylor-Powell, 2005b). Style may depend upon creativity of designer and respondent characteristics. Order of the RPT with the Post response first as the base frame of reference is recommended (Howard, Ralph, Gulanick, Maxwell, Nance, Gerber, 1979). Others indicate that ordering does not matter (Sprangers & Hoogstraten, 1979). Clear and concise instructions seem important (Klatt and Taylor-Powell, 2005b). However, Davis (2003) gave no verbal instructions with no apparent consequences. Manthei (1997) found certain items may have been confusing or respondents using alternative interpretations in their meaning.

Missing data. The RPT results in less missing data than the traditional pre-post design (Raidl et al., 2004). Unless respondents fail to complete the questionnaire appropriately, you will have pretest and posttest data for each participant. The only missing data will be from people who did not complete the questionnaire or from people who skipped items.

Other biases. Besides the frequently mentioned social desirability bias, another potential bias that may occur in the RPT design is effect of effort or effort justification (Ingram, 2004). The participant may feel the need to justify the effort s/he invested in the training (program) and underestimate her retrospective pretest response. Another bias, bearing further research, is hindsight effect as previously discussed.

3. **Instructional materials**

Besides the refereed, academic literature cited above, there are a variety of published instructional materials on the RPT design. Developed to promote the use of the design and/or to improve practice in the use of the RPT design, they are written largely for a practitioner audience. They include such topics as what the RPT is, when to use the RPT, how to create an RPT question, how to analyze and report data from an RPT question; issues and considerations in the use of the RPT design (Lamb, 2005; Klatt &Taylor-Powell, 2005a; Klatt & Taylor-Powell, 2005b; Klatt & Taylor-Powell, 2005c; Schaaf et al, 2005; Kiernan, 2001a; Kiernan 2001b)

III Summary

As we might expect, the research reviewed here is inconclusive in many respects. This might be due to the limitations of the studies as often cited by the authors themselves: small sample sizes; differences in the way individual research studies have been implemented (different sites; groups handled in differently); mandatory versus voluntary participation. Each study is answering slightly different questions, in different contexts, and using different methods. Also, as the RPT has been adopted by evaluators working in different settings, the complexities of these settings have raised additional challenges and questions to be answered.

What have we learned (or have had reinforced) based on this research

Response shift bias poses a threat to the validity of traditional pretests. Response shift can result in underestimating or overestimating program effects. In addition, pretest self-reports can be influenced by the individual's perception of what the program will cover. RPT appears to reduce response shift and preprogram assumptions, but may intensify many intensify other biases such as social desirability, effort justification or hindsight given the context.

The RPT has both strengths and limitations. As mentioned, it can reduce response shift and may provide a more accurate measure of change. It is flexible and convenient. It provides comparison data in the absence of pre-data. But, all methods have limitations. Evaluators need to appraise the appropriateness of the RPT given the context and target audience. As a self-report measure, the RPT is susceptible to biases in self-appraisal and recall. Despite its limitations, it may be the best design in many situations.

Measuring change is complex and difficult. We need to combine and use a mix of methods and not rely solely on RPT. Combining pre and RPT measures provides results that are more accurate. Plus, examining the difference in the two results provides useful information to enhance understanding and for making programming improvements.

Threats to validity vary across issue and context. We need to consider all potential biases and implement designs that help differentiate the direction and extent of bias. Aiken and West (1990) support proposal by Scriven 1976 that invokes evaluators to consider potential biases in their evaluation contexts and implement designs that can help differentiate the direction and extent of these biases. Threats to validity vary across different issues and contexts. Eckert (2000) suggests that we use a checklist of all possible threats (maturation, history, testing, instrumentation, regression, selection, mortality) and consider which threats are plausible given the setting and how viable the design might be. We suggest including those relevant to RPT: response shift, social desirability, effort justification, memory distortion, etc.

The RPT's flexibility is considered a strength, but need cognitive testing and appraisal of appropriateness given context and target audience. Not a panacea!

IV Recommendations – future research needs

Greater understanding about type of change that is most appropriate to measure with RPT: knowledge, attitude, self-efficacy, etc.

Need more understanding about the existence of response-shift in different program and population contexts, especially different cultures and diverse populations

Greater understanding of the issues related to RPT implementation, such as recall period; memory capabilities; questionnaire design; appropriateness for different types of populations

Additional research related to the effect of preprogram information on response shift.

Creative and efficient ways to check (test) for response shift

Additional research related to RPT related biases: social desirability, effort justification, memory distortion, other response style effects

Retrospective Post then Pre References

Aiken, L.S., West, S.G. (1990). Invalidity of true experiments: Self-report pretest biases. Evaluation Review, 14(4), 374-390.

Borgers, N., de Leeuw, E., & Hox, J. (2000). Children as respondents in survey research: Cognitive development and response quality. Bulletin de Methodologie Sociologique, 66, 60-75.

Bray, J. H. & Howard, G. S. (1982). Methodological considerations in the evaluation of a teacher-training program. Journal of Educational Psychology, 72(1), 62-70.

Bray, J. H., Maxwell, S. E., & Howard, G. S. (1984). Methods of analysis with response-shift bias. Educational and Psychological Measurement, 44(4), 781-804.

Breetvelt, I. S. & Van Dam, F. S. (1991). Underreporting by cancer patients: The case of response-shift. Social Science & Medicine, 32(9), 981-987.

Campbell, D. T. & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Reprinted from Handbook of research on teaching, American Educational Research Association. Chicago: Rand McNally.

Cantrell, P. (2003). Traditional vs. retrospective pretests for measuring science teaching efficacy beliefs in preservice teachers. School Science and Mathematics, 103(4), 177-185.

Davis, G. A. (2003). Using a retrospective pre-post questionnaire to determine program impact. Journal of Extension, 41(4). http://www.joe.org/joe/2003august/tt4.shtml

Deutsch, M. & Collins, M.E. (1951) Interracial housing: A psychological evaluation of a social experiment. Minneapolis: University of Minnesota Press.

Diem, K. G. (nd.) Using experimental designs for program evaluation. The State University of New Jersey, Rutgers Cooperative Extension www.rce.rutgers.edu/evaluation

Dixon, N. M. (1990). The relationship between trainee responses on participant reaction forms and posttest scores. Human Resource Development Quarterly, 1(2), 129–137.

Eckert, W. (2000). Situational enhancement of design validity: The case of training evaluation at the World Bank Institute. American Journal of Evaluation, 21(2), 185-193.

Farel, A., Umble, K., Polhamus, B. (2001). Impact of an online analytic skills course. Evaluation and the Health Professions, 24(4), 446-459.

Fuque, D.R., Newman, J.L., Scott, T.B., Gade, E.M. (1986). Variability across sources of performance ratings: Further evidence. Journal of Counseling Psychology, 33(3), 353-356.

Gamon J., Harold, N., & Creswell, J. (1994) Educational delivery methods to encourage adoption of sustainable agricultural practices. Journal of Agriculture Education, 35(1), 38-42.

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

Hawkins, S. A. & Hastie, R. (1990) Hindsight: Biased judgments of past events after the outcomes are known, Psychological Bulletin, 107(3), 311-327.

Hoogstraten, J. (1982).The retrospective pretest in an educational training context. Journal of Experimental Education, 50(4), 200-204.

Howard, G. S. (1980). Response-shift bias. A problem in evaluating interventions with pre/ post self-reports. Evaluation Review, 4(1), 93 – 106.

Howard, G.S. (1981). On validity. Evaluation Review, 5(4), 567-576.

Howard, G. S. (1990). On the construct validity of self-reports: What do the data say? American Psychologist, 45(2), 292-294.

Howard, G. S, & Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. Journal of Applied Psychology, 66(2), 144-50.

Howard, G. S., Dailey, P. R., & Gulanick, N. A. (1979). The feasibility of informed pretests in attenuating response-shift bias. Applied Psychological Measurement, 3, 481-494.

Howard, G. S., Millham, J., Slaten, S., & O'Donnell, L. (1981). Influence of subject response-style effects on retrospective measures. Applied Psychological Measurement, 5, 144-150.

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, S. W., & Gerber, S. K. (1979). Internal invalidity in pre-test-post-test self-report evaluations and a re-evaluation of retrospective pre-tests. Applied Psychological Measurement, 3, 1-23.

Howard, G. S., Schmeck, R. R., & Bray, J. H. J. (1979). Internal validity in studies

employing self-report instruments: A suggested remedy. Journal of Educational Measurement, 16, 129-135.

Howard, G. S. (1990) On the construct validity of self-reports: What do the data say? American Psychologist, 45(2), 292-294.

Ingram, M., Staten, L., Cohen, S. J., Stewart, R., & G. deZapien, J. (Accepted 11/2/2004) The use of the retrospective pre-test method to measure skill acquisition among community health workers. Internet Journal of Public Health Education. BG-1-15.

Kiernan, N.E., (2001a). Reduce bias with retrospective questions. Tipsheet #30. University Park, PA: Penn State Cooperative Extension

Kiernan, N.E., (2001b). Analyzing before-after data using excel. Tipsheet #52. University Park, PA: Penn State Cooperative Extension

Klatt, J., Taylor-Powell, E. (2005a) Using the retrospective post-then-pre design. Quick Tips #27. Program Development and Evaluation. Madison, WI: University of Wisconsin-Extension.

Klatt, J., Taylor-Powell, E. (2005b) Designing a retrospective post-then-pre question. Quick Tips #28. Program Development and Evaluation. Madison, WI: University of Wisconsin-Extension.

Klatt, J., Taylor-Powell, E. (2005c) When to use the retrospective post-then-pre design. Quick Tips #29. Program Development and Evaluation. Madison, WI: University of Wisconsin-Extension.

Koele, P. & Hoogstraten, J. (1988). A method for analyzing retrospective pretest/posttest designs: I. Theory. Bulletin of the Psychonomic Society, 26(1), 51-54.

Lam, T. C. & Bengo, P. (2003). A comparison of three retrospective self-reporting methods of measuring change in instructional practice. American Journal of Evaluation, 24 (1), 65 – 80.

Lamb, T. (2005). The retrospective pretest: An imperfect but useful tool. The Evaluation Exchange, XI(2), Summer 2005.

Lamb, T. A., & Tschillard, R. (nd) An underutilized design in applied research: The retrospective pretest. Submitted for publication to the Journal of Applied Sociology.

Le Rouzic, V., Cusick, M. C. (1998) Immediate evaluation of training

events at the economic development institute of the World Bank measuring reaction, self-efficacy, and learning in a worldwide context. Paper presented at the American Evaluation Association Conference.

Le Rouzic, V., Ouchi, F., & Zhou, C. (1999) Measuring "What People Learned" versus "What People Say They Learned": Does the difference matter? Presented at the American Evaluation Association Annual Conference.

Lee, T. M. C., Paterson, J. G., & Chan, C. C. H. (1994). The effect of occupational therapy education on students' perceived attitudes toward persons with disabilities. American Journal of Occupational Therapy, 48(7), 633-638.

Levinson, W., Gordon, G., & Skeff, K. (1990). Retrospective versus actual pre-course self-assessments. Evaluation & the Health Professions, 13, 445-452.

Mann, S. (1997). Implications of the response-shift bias for management. Journal of Management Development, 16(5). 328-336.

Manthei, R. J. (1997). The response-shift bias in a counsellor education programme. British Journal of Guidance & Counselling, 25(2), 229-237.

Mezoff, B. (1981). How to get accurate self-reports of training outcomes. Training and Development Journal, 35(9), 56-61.

Moxley, V., Eggeman, K., & Schumm, W. R. (1986). An Evaluation of the "Recovery of Hope" Program. Journal of Divorce, 10 (1/2), 241-261.

Myers, G. E. (2004). Addressing the effects of culture on the boundary-keeping practices of psychiatry residents educated outside of the United States. Academic Psychiatry, 28(1), 47-55.

Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84(3), 231-259.

Pohl, N. F. (1982). Using retrospective pre-ratings to counteract response-shift confounding. Journal of Experimental Education, 50(4), 211-214.

Pratt, C. C., McGuigan, W. M., & Katzev, A. R. (2000). Measuring program outcomes: Using retrospective pretest methodology. American Journal of Evaluation, 21(3), 341 – 349.

Raidl, M., Johnson, S., Gardiner, K., Denham, M., Spain, K., Lanting, R., Jayo, C., Liddil, A., & Barron, K. (2004). Use retrospective surveys to obtain complete data sets and measure impact in extension programs. Journal of Extension, 42(2), http://www.joe.org/joe/2004april/rb2.shtml

Rhodes, J.E., Jason, L.A. (1987). The retrospective pretest: An alternative approach in evaluating drug prevention programs. Journal of Drug Education, 17(4), 345-356).

Riley, J. L. & Greene, R. R. (1993). Influence of education on self-perceived attitudes about HIV/AIDS among human services providers. Social Work, 38(4), 396-401.

Rockwell, S. K. & Kohn, H. (1989). Post-then-pre evaluation: Measuring behavior change more accurately. Journal of Extension, 27(2), *http://www.joe.org/joe/1989summer/a5.html*.

Rockwell, S.K. 1995. "Post-activity Evaluation." Handout in Module #10 of a 15 module graduate course, ALEC826: Program Evaluation in Adult Education and Training, Agricultural Leadership, Education and Communication Department, University of Nebraska: Lincoln. Cited in Taylor-Powell, E & Renner, M. 2000. Collecting Evaluation Data: End-of-Session Questionnaires. Madison, WI: University of Wisconsin-Extension. http://www.uwex.edu/ces/pdande/Evaluation/evaluat.html

Schaaf, J., Klatt, J., Boyd, H., Taylor-Powell, E. Analysis of retrospective post-then-pre data. Quick Tips #30. Program Development and Evaluation. Madison, WI: University of Wisconsin-Extension.

Schmidt, J., Nbling, R., Lichtenberg, S., Steffanowski, A. & Wittmann, W.W. (in press): Assessment of the Outcome Quality of Inpatient Psychosomatic Rehabilitation - A Comparison between Different Strategies of Evaluation and the Development of New Measurement Instruments. In: Bengel, J., Jackel, W.H., Herdt, J. (Eds). Research in Rehabilitation - Results from a Research Network in Southwest Germany. Stuttgart, Germany: Schattauer

Sprangers, M. & Hoogstraten, J. (1989). Pretesting effects in retrospective pretest-posttest design. Journal of Applied Psychology, 74(2), 265-272.

Steckler, A., Farel, A., Breny Bontempi, J., Umble, K., Polhamus, B., & Trester, A. (2001). Can health professionals learn qualitative evaluation methods on the World Wide Web? A case example. Health Education Research, Theory & Practice, 16(6), 735-745.

Toedter, L. J., Lasker, J. N. & Campbell, D. T. (1990). The comparison group problem in bereavement studies and the retrospective pretest, Evaluation Review, 14(1), 75-90.

Umble, K., Orton, S., & Mathews, K. (June 15, 2000). Using the post-then method to assess learner change. AAHE Assessment Conference, Charlotte, North Carolina.

Umble, K., Shay, S., & Sollecito, W. (2003). An interdisciplinary MPH via distance learning: Meeting the educational needs of practitioners. Journal of Public Health Management Practice, 9(2), 123-135.

Upshaw, W.M, Umble, K.E., Orton, S., & Matthews, K. (2000). Cited in Umble et al, 2000.

Zwiebel, A. (1987). Changing educational counsellors' attitudes toward mental retardation: comparison of two measurement techniques. International Journal of Rehabilitation Research, 10(4), 383-389.