International Journal of Behavioral Development
1–9
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0165025419877973
journals.sagepub.com/home/jbd
SAGE

The retrospective pretest–posttest design redux: On its validity as an alternative to traditional pretest–posttest measurement

Todd D. Little,[1,2] Rong Chang,[1] Britt K. Gorrall,[1] Luke Waggenspack,[1] Eriko Fukuda,[1] Patricia J. Allen,[3] and Gil G. Noam[3]

Abstract

We revisit the merits of the retrospective pretest–posttest (RPP) design for repeated-measures research. The underutilized RPP method asks respondents to rate survey items twice during the same posttest measurement occasion from two specific frames of reference: “now” and “then.” Individuals first report their current attitudes or beliefs following a given intervention, and next they are prompted to think back to a specific time prior to the given intervention and rate the item again retrospectively. The design addresses many of the validity concerns that plague the traditional pretest–posttest design. Particularly when measuring noncognitive constructs, the RPP design allows participants to gauge the degree of change that they experience with greater awareness and precision than a traditional approach. We review the undesirable features of traditional designs and highlight the benefits of the retrospective approach. We offer examples from two recent, original studies and conclude with the recommendation that the RPP design be employed more broadly. We also conclude with a discussion of important directions for future examination of this design.

Keywords

Methodology, pretest–posttest design, retrospective assessment

We highlight the merits of the retrospective pretest–posttest (RPP) design for repeated-measures research and program evaluation as an alternative to traditional pretest–posttest (TPP) designs. Given the importance of validly in assessing change over time and the fact that the TPP design often fails to identify treatment effects when other methods suggest there was an effect, we revisit the RPP design as a valuable alternative to a TPP design. This alternative design, which is currently underutilized, involves asking participants at the time of the posttest to retrospectively respond to questionnaire items thinking back to a specified pretest period. In effect, participants rate each item twice within a single sitting (“then” and “now”) to measure self-perceptions of change, typically after a given intervention.

The design addresses many of the validity concerns that plague the TPP design. Particularly when measuring noncognitive constructs, empirical evidence shows that the RPP design allows participants to gauge the degree of change that they experience with greater awareness and precision than a TPP approach. To substantiate this, we provide background and address the criticisms, which are mostly unsubstantiated, that have been raised about the RPP design. We then turn to two recent, original data sets that exemplify the proper use of this design. In these examples, we address potential methodological concerns and provide an empirical basis for the validity of the RPP design as an alternative design for evaluation and longitudinal research. We conclude with suggestions for future research to further establish the RPP method as a new standard for repeated-measures research.

TPP Measurement and its Limitations

The traditional gold standard to evaluate a program or an intervention effect is to use a pretest–posttest measurement design with random assignment to groups. In the typical application of this design, respondents are asked to complete two identical self-report measures at two different points in time—pre-intervention and post-intervention. To determine the effectiveness of the program or intervention, the significance of the difference between the posttest and the pretest score is calculated. If the posttest score is significantly changed from the pretest score (and different from the control condition), the difference in the scores indicates the program/intervention impacted the participants. Quite often, even with quality interventions, this method of measurement fails to reveal a significant finding. Although a pretest has many useful features such as screening participants prior to program implementation, a number of inadequacies to this self-report design as a tool to assess change due to program implementation have been outlined: lack of

[1] Texas Tech University, USA
[2] Optentia Research Group, North-West University, South Africa
[3] Harvard Medical School and McLean Hospital, Belmont, MA, USA

Corresponding author:
Todd D. Little, Department of Educational Psychology, Texas Tech University, Lubbock, TX 79409, USA.
Email: yhat@statscamp.org

self-awareness at pretest, socially desirable responding (possibly due to lack of anonymity), retest effects, and test-reactivity (Bray, Maxwell, & Howard, 1984; Moore & Tananis, 2009).

More specifically, when using the TPP design, researchers assume that respondents have a consistent internal frame of reference for the target constructs (e.g., belief, skills, and attitudes) over time (Cronbach & Furby, 1970; Howard, Ralph, et al., 1979). One criticism of this approach, however, is that the frame of reference of the respondent is unclear (Nieuwkerk & Sprangers, 2009). The comparison referent at the pretest could be to a prior self in time, the present self, a comparison of self to others, or a mixture of ambiguous referents across the sample of participants. When others are used as the comparison standard, the others that one chooses are unclear and typically unspecified. Respondents might choose the specific group of others that is comparable to the self as the comparison referent (e.g., a White female). Such judgments can have stereotypic standards depending on the race/ethnic/gender/age/cultural group that is used for comparison. Any change in standards associated with a given group is referred to as a shifting standard (Biernat & Manis, 1994). A shifting standard refers to the idea that an individual's standard of judgment might vary from one standard for one group at the pretest to a different standard for another group at posttest. For example, female respondents may use other females as the standard of self-judgment at pretest and then use just herself as the standard at posttest.

Similarly, respondents might lack awareness about the constructs that the intervention program is intended to impact (Howard & Dailey, 1979; Howard, Schmeck, & Bray, 1979). Due to the lack of understanding or awareness, participants cannot accurately evaluate their perceptions of the constructs before the program is implemented. After completing the intervention program, participants may increase awareness and understanding about the constructs, which enable them to assess themselves more accurately. At the point of the posttest, however, participants are not able to correct their responses on the pretest. This lack of correction can lead to invalid pretest results (e.g., over- or under-estimated results). With increasing awareness due to the intervention program, respondents may be better able to accurately assess the constructs at the posttest than the pretest. Given the potentially inaccurate self-assessments at the pretest, the changes in perceptions are not correctly reflected in participants' responses. This bias can result in an invalid assessment of any "true changes" from the intervention effects (Bray et al., 1984; Howard, 1980; Howard, Ralph, et al., 1979).

Drawbacks to assessing change such as these are related to what has been termed the response shift bias that occurs between the pretest assessment and the posttest assessment (Howard, Dailey, & Gulanick, 1979; Howard, Ralph, et al., 1979). Response shift bias occurs when the comparison standard is different at pretest from at posttest, and a respondent's response is no longer a valid index of true change on the construct (Howard, Dailey, et al., 1979; Oort, 2005). Three types of response shift bias have been described: (a) recalibration (i.e., a change in a respondent's internal standards of measurement), (b) reprioritization (i.e., a change in a respondent's values), and (c) reconceptualization (i.e., redefinition of the target construct; Schwartz & Sprangers, 2000; Sprangers & Schwartz, 2000; Sprangers et al., 1999). Howard, Ralph, et al. (1979) stated the response shift bias is a highly probable threat to internal invalidity in experimental designs for understanding treatment effects. When response shift bias occurs, true changes are not adequately captured in the TPP experimental design (Howard, 1980). Assessment of true change is thereby hampered by the occurrence of response shift bias (Howard, Ralph, et al., 1979; Oort, 2005; Sprangers & Schwartz, 2000; Sprangers et al., 1999; Schwartz & Sprangers, 2000). A response shift, therefore, undermines any findings made from comparing pretest and posttest data as an index of true change (Bray et al., 1984; Howard, Ralph, et al., 1979).

The RPP Design

To overcome the significant limitations of a TPP design to capture change, Howard, Ralph, et al. (1979) recommended the RPP method. As mentioned above, the RPP design is a way to simultaneously collect retrospective pretest and (current) posttest data. If a pretest was given, the questionnaire is readministered at the posttest measurement occasion or if no pretest is given, the RPP questionnaire is administered only at the posttest period. At this posttreatment measurement occasion, respondents are requested to report their current attitudes or beliefs and, at the same time, retrospectively think back to a specific time prior to the program onset. As such, respondents are forced to focus on themselves at a specified point in time, providing a consistent frame of reference both within and across respondents (Drennan & Hyde, 2008; Howard, 1980; Sprangers, 1989a, 1989b).

In addition, exposure to the intervention program could activate awareness in the self because respondents are capable of gauging prior levels of skills, beliefs, and attitudes compared to current levels. Moreover, reactivity and retest effects may be reduced; while respondents make two distinct judgments for each item, these judgments are made within a single session time frame. These features of the RPP design are ideally suited to detect real change results from a successful intervention. Importantly, when real change does not occur, the design is also able to show lack of change as well as varying levels of change given moderating influences (e.g., dosage and fidelity; see Empirical Examples 1 and 2).

Evidence Supporting the RPP Design

A significant and growing body of empirical evidence supports the advantages of using the RPP design over the TPP design. Drennan and Hyde (2008), for example, administered a 21-item self-report tool to all students ($n = 120$) as part of a TPP design, and, in addition, 80% of the participants ($n = 96$) responded to a retrospective pretest at the time of the posttest assessment. The timing of the pretest and the posttest administrations were at the beginning of the course (Time 1) and 6 months after completing the course (Time 2). Although the results indicated that self-reported change was significant for both methodological designs, there was also evidence of a *response shift bias.* As mentioned, a response shift bias occurs when an individual's internal frame of reference about the construct being measured changes between the pretest and the posttest. In the study, students overestimated their ability at the beginning of the course when measured as a traditional pretest. Their retrospective reports, on the other hand, did not include the same magnitude of overestimation as did the reports made using a TPP design. Therefore, when using the TPP method, response shift bias was evident in that the educational program had significantly greater impact on outcomes when rated retrospectively than traditionally. Drennan and Hyde concluded that the RPP design is a valuable tool to evaluate the impact of an educational program outcome.

Other studies that have focused on the RPP design indicate that the approach can provide more sensitive estimates of true program or treatment effects than traditional approaches (e.g., Breetvelt & Van Dam, 1991; Cohen, 2016; Howard, Dailey, et al., 1979; Nakonezny & Rodgers, 2005; Pratt, McGuigan, & Katzev, 2000; Sprangers et al., 1999). Many studies from various research areas use the RPP design, especially in educational, social, and health science program outcomes. Such studies include a continuing education program for adults involved in economic development (Davis, 2002), communication skills training for medical assistants (Sprangers & Hoogstraten, 1989), leadership skill development courses (Rohs, 2002), and online analytic skills training for professionals in public health (Farel, Umble, & Polhamus, 2001).

A number of federally funded studies have utilized the RPP design. In a study funded by the National Institute on Drug Abuse (NIDA), Moberg and Finch (2007) examined program outcomes of high school students ($n = 321$) recovering from a substance use disorder across 18 high schools in seven states (California, Colorado, Minnesota, Pennsylvania, Tennessee, Texas, and Wisconsin). Moberg and Finch argued that the RPP design was the only alternative to the TPP design to use, given the nature of the outcomes (e.g., self-reported substance use). Using the RPP design, they found a significant reduction in substance use as well as in mental health symptoms. As a demonstration of differential sensitivity, students also showed greater positive attitudes about the therapeutic value of the schools but were less enthusiastic about the educational programs.

Pratt, McGuigan, and Katzev (2000) were funded by Oregon Healthy Start Evaluation to evaluate longitudinal data from mothers ($n = 307$) with first-born infants who participated in a home visitation, child-abuse prevention program. In this study, both a TPP design and an RPP design were both implemented. The data were collected when the infant was 1–7 days old (pretest) and 6 months old (posttest). A 7-item self-report measure was used to assess maternal knowledge of child development, confidence in parenting, and so on. Results indicated that all 7 items on the measure showed significant improvements with the RPP design; however, only 4 items showed improvement with the TPP design (respondents showed an underestimation of the program effect). Pratt et al. (2000) also found the presence of a response shift bias in the means of pretest items compared to the retrospective pretest. Further examination revealed that response shift bias occurred for the 3 items that failed to show significant change using the TPP design. The differential item functioning under the TPP supports the conclusion of Pratt et al. that the RPP methodology provided a legitimate and valid assessment of program outcomes compared to the TPP design.

As discussed above, the RPP design is a promising methodology to mitigate the response shift bias that seems ubiquitous in a TPP approach; however, criticisms of the RPP design also exist. These criticisms include memory-related problems (e.g., memory distortion, selective perception, and poor memory; Howard, Schmeck, et al., 1979; Lam & Bengo, 2003; Pratt et al., 2000), social desirability (Howard, Schmeck, et al., 1979), and impression management and response bias (Lam & Bengo, 2003; Taylor, Russ-Eft, & Taylor, 2009). Some researchers also doubt the appropriateness of using this design for self-reports of children and adolescents due to concerns that cognitive development in young people may limit their ability to retrospectively rate thoughts and feelings (Brossart, Clay, & Willson, 2002). Others also pointed out that regression to the mean and maturation effects can influence change over time in the RPP design (Pratt et al., 2000). Although these researchers provide theoretical criticism, they have not provided direct empirical support for any of these positions.

Table 1. Items for scale of self-regulated learning in Example 1.

Self-regulated learning
1. I make sure I know what to do.
2. I make a plan.
3. I gather what I need.
4. I check my work as I go.
5. I ask questions.
6. I follow my plan.
7. I learn from difficult situations.
8. I learn from failure.
9. I learn from my success.

Method and Results

In consideration of the previously made critiques about the RPP design (e.g., Kubota et al., 2008; Rhodes & Jason, 1987; Verrips et al., 1998), in this study, our goal is to provide evidence to answer three issues:

1. Is the RPP assessment a stable and reliable technique for program evaluations?
2. With no implementation of pretest data collection, is the retrospective pretest assessment alone able to provide sufficient information to measure changes over time?
3. Are children able to think back and retrospectively provide self-judgments to surveyed questions?

To understand the usefulness of the RPP design, we provide a demonstration of a set of analyses using two empirical longitudinal studies. Both of them are complete data sets in that all the missing values (both planned and unplanned) have been imputed using the multiple imputation technique with the auxiliary principal component scores (see e.g., Enders, 2010; Howard, Rhemtulla, & Little, 2015).

Empirical Example 1

The first data set is from a program evaluation of the changes in students' learning mind-sets and math strategies during their participation in one of the two training programs: (1) School-Year Academic Youth Development and (2) Intensified Algebra (see www.utdanacenter.org). Students in both programs provided self-report responses to a series of survey items regarding their learning mind-set (incremental/entity) and math study-strategy usage.

Description of data. Our analysis included 4,713 student responses collected across three time points (see Figure 1): Time 1 (baseline), Time 2 (6 months), and Time 3 (12 months). At all three time points, students ranged from 7th to 10th grade. Within this sample, 55% were male.

Measurement. Table 1 lists the items used for measuring students' self-regulated learning strategies of math. This scale includes 9 items which assess students' awareness of goal setting and regulatory strategies as well as perceived self-monitoring engagement and ability.

Table 2. Comparison results of latent mean differences and Cohen's effect size measures over time for self-regulated learning.

Mean						
PreT1	RetroT2	Difference	χ^2	SD_{pooled}	d	U_3 (%)
67.55	57.69	9.86	1,134.35**	13.00	.76	77.59
PreT1	RetroT3	Difference	χ^2	SD_{pooled}	d	U_3 (%)
67.55	57.40	10.15	1,166.74**	12.69	.80	78.80
RetroT2	RetroT3	Difference	χ^2	SD_{pooled}	d	U_3 (%)
57. 69	57.40	0.28	9.33*	13.414	.02	50.84

Note. $N = 4{,}713$.
**$p < .001$. *$p <.01$.

Response difference. We conducted latent variable tests of the means using χ^2 difference tests to assess statistical significance and then calculated Cohen's d and U_3 as well as the percent of positive change scores as indices of effect size between the pretest scores and the two retrospective pretest scores (i.e., at 6 months and 12 months). The results of the χ^2 tests and Cohen's effect size for the comparisons between PreT1 and RetroT2, PreT1 and RetroT3, and RetroT2 and RetroT3 are presented in Table 2.

The means of PreT1 were considerably higher than the RetroT2 and Retro T3, indicating that students estimated their self-regulated learning strategies at a quite high level at the beginning of the program. In addition, Cohen's d values showed the magnitude of the difference was noticeably large for the two comparisons of pretest to retro pretest. Here, both values Cohen's U_3 were over 77%, which indicates that 77% of the students' scores on the retro pretest are above the mean of the students' scores at the pretest.

The difference between RetroT2 and RetroT3 was trivial and nonsignificant at the .001 level. And the value of Cohen's U_3 was around 50%, meaning that the two sets of retrospective responses were essentially the same at Time 2 and Time 3 (i.e., their self-reflection ability was similar at the 6- and 12-month retrospective reports).

Analytical model. The longitudinal data were fitted consistent with the model presented below (Figure 2) to assess whether pretest self-judgments or retrospective self-judgments were more predictive of posttest self-judgments. The latent constructs were built from parceled indicators representing students' pretest, retrospective pretest, and posttest self-judgments of various facets of self-regulated learning in math. We used the effects coding scaling method to scale the means and variances in the metric of the original 100-point scale (see Little, Slegers, & Card, 2006, for details of this scale-setting technique).

Measurement invariance models (i.e., configural, weak, and strong models) were established across pretest, retrospective pretest, and posttest. Measurement invariance ensures that the psychometric properties of the instruments are comparable across the three assessment intervals. Following the measurement invariance testing, phantom constructs were added into the model to separate and standardize the variance in the measures and to evaluate the structural relationships among the constructs as assessed at pretest, retrospective pretest, and posttest. As specified in Figure 2, Response Adjustment, the phantom construct, was conceptualized as a simple (constrained) difference score. The Response Adjustment latent variable, which was modeled as a residual variable after the Retrospective Pretest construct is regressed on the Pretest construct, represents the mean difference between Pretest and Retrospective Pretest scores.

Example 1 results. Two follow-up assessments at 6 months and 12 months allowed us to examine the predictive utility of the construct in Figure 2 labeled "Response Adjustment" in two separate models. The Response Adjustment construct in the model is all the reliable variance in the retrospective pretest that is independent of the initial pretest scores collected at baseline. All models passed the measurement invariance testing (comparative fit indexes (CFIs) and Tucker-Lewis indexes (TLIs) were all over 0.97 and ΔCFIs were less than 0.01 and root mean squared error of approximations (RMSEAs) less than 0.08, see Appendix A for detailed measurement invariance testing results and interpretations). Measurement invariance indicates that the changes in the students' responses across time are meaningful in the constructs and not a measurement artifact.

Each hypothesized regression path in Figure 2 was significant at $p < 0.001$ for both 6-month and 12-month models (see Figure 3). Specifically, the Response Adjustment latent variable showed quite strong predictive effects to the Posttest latent variable. The path coefficients for Response Adjustment were significantly ($p < 0.001$) higher than path coefficients from the Pretest latent variable when predicting the Posttest scores. Response Adjustment explained more variance in the Posttest scores than Pretest (28.8% vs. 8.4% in 6-month model, and 18.2% vs. 7.1% in 12-month model). In other words, the posttest responses were better predicted by Response Adjustment (i.e., the variance in the Retrospective reports that is independent from Pretest scores) than by the "true" Pretest responses themselves. Our model results also demonstrate that a response bias exists in the pretest data that were collected at the beginning of the program.

Example 1: Interpretation and discussion. The aim of Example 1 was to evaluate change in learning mind-sets and math strategies after participating in one of two intervention programs. The retrospective findings at both 6- and 12 months support a conclusion that the interventions increased these outcomes quite substantially. The patterns of findings in this example generally replicate those found in previous research comparing these design methodologies. The means of PreT1 were considerably higher than means for the RetroT2 and RetroT3 reports, indicating that students estimated their self-regulated learning strategies at a quite high level at the beginning of the program. This finding suggests that, as students gained experience and training through the intervention program, they became aware of improvements in their knowledge and implementation of learning strategies and therefore adjusted their judgments about their pretest levels of knowledge and skill to be lower using the retrospective pretest instrument. In this case, these adjusted reports provide more sensitive assessments of student belief and ability prior to initiating the intervention program.

Further comparisons of the mean levels at each time point showed that student reports were (1) significantly higher on pretest at Time 1 than both retrospective pretests at 6- and 12 months; but (2) the means remained constant at 6- and 12 months. In other words, retrospective responses were at the same mean level over the 6-month period from the middle to the end of the intervention

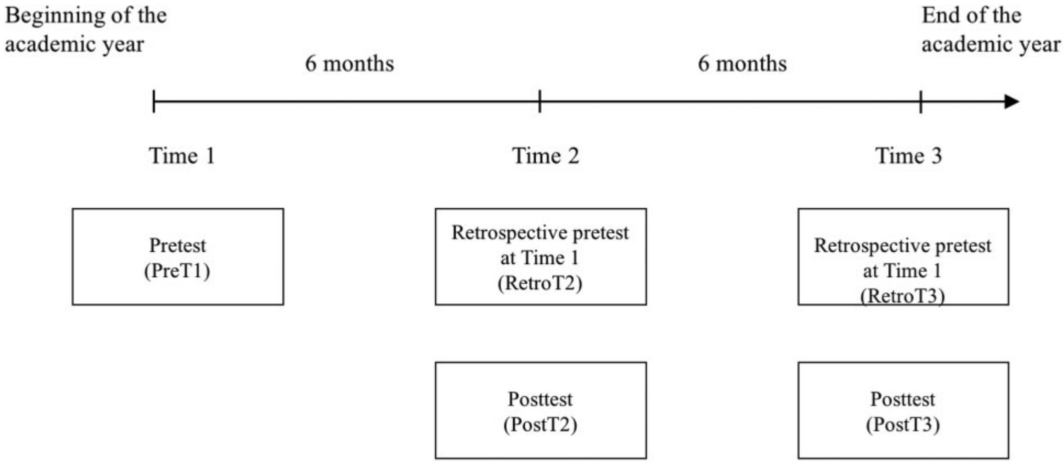
Beginning of the academic year
6 months
6 months
End of the academic year
Time 1
Time 2
Time 3
Pretest (PreT1)
Retrospective pretest at Time 1 (RetroT2)
Retrospective pretest at Time 1 (RetroT3)
Posttest (PostT2)
Posttest (PostT3)

Figure 1. An illustration of the five sets of responses over three time points. Note: PreT1 denoted the pretest at Time 1; RetroT2 and PostT2 denoted the retrospective pretest and posttest at Time 2; similarly, RetroT3 and PostT3 denoted the retrospective pretest and posttest at Time 3.

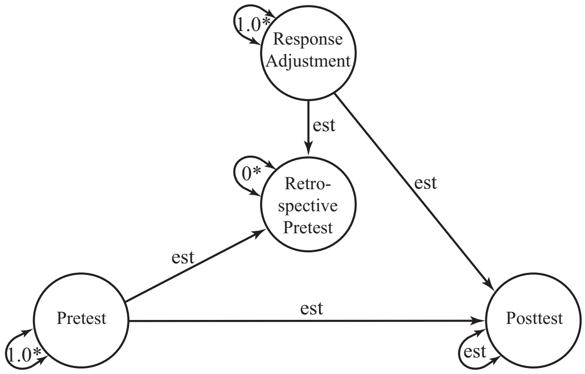

Figure 2. The analysis model.

program. These three sets of comparisons strongly suggest that respondents in this study were able to retrospect and provide recalibrated judgments of themselves prior to their participation in the training program at both 6- and 12-month intervals. The consistent pattern at 6- and 12 months can be viewed as a validity replication of the retrospection.

Follow-up assessments at 6 months and 12 months examined the predictive utility of the Response Adjustment construct. These model results showed evidence of a response bias in the pretest data that were collected at the beginning of the intervention program. Furthermore, in both the 6- and 12-month follow-up models, we found that regression estimates were significantly higher for Response Adjustment than the estimates from the Pretest scores, suggesting that Response Adjustment in the model eliminated response bias in initial responses (at baseline) and thus represented more accurate pretest self-report scores. The results indicate that, after providing an initial report, students' perceptions of the scale measurements changed sometime during the first 6 months of the program and remained consistent for the final 6 months of the intervention. Because of the different response frame of reference between pretest and retro pretest scores, we see that the data collected at the beginning of a program are less predictive of the posttest data than the independent information reflected in the Response Adjustment latent variable (i.e., Response Adjustment is estimated as orthogonal to the pretest construct).

Empirical Example 2

In the second empirical example, we were interested to extend our investigation to study the sensitivity of retrospective pretest scores to detect differences in a longitudinal program evaluation. In this additional example, the evaluation relied on a RPP design, similar to that for Empirical Example 1, but without a pretest survey.

Description of the data. Our second example is based on recent data collected through a national evaluation of Science, Technology, Engineering, and Math (STEM) attitudes among youth participating in after-school STEM programming. In the Afterschool & STEM System-Building Evaluation (see Allen et al., in press), a total of 1,599 students (Grades 4–12) enrolled in an STEM-focused after-school program in 1 of 11 participating states and completed a retrospective pre–post survey on perceived change in STEM-related attitudes. In this sample, 733 (46%) were female students and 866 (54%) were male. These students were mostly from fourth to eighth grades. The data for current use contains two sets of responses—retrospective pretest and posttest, both collected at the end of the program.

Materials. The Afterschool & STEM System-Building Evaluation (see Allen et al., in press) student survey was created using Qualtrics (Qualtrics.com) and administered electronically using tablet devices. Students participating in a STEM-focused after-school program (supported by 1 of 11 participating state after-school networks) completed The PEAR Institute's *Common Instrument Suite* (CIS). The CIS is a self-report survey comprised of a battery of items that measure STEM-related attitudes (i.e., STEM engagement, STEM career knowledge) and 21st-century skills (i.e., quality of relationships with peers and adults, critical thinking, and perseverance). A visual analog scale was utilized as the response scale format, scored from 0 (*strongly disagree*) to 99 (*strongly agree*), with 49 (*neutral*) representing the midpoint.

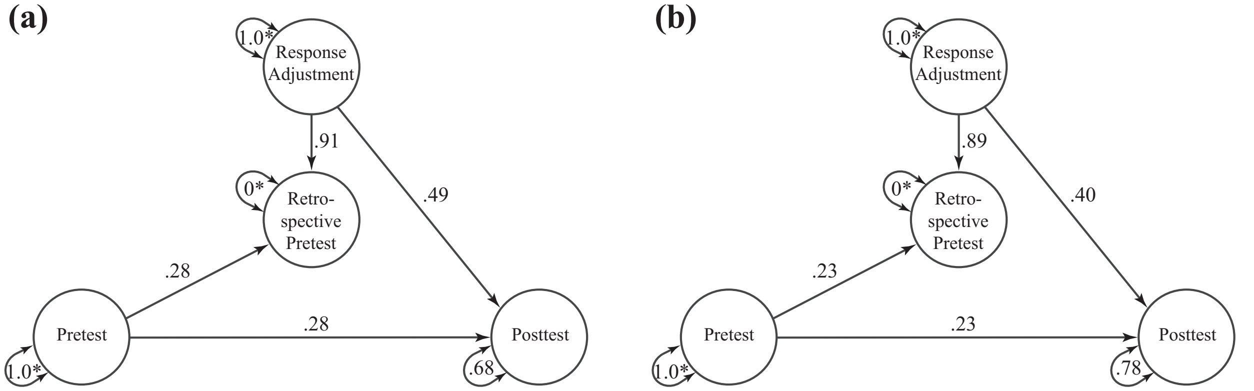

Figure 3. The results of analysis models for self-regulated learning scale at (a) 6- and (b) 12-month retrospective pretest periods. *Note:* All regression paths significant at $p < .001$. Confidence intervals are as follows: Phantom Pretest–Retrospective Pretest (6 month) [3.20, 4.26], (12 Month) [2.33, 3.15]. Response Adjustment–Retrospective Pretest (6 month) [11.42, 15.18], (12 month) [10.94, 14.78]. Phatom Pretest–Posttest (6 month) [2.82, 3.88], (12 month) [2.54, 3.68]. Response Adjustment–Posttest (6 month) [5.2, 7.16], (12 month) [4.07, 5.63].

Table 3. After-school programming results across Grade levels 4–12.

Grade level (*n*)	Retro $\bar{X}$ (*SD*)	Post $\bar{X}$ (*SD*)	Latent $\Delta\bar{X}$	95% CI	*d*	β	R^2
4 (364)	60.01 (19.47)	69.44 (18.46)	9.43	[6.67, 12.18]	.50	.73	.53
5 (353)	60.65 (20.33)	71.18 (19.94)	10.53	[7.56, 13.50]	.52	.64	.41
6 (411)	59.91 (20.73)	69.44 (20.08)	9.53	[6.74, 12.32]	.46	.75	.56
7 (271)	59.56 (19.17)	69.47 (18.18)	9.91	[6.76, 13.06]	.53	.69	.48
8 (115)	60.08 (19.36)	68.93 (19.69)	8.85	[3.80, 13.90]	.45	.85	.72
9–12 (85)	62.25 (18.49)	73.78 (18.18)	11.53	[6.02, 17.04]	.63	.63	.40

Note. Scores were measured on a scale from 0 (*strongly disagree*) to 99 (*strongly agree*), with 49 (*neutral*) representing the midpoint.

Analytic model. Latent variable analyses were conducted on the sample of STEM after-school programs receiving resources and training support from 11 state after-school networks (Florida, Indiana, Iowa, Kansas, Massachusetts, Maryland, Michigan, Nebraska, Oregon, Pennsylvania, and South Carolina). Programs varied in terms of their duration ranging from less than 1 week to greater than 8 weeks in length.

We first examined whether fourth and fifth graders were able to respond retrospectively in a manner similar to older students. Prior literature has asserted that the RPP may not be valid to use with younger students, yet no empirical evidence has supported this statement. Next, we examined the variability across states and program duration to examine the sensitivity of the RPP self-report design to detect the expected differences in state programming and program duration. Measurement invariance across time and states was conducted prior the analyses of the structural regression paths of the posttest on the retrospective pretest. The effects coding scaling method was used. After strong invariance was established, phantom constructs were added into the strong invariance model to evaluate the standardized relationships between the retrospective pretest and the posttest across states.

Cohen's *d* was used for latent mean comparisons. Following Hattie (2009; see also Valentine & Cooper, 2003), we considered a Cohen's *d* value of .2 or greater to represent a positive and important program effect, whereas *d* values between .1 and .2 indicate that an intervention shows "promise."

The effect of program duration on *STEM engagement* posttest scores was analyzed by evaluating the strength of the linear association. Linear associations with sizable magnitudes would indicate the great amount of time a student spent in an after-school program positively influenced their self-reported *STEM engagement* posttest score. Pearson's *r* was used for comparisons.

Example 2 results. Strong measurement invariance was established across retrospective pretest and posttest scores and multiple groups (see Appendix B for measurement invariance results) for all reported results.

We turn first to the results from the grade-level comparison across time are presented in Table 3. Latent mean differences and Cohen's *d* for students across the six grade levels were similar in magnitude. More importantly, the estimated parameters for students in Grades 4 and 5 demonstrated similar mean differences and longitudinal associations between retrospective pretest and posttest scores as older students, supporting the idea that youth as young as Grade 4 can reliably use the design.

Confirmatory factor analysis on the RPP scores detected varying differences between states and program duration on *STEM engagement*. Pronounced mean differences related to *STEM engagement* highlight the sensitivity of the design to variations in after-school program quality across the 11 states. Latent mean-level comparisons revealed numerous and variable differences among all states on *STEM engagement* posttest scores. For brevity, we focus on the results of two states "State 1" and "State 2" (see Table 4 for selected results).

The latent mean difference from the retrospective pretest to posttest for State 1 (n = 122) was 5.04. The effect size was d =.26, and this demonstrated a modest program effect. The latent correlation between the retrospective pretest and the posttest on

Table 4. After-school programming results for State 1 and State 2.

State			Latent				
program	Retro $\bar{X}$ (*SD*)	Post $\bar{X}$ (*SD*)	$\Delta\bar{X}$	95% CI	*d*	β	R^2
State 1	65.78 (16.90)	70.57 (19.20)	5.04	[.53, 9.59]	.26	.75	.57
State 2	61.81 (19.00)	73.41 (15.72)	11.60	[8.0, 15.2]	.65	.66	.43

Note: State 1: $n = 122$, State 2: $n = 179$. Scores were measured on a scale from 0 (*strongly disagree*) to 99 (*strongly agree*), with 49 (*neutral*) representing the midpoint.

STEM engagement was .75 for State 1. The latent mean difference for State 2 ($n = 179$) between the retrospective pretest and posttest was 11.60. The effect size was $d = .65$ and reflected a large positive visible program effect. On *STEM engagement*, the latent correlation between the retrospective pretest and posttest was .66 for State 2. Although State 2 had a larger latent mean difference score compared to State 1, the regression coefficient (β) on the posttest on the retrospective pretest for State 1 was greater in magnitude versus State 2. Furthermore, the R^2 for State 1 was greater in magnitude compared to the R^2 for State 2.

The programs also varied in terms of their duration. Across all states, we found that the retrospective evaluations had a nonsignificant association with program duration but the posttest scores had moderate positive correlations with program duration, with most correlations above .26 (.39 was the highest, but 3 of the 11 states had a correlation below .21). Across all states, the correlation with duration was .26. The variability across states in the strength of the correlation was also associated with overall program quality, indicating differential sensitivity of the RPP design to program characteristics.

Interpretation and discussion. The goal of this example was to demonstrate the sensitivity of the RPP to expected differences in program characteristics and to show that children as young as fourth grade could respond retrospectively in a manner that was consistent with older children. We found considerable variability between states in the degree to which STEM *engagement* scores changed. These state-level differences can be interpreted as differential implementation sensitivity. Similarly, the association with duration of the programs within each state all showed duration effects that also varied by state. Lastly, we also showed that children as young at fourth grade are able to retrospectively report on these effects in a manner that was not significantly different from the older youth.

General Discussion

The goals of this article are to (a) review the numerous problems of using a TPP design, (b) reintroduce the logic of the RPP as a viable alternative to the TPP, and (c) provide evidence of the empirical behavior of the RPP to age differences and differences in various program characteristics. Our review of the problems associated with the TPP design (e.g., response shift bias, weak effect sizes) and the merits of the RPP design (e.g., sensitivity to change), as well as the analyses reported above highlight the utility and the distinct advantages of the RPP design over a TPP design.

Our first example revealed that traditional pretest data collected at the start of program were less predictive and less sensitive to change than data from the retrospective pretest. In addition, our second example showed that without collecting students' pretest scores, the retrospective pretest data were still sensitive to the effects of both after-school program quality between states and program duration within states. Additionally, individuals as young as Grade 4 (approximately 9 years old) exhibited the ability to retrospect about their initial status before entering a program. The similar mean differences and strength of longitudinal associations provide evidence to refute the assertions that RPP designs are not appropriate in studies with preadolescent children.

As mentioned, the retrospective pretest data were sensitive to differences in program duration and program quality. This sensitivity to expected differences provides strong validity evidence to support using the design as a retrospective pretest. The RPP design provides an economical and efficient means to collect quality evaluation data when it may not be feasible to collect data at multiple occasions. Moreover, the time commitment of the respondent is reduced since each item is only asked once but rated twice. In our view, the RPP design might be a preferred approach to collecting program evaluation data, particularly for many noncognitive constructs from beliefs, preferences, and conceptions to attitudes, skills, and values.

Although our study generated several important findings, some limitations exist and need future attention. The most important of these is that data from both examples were initially collected for the purpose of program evaluation, which did not have an experimental focus. In addition, participants were in academic and after-school program settings with STEM focuses. As such, the results generalized from this evaluation using RPP design might be different in other settings. Second, our participants were children in the fourth grade or above, generalizing to younger populations may not be appropriate. Third, in the methodology of Example 2, a calendar was provided to students in the instruction block. The calendar was designed to help anchor students to a predetermined retrospective point in time. A control group that did not see a calendar was not included; therefore, future studies should evaluate the use of including novel retrospective anchors. Finally, even though the response shift bias can be measured by the research design, the results hinge on the accuracy of self-reported data from the surveyed participants.

Based on the above, this study calls for future research that not only targets broader contexts but also extends to experimental designs. Several directions may be considered. One of these is to include participants at lower ages to evaluate the age-related generalizability of the RPP design. Although we emphasized the sensitivity of the RRP to known differences in program characteristics, we still do not have a full picture of the accuracy of the ratings. That is, relative change is well captured by the design but we do not know how well absolute change is captured. Clearly, designing a study that can use some form of gold-standard for accuracy such as behavioral performance data, rigorous observations, or interviews would establish the degree of absolute change that the design brings. Other directions for further validating the RPP design include refining techniques to aid recall and examining the impact of social desirability, recall/memory biases, and related phenomena.

Conclusion

Measuring change is extremely important for field development and securing funding. The most widely used evaluation design, the TPP, often does not detect true change or has weak effect sizes if found to be significant, even when other methods such as behavioral tests, observations, focus groups, and interviews suggest otherwise. The voice of the individual, through self-report feedback, is essential, but so is the need to accurately capture change in response to a given intervention or treatment. Especially in the field of youth development, we need the self-evaluation of the young people regarding their perceptions of change. Based on the extensive literature review provided, as well as the evidence presented from two original evaluation data sets, we have shown that the RPP method is both psychometrically and practically a strong alternative to the TPP. We strongly advocate for a paradigm shift from the TPP to the RPP for the assessment of noncognitive constructs (from beliefs, preferences, and conceptions to attitudes, skills, and values). In our view, the RPP design is ideally suited to reduce bias and to capture true change effects. The design has great promise to provide researchers with an effective design to identify program effects that can inform practice and shape policy.

Acknowledgments

We would like to thank the Dana Center at University of Texas for sharing data to use as an example and we thank Pavel Panko at Texas Tech University for assisting with data preparation. For the second data set, we thank Hye Ran Park at Texas Tech University and Karene Thomas with The PEAR Institute: Partnerships in Education and Resilience at McLean Hospital and Harvard Medical School for their assistance with the evaluation project.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by grants from the Mott and Noyce Foundations and the STEM Next Opportunity Fund (Dr. Gil G. Noam, PI) The contribution to this work by the first three authors was equivalent.

ORCID iD

Todd D. Little https://orcid.org/0000-0002-4146-4712

References

Allen, P. J., Chang, R., Gorrall, B. K., Waggenspack, L., Fukuda, E., Little, T. D., & Noam, G. G. (in press). From quality to outcomes: A national study of afterschool STEM programming. *International Journal of STEM Education*.

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology*, *66*, 5–20.

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

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

Brossart, D. F., Clay, D. L., & Willson, V. L. (2002). Methodological and statistical considerations for threats to internal validity in pediatric outcome data: Response shift in self-report outcomes. *Journal of Pediatric Psychology*, *27*, 97–107.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit Indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255.

Cohen, E. H. (2016). Self-assessing the benefits of educational tours. *Journal of Travel Research*, *55*, 353–361.

Cronbach, L. J., & Furby, L. (1970). How we should measure ‘change’: Or should we?. *Psychological Bulletin*, *74*, 68–80.

Davis, G. A. (2002). *Using a retrospective pre-post questionnaire to determine program impact*. Paper presented at the annual meeting of the Mid-Western Educational Research Association, Columbus, OH.

Drennan, J., & Hyde, A. (2008). Controlling response shift bias: the use of the retrospective pre-test design in the evaluation of a master’s programme. *Assessment & Evaluation in Higher Education*, *33*, 699–709.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

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

Hattie, J. (2009). *Visible learning*. London, UK: Routledge.

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

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

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

Howard, G. S., Schmeck, R. R., & Bray, J. H. (1979). Internal invalidity in studies employing self-report instruments: A suggested remedy. *Journal of Educational Measurement*, *16*, 129–135.

Howard, W., Rhemtulla, M., & Little, T. D. (2015). Using principal components as auxiliary variables in missing data estimation. *Multivariate Behavioral Research*, *50*, 285–299.

Kubota, Y., Yoneda, K., Nakai, K., Katsuura, J., Moriue, T., Matsuoka, Y., & . . . Ohya, Y. (2008). Effect of sequential applications of topical tacrolimus and topical corticosteroids in treatment of pediatric atopic dermatitis: An open-label pilot study. *Journal of American Academic Dermatology*, *60*, 212–217.

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

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*, 59–72.

Moberg, D. P., & Finch, A. J. (2007). Recovery high schools: A descriptive study of school programs and students. *Journal of Groups in Addiction & Recovery*, *2*, 128–161.

Moore, D., & Tananis, C. A. (2009). Measuring change in a short-term educational program using a retrospective pretest design. *American Journal of Evaluation*, *30*, 189–202.

Nakonezny, P. A., & Rodgers, J. L. (2005). An empirical evaluation of the retrospective pretest: Are there advantages to looking back? *Journal of Modern Applied Statistical Methods*, *4*, 240–250.

Nieuwkerk, P. T., & Sprangers, M. A. G. (2009). Each measure of patient-reported change provides useful information and is susceptible to bias: The need to combine methods to assess their relative validity. *Arthritis & Rheumatism*, *61*, 1623–1625.

Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, *14*, 587–598.

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

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

Rohs, F. R. (2002). Improving the evaluation of leadership programs: Control response shift. *Journal of Leadership Education*, *1*, 1–12.

Schwartz, C. E., & Sprangers, M. A. G. (2000). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. In C. E. Schwartz & M. A. G. Sprangers (Eds.), *Adaptation to changing health: Response shift in quality-of-life research* (pp. 81–107). Washington, DC: American Psychological Association.

Sprangers, M. A. G. (1989a). Response-shift bias in program evaluation. *Impact Assessment Bulletin*, *7*, 153–166.

Sprangers, M. A. G. (1989b). Subject bias and the retrospective pre-test in retrospect. *Bulletin of the Psychonomic Society*, *27*, 11–14.

Sprangers, M. A. G., & Hoogstraten, J. (1989). Pretesting effects in retrospective pretest-posttest designs. *Journal of Applied Psychology*, *74*, 265–272.

Sprangers, M. A. G., & Schwartz, C. E. (2000). Integrating response shift into health-related quality-of-life research: A theoretical model. In C. E. Schwartz & M. A. G. Sprangers (Eds.), *Adaptation to changing health: Response shift in quality-of-life research* (pp. 11–23). Washington, DC: American Psychological Association.

Sprangers, M. A. G., Van Dam, F. S., Broersen, J., Lodder, L., Wever, L., Visser, M. R., & . . . Smets, E. M. (1999). Revealing response shift in longitudinal research on fatigue—the use of the thentest approach. *Acta Oncologica (Stockholm, Sweden)*, *38*, 709–718.

Taylor, P. J., Russ-Eft, D. F., & Taylor, H. (2009). Gilding the outcome by tarnishing the past: Inflationary biases in retrospective pretests. *American Journal of Evaluation*, *30*, 31–43.

Valentine, J. C., & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes*. Washington, DC: What Works Clearinghouse.

Verrips, G. H., Hirasing, R. A., Fekkes, M., Vogels, T., Verloove-Vanhorick, S. P., & Delemarre-Van de Waal, H. A. (1998). Psychological responses to the needle-free Medi-Jector® or the multidose Disteronic® injection pen in human growth hormone therapy. *Acta Paediatrica*, *87*, 154–158.

Appendix A

Table A1. Factorial invariance for the analytical model of 6 months in Example 1.

Model	χ^2	*df*	*p*	RMSEA	TLI	CFI	ΔCFI	Held?
Configural	723.660	18	<.001	.071	.953	.976		—
Weak	838.971	22	<.001	.069	.954	.972	.004	Yes
Strong	1,048.219	26	<.001	.071	.952	.966	.006	Yes

Table A2. Factorial invariance for the analytical model of 12 months in Example 1.

Model	χ^2	*df*	*p*	RMSEA	TLI	CFI	ΔCFI	Held?
Configural	1,066.821	18	<.001	.079	.934	.967		—
Weak	1,039.711	22	<.001	.074	.943	.965	.002	Yes
Strong	1,361.407	26	<.001	.075	.937	.957	.008	Yes

Note. Both configural models across the time demonstrated good fit to the data, with both CFI and TLI above .90, RMSEA below .08. As the configural factorial invariance models were satisfied, the weak and strong invariance (i.e., the corresponding item loadings and intercepts were constrained equivalent) were continually assessed in a sequence. Following the Cheung and Rensvold (2002) criteria for factorial invariance, if the change between each level of constraint model is less than .01, then invariance holds. The fact that the weak and strong invariant models held indicated that students' interpretations of the items are equivalent and perform on the same metric over the time.

Appendix B

Table B1. Measurement invariance across time and grades in Example 2.

Model	χ^2	*df*	*p*	RMSEA	TLI	CFI	ΔCFI	Held?
Configural	66.795	30	<.001	.068	.991	.997	—	—
Weak	168.384	52	<.001	.092	.984	.991	.006	Yes
Strong	268.525	74	<.001	.099	.981	.985	.006	Yes

Table B2. Measurement invariance across time and states in Example 2.

Model	χ^2	*df*	*p*	RMSEA	TLI	CFI	ΔCFI	Held?
Configural	106.936	55	<.001	.081	.988	.996	—	—
Weak	254.618	97	<.001	.106	.979	.988	.008	Yes
Strong	355.097	139	<.001	.103	.980	.983	.005	Yes