American Journal of Evaluation
Volume 30 Number 2
June 2009 189-202

10.1177/1098214009334506
http://aje.sagepub.com
hosted at
http://online.sagepub.com

Measuring Change in a Short-Term Educational Program Using a Retrospective Pretest Design

Debra Moore
Cynthia A. Tananis
University of Pittsburgh

The Pennsylvania Governor's School for International Studies is an intensive summer program designed to give talented high school students a challenging introduction to the study of international affairs. One focus of the evaluation seeks to understand the effect of the program on the students' perception of their knowledge concerning core issues. Across the long history of the program, a variety of measures were used (and subsequently discarded) to assess changes in knowledge and perception of competence. Four years ago the program instituted a *retrospective pre-post design*. Results from these years, indicate that these students have consistently overestimated their pre-test understanding of core competencies emphasized in the program and that they seem better able to assess their knowledge gains and their initial inflated sense of knowledge as a result of the program. This article offers an overview of the development, application, use and analysis of a retrospective pre-post instrument to address response shift bias.

Keywords: *pre–post test; response-shift bias; retrospective pre–post test; gifted education; perceived competence*

Introduction

Evaluation of short-term intensive intervention programs is often problematic (Bamberger, Rugh, Church, & Fort, 2004; D'Eon, Sadownik, Harrison, & Nation, 2008; Pratt, McGuigan & Katzev, 2000). This is especially true when specific measures of content knowledge are not available or when resources and expertise do not allow an appropriate instrument to be constructed. In these cases, program evaluation relies on self-report measures of a participant's perceived change as a measure of program effectiveness. Although self-report measures have their own documented limitations (i.e., Krosnick, 1999), evaluators also are hindered by a lack of design options (Hill & Betz, 2005), which in turn can be exacerbated by insufficient time and money as well as restrictive situations which force trade-offs in reliability and validity.

Authors' Note: Cynthia A. Tananis, 4316 WWPH, School of Education, Pittsburgh, PA 15260; e-mail: tananis@pitt.edu.

One documented limitation of the popular pretest–posttest design is a phenomenon that Howard, Schmeck, & Bray (1979) termed *response-shift bias*. This phenomenon occurs when a participant uses a different internal understanding of the construct being measured to complete the pretest and the posttest. The resulting shift in understanding introduces bias into the attempt to accurately measure program effectiveness. Howard (1980) proposed the use of a retrospective pretest to control for this source of measurement error. This article demonstrates the response-shift phenomenon among a group of gifted high school students in an educational setting and discusses the importance of validity studies in an effort to give evaluators confide nce in the accuracy of their results when using the retrospective pre–post test design.

Background and Rationale

Obtaining accurate and meaningful data to assess the impact of program effectiveness can be difficult and may involve trade-offs between collecting data efficiently and inexpensively and using instruments that have good reliability and provide valid interpretations of scores (validity). Ideally, to collect the best data, the evaluation instrument should have good psychometric properties (Hill & Betz, 2005). Principally, the instrument should accurately measure the construct it intends to measure for the purpose of judging program effectiveness. This requires collecting validity evidence for the inferences from the results obtained from the instrument. Additionally, the instrument should consistently measure the desired construct. This requires collecting reliability evidence for the instrument in question. Evaluators, however, often work under tight time and budget constraints, which may preclude some of these efforts (Bamberger et al., 2004).

In addition, evaluation should take a minimum of program time, be inexpensive and easy to administer, analyze, and report (Cooke, 1998). Because of this, design options for evaluators may be limited (Hill & Betz, 2005), often including the lack of easily available control groups. Time and budget restraints typically do not allow for the expense and effort needed to find and administer the evaluation instrument to an adequate control group (Bamberger et al., 2004). Because self-assessments meet the time and expense criteria well, even if they do not provide for a control group, they are frequently used in program evaluation (D'Eon et al., 2008).

Within the context of self-assessments, design options have been limited to a Posttest Only design or a One-Group Pretest–Posttest design (Campbell & Stanley, 1963). As discussed in Stanley and Campbell, the Posttest Only design suffers from a lack of a baseline for comparison. In this regard, the One-Group Pretest–Posttest design is more desirable because it provides a baseline for comparison (Campbell & Stanley, 1963). From an evaluation perspective, the assumption within this design is that the difference between pretest and posttest scores reflects the amount of change because of the program (Pratt et al., 2000). In order for this to be an accurate reflection of change because of the intervention in question, however, the researcher assumes that the participant is using the same internal standard to judge attitude, behavior, or perception as it applies to the items on the pretest and the posttest (Howard, Schmeck, et al., 1979). If this internal metric used to complete the pretest shifts before completing the posttest, the posttest scores will reflect not only the program effect but also this shift in understanding because of the intervention (Howard, Schmeck, et al., 1979). This phenomenon would bias the attempt to accurately measure program effectiveness based on pretest and posttest scores.

Response-Shift Bias—Considering the Literature

Howard, Ralph, et al. (1979) renewed interest in this phenomenon of a change in a participant's frame of reference, which he termed "response shift." An evaluation for an Air Force

communication skills training program suggested that participants had become more dogmatic as a result of the workshop. Interviews with the participants afterward revealed that they had changed their original perceptions about their initial levels of dogmatism. In another study (Howard et al., 1979) designed to investigate this phenomenon further using the same training program, the participants were divided into two groups. One group was given the Rokeach Dogmatism Scale (RDS) as a pretest and again as a posttest, whereas the other group was given the RDS as a posttest and a retrospective pretest. Analysis of the pre–post test scores for the first group revealed no difference in the participants' level of dogmatism following the workshop. However, analysis of the retrospective pre–post test scores for the second group revealed they were significantly less dogmatic after the training session.

Because an objective measure of dogmatism was not available at the time, Howard, Ralph, and his associates (1979) decided to design a study that would also include an objective measure of the construct being measured in addition to the self-report indices. He chose a workshop on assertiveness for females. As before, analysis of the pre–post test scores revealed no significant gains in assertiveness for this group, whereas analysis of the retrospective pre–post test group revealed significant gains in assertiveness. Additionally, scores on the objective measure of assertiveness were more highly correlated with the retrospective pre–post test scores than with the pre–post test scores, although these correlations were not significantly different.

One final study in this series of studies by Howard, Ralph, et al. (1979) focused on changes in helping skills over a semester-long course designed to improve those skills. For this study, interviews of the participants interacting with clients were conducted and tape-recorded before and after the course. These interviews were rated by judges for the level of helping skills. In addition, self-report measures in the pretest, posttest, and retrospective pre–post test formats were given. Judges' ratings and scores from the retrospective pre–post test format showed significant changes in the level of helping skills, whereas scores from the pre–post test format showed no significant treatment effects. When participants were asked to remember their pretest scores, interestingly, the remembered pretest scores were almost identical to the actual pretest scores and still significantly different from the retrospective pretest scores.

To determine whether a similar response-shift effect was present for cognitive variables, Bray, Maxwell, and Howard (1984) used a retrospective pre–post test design, along with an objective measure, to determine perception of change in knowledge of learning theory for 33 students enrolled in an educational psychology class. Although the authors do not advocate using a self-reported measure of change in place of objective measures of knowledge gain, they were intrigued to find that response-shift bias existed in a cognitive context. In addition, the objective measure of knowledge change, a traditional essay exam, was more highly correlated with the retrospective pre–post test scores than with the pretest and posttest scores.

Since this series of studies by Howard and his colleagues, several other studies have been performed in a variety of settings which also provide evidence of response-shift bias. Gutek and Winter (1992), investigating job satisfaction, concluded that response-shift bias is a threat to validity across time and further suggest including retrospective pretests in studies which investigate attitudinal changes over time. Hoogstraten (1982) tested the idea of sensitization by the pretest on a retrospective pre–post design in psychology students. He found that the self-reported retrospective pre–post test scores showed more improvement as a result of treatment than the traditional pre–post test scores, and that the retrospective pre–post test scores agreed more similarly with the objective measure of performance then the traditional pre–post test scores.

Manthei (1997) used a retrospective pre–post test design to evaluate 31 master's level counselors in training that further illuminates the response-shift bias phenomenon. In this study, when the data were analyzed, three patterns of student responses emerged. One group had

pretest scores approximately equal to their retrospective pretest scores. Individuals from this group were interviewed, and it was determined that they felt they already possessed a high level of skill in the area and the class only reinforced their belief. The second group of students had pretest scores lower than their retrospective pretest scores. Individuals from this group were also interviewed, and it was determined that the training had given these students a greater appreciation of their beginning skill level. The final group of students had pretest scores higher than their retrospective pretest scores. Interviews with these students revealed that they had systematically overestimated their beginning knowledge and skill level and that the training helped them see their deficiencies more clearly.

Pratt et al. (2000) used a retrospective pretest as one measure in a battery of assessments to determine the effectiveness of the Oregon Healthy Start (OHS) program. Comparing posttest scores to pretest scores revealed a significant improvement on four of the seven items. However, comparing the scores from the retrospective pretest and the posttest revealed a significant difference on all seven items. Cantrell (2003) investigated the impact of methods and practicum classes on the teacher self-efficacy beliefs of preservice science teachers using a retrospective pre–post test design. Students reported that in the beginning they thought they could explain concepts to students, but as they were asked to do it in a teaching situation, they found it more difficult to do than they first imagined.

In summary, there is substantial empirical evidence to show that response-shift bias occurs when self-report instruments are used to measure differences in a participant's perception and that this bias can mask program effectiveness. Response shift is most likely to occur in contexts when the training or educational program being evaluated is designed to increase a participant's awareness of the specific construct that is being measured. In these cases, the retrospective pre–post test design can help control for response-shift bias by collecting self-reported change more accurately (Howard, 1980). Because the posttest and the retrospective pretest are administered simultaneously at the end of the program, participants are more likely to use the same understanding of the construct to complete both the posttest measure and the retrospective pretest measure.

Despite the apparent strength of this design in controlling for response-shift bias, it is not without limitations. Klatt and Taylor-Powell (2005) note that the retrospective pre–post test design still uses a self-report methodology and, as such, can suffer from the same types of biases. One well-documented source of bias in self-evaluation is social desirability bias (Krosnick, 1999). This phenomenon refers to the systematic underreporting of behaviors or attitudes that the participant considers not socially respected and a corresponding overreporting of behaviors or attitudes that the participant considers socially respected. This can be deliberate or unconscious (Krosnick, 1999) but introduces error in the measurement of program effectiveness by inflating or deflating responses and overall results. Related to this phenomenon is the problem of acquiescence (Krosnick & Fabrigar, 1998). Acquiescence refers to a participant's tendency to answer affirmatively, or in agreement, with any assertion put forth in a question. This phenomenon would also tend to inflate responses and overall results. Another well-documented source of bias in self-report measures is recall accuracy (Schwarz, 2007). Research has shown that respondents tend to use estimation strategies when applicable rather than a strict recall-and-count strategy, and that even seminal events in participant's lives tend to be underreported (for a comprehensive review see Schwarz, 2007). In addition, Schwarz (2007) provides a thorough discussion of the literature regarding minor changes in wording, word order, and question format, and the differential effect these changes have on respondents' answers.

All these limitations are probably best understood in the context of the cognitive processes that respondents undergo to complete a questionnaire. Krosnick (1999) and Schwarz (2007) both provide a review of the relevant literature on what has become the generally accepted

stepwise process that participants use to respond to a survey item. First, participants must interpret the item and figure out what is being asked. Second, they must search their memories for relevant information and then, third, integrate the retrieved information into a single judgment about what they are being asked. Finally, the participants must select the response that best reflects the single judgment they have formulated. Krosnick (1999) has extended this paradigm to include a continuum of thoroughness, optimizing versus satisficing, in responding to an item. Optimizing refers to the optimal way that a researcher would like for participants to respond and includes all four of the steps performed as outlined above. Satisficing refers to a less than optimal way of responding to a question in which the participant estimates their answer because they do not or cannot expend the cognitive energy required to provide a thorough answer.

Ross (1989) offers specific caution about the flawed nature of recall and how an implicit theory of change (p. 342) can lead respondents to assume a positive change because of program intervention. Taylor, Russ-Eft, and Taylor (2009) further summarize this phenomenon, indicating that respondents start with an inflated sense of posttest competence, and from that point indicate a lower level of competence, retrospectively, to maintain consistency with their theory of a positive change because of the intervention.

Klatt and Taylor-Powell (2005) also note two limitations that are specific to the retrospective pre–post test design. They state that little is known about how a participant's culture, age, stage in life, or literacy level will affect their ability to complete a retrospective pretest and what accommodations would be needed in different situations. This study addresses one aspect of this limitation by demonstrating the phenomenon among a group of academically gifted high school students in an educational setting and the results obtained from using the methodology. In doing so, it generalizes the methodology to a different population. The second limitation noted by Klatt and Tayor-Powell (2005) is the lack of information about what are best practices when using a retrospective pre–post test design. It is hoped that this study will provide additional information about the methodology, given the different situation and population to which it is applied.

Method

Program Background

The Pennsylvania Governor's School for International Studies (PGSIS) serves the primary purpose of providing a select group of academically talented and highly motivated high school students with a challenging introduction to the study of international affairs and global issues. In addition, as is important to any global education program, students are exposed to a world language unfamiliar to them and study cultures connected to that language. Created in 1984 by the Pennsylvania Department of Education in cooperation with the Pennsylvania Council for International Education and Center for International Studies at the University of Pittsburgh, PGSIS has, for the past 23 years, included formal coursework, a variety of co-curricular activities, and a residential setting designed to foster personal and intellectual development. The school occurs during summer months for a period of 5 weeks.

The school offers a core curriculum designed to provide the students with a challenging, integrated introduction to global issues, intercultural communication, Japanese language and culture, Portuguese language and Brazilian culture, global citizenship, negotiation and diplomacy, and international political economy. The program includes formal course work, independent and collaborative research, experiential learning through simulations and fieldwork

Table 1
Cohort Characteristics for All Years

Cohort Characteristics	2004, $n = 100$	2005, $n = 100$	2006, $n = 100$
Gender			
Male	40	36	36
Female	60	64	64
Native language			
English	83	86	73
Asian derivative	3	8	15
Romance	3	3	4
Other	11	3	8
Racial background			
Caucasian	72	73	73
Asian American	20	13	15
Hispanic Latino	3	6	2
African American	0	3	3
Mixed	1	1	3
Native American	1	0	0
Other	3	4	4

as well as special events and cultural activities with a heavy emphasis on interdisciplinary and multidisciplinary learning. Students participate in an ongoing, country-level simulation experience, where they act as key government officials that negotiate treaties and agreements, consider resources and make decisions that ultimately affect themselves and others. This simulation, International Communication and Negotiation Simulations (ICONS), is integrated into coursework and other discussions as well.

Participants

Participants for this study were 100 high school junior students who were selected to attend the PGSIS at the University of Pittsburgh from across the state of Pennsylvania in three consecutive years; 2004, 2005, 2006.[1] All the three years were similar in gender composition with females outnumbering males approximately 3:2 (see Table 1). The racial background of the students across the three years was also very similar. The sample was predominately White, with representation by Asian American, Hispanic Latino, mixed ethnic backgrounds, Native American, African American, or other depending on the year (see Table 1). Most of the participants were born in the United States and identified English as their native language. Other languages represented depending on the year were an Asian derivative, a Romance language, or other languages (see Table 1).

Instrument Development

All evaluation activities were planned in conjunction with the PGSIS administrative staff and faculty and conducted by the Collaborative for Evaluation and Assessment Capacity (CEAC), at the University of Pittsburgh's School of Education.[2] The primary purposes of the evaluation were to document the impact of PGSIS on the participating students and to gather formative information that would be used for planning activities for subsequent programs, and for long-range planning purposes. Two instruments, an *Incoming* (pretest) and *Exiting* (posttest) *Questionnaire*, were administered to collect data to answer the formal evaluation

questions. In addition, students were provided with various opportunities to voice their opinions and give feedback during discussions with the school administration and faculty. A formal focus group process involving all students is also conducted toward the end of the school session. These additional evaluation activities were included to provide a deeper context for the data collected via surveys and also afforded students the opportunity to express formative reflections throughout the experience.

The *Incoming Questionnaire* is a survey that collects demographic information and a preassessment of core competencies, as designated by PGSIS faculty and staff, integral to a global education program as well as preassessment data regarding students' knowledge of issues they will study throughout the program and, especially, as part of the ICONS experience. This survey consists of close-ended items, using a 4-point Likert scale, and is completed on the first full day of the program. The *Exiting Questionnaire* provides students the opportunity to rate each course within the program, using a 5-point Likert scale, on the following variables: intellectual challenge, classroom environment, use of class time, readings and materials, and teaching techniques. Originally, this questionnaire consisted of both closed- and open-ended questions to elicit ratings and feedback regarding specific academic activities, the residency program, school administration, career and educational plans, and overall reflections on the conclusion of the 5-week program. In addition to the formal instruments, school administrators and the evaluators observed a variety of class sessions and activities throughout the program. However, it was observed that the magnitude of increases reflected in a comparison of scores on the close-ended items did not seem to mirror the magnitude of increases expressed on the open-ended items and the in-class exercises.

The apparent disconnect between the data collected on the close-ended portion and open-ended portion of the *Exiting Questionnaire*, along with the information gathered during the other evaluation activities, led to the inclusion of a retrospective pretest in the *Exiting Questionnaire* starting in 2004. The retrospective pretest was administered on the same page as the posttest except, on this portion of the questionnaire, students were asked to *reassess* their prior knowledge and skills by reflecting back and reassessing themselves on the same core competencies as those items included on the *Incoming Questionnaire* (see appendix). This was an attempt to capture how much students actually perceived they have changed or developed within each area and account for any response-shift bias that could have occurred. Data collected across three consecutive years since the introduction of the retrospective pretest were then examined for this possible confounding factor.

Instrument

There were 11 items common to the *Incoming* and *Exiting Questionnaires* across the three years that were designed to measure participants' perception of their knowledge of core competencies addressed in the program with the retrospective pretest. The incoming data set and exiting data set from each respective year was merged using the ID variable as the matching variable. Only students completing both the *Incoming* and *Exiting Questionnaire* were included in the analysis. A composite pretest score, composite posttest score, and composite retrospective pretest score were computed by summing the responses across the 11 items (see Table 2 for Cronbach's [1951] α reliabilities). Once the data set for each year was created, a year variable was added to each. The three data sets were then transposed and merged into one data set for analysis.

Gender and Self-Efficacy

Because a strong link between gender and self-efficacy, especially concerning academic domains, is well documented (Bandura, 1977; Choi, 2004; Lent, Brown, & Gore,

Table 2
Reliability Estimates for Composite Scales by Year and Overall

	Pre	Post	Retro
2004	0.78	0.90	0.88
2005	0.82	0.87	0.85
2006	0.73	0.84	0.80
Overall	0.78	0.87	0.85

1997) and the evaluation instruments asked for participants to rate their own level of competency with regard to the prompts, it was possible that the gender of the participants could confound the results of any analysis. For this reason, interaction effects were of concern with these data.

Results

A 3 × 3 × 2 mixed analysis of variance (ANOVA) was performed using the SAS MIXED procedure adjusting for the violation of compound symmetry (Wolfinger & Chang, 2007). The within-participants independent variable was time with three levels; pretest, posttest, and retrospective pretest. The two between-participants independent variables were gender (male, female) and year (2004, 2005, 2006). The assumption of normality was violated for males on the posttest for 2004 and for females on the posttest for 2004.[3] The assumption of normality was met in all other cases. Because of the large sample size ($n = 295$),[4] even within the disaggregated data analysis, effect sizes using Cohen's d were calculated for results to gauge the practical significance of the results (Cohen, 1988).

Overall Results

The pattern of differences on test scores between genders was significantly different among the three test times, $F(2, 289) = 10.72$, $p < .001$. Additionally, the pattern of differences on test scores among the three years was significantly different among the three test times, $F(4, 289) = 4.24$, $p = .002$. There was also a significant difference on test scores among the three test times averaged across gender, $F(2, 289) = 296.91$, $p < .001$. Males had significantly higher pretest scores and retrospective pretest scores than females; $F(1, 289) = 4.97$, $p = .027$, Cohen's $d = .130$ and $F(1, 289) = 14.06$, $p < .001$, Cohen's $d = .218$, respectively. Females had marginally significantly higher posttest scores than males; $F(1, 289) = 3.62$, $p < .058$, Cohen's $d = .111$. There was also a significant difference on test scores in 2004, 2005, and 2006; $F(2, 289) = 99.55$, $p < .001$, $F(2, 289) = 113.91$, $p < .001$, and $F(2, 289) = 91.09$, $p < .001$, respectively. There were no other significant effects.

Simple Main Effect of Gender

To determine the pattern of differences on scores between genders among the three test times, simple main effects were performed using a Scheffé adjustment. Males had pretest scores and retrospective pretest scores that were significantly lower than their posttest scores, $t(289) = -6.45$, $p < .001$, Cohen's $d = .833$ and $t(289) = 12.34$, $p < .001$, Cohen's $d = 1.38$, respectively. Retrospective pretest scores were also significantly lower than pretest

Table 3
Descriptive Statistics for Test Scores Between Genders Among Test Times

	Mean	*SD*
Female, $n = 184$		
Pre	28.94	0.351
Post	35.21	0.420
Retro	22.42	0.446
Male, $n = 111$		
Pre	30.21	0.456
Post	33.89	0.548
Retro	25.17	0.583
Overall, $n = 295$		
Pre	29.58	0.286
Post	34.55	0.345
Retro	23.79	0.367

scores, $t(289) = 9.38$, $p < .001$, Cohen's $d = .922$. Likewise, females had pretest scores and retrospective pretest scores that were significantly lower than their posttest scores, $t(289) = -14.23$, $p < .001$, Cohen's $d = .315$ and $t(289) = 23.69$, $p < .001$, Cohen's $d = .718$, respectively. Retrospective pretest scores were significantly lower than pretest scores, $t(289) = 15.87$, $p < .001$, Cohen's $d = .545$. Overall, this same pattern of results was seen averaged across genders; pretest scores and retrospective pretest scores were significantly lower than posttest scores, $t(289) = -13.83$, $p < .001$ and $t(289) = 24.19$, $p < .001$, respectively, and retrospective pretest scores were also significantly lower than pretest scores, $t(289) = 17.08$, $p < .001$. Descriptive statistics for test scores between genders among the three test times are reported in Table 3.

Simple Main Effect of Time

To determine the pattern of differences on scores among years among the test times, simple main effects were performed using a Scheffé adjustment. For 2004, pretest scores and retrospective pretest scores were significantly lower than posttest scores, $t(289) = -9.66$, $p < .001$, Cohen's $d = .562$ and $t(289) = 14.10$, $p < .001$, Cohen's $d = .826$, respectively. Retrospective pretest scores were significantly lower than pretest scores, $t(289) = 8.26$, $p < .001$, Cohen's $d = .481$. For 2005, pretest scores and retrospective pretest scores were significantly lower than posttest scores, $t(289) = -5.83$, $p < .001$, Cohen's $d = .343$ and $t(289) = 14.30$, $p < .001$, Cohen's $d = .831$, respectively. Retrospective pretest scores were significantly lower than pretest scores, $t(289) = 12.55$, $p < .001$, Cohen's $d = .730$. For 2006, pretest scores and retrospective pretest scores were significantly lower than posttest scores, $t(289) = -8.51$, $p < .001$, Cohen's $d = .494$ and $t(289) = 13.48$, $p < .001$, Cohen's $d = .784$, respectively. Retrospective pretest scores were significantly lower than pretest scores, $t(289) = 8.66$, $p < .001$, Cohen's $d = .503$. However, there was no significant difference on pretest scores among the years; $t(289) = -1.66$, $p = .999$; $t(289) = .37$, $p = 1.00$; and $t(289) = 2.01$, $p = .854$. There was also no significant difference on posttest scores among the years; $t(289) = 1.12$, $p = .996$; $t(289) = 1.03$, $p = .998$; and $t(289) = -.11$, $p = 1.00$. Additionally, there was no significant difference on retrospective pretest scores; $t(289) = 1.85$, $p = .903$; $t(289) = .63$, $p = .999$; and $t(289) = -1.22$, $p = .993$. Descriptive statistics for test scores among years among the

Table 4
Descriptive Statistics for Test Scores Among Years Among Test Times

	Mean	*SD*
2004, $n = 100$		
Pre	29.28	0.486
Post	35.15	0.584
Retro	24.54	0.622
2005, $n = 96$		
Pre	30.44	0.505
Post	34.20	0.614
Retro	22.86	0.653
2006, $n = 99$		
Pre	29.02	0.497
Post	34.30	0.595
Retro	23.97	0.632

three tests are reported in Table 4. In addition, a trend toward increasing reliabilities between the pretest and retrospective pretest were noted for all the three years.

Discussion

The pattern of pretest scores, posttest scores, and retrospective pretest scores was consistent across years, between genders, and overall. Pretest scores and retrospective pretest scores were *significantly lower* than posttest scores for 2004, 2005, 2006, and across all the three years. Retrospective pretest scores were *significantly lower* than pretest scores for 2004, 2005, 2006, and across all the three years as well. This pattern was also true for male pretest, posttest, and retrospective pretest scores as well as female pretest, posttest, and retrospective pretest scores whether they were analyzed separately or aggregately. This pattern of results is consistent with previous studies using the retrospective pretest and suggests that participants are overestimating their initial level of competency. In this case, overall program effectiveness would be underestimated when measured by pretest–posttest scores.

Gender differences did appear, however. Males had a significantly higher mean pretest score and retrospective pretest score then females. These results appear to be consistent with the research on gender and self-efficacy. However, even though only marginally significant, females had a higher mean posttest score than males. These gender differences appear to be separate from and operating under some influence other than the change in internal metric taking place because the same overall pattern of scores was seen despite gender. The effect sizes for the results clearly show a strong impact of time (pretest, posttest, or retrospective pretest), rather than gender, on scores for all participants across all years. Although gender may be influencing the results, time is exhibiting more of an influence than gender.

The three years of data show a tendency for the reliability estimates to increase between pretests and retrospective pretests (0.10, 0.03, and 0.07, respectively) even though the change in reliability is more dramatic for some years. Because Cronbach's α was used as a measure of internal consistency, by definition, this trend suggests that students are answering the retrospective pretest items more consistently than the pretest items. This could perhaps mean that the students have a more coherent understanding of the construct after the program compared to before the program; however, it is difficult to conclude this based on the limited information

available in this study. It is interesting to note that D'Eon et al. (2008) also found this pattern of increased reliability in retrospective pretest scores compared to pretest scores. These results suggest using the retrospective pretest to improve reliability and consistency in responding.

There are, however, limitations to the study. It is possible that recall accuracy could be playing a role in the results seen in this study. Students were involved in an intense immersion setting for 5 weeks and then asked to reflect back on before the program started. Students could have trouble reflecting back to before the program and the intensity of the program could cloud the memory. However, recall accuracy is normally associated with frequencies, especially frequencies of behavior, which is not the case for this particular evaluation. Acquiescence could be an alternative explanation for the results seen, as well. In this case, however, the tendency to agree with any statement would not explain the lower retrospective pretest scores compared to pretest scores and posttest scores. Acquiescence would tend to make the retrospective pretest scores higher.

The most plausible alternative explanation for the results seen in this study would be some form of social desirability bias. Hill and Betz (2005) suggest that, in this situation, students would underestimate retrospective pretest scores and/or overestimate posttest scores in an effort to justify the amount of effort they put into the course. Alternately, Paulhus (2002) proposes a source of bias he calls *impression management*. Impression management is defined as consistently answering in such a way as to always present oneself in a positive way. Impression management could explain inflated pretest scores, inflated posttest scores, and the tendency to answer in socially desirable ways. All these response sets would be an effort by the respondent to present themselves in a favorable way. This is further suggested by Taylor, Russ-Eft, and Taylor (2009) who found evidence of bias from respondents' implicit theories of positive change as well as self-enhancing efforts. In our case, if the students felt that learning a lot during the program was the favorable answer, it could explain the deflated retrospective pretest scores in relation to the posttest scores. This would mean that the relationship between the pretest scores and retrospective pretest scores in this study was a happy coincidence.

Considering the relationship between all the three scores in the current study, for an alternative hypothesis to be the case, students would have to overestimate their posttest scores and underestimate their retrospective pretest scores *at the same time* while remembering their pretest scores, so they could adjust their posttest and retrospective pretest scores accordingly. In addition, just as interview evidence in previous retrospective pretest studies shed light on the response-shift bias phenomenon, the open-ended comment evidence from the *Exiting Questionnaire* tends to dispute alternative explanations for the current study. Students overwhelmingly mentioned never knowing there was so much to learn or having learned so much during the program when asked to "describe the three most important ideas or skills you learned as a student." To explore this notion further, students offered numerous specific examples in open-ended survey responses, focus groups, and whole group debriefings, indicating that they had overestimated their knowledge and skills prior to the program, and had come to better understand the complexities and depth of knowledge required for competence as they progressed in the program. One student comment summarizes a sentiment gathered from many of the students across the three years of data we explored:

> I really thought I knew a lot about [global issues] before coming here. Maybe I did know more than some other students back home because I have special interests and want to work internationally, but when I got here and started going to classes and doing [the simulations] I found out just how little I really did know. I had to really work hard to catch up and be able to do [the simulations] well. I look back now and realize how much more I know because of [the program] and I'm amazed.

Summary and Future Directions

The data from this study support previous research suggesting that before beginning a program, participants can *overestimate* their initial level of competency, but after completing the program, will change their perception of that initial level of competency and reflect this change in the retrospective pretest scores. Therefore, the difference between the retrospective pretest score and the posttest score may give a better estimate of the magnitude of change in these situations. Of particular importance, in this study, is the documentation of the phenomenon in a group of academically gifted high school students which did not exist in the literature to date. In this study, the statistical analysis and the qualitative data align and support the use of the retrospective pretest design as a more accurate measure of program effectiveness.

Regardless of design used, however, measurement error is present. Hill and Betz (2005) suggest that evaluators could be trading one type of bias for another by using the retrospective pre–post test design. They found that retrospective pretest items were more biased than pretest items in certain contexts. This is an important reminder that one way to help reduce measurement error is to construct good instruments paying particular attention to wording, item context, and recall context. However, if the retrospective pretest methodology truly means trading one type of bias for another, the question becomes which of the two biases is less desirable of the two.

Although it might seem feasible to use the pre–post test score because it represents a conservative estimate of program effectiveness, the issue of accuracy still remains. In cases where the response-shift bias is greater than any bias introduced in using the retrospective pretest, the retrospective pre–post test score becomes a less-biased measure of program effectiveness. In cases where socially desirable responding would be a greater source of bias than the response-shift bias, the pre–post test score would be a less-biased measure of program effectiveness. One issue of concern is which design is more appropriate in cases where the pre–post test scores did not show significant improvement, but the retrospective pre–post test scores did show significant improvement. Another important concern is in situations when evaluators are forced to use a posttest only design, whether because of time and money constraints or because of pretest sensitivity issues. In these cases, can the retrospective pretest provide an accurate baseline by which to eventually judge program effectiveness?

From this perspective, decisions must be made as to which method of pretest, posttest, or retrospective pretest/posttest might be most effective and efficient to use. Our study focused on issues of validity: how to best represent the participant's assessment of preprogram competence compared to postprogram competence.[5] These findings were then triangulated with other data sources to explore areas of potential program impact. A few retrospective pre–post test studies (Bray et al., 1984; Hoogstraten 1982; Howard et al., 1979) have examined the validity of retrospective pre–post test scores versus pre–post tests scores. These studies found that the self-report retrospective pretest scores correlated more highly with scores on objective pretest measures of skill development or knowledge than the self-report pretest scores. These studies suggest that the retrospective pretest may be capturing a more accurate measure of preintervention function than a pretest given before the program begins. Additional validity studies involving the retrospective pre–post test methodology in conjunction with objective measures of the same constructs could shed light on which methodology will yield less-biased measures of program effectiveness. In the meantime, used with the cautions identified in the literature, the retrospective pre–post test design seems a promising alternative to the typical pre–post test design in settings where perception of knowledge (both pre and post) serves to evaluate program effectiveness.

Notes

1. Students must apply for admissions for this highly competitive placement within a set of specialized "schools" for gifted students. All students are rising seniors in a Pennsylvania high school.

2. Further information is available via the Collaborative for Evaluation and Assessment Capacity.

3. A series of nonparametric tests were run on the data because of the violation of normality in one cell, but results were the same. Because the parametric test has more power, those results were reported.

4. SAS deleted five cases because of missing data. See Table 1 for participant details.

5. Response-shift bias may well be due to any number of factors; however, detailing the specific source of this bias was not the primary intent of this study. These issues are especially well detailed by Taylor, Russ-Eft, and Taylor (2009).

Appendix

Exiting Questionnaire Retrospective Pre/Posttest Items

You have examined many different topics and skills across your experiences at PGSIS. We would like to get a sense of how competent you feel you are in the areas listed below. Below, we ask you to consider each area and rate your level of competence NOW, after attending PGSIS, and also rate your level of competence BEFORE experiencing PGSIS. Your responses allow us to gauge the growth you perceive in your competence as a result of participation in PGSIS. Please circle one number for NOW and one number for BEFORE, according to the following scale:

1. Not very competent
2. Somewhat competent
3. Reasonably competent
4. Extremely competent

1. I am/was knowledgeable about how history has shaped the global problems and issues of today.
2. I am/was knowledgeable about contemporary international and global issues.
3. I can/could place myself in the shoes of someone who has had very different life experiences than me.
4. I understand/understood how economic, political, cultural, technological and environmental forces impact current global issues and problems.
5. I am/was knowledgeable about other languages and cultures.
6. I understand/understood how the process of globalization (global interdependence) affects the national interests of the United States and those of other countries.
7. I understand/understood the complexities of intercultural relationships and communication.
8. I have been/had been exposed to ideas about how the world could be organized in the future (differently or "alternatively") in order to better address some of the world's major global problems and issues.
9. I understand/understood how policy decisions on international issues are made.
10. I am/was good at seeing issues from another person or group's perspective.
11. I am/was ready for college, in terms of my ability to do academic research and writing.

References

Bamberger, M., Rugh, J., Church, M., & Fort, L. (2004). Shoestring evaluation: Designing impact evaluations under budget, time, and data constraints. *American Journal of Evaluation*, *25*, 5-37.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*, 191-215.

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

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Boston: Houghton Mifflin Co.

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

Choi, N. (2004). Sex role group differences in specifying academic and general self-efficacy. *The Journal of Psychology*, *138*, 149-159.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cooke, B. (1998). Beyond basic training. *Evaluation Exchange*, *4*, 8-9.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

D'Eon, M., Sadownik, L., Harrison, A., & Nation, J. (2008). Using self-assessments to detect workshop success: Do they work? *American Journal of Evaluation*, *29*, 92-98.

Gutek, B. A., & Winter, S. J. (1992). Consistency of job satisfaction across situations: Fact or framing artifact? *Journal of Vocational Behavior*, *41*, 61-78.

Hill, L. G. & Betz, D. L. (2005). Revisiting the retrospective pretest. *American Journal of Evaluation*, *26*, 501-517.

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

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., 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., & Bary, J. H. (1979). Internal invalidity in studies employing self-report instruments: A suggested remedy. *Journal of Educational Measurement*, *16*, 129-135.

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.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537-567.

Krosnick, J. A., & Fabrigar, L. R. (1998). *Designing good questionnaires: Insights from psychology.* New York: Oxford University Press.

Lent, R. W., Brown, S. D., & Gore, P. A. (1997). Discriminant and predictive validity of academic self-concept, academic self-efficacy and mathematics specific self-efficacy. *Journal of Counseling Psychology*, *44*, 307-315.

Manthei, R. J. (1997). The response-shift bias in a counselor education programme. *British Journal of Guidance and Counseling*, *25*, 229-237.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49-69). Mahwah, NJ: Erlbaum.

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

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review*, *96*, 341-357.

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology*, *21*, 277-287.

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

Wolfinger, R., & Chang, M. (2007). *Comparing the SAS GLM and MIXED procedures for repeated measures*. SAS Institute). Retrieved March 6, 2008, from http://support.sas.com\rnd\app\papers\mixedglm.pdf