A comparison study of pre/post-test and retrospective pre-test for measuring faculty attitude change

Stephanie Chasteen
Chasteen Educational Consulting, UCB 390, Boulder CO 80029

Rajendra Chattergoon
School of Education, University of Colorado Boulder, UCB 249, Boulder CO 80029

We report on our investigation of a *retrospective pre-test* to measure faculty attitude change towards the use of active learning after the Physics and Astronomy New Faculty Workshop (NFW). The purpose of the study is to explore alternative methods of evaluating the effectiveness of educational interventions aimed at attitude change. In the current study, we focus on faculty attitudes that would support change in teaching practice. Using traditional pre/post surveys, we find that only *knowledge* of and *skill* using active learning are substantively increased by the workshop. We administered a retrospective pre-test, where participants retrospectively rate their pre-workshop attitudes on the post-workshop survey. The rationale for this approach is that participants do not start with a common understanding of what "active learning" entails, and the workshop provides a normalizing experience so participants shift their understanding of active learning (termed *response shift bias*) as well as potentially generating gains in positive attitudes towards active learning. Using the retrospective pre-test, we see attitudinal gains for most items, but pre-test and retrospective pre-test results are poorly and inconsistently correlated. Preliminary interviews are suggestive of response shift bias, but only for some items. We can conclude that the validity of pre-workshop attitude ratings is questionable, but because of a conflation of response shift bias with other reporting biases (such as social desirability) and respondent characteristics, further research is needed to indicate whether retrospective pre-testing is an improved approach.

2019 PERC Proceedings edited by Cao, Wolf, and Bennett; Peer-reviewed, doi.org/10.1119/perc.2019.pr.Chasteen

I. INTRODUCTION

The mission of the New Faculty Workshop in Physics and Astronomy (NFW) is to improve student learning in physics and astronomy by helping all faculty become long-term users of evidence-based instructional practices. The workshop is effective at supporting faculty knowledge and initial use, but not necessarily the sustained implementation of such practices [1,2].

In previous work [3,4] we have proposed a set of post-workshop participant outcomes which may lead to increased, sustainable use of effective teaching practices, based on two theoretical perspectives: Self-Determination Theory and Theory of Planned Behavior. Self-Determination Theory [5] posits that intrinsic motivation is supported by feelings of *competence* and mastery, *relatedness* (or peer support), and *autonomy* (or sense of choice). To consider how this motivation might connect to actual behavior, we use the Theory of Planned Behavior [6] (a.k.a. the Reasoned Action Approach [7]), which indicates that intention is translated to behavior when people have positive *attitudes* towards the behavior, perceive *subjective norms* that peers approve of the behavior, and have *perceived control* over their behavior and outcomes. We combine these two theories to identify a Theory of Action for the workshop [4], outlining how it is intended to generate sustained use of evidence-based instructional practices (EBIPs). Below are the general areas of participant workshop outcomes that we suggest would lead towards initial and sustained use of EBIPs:

- **Competence** and the ability to use EBIPs.
- **Autonomy** and control over their choices.
- **Positive attitudes** towards EBIPs.
- **Subjective norms;** peer approval of the use of EBIPs.

To evaluate these outcomes, we developed a set of pre- and post-workshop survey questions to measure participant attitudes. As we will describe in this paper, the traditional pre/post-tests on these questions showed few gains, but we had reason to suspect that some workshop outcomes were not being fully brought to light by these measures.

One challenge in measuring impacts of training programs is that participants often do not have a good initial understanding of the topic that the program intends to address, and thus are poorly equipped to report their knowledge or skills of this topic prior to the intervention [8-13]. Thus, the treatment can affect not only participants' ability but also their understanding of the construct being measured; participants do not know what they do not know. This change in the understanding of the underlying construct is called *response shift bias* [8,9]. As a hypothetical example, in a workshop on web-based communication systems, participants ratings of their skill before and after the workshop may be affected by response-shift bias because the workshop made the participant realize that web-based instruction is harder than they thought [14]. Response shift bias typically results in participants *over*-estimating their pre-existing knowledge, skill, confidence, etcetera, making the effects of training seem less impressive [11-15].

To mitigate the effects of response shift bias, several program evaluators have made use of a *retrospective pre-test* design, where participants are asked to rate (post-workshop) what their behavior or thoughts were before the workshop [10-15]. This design also has some practical advantages in terms of time and resources. The known risks in using retrospective pretests is that they can introduce or intensify other biases (which we term "reporting biases") like social desirability (trying to please the workshop organizers), effort justification (it was worth the time to attend the training), hindsight bias (where memories are distorted based on new information), and availability heuristics (where judgments of behavior are biased by recent events) [15-18].

This paper compares traditional pre/post gains and the retrospective pre-test gains in the NFW, to inform others seeking methods to evaluate attitude change.

II. METHODS

The methods of this study rely primarily on survey responses from participants across the past 4 years of the NFW, collected as part of the external evaluation effort.

A. The New Faculty Workshop and participants

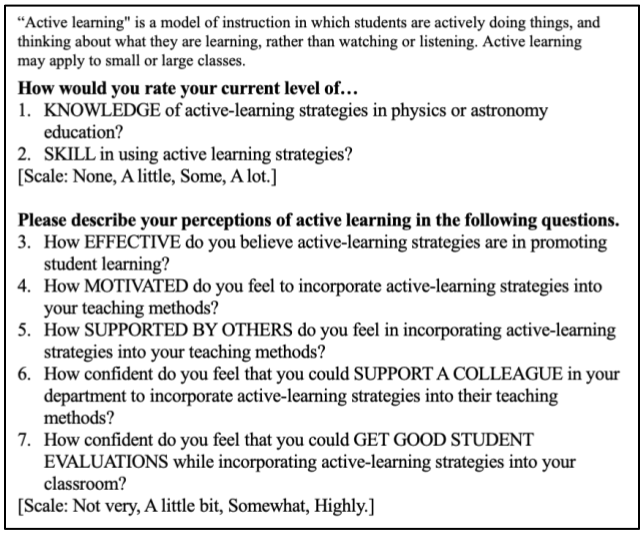
The NFW is offered twice a year, typically in June and November, to a cohort of approximately 50-70 faculty. Workshop attendance represents about 40% of new faculty hires in physics. The full data set from this evaluation covers 8 cohorts of faculty participants from June 2015 – October 2018, for a total of 478 respondents.

Workshop attendees are generally representative of new faculty in physics: There are more physics faculty (85% of registrants) than astronomy faculty, and most survey respondents identify as male (70%) and White (67%). A sizeable fraction of attendees identifies as Asian (23%), which exceeds the national representation in physics (14%) [19]. A greater number of respondents are from Bachelor's granting institutions (53%) than Ph.D. granting institutions compared to national representation of physics departments (31% Bachelor's granting). However, data on institution type is somewhat inconsistent between participants registration and their survey responses, and so results related to institution type should be interpreted with caution.

B. Survey instruments and response rates

As part of the external evaluation, we give an online pre-workshop and post-workshop survey to each cohort, and a one-year survey a year later. This paper focuses on a series of 7 Likert-scale questions (see Fig. 1) included on all 3 surveys that tap into participants attitudes and beliefs about active learning (inspired by similar questions in mathematics workshops [20]). The pre-test and post-test versions appeared as shown, and were administered on the pre- and post-workshop surveys respectively.

For the retrospective pre-test (administered on the post-workshop survey), the initial stems (e.g., "How would you rate your current level of…") were removed, and participants were asked to rate their levels "Prior to the conference" and "After attending the conference." The retrospective pre-test questions were administered to three workshop cohorts: November 2017, June 2018, and October 2018. Within these cohorts, out of 186 workshop participants, a total of 166 (89%) responded to the pre-workshop survey, 137 (73%) responded to the post-workshop survey (which includes the retrospective pre-test), and 113 (61%) usable responses are available within the matched pre/post sample. For the retrospective pre-test questions, *n* ranged from 109-113 responses (depending on the question), except for Q7, with *n* = 67-72 responses.



"Active learning" is a model of instruction in which students are actively doing things, and thinking about what they are learning, rather than watching or listening. Active learning may apply to small or large classes.

How would you rate your current level of…

1. KNOWLEDGE of active-learning strategies in physics or astronomy education?
2. SKILL in using active learning strategies?

[Scale: None, A little, Some, A lot.]

Please describe your perceptions of active learning in the following questions.

3. How EFFECTIVE do you believe active-learning strategies are in promoting student learning?
4. How MOTIVATED do you feel to incorporate active-learning strategies into your teaching methods?
5. How SUPPORTED BY OTHERS do you feel in incorporating active-learning strategies into your teaching methods?
6. How confident do you feel that you could SUPPORT A COLLEAGUE in your department to incorporate active-learning strategies into their teaching methods?
7. How confident do you feel that you could GET GOOD STUDENT EVALUATIONS while incorporating active-learning strategies into your classroom?

[Scale: Not very, A little bit, Somewhat, Highly.]

FIG. 1. Attitude survey questions.

C. Interviews

As an initial exploration of the results, we interviewed participants from the November 2017 workshop who had demonstrated large shifts in their responses from pre-test to retrospective pre-test. Thirteen people were contacted, and 6 interviews were conducted (the others declined or did not respond) via video conference. The survey item was read to the participant, they were told how their response changed from pre-workshop to post-workshop, and they were asked to reflect on this difference.

III. FINDINGS

A. Traditional pre/post-test results

Table I shows the mean and median results from each administration of the attitude questions. Because respondents cannot select ratings between two scale points, we find the median to be a somewhat more honest representation of the distribution of participant responses. Standard deviations range from 0.5-1.0; most are 0.7. We see evidence of ceiling effects with our 4-point scale on most items, except Q1 and Q2 (*knowledge* and *skill*).

TABLE I. Mean, median and traditional gain on a 4-point scale.

Item	Pre		Post		Gain		Effect
	M	*Med.**	M	*Med.*	M	*Med.*	Size
Q1	2.7	*3*	3.8	*4*	1.1	*1*	1.4
Q2	2.2	*2*	3.0	*3*	0.8	*1*	1.0
Q3	3.7	*4*	3.8	*4*	0.1	*0*	0.3
Q4	3.6	*4*	3.8	*4*	0.2	*0*	0.4
Q5	3.4	*3*	3.4	*4*	0.1	*0*	0.1
Q6	2.8	*3*	3.2	*3*	0.4	*0*	0.4
Q7	3.0	*3*	3.1	*3*	0.1	*0*	0.1

*"Med." = median

Gains are computed by subtracting post-survey from pre-survey responses. The mean (median) gain is the average (median) of the individual gain scores for each item. Effect size is computed by dividing the mean gain by the pooled standard deviation across all pre- and post-survey responses.

For Q1 and Q2 (and only these items) we see post-workshop gains, with effect sizes of 1.4 and 1.0 respectively. Recall that these results are from only 3 out of the 8 cohorts for which we have data: Results are fairly similar to those for the full 8 cohorts, except that these 3 cohorts reported higher post-workshop means and gains for Q1 and Q2 (Table 1) compared to historical values (Q1 gain 0.7, effect size 0.8; Q2 gain 0.4, effect size 0.5). For the full data set, we investigated pre/post gains by various background factors and found no strong effects by the level of use of active learning prior to the workshop, but did find a modest effect of institution type such that those at Ph.D.-granting institutions had slightly larger gains on the *effectiveness* item (Q3), and the *good student evaluations* item (Q7) (effect size difference of 0.4 and 0.6, respectively), compared to those at institutions that do not grant a Ph.D.

Thus, using traditional gain scores we mainly see a difference in knowledge of active learning techniques pre-workshop, and less consistent increase in reported skill.

B. Retrospective pre-test

Table II shows results for the retrospective pretest. Standard deviations range from 0.4-1.0; most are 0.7-0.8.

TABLE II. Mean, median and alternative gain on a 4-point scale.

Item	Retro Pre.		Post		Alt. Gain*		Effect
	M	*Med.*	M	*Med.**	M	*Med.*	Size
Q1	2.5	*3*	3.8	*4*	1.3	*1*	1.5
Q2	2.1	*2*	3.0	*3*	0.9	*1*	1.1
Q3	3.2	*3*	3.8	*4*	0.6	*0*	0.9
Q4	3.0	*3*	3.8	*4*	0.8	*1*	1.0
Q5	2.8	*3*	3.4	*4*	0.7	*0*	0.8
Q6	2.1	*3*	3.2	*3*	1.1	*1*	1.0
Q7	2.6	*2.5*	3.1	*3*	0.6	*0.5*	0.6

*"Med." = median, "alt. gain" = alternative gain.

The alternative gain is computed entirely from post-survey responses by subtracting the post-survey retrospective pre-test response from the post-survey response. Mean, median, and effect sizes are computed as described above. We see sizable alternative gains for all items except the *good student evaluations* item (Q7), likely because this item was omitted from the November 2017 survey and thus includes a smaller number of participants. Thus, for Q3-6, the alternative gain is noticeably larger than the traditional gain. There were no notable impacts of respondents' institution type, or degree of active learning use, on the effect size of the alternative gain.

How do the retrospective pre-test results compare to the traditional pre-test results? Table III provides summary statistics comparing the two measures. The difference ("Comparison score") is calculated by subtracting the retrospective pre-survey responses from the actual pre-survey responses. The effect size is computed by dividing the mean gain by the pooled standard deviation across all actual pre- and retrospective pre-survey responses. On the four items with a larger alternative than traditional gains (Q3-6), respondents tended to select an option on the retrospective pre-survey that is one point lower than their actual pre-survey response. Thus, respondents typically rated their pre-survey attitude levels as less favorable after the workshop than they did before engaging in the workshop. These results are consistent with response shift bias [8-15], particularly for Q3-6, but without additional data, we cannot know if this is due to response shift bias or other reporting biases such as social desirability.

We also investigated correlations between the different measures; Table IV. The correlations between the actual pre-test and retrospective pre-test responses, and between traditional and alternative gains, are modest (about 0.5) for many items. However, these correlations are quite low (<0.5) for Q4, Q5, and Q7. If all participants were impacted by response shift bias equally for all questions, we would expect higher correlations between the types of pre-survey responses, and between the types of gains, since the response shift would result only in a uniform shift of the overall baseline of pre-survey beliefs. It is clear that different questions are differentially impacted by the use of the retrospective pre-test, in line with prior research [13].

TABLE III. Comparison of pre- and retro. pre-survey responses.

Item	Pre		Retro. Pre*		Comp.*		Effect
	M*	Med.*	M	Med.	M	Med.	Size
Q1	2.7	3	2.5	3	-0.2	0	-0.3
Q2	2.2	2	2.1	2	-0.2	0	-0.2
Q3	3.7	4	3.2	3	-0.4	0	-0.6
Q4	3.6	4	3.0	3	-0.6	-1	-0.7
Q5	3.4	3	2.8	3	-0.6	-1	-0.7
Q6	2.8	3	2.1	3	-0.7	-1	-0.6
Q7	3.0	3	2.6	2.5	-0.4	0	-0.4

* "M" = mean, "Med." = median, "Retro. Pre" = retrospective pre-test score, "Comp." = Comparison score.

TABLE IV. Correlations between items and between gains.

Item	Correlation: Pre vs. Retro. Pre	Correlation: Gain vs. Alt. Gain
Q1	0.59	0.55
Q2	0.58	0.49
Q3	0.48	0.39
Q4	0.33	0.31
Q5	0.23	0.26
Q6	0.49	0.38
Q7	0.37	0.15

A few explanations may be posited for these results. It is possible that the relationship between pre-survey and retrospective pre-survey responses for each participant is affected by a covariate (such as teaching experience or ethnicity), such that some groups are more or less affected by response shift bias, or reporting biases (such as social desirability). For example, those with high levels of use of active learning may not experience response shift bias, less experienced teachers may be more prone to anchoring effects, and some ethnicities may be more influenced by social desirability. Such effects could result in poor correlations as the types of scores/gains are differentially correlated for different types of respondents. It is also possible that the pre-survey and retrospective pre-survey items measure very different constructs (e.g., how confident they are feeling in their teaching while taking the pre-survey, versus how positive of a reaction they had to the workshop on the retrospective pre-survey).

Regardless of the reason for the difference, the alternative and traditional gain scores are not equivalent or interchangeable measures and depend upon the item.

C. Interviews

To explore the reason behind the response shift, we interviewed 6 participants from November 2017 who had large response shifts. Respondents were interviewed only on questions for which they had a demonstrated shift from pre-test to retrospective pre-test; these shifts ranged from 0.8-1.2. Q7 was not included in this questioning as it was inadvertently removed from the survey for this cohort.

Results differed by question. For Q1, Q2, Q4, and Q6 there was fairly clear evidence of response shift bias in that the experience of being at the workshop changed respondents' understanding of the question: Most indicated that their retrospective pre-test result was a better representation of their learning from the workshop.

For Q1 (*knowledge*), all 3 respondents with shifts indicated that their views of active learning had been expanded by the workshop, and hadn't fully appreciated the scope of active learning strategies or their use. For Q2 (*skill*), all 4 with a response shift indicated that they now judged their skill in using active learning more harshly. For Q4 (*motivation*), 4 out of 5 with shifts realized that they needed greater motivation, or that they now felt more motivated and downgraded their original response. For Q6 (*support a colleague*), 4 out of the 6 respondents with a shift indicated that they had a more realistic idea of their knowledge and what support is needed to use active learning well.

For other questions, the evidence was more mixed. For Q3 (*effectiveness*), some responses indicated that they felt even less skeptical post-workshop and so downgraded their original responses; for others, the reasons were contextual and idiosyncratic. For Q5 (*supported by others*) 2 out of 5 with shifts indicated that they now more fully realized the support needed to undertake active learning. These interview results are intriguing, though inconclusive.

IV. CONCLUSIONS

We report on our investigation of the use of a *retrospective pre-test* to measure faculty attitude change towards the use of active learning. For the past 3 out of 8 workshops, we asked participants (post-workshop) to retrospectively assess their pre-workshop attitudes. Using traditional pre/post gains, we find that only *knowledge* of active learning is consistently increased – although *skill* was also improved in the 3 study cohorts. With the retrospective pre-test, however, we find that participants report gains for almost all items, with effect sizes of 0.8 or greater.

We investigated correlations between retrospective pre-test and pre-test results, and find that these correlations are modest and differ by item, suggesting that respondents pre-workshop attitudes do not uniformly or fully predict their retrospective pre-workshop attitudes. Similar results are observed for the two gain measures.

Two main hypotheses are posed for these results. *Response shift bias* would describe a shift in participants' understanding of the ideas that the survey items are testing; participants may leave the workshop with more conservative estimates of their prior attitudes once they fully appreciate the range of active learning techniques and their associated challenges. *Reporting bias,* on the other hand, is a range of biases (such as social desirability, effort justification, hindsight bias, and availability heuristics) that are more likely to become salient when using retrospective pre-tests than with a traditional pre/post-test [15-18]. Preliminary interviews provide evidence for response shift bias for some questions (*knowledge*, *skill*, *motivated*, and *support a colleague*) though reporting bias could also be at play.

Based on the current results, *effective*, *motivated*, *supported by others*, and *support a colleague* demonstrated gains not observed with traditional pre/post-testing, but of these, only *motivation* and *support a colleague* showed evidence of response shift bias during interviews, and most (except *effective* and *support a colleague*) were strongly confounded by participant characteristics, as judged by correlations. Thus, to date, *support a colleague* shows the most promise as an item measuring retrospective pre/post gains which may be consistent across participants.

Additional research is needed to be able to adequately interpret these results. First, the comparison between traditional and retrospective pre-test results needs to be made across different types of participants (such as those with different educational backgrounds or teaching practices), with adequate sample sizes. Second, interviews across more participants, more diverse participants, and immediately after the workshop, would yield greater insight. Such research would also allow us to identify the questions for which retrospective pre-testing might be valid and useful.

If it were found to be adequately valid, retrospective pre-testing could be an efficient method for its' intended purpose of providing advice to conference organizers and judging program effectiveness [11, 12]. Survey methodologists (including author RC) would question such a recommendation, however, as the retrospective pre-test design does not measure a true gain (as no time elapses between administration of the questions), and workshop effects are entangled with the introduction of reporting bias. Past research has indicated concern about biases introduced by retrospective pre-tests [15-18]. One such concern is that respondents use a general heuristic that pre-test results should be lower than post-test results, and so anchor their retrospective pre-test response by their post-test results [16, 17]; attitude questions of the type that we have used are particularly susceptible to such heuristics because they are broad questions requiring estimation [16]. Interviewee reports of "downgrading" pre-workshop responses may be indicative of such anchoring.

Some survey modifications are suggested by these results. For one, the scale options could be expanded to reduce ceiling effects within the Likert scale. Adding an open-ended question reflecting on pre/retrospective pre-test differences [11] could provide insight. Anchoring effects could be mitigated by providing separate surveys with post-test items and retrospective pre-test items [17, 18], though this reduces survey efficiency. We might use retrospective pre-tests only for some items, such as for measuring subjective experiences and not program effects [18], to avoid over-estimating program effects due to social desirability bias, but allow for measurement of other types of gains. We might also provide a clearer description of active learning on the pre-test (see Fig. 1), though this may not be sufficient to provide adequate knowledge [15]; perhaps a short tutorial on active learning would be more useful.

In conclusion, we have learned that traditional pre/post attitude gains may be threatened by response shift bias, reducing apparent gain due to the workshop. Our results are inconclusive as to whether the retrospective pre-test is an improvement, however, due to the entanglement of response shift bias with respondent characteristics and reporting bias. Further research is recommended.

ACKNOWLEDGMENTS

We thank Alice Olmstead for conducting faculty interviews, and Alistair McInerny for introducing us to the idea of retrospective pre-tests. This work is supported by NSF DUE #1431638 and Chasteen Educational Consulting.

[1] C. Henderson, Promoting instructional change in new faculty: An evaluation of the physics and astronomy new faculty workshop, Am. J. Phys., **76**, 179 (2008)

[2] C. Henderson, M. Dancy, and M. Niewiadomska-Bugaj, Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process? Phys. Rev. ST Phys. Ed. Res., **8**, 020104 (2012).

[3] S. Chasteen, R. Chattergoon, E. Prather, and R. Hilborn, Evaluation Methodology and Results for the New Faculty Workshops, Proceedings of the Physics Education Research Conference 2016, Sacramento, CA, 2016.

[4] S. Chasteen and R. Chattergoon, Supporting teaching autonomy in the New Faculty Workshop, poster presented at the 2018 American Association of Physics Teachers conference, Washington, DC, 2018. https://www.chasteenconsulting.com/wp-content/uploads/2019/07/NFW-AAPT-2018-poster-v2.pdf

[5] E.L. Deci and R. M. Ryan, Self-determination theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology: Vol. 1* (pp. 416-437). (Sage Publications, Thousand Oaks, CA, 2012).

[6] I. Azjen, The theory of planned behavior, Organizational Behavior and Human Decision Processes, **50**, 179 (1991).

[7] M. Fishbein and I. Azjen, *Predicting and changing behavior: The Reasoned Action Approach.* (Taylor & Francis, New York, 2010).

[8] G.S. Howard and P.R. Dailey, Response-shift bias: A source of contamination of self-report measures, J. Appl. Psych., **64**, 144 (1979).

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

[10] J. Klatt and E. Taylor-Powell, Using the retrospective post-then-pre design, Quick Tips 27, Program Development and Evaluation. (University of Wisconsin-Extension, Madison, WI, 2005), https://fyi.extension.wisc.edu/programdevelopment/files/2016/04/Tipsheet27.pdf

[11] J.M. Allen and K. Nimon, Retrospective Pre-test: A Practical Technique for Professional Development Evaluation, J. of Industrial Teacher Education, **44**, 27 (2007).

[12] D. Moore and C.A. Tananis, Measuring change in a short-term educational program using a retrospective pretest design, Am. J. Eval., **30**, 189 (2009).

[13] C.C. Pratt, W.M. McGuigan, and A.R. Katzev, Measuring program outcomes: Using a retrospective pretest methodology, Am. J. Eval., **21**, 341 (2000).

[14] T.C.M. Lam and P. Bengo, A comparison of three retrospective self-reporting methods of measuring change in instructional practice, Am. J. Eval., **24**, 65 (2002).

[15] J. Klatt and E. Taylor-Powell, Synthesis of literature relative to the retrospective pre-test design, panel presentation for 2005 Joint CES/AEA Conference, Toronto, Canada, 2005. http://comm.eval.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=31536e2f-4d71-4904-ae5d-056e3280c767

[16] P.J. Taylor, D.F. Russ-Eft, and H. Taylor, Gilding the outcome by tarnishing the past: Inflationary biases in retrospective pre-tests, Am. J. Eval., **30**, 31 (2009).

[17] K. Nimon, D. Zigarmi, and J. Allen, Measures of program effectiveness based on retrospective pretest data: Are all created equal? Am. J. Eval., **32**, 8 (2011).

[18] L.G. Hill and D.L. Betz, Revisiting the retrospective pretest. Am. J. Eval., **26** (2005).

[19] R. Ivie, G. Anderson and S. White, African Americans and Hispanics among Physics and Astronomy faculty, AIP Focus On, July 2014, accessed at https://www.aip.org/statistics/data-graphics/race-and-ethnicity-physics-faculty-0.

[20] C.N. Hayward, M. Kogan and S.L. Laursen, Facilitating instructor adoption of inquiry-based learning in college mathematics, Int. J. Res. Undergrad. Math. Ed., **2**, 59 (2016).