JOURNAL OF BALTIC SCIENCE EDUCATION
ISSN 1648-3898

A COMPARISON OF STANDARD AND RETROSPECTIVE PRE-POST TESTING FOR MEASURING THE CHANGES IN SCIENCE TEACHING EFFICACY BELIEFS

Murat Bursal

Abstract. *Thirty-nine American and 78 Turkish preservice elementary teachers' personal science teaching efficacy (PSTE) beliefs were investigated during science methods courses with standard and retrospective pre-post testing methods. Significant differences in the PSTE gain scores, which indicate the changes in the mean PSTE scores from standard/retrospective pretests to the posttest, were found between the standard and retrospective measurements in both samples. Significant differences between the standard and retrospectively measured gain scores were detected among all subgroups under study, which were formed by participants' PSTE levels and gender. It has been concluded that the differences between the standard and retrospectively measured PSTE gain scores are due to the difference in the nature of these measurement methods and can be seen in most research samples in educational studies around the world. The findings of this study suggest that the response-shift bias should be considered as a common threat to validity for research studies measuring self-efficacy beliefs with the standard pre-post testing method.*

Key words: *personal science teaching efficacy, preservice elementary teacher, response-shift bias, retrospective pretest.*

Murat Bursal
Cumhuriyet University, Sivas, Turkey

Introduction

In the late 1970s, the term *self-efficacy* has been introduced to educational literacy with the social cognitive theory of Albert Bandura (1977). Bandura (1997) stated that "Efficacy expectation is a major determinant of people's choice of activities, how much effort they will expend, and how long they will sustain effort in dealing with stressful situations" (p. 194). Exploring the various aspects of self-efficacy beliefs in his pioneer book, he has defined self-efficacy beliefs as "the beliefs in one's capability to organize and execute the courses of action required to produce given attainments" (p. 3). With the widely reception of the concept, the efficacy belief definitions were adapted to different subject areas. In the area of science teaching, personal science teaching efficacy (PSTE) and science teaching outcome expectancy (STOE) beliefs are defined (Ashton & Webb, 1986). PSTE refers to beliefs in personal ability to teach science effectively and STOE beliefs indicate that effective teaching will positively affect student learning. As Ashton and Webb (1986) described, PSTE and STOE may be independent from each other, therefore it is possible for a teacher to possess low PSTE beliefs, but high STOE beliefs, or vice versa (Cantrell, Young, & Moore, 2003; Moore & Watson, 1999).

Exploring the self-efficacy beliefs of elementary science teachers drew the attention of many science researchers because it has been well documented that holding high PSTE beliefs would lead positive teacher attitudes about science teaching (Brigido, Borrachero, Bermejo, & Mellado, 2013; de Laat & Watters, 1995; Eshach, 2003; Taştan-Kırık, 2013; Yürük, 2011), more student-centered science instruction (de Laat & Watters; Marshall, Horton, Igo, & Switzer, 2009; Palmer, 2001) and improvement of the quality of science instruction at schools (Appleton, 2003; Bahcivan, 2014; Braund & Leigh, 2013; Cantrell et al., 2003; Plourde, 2002; Richardson & Liang, 2008; Utley, Moseley, & Bryant, 2005). On the other hand, many studies have reported that a vast amount of elementary teachers lacks the sufficient level of efficacy beliefs

for teaching science (Appleton; Bayraktar, 2011; Huinker & Madison, 1997; Jarrett, 1999; McKinnon, Moussa-Inaty, & Barza, 2014; Nilsson & van Driel, 2011; Tekkaya, Cakiroglu, & Ozkan, 2004; Wenner, 1993; 1995; 2001). Therefore, it has become an important research topic to investigate the sources that build and shape the self-efficacy beliefs of teachers and the attention has shifted toward preservive teacher education programs.

With an increasing emphasis over the years on self-efficacy beliefs, numerous studies have been conducted to measure preservice teachers' self-efficacy beliefs before, during, or after science content and/or methods courses. The changes in the science teaching efficacy scores of preservice teachers have usually been investigated with standard pretest-posttest design studies (Bursal, 2008; Eshach, 2003; Ginns, Tulip, Watters, & Lucas, 1995; Huinker & Madison, 1997; Morrell & Carroll, 2003; Plourde, 2002; Soprano & Yang, 2013; Yılmaz & Huyugüzel-Çavaş, 2008). Many of these studies used Likert scale instruments, such as the Science Teaching Efficacy Belief Instrument (STEBI) developed by Enochs and Riggs (1990). However, a common problem encountered in the standard pretest-posttest designs, especially measurements with Likert scale instruments, is the *response-shift bias* (Cantrell, 2003; Howard, 1980; Rohs, 1999). For example, Howard and Dailey (1979) asked their participants to express their opinions before and after a training program and have found that participants tended to use an altered set of units in their self-reports. This change in the self-report metric from pre-test to post-test has been labelled as response-shift bias.

To be able to compare the scores from a pretest and a posttest, a fundamental assumption is that the metric the respondents use in both tests is the same (Cantrell, 2003). However, compared to posttest, respondents usually have only a partial understanding of the construct to be measured at the time of the pretest. Furthermore, in course evaluation studies, participants may be often unaware of accurate estimates of their knowledge and skills on a pretest (Lamb & Tschillard, 2005). Therefore, it is almost always unrealistic to expect that respondents will use the same metric both in the pretest and the posttest (Aiken & West, 1990; Cantrell; Howard, 1980). For this reason, response-shift bias poses a threat to validity of the measures in the standard pretest-posttest designs, especially in studies that employ Likert scale instruments based on self-reports (Cantrell; Hoogstraten, 1985; Howard & Dailey, 1979; Howard, Schmeck, & Bray, 1979; Rohs, 1999).

An alternative method to overcome this problem is allowing participants to evaluate their prior and present beliefs together (Allen & Nimon, 2007; Cantrell, 2003). In this method, people are given the measurement instrument at the end of the study and are asked to evaluate their beliefs prior to the study and after the study at the same time. The posttest format in this alternative method is not different from the standard format. However, the alternative pretest that is given after the study to evaluate the prior beliefs is called *retrospective pretest* (Cantrell; Howard, 1980; Rohs, 1999). The proponents of the retrospective testing argue that people are more likely to use the same metric for evaluating their beliefs in the retrospective pretest and the posttest and thus retrospective pretests provide a more accurate measure of the intervention during the research study (Allen & Nimon; Cantrell; Hoogstraten, 1985). For the same reason, compared to the standard pretest-posttest design, the response-shift bias has been reduced in retrospective designs. Howard, Ralph et al. (1979) discussed five experimental studies in their paper and cited that retrospective pretests yield higher statistical power than their respective standard pretests.

Based on the previous studies of Howard (1980) and Howard, Ralph et al. (1979), Hoogstraten (1985) measured 67 psychology freshman students' self-reports via standard and retrospective testing methods and concluded that retrospective testing approach yields more accurate data about the improvement of performance than the standard pretest-posttest approach. Similarly, after a workshop evaluation study, Lamb and Tschillard (2005) concluded that retrospective pretests provide more valid pretest scores than the standard pretests to determine impact of the workshop because standard pretest underestimated the impact of the workshop on participants' learning. Lamb and Tschillard recommend that retrospective testing is useful when there may be a response shift effect.

In the area of science teaching, very limited number of studies has been conducted on the difference between these testing methods. Cantrell (2003) compared the standard and retrospectively measured STEBI scores of 37 American preservice teachers during science methods and practicum courses. She has found that participants tended to rate themselves much higher on the standard pretest than on the retrospective pretest. Therefore, the gain scores calculated for the standard pretest-posttest design was larger than of retrospective pretest-posttest design. Cantrell has concluded that "Follow-up interviews with the participants provided evidence for greater internal validity for the retrospective pretest. Findings support the notion that retrospective pretests may produce gain scores with greater validity and greater statistical power." (p. 177). Hechter (2011) designed a similar study with 69 Canadian preservice elementary teachers and consistent with Cantrell's findings, his participants reported a lower level of science teaching self-efficacy on the retrospective pretest than on the standard pretest.

There are numerous examples of studies investigating the science teaching self-efficacy gains of preservice

teachers with standard pretest-posttest research designs (Bleicher, 2009; Bleicher & Lindgren, 2005; Bursal, 2008; Ginns et al., 1995; Ginns & Watters, 1999; Huinker & Madison, 1997; Morrell & Carroll, 2003; Palmer, 2006a, 2006b; Plourde, 2002; Richardson & Liang, 2008; Soprano & Yang, 2013; Utley et al., 2005; Yılmaz & Huyugüzel-Çavaş, 2008), but no study –except Cantrell (2003) and Hechter (2011) – is located in the literature using retrospective tests to measure the same construct. Also, it is not investigated that whether the differences between the standard and retrospective test scores are seen in samples from different contexts. Furthermore, the impact of preservice teachers' gender or levels of self-efficacy beliefs on the differences in these test scores have not been considered in a research study yet.

For the reasons discussed above, this study will contribute to the literature not only by investigating the differences between the preservice elementary teachers' PSTE gain scores from standard and retrospective tests, but also investigating this phenomenon in two countries with distinct educational and cultural structures. Moreover, the impact of participants' self-efficacy levels and gender on the differences between gain scores are also studied.

Problem of Research

The research question and related sub-questions investigated in this study are:

- Does the testing method affect the changes in the PSTE scores of preservice teachers in a pretest-posttest design study?
    i. Do PSTE gain scores of American and/or Turkish preservice elementary teachers from pretest to posttest differ between the standard and retrospective pretest-posttest designs?
    ii. Do participants' PSTE level and/or gender cause any significant difference between the standard and retrospectively measured PSTE gain scores of preservice elementary teachers?

Methodology of Research

General Background of Research

This study has been designed as a survey study, which employs a retrospective pretest in addition to standard pretest-posttest, with two distinct samples of preservice elementary teachers. Samples were selected from two countries (United States and Turkey), which have distinct educational and cultural structure to explore the same phenomenon in two distinctly different contexts. The samples were selected via purposive sampling method, where the criteria for selection were; completing all science courses required for elementary teaching licensure, attending a science methods course, and also participating in three measurements required for this study.

This study was approved by the ethical committees of both universities, where the study was conducted. Consent forms, explaining the outline of the study and the data collection process, were given to the all participants and they were informed that the participation in the study was completely voluntary. Only the volunteers who had signed the consent forms were included in the study.

Sample of Research

The American (US) sample consisted of 55 preservice elementary teachers from a midwestern American university and the Turkish (TR) sample consisted of 163 preservice elementary teachers from a mid-Anatolian Turkish university. The US sample was enrolled in a master of education licensure program in elementary education, whereas the TR sample was enrolled in an undergraduate elementary education program, which is sufficient for receiving elementary teaching licensure in Turkey.

The age range was between 22 and 31 for the US sample, and 19 and 30 for the TR sample. 96% of the US sample was between 22 and 24 and 92% of the TR sample was between 20 and 23 years of age. All participants from both samples had taken college-level science courses required by their programs prior to the study and also were enrolled in a 3 credit science methods course. The course titles were; Science Instruction in the Elementary Grades for the US sample and Science Teaching II for the TR sample.

Among the participants, 39 of the US sample (34 females and 5 males) and 78 of the TR sample (25 females and 53 males) have attended all three measurements (standard pretest, posttest, and retrospective pretest) and responded all items. Therefore, statistical analyses are conducted only for these participants.

Instrument and Procedures

The Science Teaching Efficacy Belief Instrument (STEBI-B) (Enochs & Riggs, 1990) was administered at the beginning of the science methods course as standard pretest and at the end of the course semester as standard posttest and retrospective pretest. STEBI-B instrument consists of two subsets; personal science teaching efficacy (PSTE) and science teaching outcome expectancy (STOE). Since the objective of this study is to compare the findings from two different testing methods in measuring personal efficacy beliefs, only the PSTE scale of the STEBI-B was relevant to the study and therefore only this subset was administered to the participants.

The PSTE subset of the STEBI consists of 13 items (5 positive and 8 negative items), each to be rated by the respondent on a one (strongly disagree) to five (strongly agree) rating scale. The negatively-worded item scores are recoded as "5=1, 4=2, 3=3, 2=4, 1=5". The score range for the PSTE subset is 13–65, and the higher the score, the higher the level of PSTE. The internal reliability alpha coefficient of the PSTE scale of the STEBI-B was calculated to be .90 (Enochs & Riggs, 1990).

The PSTE scale of the STEBI-B was translated into Turkish by the author and a Turkish graduate student majoring in elementary science education at an American university. Turkish version of the instrument was then back-translated into English with the help of a different Turkish graduate student majoring in secondary science education at the same American university. Examination of the original version and the back-translated version by these experts indicated that the Turkish translation of the STEBI-B was parallel to the original research instrument.

The Cronbach's α coefficients were found to be .86 and .89 for the standard pretest, .80 and .82 for the standard posttest, and .88 and .85 for the retrospective pretest for the US and TR samples respectively. Also, exploratory factor analysis has been run to check the structural validity of the Turkish version of STEBI-B. It is confirmed for all cases that all PSTE items were grouped under a single factor, thus the PSTE items showed uni-dimensional behavior in all measurements. The findings revealed that all PSTE items had factor loadings higher than .31 in the standard pretest, .36 in the standard posttest, and .43 in the retrospective pretest for the Turkish version of the STEBI-B. Based on these data, Turkish translation of the instrument is found to be a valid and reliable instrument to measure Turkish participants' PSTE beliefs.

Data Analysis

As the first step of the data analysis, the normality assumptions for the data distributions are tested with Kolmogorov-Smirnov and Shapiro-Wilk normality tests. Based on the sample sizes, the Kolmogorov-Smirnov test was used for the TR sample (TR sample size > 50) and the Shapiro-Wilk test was used for the US sample (US sample size < 50). Statistically insignificant results of these normality tests (p>.05) for all subgroups, confirmed that the standard pretest, standard posttest, and retrospective pretest distributions of both US and TR samples can be considered normal. For this reason, parametric tests are decided to be used to investigate the research questions (Pallant, 2007).

To investigate the research problems; firstly, the gain scores from the standard pretest-posttest and retrospective pretest-posttest measurements are calculated. The comparison of these two sets of gain scores (dependent variables) with the inclusion of the interaction effects of the dependent and independent variables (PSTE level and gender) are done with two-factor repeated measures analysis of variance test. The PSTE level groups (low/high PSTE) are determined by median splits in both groups, where the median PSTE scores were 50 for both groups. By using the two-factor repeated measures ANOVA design, the impact of the main effect (retrospective vs. standard measurement method effect) and the interaction effects (PSTE level*measurement method and gender*measurement method) on the differences between the standard and retrospectively measured gain scores are tested together. In this way, the significance alpha level is secured to be .05 and a high statistical power is produced for the analysis (Pallant, 2007). Additionally, partial eta-squared (η^2) effect sizes are calculated to explore the practical significance of the main effect and interaction effects investigated in this study.

Results of Research

Participants' standard pretest, retrospective pretest, and posttest mean scores for the PSTE subset of the STEBI-B and their gain scores from standard/retrospective pretests to the posttest are given in Table 1. From the comparison of the standard and retrospectively measured pretest scores, the mean scores from the retrospective

measurements are lower than the mean scores from the standard measurements in both samples. Since the same posttest is used in both measurement methods, the gain scores from the retrospective method are found to be higher than those from the standard method in both samples due to lower retrospective pretest scores.

As seen in Table 1, the mean differences between the retrospective and standard gain scores are 3.92 for the US and 4.60 for the TR samples. The repeated measured ANOVA results for the main effect (measurement method effect) indicated that the PSTE gain scores were significantly higher in the retrospective pretest-posttest method than the standard pretest-posttest method both for the US [F (1, 37) = 15.14; $p < .001$; $\eta^2 = .29$] and TR [F (1, 76) = 23.15; $p < .001$; $\eta^2 = .23$] samples. Also, the partial eta-squared effect size values indicated that main effect is accounted for more than %20 of the variances of the gain scores in both samples and therefore the measurement method (standard/retrospective) has a very large effect (Pallant, 2007) on the PSTE gain scores.

Table 1. US and TR samples' pretest/posttest PSTE scores and gain scores from standard and retrospective tests.

	US Sample (n=39)		TR Sample (n=78)	
	Mean	S. D.	Mean	S. D.
Standard Pretest	47.36	6.46	48.27	8.15
Retrospective Pretest	43.44	7.90	43.67	7.47
Posttest	50.23	6.00	48.78	6.90
Gain (Standard)	2.87	5.14	0.51	7.22
Gain (Retrospective)	6.79	7.40	5.11	6.12

The gain scores of the US and TR samples according to their PSTE level and gender are summarized in Tables 2 and 3. When participants are grouped into low and high PSTE groups, the median of the PSTE posttest scores are used. Since the median PSTE scores were 50 for both US and TR samples, participants with PSTE scores of lower than 50 are put into the low-PSTE group and participants with PSTE scores of 50 and above are put into the high-PSTE group.

Table 2. US Sample's (n=39) PSTE gain scores in standard and retrospective tests by their PSTE levels and gender.

	Gain (Standard)		Gain (Retrospective)		
	Mean	S. D.	Mean	S. D.	n
Low PSTE	1.62	4.36	4.86	5.41	21
High PSTE	4.33	5.69	9.06	8.84	18
Male	1.20	4.92	1.60	2.30	5
Female	3.12	5.19	7.56	7.60	34

Table 3. TR sample's (n=78) PSTE gain scores in standard and retrospective tests by their PSTE levels and gender

	Gain (Standard)		Gain (Retrospective)		
	Mean	S. D.	Mean	S. D.	n
Low PSTE	-1.12	7.19	3.30	3.84	40
High PSTE	2.24	6.94	7.03	7.42	38
Male	0.66	7.66	4.94	6.06	53
Female	0.20	6.33	5.48	6.37	25

When the participants' gain scores from standard and retrospective tests in Tables 2 and 3 are compared according to their PSTE levels, interesting shifts can be seen. For example, compared to a slight increase in the US low-PSTE group (ΔPSTE = 1.62) and a negative change in the TR low-PSTE group (ΔPSTE = -1.12) in the standard pretest-posttest method, the gain scores increased to 4.86 for the US and 3.30 for the TR low-PSTE groups in the retrospective pretest-posttest method. While the main effect of the measurement method was found to be significant, no significant interaction of the PSTE levels and measurement method is detected neither in US [F (1, 37) = 0.53; p = .47; η^2 = .01] nor in TR [F (1, 76) = 0.04; p = .85; η^2 < .001] samples. Therefore, no significant difference between the low and high PSTE groups has been observed for the difference of the PSTE gains from standard and retrospective measurements. Compared to very large effect sizes calculated for the main effect of the measurement method, partial eta-squared values for the interaction effect of the PSTE levels and the measurement method indicated that the effects of this interaction on the gain scores are very small in both samples.

As seen in Tables 2 and 3, consistent with the results from Table 1, both males and females from the US and TR samples experienced similar changes between the standard and retrospective measurements. Both genders experienced higher gains in the retrospective pretest-posttest method, compared to the standard pretest-posttest method. The results of the repeated measures ANOVA showed that there was no significant difference between the males and females in terms of the difference between their gain scores from standard and retrospective tests, both for the US [F (1, 37) = 1.82; p = .19; η^2 = .05] and TR [F (1, 76) = 0.24; p = .63; η^2 = .003] samples. Therefore, no statistically significant interaction effect of gender and measurement method has been observed on the difference between the gain scores. The partial eta-squared effect sizes also indicate that gender and measurement method interaction has a lower than moderate effect (Pallant, 2007) on the US sample's and a very small effect (Pallant, 2007) on the TR sample's PSTE gain scores.

Discussion

This research study investigated the PSTE changes of preservice elementary teachers during science methods courses, which are measured with standard and retrospective tests. Based on the findings from the first research problem, in agreement with the results from the similar studies of Cantrell (2003) and Hechter (2011), the participants from two distinctly different samples reported lower scores in the retrospective pretest than their scores in the standard pretest. For this reason, compared to the standard pretest-posttest method, participants in both US and TR samples experienced significantly higher PSTE gains in the retrospective pretest-posttest method.

This finding raises the concern that if one was to compare the pretest and posttest scores of these participants with the standard and retrospective tests, the conclusions would be drastically different. For example, suppose that a researcher is comparing the pretest and posttest scores of the participants of this study with the standard method. The conclusion would be a statistically insignificant change (p = .068). Therefore, the science methods courses that participants had been attending would be labelled as ineffective. Moreover, the methods course at the Turkish university would have been reported to negatively impact the self-efficacy beliefs of more than half of the Turkish participants (low-PSTE group). However, retrospective test scores tell a different story. When students are allowed to evaluate themselves with the same metric for their self-efficacy beliefs before and after the science methods course, the differences between their mean scores becomes higher and therefore the pretest-posttest comparison test result provides a significant value (p < .001). These differences in test results would certainly alter the evaluations about the science methods course.

The second research problem was about the impact of two independent variables on the differences between the gain scores from standard and retrospective tests. The interaction effects for these variables (PSTE level*measurement method and gender*measurement method) have been both found as statistically insignificant. Therefore, unlike the significant impact of the measurement method, no significant impact of participants' PSTE level or gender has been detected on the change of their gain scores between standard and retrospective measurements. These findings indicate that the differences between the standard and retrospective measurements are due to the nature of the measurement methods and not specific to any gender, PSTE level, or nationality. Combined with the results of Cantrell (2003) and Hechter (2011) in science education, as well as with the results of studies in other areas (Aiken & West, 1990; Hoogstraten, 1985; Howard, 1980; Howard & Dailey, 1979; Howard, Ralph et al., 1979; Lamb & Tschillard, 2005; Rohs, 1999), one can conclude from the findings of this study that the differences between the retrospective and standard measurements can be seen in almost all educational studies around the world.

In light of the previous literature, the main reason for observing a significant difference between the standard and retrospective measurement methods is due to the change of metric by time. As was discussed in a limited number of previous studies (Cantrell, 2003; Howard & Dailey, 1979; Howard et al., 1979; Rohs, 1999), the change of metric from pretest to posttest (response-shift bias) in standard pretest-posttest design studies, poses a significant threat to validity, especially in studies that use self-report Likert scale instruments. Therefore, a significant outcome of this study is to draw attention on the use of retrospective pretests, instead of standard pretests, to enhance the validity by allowing participants to evaluate themselves with the same metric in both pretest and the posttest. One of the participants of Cantrell's (2003) study comments about this fact as:

> Before, I assumed I didn't need to know as much to be a good teacher. I can't believe I strongly agreed in the beginning [standard pretest]. I gave myself more credit than I think I really had. In looking back, [retrospective pretest] I was more realistic because of learning so much in class. My whole perspective has changed (p. 181).

On the other hand, there are serious critics of the retrospective testing as well. For example, Nimon (2014) noted that retrospective measurements in workforce education may also be subject to bias. In her paper, where she addressed the critics about the standard and retrospective measurements, Nimon stated that "While response shift theory justifies the retrospective judgement as more valid, implicit theories and the concept of impression management support the standard judgement as more valid (p. 267)." Similarly, Hill and Betz (2005) argued that using retrospective pretests does not necessarily eliminate bias. Instead, they defined that emotion related biases and recall bias may appear in retrospective testing approach, especially when the study duration is long. Nimon (2014) suggests that additional measures are needed in evaluation research studies to validate the retrospective data, which can be achieved by using control groups or concurrent validity measures.

After a critical comparison of standard and retrospective testing methods, Nimon (2014) argued that differences between the ratings from these tests may not always be related to response-shift bias; however, retrospective pretest can be considered to be a more valid assessment tool, especially when participants cannot to be expected to know their initial levels before an intervention. Based on the critics on both sides, one can conclude that investigating the change of preservice teachers' PSTE beliefs during a science methods course would fit the situation described above by Nimon (2014) and therefore retrospective testing deserves more attention from science educators in studying the changes in their students' beliefs.

Conclusions

This study joins the related literature on measuring PSTE beliefs (Cantrell, 2003; Hechter, 2011) that there is a statistically significant impact of the testing method on the preservice teachers' personal efficacy levels. Although same participants were surveyed among the same time period, two different scenarios were obtained from standard and retrospective testing methods. This study does not claim that retrospective testing method is more superior, or more reliable than the standard testing method. However, it serves as an alarming signal that response-shift bias would seriously impact the results of studies that employ the standard pre-post testing method and reminds an alternative testing method of retrospective testing, which would be more suitable for the nature of measuring self-reported variables such as the personal self-efficacy beliefs.

From the findings of this study and the further discussions in the previous literature on the comparisons of standard and retrospective testing methods (Cantrell, 2003; Nimon, 2014), it can be concluded that both standard and retrospective testing methods have their own limitations; however educational researchers should pay more attention on how to eliminate the response-shift bias when using pretest-posttest comparisons. When a control group or additional validity measures are not available, retrospective testing methods seem to be a better alternative for measuring the longitudinal changes in participants' self-reported measures. So far, research shows that response-shift bias is due to human nature and it is a common universal threat to validity in educational research. However, further studies are needed from different research areas that use Likert scale instruments to draw attention on this important topic. By this way, it can be investigated in detail how the metric changes occur in educational contexts from pretest to posttest and how this significant threat can be eliminated from research studies.

References

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

Allen, J. M., & Nimon, K. (2007). Retrospective pretest: A practical technique for Professional development evaluation. *Journal of Industrial Teacher Education, 44* (3), 27–42.

Appleton, K. (2003). How do beginning primary school teachers cope with science? Toward an understanding of science teaching practice, *Research in Science Education, 33*, 1–25.

Ashton, P. T., & Webb, R. B. (1986). *Making a difference: Teachers' sense of efficacy and student achievement*. White Plains, N.Y.: Longman.

Bahcivan, E. (2014). Examining relationships among Turkish pre-service science teachers' conceptions of teaching and learning, scientific epistemological beliefs and science teaching efficacy beliefs. *Journal of Baltic Science Education, 13* (6), 870–882.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84* (2), 191-215.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman.

Bayraktar, S. (2011). Turkish preservice primary school teachers' science teaching efficacy beliefs and attitudes toward science: The effect of a primary teacher education program. *School Science and Mathematics, 111* (3), 83–92.

Bleicher, R. E. (2009). Variable relationships among different science learners in elementary science-methods courses, *International Journal of Science and Mathematics Education, 7*, 293–313.

Bleicher, R. E., & Lindgren, J. (2005). Success in science learning and preservice science teaching self-efficacy. *Journal of Science Teacher Education, 16*, 205–225.

Braund, M., & Leigh, J. (2013). Frequency and efficacy of talk-related tasks in primary science. *Research in Science Education, 43*, 457–478.

Brigido, M., Borrachero, A. B., Bermejo, M. L., & Mellado, V. (2013). Prospective primary teachers' self-efficacy and emotions in science teaching. *European Journal of Teacher Education, 36* (2), 200–217.

Bursal, M. (2008). Changes in Turkish pre-service elementary teachers' personal science teaching efficacy beliefs and science anxieties during a science method course. *Journal of Turkish Science Education (TUSED), 5* (1), 99–112.

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

Cantrell, P., Young S., & Moore A. (2003). Factors affecting science teaching efficacy of preservice elementary teachers. *Journal of Science Teacher Education, 14* (3), 177–192.

de Laat, J., & Watters, J. J. (1995). Science teaching self-efficacy in primary school. *Research in Science Education, 25* (4), 453–464.

Enochs, L. G., & Riggs, I. M. (1990). Further development of an elementary science teaching efficacy instrument: A preservice elementary scale. *School Science and Mathematics*, *90* (8), 694–706.

Eshach, H. (2003). Inquiry-events as a tool for changing science teaching efficacy belief of kindergarten and elementary school teachers. *Journal of Science Education and Technology, 12* (4), 495–501.

Ginns, I. S., Tulip, D. F., Watters, J. J., & Lucas, K. B. (1995). Changes in preservice elementary teachers' sense of efficacy in teaching science. *School Science and Mathematics, 95* (8), 394–400.

Ginns, I. S., & Watters, J. J. (1999). Beginning elementary school teachers and the effective teaching of science. *Journal of Science Teacher Education, 10* (4), 287–313.

Hechter, R. P. (2011). Changes in preservice elementary teachers' personal science teaching efficacy and science teaching outcome expectancies: The influence of context. *Journal of Science Teacher Education, 22*, 187–202.

Hill, L. G., & Betz, D. L. (2005). Revisiting the retrospective pretest. *American Journal of Evaluation, 26* (4), 501–507.

Hoogstraten, J. (1985). Influence of objective measures on self-reports in a retrospective pretest–posttest design. *Journal of Experimental Education, 53* (4), 207–210.

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

Howard, G. S., & Dailey, P. R. (1979). Response shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology, 64* (2), 144–150.

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 evaluation and a re-evaluation of retrospective pretests. *Applied Psychological Measurement, 3* (1), 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* (2), 129–135.

Huinker, D., & Madison, S. K. (1997). Preparing efficacious elementary teachers in science and mathematics: The influence of methods courses. *Journal of Science Teacher Education, 8* (2), 107–126.

Jarrett, O. S. (1999). Science interest and confidence among preservice elementary teachers. *Journal of Elementary Science Education, 11* (1), 47–57.

Lamb, T. A., & Tschillard, R. (2005). Evaluating learning in Professional development workshops: Using the retrospective pretest. *Journal of Research in Professional Learning*. Accessed October 10, 2014. http://olms1.cte.jhu.edu/olms/data/resource/6794/Evaluating%20Learning%20in%20PD%20Workshops_OST-PD.pdf

Marshall, J. C., Horton, R., Igo, B. L., & Switzer, D. M. (2009). K-12 science and mathematics teachers' beliefs about and use of inquiry in the classroom. *International Journal of Science and Mathematics Education, 7*, 575–596.

McKinnon, M., Moussa-Inaty, J., & Barza, L. (2014). Science teaching self-efficacy of culturally foreign teachers: A baseline study in Abu Dhabi. *International Journal of Educational Research, 66*, 78–89.

Moore, J. J., & Watson, S. B. (1999). Contributors to the decision of elementary education majors to choose science as an academic concentration. *Journal of Elementary Science Education, 11*, 37–46.

Morrell, P. D., & Carroll, J. B. (2003). An extended examination of pre-service elementary teachers' science teaching self-efficacy. *School Science and Mathematics, 103* (5), 246–251.

Nilsson, P., & van Driel, J. (2011). How will we understand what we teach? Primary student teachers' perceptions of their development of knowledge and attitudes towards physics. *Research in Science Education, 41*, 541–560.

Nimon, K. (2014). Explaining differences between retrospective and traditional pretest self-assessments: competing theories and empirical evidence. *International Journal of Research and Method in Education, 37* (3), 256–269.

Pallant, J. (2007). *SPSS survival manual*. New York: Mc Graw Hill.

Palmer, D. H. (2001). Factors contributing to attitude exchange amongst preservice elementary teachers. *Science Education, 86* (1), 122–138.

Palmer, D. H. (2006a). Durability of changes in self-efficacy of preservice primary teachers. *International Journal of Science Education, 28*, 655–71.

Palmer, D. H. (2006b). Sources of self-efficacy in a science methods course for primary teacher education students. *Research in Science Education, 36*, 337–353.

Plourde, L. A. (2002). The influence of student teaching on preservice elementary teachers' science self-efficacy and outcome expectancy beliefs. *Journal of Instructional Psychology, 29* (4), 245–253.

Richardson, G. M., & Liang, L. L. (2008). The use of inquiry in the development of preservice teacher efficacy in mathematics and science. *Journal of Elementary Science Education, 20* (1), 1–16.

Rohs, F. R. (1999). Response shift bias: A problem in evaluating leadership development with self-report pretest-posttest measures. *Journal of Agricultural Education, 40* (4), 28–37.

Soprano, K., & Yang, L. (2013). Inquiring into my science teaching through action research: A case study on one pre-service teacher's inquiry-based science teaching and self-efficacy. *International Journal of Science and Mathematics Education, 11* (6), 1351–1368.

Taştan-Kırık, Ö. (2013). Science teaching efficacy of preservice elementary teachers: Examination of the multiple factors reported as influential, *Research in Science Education, 43*, 2497–2515.

Tekkaya, C., Cakiroglu, J., & Ozkan, O. (2004). Turkish pre-service science teachers' understanding of science and their confidence in teaching it. *Journal of Education for Teaching, 30* (1), 57–66.

Utley, J., Moseley, C., & Bryant, R. (2005). Relationship between science and mathematics teaching efficacy of preservice elementary teachers. *School Science and Mathematics, 105* (2), 82–87.

Wenner, G. (1993). Relationship between science knowledge levels and beliefs toward science instruction held by preservice elementary teachers. *Journal of Science Education and Technology, 2* (3), 461–468.

Wenner, G. (1995). Science knowledge and efficacy beliefs among preservice elementary teachers: A follow-up study. *Journal of Science Education and Technology, 4* (4), 307–315.

Wenner, G. (2001). Science and mathematics efficacy beliefs held by practicing and prospective teachers: A 5-year perspective. *Journal of Science Education and Technology, 10* (2), 181–187.

Yılmaz, H., & Huyugüzel-Çavaş, P. (2008). The effect of the teaching practice on pre-service elementary teachers' science teaching efficacy and classroom management beliefs. *Eurasia Journal of Mathematics, Science & Technology Education, 4* (1), 45–54.

Yürük, N. (2011). The predictors of pre-service elementary teachers' anxiety about teaching science. *Journal of Baltic Science Education, 10* (1), 17–26.

Received: *March 23, 2015* Accepted: *April 26, 2015*

Murat Bursal Ph.D., Associate Professor, Cumhuriyet University, Faculty of Education, Elementary Science Education, 58140 Sivas, Turkey.
E-mail: mbursal@cumhuriyet.edu.tr, mbursal@gmail.com
Website: http://mbursal.cumhuriyet.edu.tr/