



An experimental study of a museum-based, science PD programme's impact on teachers and their students

C. Aaron Price & A. Chiu

To cite this article: C. Aaron Price & A. Chiu (2018): An experimental study of a museum-based, science PD programme's impact on teachers and their students, International Journal of Science Education, DOI: [10.1080/09500693.2018.1457816](https://doi.org/10.1080/09500693.2018.1457816)

To link to this article: <https://doi.org/10.1080/09500693.2018.1457816>



© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 09 Apr 2018.



[Submit your article to this journal](#)



[View related articles](#)



[View Crossmark data](#)

An experimental study of a museum-based, science PD programme's impact on teachers and their students

C. Aaron Price  and A. Chiu

Museum of Science and Industry, Chicago, IL, USA

ABSTRACT

We present results of an experimental study of an urban, museum-based science teacher PD programme. A total of 125 teachers and 1676 of their students in grades 4–8 were tested at the beginning and end of the school year in which the PD programme took place. Teachers and students were assessed on subject content knowledge and attitudes towards science, along with teacher classroom behaviour. Subject content questions were mostly taken from standardised state tests and literature, with an 'Explain:' prompt added to some items. Teachers in the treatment group showed a 7% gain in subject content knowledge over the control group. Students of teachers in the treatment group showed a 4% gain in subject content knowledge over the control group on multiple-choice items and an 11% gain on the constructed response items. There was no overall change in science attitudes of teachers or students over the control groups but we did find differences in teachers' reported self-efficacy and teaching anxiety levels, plus PD teachers reported doing more student-centered science teaching activities than the control group. All teachers came into the PD with high initial excitement, perhaps reflecting its context within an informal learning environment.

ARTICLE HISTORY

Received 23 August 2017
Accepted 23 March 2018

KEYWORDS


Professional development;
informal science; museum;
teacher education

Introduction

Teacher professional development, sometimes referred to as teacher education, teacher learning and in-service education (hereafter: PD), is a vital element of science education reform and innovation (National Academies of Sciences, Engineering and Medicine, 2015). PD leads to change in teacher knowledge, confidence and awareness. In turn, that can lead to behavioural change and increased student learning (Desimone, 2009; Wilson, 2013). Perceptions of the importance of teacher education are increasing and it is now included at the top of international education policy discussions (Darling-Hammond, 2017).

Lately, there have been calls for more studies that directly link teacher and student performance using experimental designs that allow causal connections (National Research

CONTACT C. Aaron Price  aaron.price@msichicago.org

 Supplemental data for this article can be accessed at <https://doi.org/10.1080/09500693.2018.1457816>

© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

Council, 2010; Wilson, 2013; Yoon, Duncan, Lee, Scarloss, & Shapley, 2007). Others have also suggested more natural research designs that include measures of teacher instruction behaviour (Blank, de las Alas, & Smith, 2008) and instruments that are aligned with what teachers and students see in everyday classrooms (Tierney, 2013).

This study looked at the impact of a science museum-based PD programme on subject content knowledge and attitudes towards the science of its teacher participants and their students, along with behaviours of the teachers. The studied PD programme is located in a museum serving a wide, mostly urban community that is heterogeneous in terms of demographic and socio-economic backgrounds. With generalisability in mind, this study includes an experimental design while incorporating assessments commonly used in the classroom. Our research question was, 'What impact does participation in a science-museum based PD programme have on its teacher participants' and their students' scientific content knowledge and attitudes towards science?'

We begin with a literature review of the PD field focused on research design and PD within informal science institutions. We then discuss the PD programme being studied followed by a description of the methodology and population. Finally, we discuss results and interpret them through the field of science-themed PD. Implications are discussed with a focus on what can and cannot be generalised to other PD programmes and implications for practitioners and researchers.

Literature review

Teacher in-service professional development is characterised as a systematic effort to bring about change in the classroom practices of teachers, their attitudes and beliefs and, ultimately, increased or improved student learning (Desimone, 2009; Guskey, 2002; Wilson, 2013). This process can be complex and slow (Osborne, Simon, Christodoulou, Howell-Richardson, & Richardson, 2013), involving many causal links that take time to transverse (Wayne, Yoon, Zhu, Cronen, & Garet, 2008). Most PD programmes are formal learning experiences for teachers that take place in classrooms with relatively broad topics and goals (National Academies of Sciences, Engineering and Medicine, 2015). But, they can also be less formal, more discrete activities such as workshops, conferences, communities of practice and institutes focused on more specific topics (Desimone, 2009; Stokes, Evans, & Craig, 2017; Vangrieken, Meredith, Packer, & Kyndt, 2017). There are also PD programmes embedded in practice, including co-teaching, mentoring, observation and/or reflective practice (Cohen & Ball, 1999; Putnam & Borko, 2000). Postholm (2012)'s review of recent PD scholarship found that some of the most effective PD models involve in-school learning processes. More recently, some programmes have begun to incorporate the whole school improvement process (Chiu, Price, & Ovrachim, 2015). Teachers can be overwhelmed by the PD options. NRC (2010) cites a school district that had more than 1000 PD opportunities listed in its catalogue for one year. Due to the differing educational environments within each country, international consensus on what constitutes teacher education is even more fragmented (Musset, 2010). PD assessment is one area in particular where international PD programmes can learn more from one another (Darling-Hammond, 2017).

Yet, among this complex landscape, a core set of established best practices in teacher PD may be emerging. Desimone (2009) suggests five aspects of effective professional

development – focused content, active learning, coherence, sufficient duration and collective participation. In a literature review of 44 research studies, Van Driel, Meirink, van Veen, & Zwart (2012) recommended adding a sixth aspect – school organisational conditions. Internationally, the most common components of teacher PD found in OECD nations include a focus on content knowledge and pedagogical technique (Musset, 2010). While the literature is coming to a consensus on the broad features of effective PD, the evidence is weak about the level of impact (Avalos, 2011) and the specific features that make the largest differences (Wayne et al., 2008).

Science-focused PD has its own unique needs. NRC (2010) calls for science PD opportunities that are rich in scientific and engineering practices, crosscutting concepts and disciplinary core ideas – central pillars of the Next Generation Science Standards (NGSS), a new set of science standards in the process of adoption in the United States. Science-specific PD models tend to focus on the problem- and inquiry-based learning and pedagogical techniques as specific vehicles for teaching science (Akerson & Hanuscin, 2007; Asghar, Ellington, Rice, Johnson, & Prime, 2012; Capps & Crawford, 2013; Nadelson, Seifert, Moll, & Coats, 2012). Globally, policy-makers are increasing their calls for more science PD programmes focused on knowledge and practice (Luft & Hewson, 2014).

In response to the active growth of the field and a wide variety of experiences, researchers have called for more complex, exhaustive and rigorous research in PD (Huber, 2011). In particular, there have been calls for more experimental designs (Wayne et al., 2008; Whitcomb, Borke, & Liston, 2009; Wilson, 2013) and studies of large cohorts of teachers across multiple school districts (NRC, 2010). A 2009 review of 1343 teacher PD studies found only 9 that met the What Works Clearinghouse, the U.S. Department of Education's Institute of Education Sciences repository of educational intervention effectiveness, criteria of acceptable study design, *none* of which included middle or high school teachers (Guskey & Yoon, 2009; Yoon et al., 2007) and only *one* was about science. While that criteria discounts qualitative (Adams St Pierre & Roulston, 2006) and mixed-methods research (Chatterji, 2005), which are vital in studying the complexity of PD programmes (James & McCormick, 2009), the criteria can be useful to evaluate studies that are intentionally quantitative and whose primary goals include generalisability. Blank et al. (2008) suggest research on PD should be focused on four programmatic aspects: programme quality, teacher content knowledge, teacher instruction and student learning. Measures used in PD studies can be categorised as proximal or distal (cited by Kennedy, 2016). Proximal measures are designed for the study and may report greater impact. Distal measures, such as state-sponsored standard instruments, are often used in studies of PD impact on student achievement (Akiba & Liang, 2016; Desimone, Smith, & Phillips, 2013; Martin et al., 2010; Ross, Bruce, & Hogaboam-Gray, 2006).

Together, this literature led to our adoption of an experimental design that measures both teacher and student learning along teacher classroom activity while using a large, population sample spanning dozens of schools and districts.

Large-scale studies of science-focused PD employing experimental designs are difficult to find in the literature (Whitcomb et al., 2009). Penuel, Gallagher, and Moorthy (2011) found that using models of teaching and assessment to prepare teachers to design sequences of instructional experiences for students led to increased student learning

about earth science topics. Another found differential effects addressing the same elementary science content using three different PD models (Heller, Daehler, Wong, Shinohara, & Miratrix, 2012). A recent experimental study found an attitude-focused PD programme had a positive impact on lowering teacher anxiety and reliance on contextual factors, but no impact on other attitude traits such as self-efficacy, beliefs about relevance or teaching behaviour (van Aalderen-Smeets, Walma van der Molen, van Hest, & Poortman, 2017). Kyriakides, Christoforidou, Panayiotou, and Creemers (2017) found a strong impact on teacher skills over a control group when looking at a three-year PD programme that was individualised to teacher needs.

Professional development by informal education institutions

Broadening our understanding of the context in which teachers both teach and learn is one of the fundamental challenges of modern PD research (Luft & Hewson, 2014). Among those contexts are museums and other informal science education (ISE) institutions, who are becoming increasingly active in providing teacher PD, but are rarely included in research studies (NRC, 2010). They can help build capacity and take advantage of community expertise and resources (Traphagen & Traill, 2014). ISE programmes tend to be more object-based, student-centred and have a broader content focus, while more formal PD settings tend to be more expert-based, teacher-centred and with focused content (Astor-Jack, McCallie, & Balcerzak, 2006). But successful informal PD programmes also consider formal aspects of education, such as policy, theories of learning, programme design and assessment (Bevan et al., 2010). Museum and science centre-based PD programmes often collaborate with other informal education organisations such as libraries, afterschool clubs, youth programmes and cultural institutions (Bevan et al., 2010) or universities (Gupta, Adams, Kisiel, & Dewitt, 2010). Extensive teacher education programmes offered by informal science centres in the United States include those hosted at the American Museum of Natural History (Nadeau et al., 2013), Exploratorium (Heredia & Yu, 2015), Museum of Science, Boston (Cunningham, 2009) and Museum of Science and Industry, Chicago (Wunar & Kowrach, 2017) among many others. A search of Villegas-Reimers (2003) seminal review of the global PD field found no mentions of the words ‘museum’, ‘zoo’, ‘aquarium’ or ‘science centre’, suggesting that teacher PD could be an area of growth for ISEs in all nations.

Many studies have looked at unique aspects ISEs can offer teacher education programmes, such as learning in a low-stakes, supportive environment instruction and making connections with other STEM-rich institutions (Anderson, Lawson, & Mayer-Smith, 2006; Buczynski & Hansen, 2010; Çil, Maccario, & Yanmaz, 2016; Gupta & Adams, 2012; Setioko & Irving, 2017; Yu & Yang, 2010). ISEs can make use of their unique resources by leveraging them in their PD models and showing teachers how they can integrate them into their classroom curriculum (Holliday, Lederman, & Lederman, 2014; Phillips, Finkelstein, & Wever-Frerichs, 2007). One study found a museum-based PD programme showed gains in teachers up to 2 years after they finished the programme and attributed it specifically to the field work the teachers were able to do in the programme (Melber & Cox-Petersen, 2005). A case study of another museum programme found that the excitement of the teacher being in a

highly engaging environment helped motivate them to apply what they learned in the programme (Grenier, 2010).

Methods

Study context: a middle school PD programme at a large, urban science museum

The studied teacher PD programme takes place at a large urban science museum in the United States. It typically runs four courses per year which rotate among five overall topics. Two courses run in the summer and two during the academic school year. This study focuses on two of the academic year courses on the topics of ecological science (Expedition Green, hereafter EG) and physical science (Get Re-Energized, hereafter GRE). Enrolled educators mostly teach children aged 8–12 (typically grades 4–8 in the United States). A typical course schedule includes six day-long sessions spread across the school year. The 8-hour day is separated into about eight sessions focused on a unifying theme and includes content/pedagogical lessons along with breakfast and lunch. Teachers are asked to actively participate, keep a portfolio (for which they receive written feedback), complete homework, and collaborate with their partner teachers in person and online. Participants receive funding for substitute teachers, a bus ride for their students to visit the museum and instructional material support. They also join a professional learning community, earn state-certified clock hours toward their continuing professional development units and have the opportunity to leverage their participation in the course to earn graduate credit at local universities. They are requested to attend with a partner from their school. The selection process emphasises teachers new to teaching science and those from lower-resourced schools. Each course consists of two cohorts of about 32 teachers for a total of 128 participants per school year (not counting the summer courses which include an additional 64 participants).

The programme model's overarching goal is to better prepare teachers for teaching STEM concepts to increase student learning and is aligned with five of the six key design principles proposed by Desimone (2009) and Van Driel et al. (2012) (Table 1). Together, the principles lead to effective PD. The programme focuses on teacher subject and pedagogical content knowledge, increasing confidence, fostering stewardship among teachers and creating a community of practice to provide stability and support. There is a heavy emphasis on the NGSS's aspects of three-dimensional learning and cross-cutting concepts, presented in a constructivist manner with the programme educators modelling NGSS-aligned practices. Other key aspects of the model include its location within a museum and its enrolment size. PD staff often make connections between content and exhibits, which has been shown to help teachers make connections to their own classrooms (Holliday et al., 2014).

This research study design is informed by Blank et al.'s (2008) suggestion to measure teacher and student content knowledge along with teaching instruction. In so doing, we employed an experimental, pre-/post-test design of both teachers and their students. With Brewer and Crano's (2000) definition of ecological validity in mind, our assessments were chosen from those widely used in schools to better represent what the population encounters in everyday life. Using research design categories described in a recent commentary on PD research in Education Researcher, the study would qualify as a Stage 3 project – generally

Table 1. Alignment of studied programme with key design features of effective teacher PD.

Design feature	Aligned programme characteristics
Content focus (Desimone, 2009)	Each course is focused on one of five topics (anatomy and life science, earth systems science, environmental science, physical science, space science and engineering)
Active learning (Desimone, 2009)	Courses are run with teachers treated as learners and course staff modelling the inquiry and NGSS-aligned behaviour. Authentic resources are incorporated, such as ongoing classroom portfolio reviews
Coherence (Desimone, 2009)	The staff aligns the programme to address many of the area-specific challenges that teachers experience. This includes school, district and community-related issues. For example, the largest local school district often places K-8 grades within the same building. This means many sixth, seventh and eighth grade science teachers do not have access to science classrooms or materials. The programme is designed to provide them with <i>all</i> the materials needed to implement the lessons which are designed to not require pre-installed laboratory equipment
Duration (Desimone, 2009)	With over 56 contact hours across an entire school year, the programme has time to dive deep. Since the hours are spread across the year, it sustains momentum and allows teachers to reflect and staff to iterate as needed
Collective/collaborative participation (Desimone, 2009)	Teachers are requested to participate with a partner teacher from the same school. The building of a community of practice among all teachers is a key programme element through considerable group work, professional networking time and online community resources that are available beyond the course period
School organisational conditions (Van Driel et al., 2012)	The staff maintains awareness of school conditions, but this is the one design feature that is not heavily addressed in the programme. One reason is that on average over 40 schools are represented in a course. The Museum does offer a separate PD programme that is focused on the whole school, in which many course teachers also participate. However, none of the teachers in this study were participating at the time of data collection

reflecting that it is an experimental study of the moderate size of both heterogeneous teacher and student populations in real-world settings (Hill, Beisiegel, & Jacob, 2013).

Participant selection

During the 2015–2016 school year, 198 teachers applied and were accepted into the PD programme. The pool of teachers was then randomly divided into two groups by a third party and weighted so that the treatment group equalled the capacity of the programme ($N = 128$, 64 in each course). The treatment group was given a US\$50 financial incentive to participate in the study. The control group was guaranteed acceptance into the following year's programme and provided with the same \$50 financial incentive to participate as the treatment group, an additional \$50 gift card to a popular teacher curriculum website and a free bus reservation for a student field trip to the museum. The final numbers, which reflect those who turned down participation, included 63 accepted into the EG treatment group, 37 into the EG control group, 64 into the GRE treatment group and 33 into the GRE control group. Teachers were asked to give the student tests to their first and last classes of the day to prevent preferential selection of high achieving classes.

Teacher instruments

The teacher pre-test was given online and took approximately 20 minutes to complete. It was composed of 40 questions in four sections addressing science attitudes, science

behaviours, subject content knowledge and demographics. The attitude and behaviour items were taken from the Dimensions of Attitude toward Science (DAS) Instrument (van Aalderen-Smeets & Walma van der Molen, 2013). The DAS was designed to measure attitudes of pre- and in-service teachers at the primary school level in the Netherlands and has shown strong validity and reliability in international studies. The DAS attitude scale has 28 items corresponding to seven subscales in a five-point Likert format. Their original scale categories ranged from Totally Disagree to Totally Agree, but we changed the word 'Totally' to 'Strongly'. Van Aalderen-Smeets et al. also report seven DAS subscales – Anxiety, Contextual Factors or Enjoyment, Difficulty of Science Teaching, Gender-Stereotypical Beliefs, Perceived Dependency, Perceived Relevance and Self-efficacy. The DAS also includes a separate instrument in which in-service teachers can report how often they engage in science teaching behaviours (Behavioural Intention Scale), which we included as well. This includes seven items using a five-point scale with categories labelled as: 'Seldom or Never,' 'Couple Times a Year,' '1-3 times a month,' 'Weekly' and 'Daily.' We modified several items on the DAS Instrument slightly, substituting the word 'primary' with 'K-8' and omitting mention of specific Dutch learning curricula.

Two different subject content knowledge sections were created – one for each course topic with 17–18 items each. The items were taken from state teacher certification tests, online curriculum web sites and published academic literature (see Online Supplemental Material for the full list of sources per test). Reliability for the teacher content knowledge sections was $\alpha = .77$ for the EG content and $\alpha = .85$ for the GRE content.

All subject content questions taken from these sources were in a multiple-choice format. Studies have found that multiple-choice format items have been less sensitive to extreme ranges of ability (Ercikan et al., 1998; Rauch & Hartig, 2010) and show increased bias when used in large scale, across programme assessments (Kim, Walker, & McHale, 2010). To account for this, we extended four of the items on each instrument by appending an 'Explain:' prompt to generate a constructed response. This has been shown as an effective way to increase sensitivity and discrimination of multiple-choice-based assessments (Chen, Gotwals, Anderson, & Reckase, 2016; Lee, Liu, & Linn, 2011). Conceptually, combining item structures in this fashion can also turn the assessment process into a learning experience with the multiple-choice options acting as a scaffold for the constructed response (Cooper, 2015).

Student instruments

The student instrument had three major sections: science attitudes, subject content knowledge questions and demographics. Our attitude items were taken from a questionnaire developed by Barmby, Kind & Jones (2008), designed to measure the change in attitudes towards science in students ages 11–14. It is composed of 37 items that measure six factors of science attitudes. Because of length, we only included the factors of learning science in school, practical work in science and science outside of school, which most closely corresponded to the learning goals of the PD programme. We omitted the factors of self-concept in science, future participation in science and importance of science. We made some minor linguistic modifications. Reliability for the student attitude section was $\alpha = .91$.

Two subject content knowledge sections were created for each course topic – one for elementary (grades 4 and 5 in the United States) and one for middle school (grades 6, 7 and 8). The vast majority of these items were taken from standardized tests from various states within the United States with a few taken from AP tests, curriculum web sites and published academic literature (see Online Supplemental Material for item sources). There were 14–17 subject content items on each student instrument. As with the teacher tests, we added an open-ended ‘Explain.’ prompt to four of the items on each test. Reliability for the subject content sections were $\alpha = .71$ for the EG elementary test, $\alpha = .79$ for the EG middle school test, $\alpha = .66$ for the GRE elementary test and $\alpha = .67$ for the GRE middle school test. Lower reliability estimates for the GRE tests could be due to having fewer items (14 and 15) than the EG tests (17 items each).

Study logistics

After acceptance into their assigned groups, teachers in all groups were sent information about the research study via email. Teacher tests were taken online. Student tests were shipped to teachers to hand out in class for completion at home. Teacher and student pre-tests were sent to teachers of both the treatment and control groups prior to the first scheduled PD session. Post-tests were sent to both groups after the last PD session.

Population

The average age of teachers was 38 (SD = 11) and they had an average of 10 (SD = 8) years teaching experience. Gender questions were asked in the open-ended, ‘What is your gender?’ format (Human Rights Commission, 2016). Their gender makeup was 86% female and 14% male with no responses that could be categorised as neither female nor male. The top three race/ethnic groups were White (75%), Black/African American (16%) and Chinese (4%). Asked separately, 11% identified as of Hispanic, Latino, or Spanish origin. When asked to classify the population that their school serves on a five-point scale, teachers reported 50% Lower Class, 15.4% Lower Middle Class, 3.1% Middle Class. Less than 1% reported Upper Middle Class and none reported Upper Class. Teachers represented 93 schools in 50 different school districts (when counting independent schools as distinct school districts). Eighty were public schools, five private, five parochial, three charters. Students self-reported as 55% female and 45% male. The top three student racial groups were White (57%), Black or African American (30%) and American Indian or Alaskan Native (9%). About 62% reported as being of Hispanic, Latino or Spanish origin.

Analysis

Quantitative data were analysed with SPSS 19. All Likert responses from the attitudes and behaviours sections were converted into a numerical ascending scale from 1 to 5. Responses to the subject content sections were coded as correct (1) and incorrect (0). The Explain items were coded as either correct (1), incorrect (0) or missing data (blank or irrelevant responses). We chose not to include partial credit because responses were very succinct making an analysis of nuance difficult. A rubric was developed by two

researchers using an iterative, cycling process (see Online Supplemental Material for the full rubric) that involved coding a sample of the data independently, comparing results and revising the rubric until they were at 80% or better agreement, a common threshold for acceptable inter-rater percent agreement (McHugh, 2012). A third researcher adjudicated any final disagreements. The rest of the data was coded by both researchers using that final rubric.

A composite score based on the mean of correct multiple choice answers on each subject content test was computed. A separate composite score consisting of only the mean scores to the constructed response answers was also computed. These data was not combined because the added variance involved in converting qualitative data into quantitative measures would be lost, leading to overconfidence in the final result (Hammer & Berland, 2014). Also, the items were not designed with constructed responses in mind. Mean scores were also computed for each of the Likert scales and subscales. Data were then compared using functions from the General Linear Model to investigate differences and relationships between groups, with the p value set at .05.

Results

About 83% of the teachers accepted into the programme took the pre-test and 68% took both the pre- and post-test. For the control group, 68.5% of the teachers recruited took the pre-test and 57.1% took both the pre- and post-test. Attrition was attributed mostly to teachers dropping out of the course, leaving the teaching profession and incomplete or invalid student/family consent. A total of 125 teachers (85 treatment, 40 control) and 1676 students (1121 treatment, 551 control) completed both the pre- and post-tests.

Comparison of control and treatment group demographics

Using t -tests, we found no statistical differences between the teacher control and treatment groups in terms of age, gender and years of prior teaching experience, school SES, history of prior teacher PD courses at the museum or whether they are currently taking other PD courses taught by any organisation. Also using t -tests, we found no statistical differences in gender between the student control and treatment groups. We found no statistical difference in the student racial and ethnic groups between the control and treatment groups, except in one case. The control group had 11% more students who identified as Hispanic, which was a significant difference according to a t -test.

Teacher scores

On the subject content sections, pre-test scores of teachers in the control group were higher than teachers in the treatment group. We did not find any differences in demographics or teaching/PD experience that could explain this difference. However, 83% of teachers in the treatment group fully participated in the study while 69% in the control group fully participated. So it is possible that a selection effect exists, in that teachers who participated in the control group portion of the study may have been more intrinsically motivated. However, pre-test attitude scores between the two groups were the same (see Tables 2 and 3).

Table 2. Teacher mean scores.

	Treatment						Control					
	Pre			Post			Pre			Post		
	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>
Environment test	50	56%	11	43	61%	16	22	64%	14	19	65%	12
Physics test	53	58%	16	42	72%	16	26	69%	16	21	70%	15
DAS – attitudes	103	3.23	26	88	3.46	0.45	47	3.22	0.22	40	3.41	0.44
DAS – behavioural	106	3.17	66	88	3.49	0.51	48	3.36	0.53	40	3.41	0.44

Note: Test scores are the percentage of correct answers. DAS scores are on a 5pt. ascending scale.

Overall, subject content test scores of teachers in the treatment group increased by an average of 9% while scores of teachers in the control group increased by an average of 2% (Table 2). This means teachers in the PD showed an increase in subject content knowledge over the control group. This difference in the increase was statistically significant according to a repeated-measures ANCOVA, with the treatment condition as the covariate. The effect size can be classified as medium according to guidelines by Cohen (1988). When included as a covariate, prior teaching experience was not significant at the $p = .05$ level (Table 3). Teachers in the physical science course showed greater gains than the environmental science course (Table 4).

Entering the study, the mean attitude score of teachers of both conditions was 3.2 (0.25) (Table 4). This means both groups came into the PD with the same attitudes toward science. The mean scores of factors defined by van Aalderen-Sweets et al. (2013) ranged from a low of 1.9 (0.72) for the Gender-Stereotypical Beliefs subscale to a high of 4.5 (0.40) for the Perceived Relevance subscale. Between the pre- and post-tests, teacher mean attitude scores increased by 0.23 points for the treatment group and 0.19 points for the control group. The difference was not statistically significant according to a repeated-measures ANOVA. This means teachers in the PD did not show an increase in overall science attitudes over that of the control group. We found no statistically significant difference between the control and treatment groups on the Perceived Relevance, Gender-Stereotypical Beliefs, Difficulty of Science Teaching, Perceived Dependency on Contextual Factors or Enjoyment subscales. However, we did find differences on the Self-efficacy and Anxiety subscales. For Efficacy, scores for the treatment group increased from 3.53 to 3.96 while scores for the control group remained the same at 3.89. This difference was significant according to a repeated-measures ANOVA, $F(2,117) = 9.38, p = .003, \eta_p^2 = .075$, meaning teachers in the treatment group showed more growth in self-efficacy than teachers in the control group. For Anxiety, scores for the treatment group decreased from 2.2 to 1.8 while scores for the control group remained the same at 1.9. This difference was significant according to a repeated-measures ANOVA, $F(2,117) = 4.53, p = .035, \eta_p^2 = .038$. This means

Table 3. Repeated-measures analysis of variance of teacher subject content scores.

Source	df	<i>F</i>	η_p^2	<i>P</i>
Fixed effects				
Time	1	8.46	.069	.004**
Condition	1	2.94	.025	.423
Time × condition covariates	1	6.14	.051	.015*
Teaching experience	1	.648	.006	.42

* $p < .05$. ** $p < .01$.

Table 4. Teacher mean scores for the dimensions of attitude towards science subscales.

	Treatment (N = 78)				Control (N = 39)			
	Pre		Post		Pre		Post	
	M	SD	M	SD	M	SD	M	SD
Self-efficacy**	3.53	0.70	3.96	0.58	3.86	0.66	3.88	0.66
Perceived relevance	4.48	0.44	4.62	0.37	4.66	0.35	4.77	0.32
Gender-stereotypical beliefs	1.97	0.77	1.76	0.68	1.87	0.61	1.71	0.54
Anxiety*	2.18	0.90	1.82	0.79	1.89	0.86	1.89	0.76
Difficulty of science teaching	3.53	0.78	3.55	0.79	3.38	0.69	3.49	0.66
Perceived dependency on context factors	2.75	0.83	2.79	0.88	2.39	0.79	2.72	0.76
Enjoyment	4.18	0.80	4.26	0.87	4.46	0.55	4.48	0.54

* $p < .05$. ** $p < .01$.

teachers in the treatment group reported a slightly lower drop in anxiety than the control group, although the control group began with lower anxiety overall.

Teacher behavioural scores increased by 0.32 points for the treatment group and 0.05 points for the control group. The difference is significant according to a repeated-measures ANOVA comparing the effect of the treatment condition on the behavioural mean score, $F(2, 121) = 10.86, p = .001, \eta_p^2 = .082$. This means treatment group teachers reported more science teaching activities than the control group. Of the subscales identified by van Aalderen-Smeets et al. (2013), we did not find significant differences in change for the Teach Science, Personal, Excursion or Investigate Freely subscales. However, we did find differences in the Activities subscale, $F(1, 121) = 12.24, p < .001, \eta_p^2 = .10$ and the Investigating Together subscale, $F(1, 119) = 4.55, p = .035, \eta_p^2 = .04$. Based on definitions of the subscales as defined by the scales' authors, this means that teachers in the PD programme reported a greater increase in doing more hands-on science activities alongside their students than those in the control group.

Student scores

On the multiple choice subject content tests, scores of students in the treatment group increased by an average of 7.6% while scores of students in the control group increased by an average of 4.2% (Table 5). This difference in the increase was statistically significant with a medium effect size according to a repeated-measures ANCOVA (Table 6).

Mean scores of the constructed response items were lower than for the multiple choice items. The mean pre-test score for the multiple choice test (across groups) was 52% while it was 22% for the constructed response items. One reason for this could be that multiple-choice items coded dichotomously can be subject to affects caused by guessing (Lesage,

Table 5. Student mean scores for the multiple choice subject content and attitude tests.

	Treatment						Control					
	Pre			Post			Pre			Post		
	N	M	SD	N	M	SD	N	M	SD	N	M	SD
Env. test – elementary	308	0.47	0.20	221	0.55	0.20	84	0.42	0.16	81	0.49	0.23
Env. test – middle	592	0.46	0.19	297	0.54	0.22	189	0.51	0.21	105	0.51	0.25
Phy. test – elementary	470	0.47	0.19	325	0.58	0.20	62	0.41	0.18	31	0.61	0.19
Phy. test – middle	1188	0.52	0.17	278	0.60	0.20	448	0.51	0.18	334	0.55	0.19
Composite – multiple choice												
Attitude scale	2573	3.81	0.60	1126	3.82	0.72	790	3.85	0.62	564	3.80	0.68

Note: Test scores are the ratio of correct answers. Attitude scores reflect a 5pt. ascending scale.

Table 6. Repeated-measures analysis of variance of student multiple choice subject content subject content scores.

Source	df	<i>F</i>	η_p^2	<i>p</i>
Fixed effects				
Time	1	87.1	.075	<.000***
Condition	1	1.17	.001	.279
Time × condition	1	7.56	.007	.006**

p*<.01. *p*<.001.

Valcke, & Sabbe, 2013), thus inflating the multiple choice scores and making the test less sensitive to lower achieving students. Overall, the treatment group showed an increase of 15% in their constructed response scores while the control group saw an increase in 4% in constructed response scores. This difference was statistically significant according to a repeated-measures ANOVA comparing the effect of the treatment condition on the constructed response mean scores, $F(2,1028) = 15.85$, $p < .001$, $\eta_p^2 = .02$. Interestingly, the control group gain for constructed response items was the same as the control group gain for the multiple-choice items, but the treatment group gain for the constructed response items was much higher than that for the multiple-choice items, suggesting that course impact may be larger for higher achieving students whose gains may be muted in the multiple-choice data.

On the attitude tests, mean scores of students in the treatment group increased by an average of 0.01 points while mean scores of the control group decreased by an average of 0.05 points. The difference was not statistically significant according to a repeated-measures ANOVA. We also found no statistically significant difference on any of the specific items.

Discussion

This study investigated the impact of a PD programme on teachers' and students' scientific content knowledge and attitudes, along with classroom behaviour for teachers. It was unique in that it used a large sample size from many different school districts and demographics, included a randomly selected control group and used assessments taken from real world instruments encountered by students and teachers. Results show small but significant increases in subject content knowledge of both teachers and their students over that of the control group. However, we did not find a change in overall attitudes towards science for either group. Usually, effecting attitude change is considered the first step before the change in achievement (Osborne, Simon, & Collins, 2003; Summers & Abd-El-Khalick, 2017). This is an instance where achievement increases precede attitude change. Conversely, behaviour change is often considered the last step in models of teacher education, yet we see the almost instant impact in this study. Teachers in the treatment group reported doing more science activities with their students and also making those activities student-centered more than those in the control group, all during the same year the PD took place. It may be time to think about PD in less of a linear process and as more complex, parallel processes that interact with each other.

Despite the lack of overall science attitude change among teachers, we did find a positive impact on specific attitudes related to self-efficacy and anxiety. Few large-scale studies

have explored PD programme impact on self-efficacy and those with control groups are especially rare (Ross & Bruce, 2007), which is important because a certain level of self-efficacy is needed to even participate in PD opportunities. Many studies have found long duration programmes can have a strong impact on self-efficacy (Blonder, Benny, & Jones, 2014; Duran, Ballone-Duran, Haney, & Beltyukova, 2009; Sandholtz & Ringstaff, 2014), but this study shows some immediate impact. Our finding of lower teacher anxiety is in agreement with other PD studies (Cox & Carpenter, 1989; van Aalderen-Smeets et al., 2017). Anxiety is often the first thing to be affected by PD, as it can be a prerequisite for the embracement of new material being presented to teachers (Guskey, 2002). But, sometimes PD can lead to *increased* anxiety due to concerns associated with change and being challenged with new ideas. However, the anxiety-related items used in this study were written to be directly linked to *classroom teaching* and not feelings of *overall* anxiety. For example, all items have the phrase ‘teaching science’ in them (‘I feel nervous about teaching science.’). So we feel our measured impact is directly related to teaching activities and not about emotional attitudes about teaching overall. Thus, our results show that teachers are feeling more confident and less anxious about classroom teaching about science, but not perhaps about science education as a whole. PD programs may want to make what they teach more salient to other aspects of the non-classroom teaching experience (career confidence, leadership, peer support, etc.).

The studied PD programme is highly resourced, but still constituted only about 42 hours of contact time over the course of an academic year. We were surprised that such a direct impact on teachers and their students was seen within that time frame. The link between teacher PD and student impact is long and complex, and other science and mathematics PD studies have found more impact in the follow-up years than in the PD year (Harris & Sass, 2011; Johnson & Fargo, 2010). The rapid impact could be due to the teacher participants being asked to implement change in their classes during the school year so that they can discuss their experiences with other teachers in subsequent PD session days. This may motivate teachers to quickly apply what they picked up in the PD to the classroom. This may also explain the increase in science teaching behaviour reported by teachers.

The study was designed with ecological validity in mind. All of the items chosen on the content knowledge sections came from instruments taken from the literature and sources commonly used in the classroom, hopefully limiting bias caused by researchers purposefully designing items to match the programme. Choosing appropriate outcomes measures can be difficult, since researchers do not want to be cherry picking topics to assess, yet want to be focused on the same overall goals of the PD. We also avoided using self-report measures about content knowledge gains, which are too often used in PD studies (Van Driel et al., 2012). Wayne et al. (2008) suggests a strategy of choosing ‘multiple instruments that are more or less closely aligned with the specific focus of the PD’, which is the process we followed.

Our effect sizes are not large, but in line with that found in the Yoon et al. (2007) meta-analysis of the nine PD studies they classified as rigorous. Their average student achievement effect size was 0.54 while the effect size of the differences between our conditions is 0.6. Our results expand on the Yoon analysis in that our study included middle school teachers (as opposed to elementary) cover a more recent time period (their analysed studies took place from 1983 to 2003) and our study population was broader in terms of numbers

of teachers, schools and districts. Our effect sizes should be interpreted in light of our avoidance of researcher-development assessments and a quasi-experimental design, which have both been shown to lead to overstated effect sizes in educational research (Cheung & Slavin, 2016). Our effects are also in line with those of another recent museum-based PD study (Schmidt & Cogan, 2014), with the main difference being our study was broader in terms of population and content areas covered. Melber and Cox-Petersen (2005)'s study of museum PD showed perceived higher content knowledge impact, but relied on self-report data.

Teachers may be more excited about PD in ISEs because of the inherent association of ISEs with engagement and fun. Indeed, when measured using the same instrument we used, teachers entering ISE-based PD often show higher overall science attitudes compared to teachers entering non-ISE-based PD (Korur, Vargas, & Serrano, 2016; Riegle-Crumb et al., 2015; Rouweler, 2016; van Aalderen-Smeets & Walma van der Molen, 2013). This suggests that ISE PD could be used to recruit teachers who ordinarily may be less inclined to attend PD. This may also introduce a ceiling effect and explain the lack of change we see on most of our attitude subscales. One challenge informal institutions have is fostering the transfer of PD experiences to the classroom, since they sometimes involve aspects and resources not accessible in a classroom (Astor-Jack et al., 2006; Buczynski & Hansen, 2010). In this programme, using museum exhibits as models of the phenomenon and not as direct teaching aids may have minimised this tendency. Most of the change we found in teaching behaviour was around student-centred teaching. It was mostly in the form of spending more time working side-by-side with their students which is not an ISE-specific behaviour. More studies about the specific impact of museum exhibits on PD are needed.

Blank et al. (2008) called for PD research to be focused on programme quality, teacher content knowledge, teacher instruction and student learning while the National Academies of Sciences, Engineering and Medicine's (2015) report on science teacher learning's first research recommendation is to link PD experiences with changes in teacher practice and student learning. This study focused on each of those areas except for programme quality. It found gains in teacher content knowledge and student learning and explored changes in self-reported teacher attitudes and behaviour. We hope the findings of this study will add to the discussion about ISE-supported PD programmes both within and outside of the United States.

Limitations and further research

Reliability of our subject content sections was limited due to the large number of sources for our items. However, we found it challenging to identify single publicly available assessments that covered these topics at these age ranges (especially for the environmental science student and teacher instruments). This shows the need for more publicly accessible assessments of environmental science and energy at age ranges outside of high school. Also, students did the tests as homework, so approximately half of the students opted not to participate or had incomplete guardian consents meaning those who did participate may have been more motivated and/or have more motivated guardians. Students filled out their surveys as homework due to district-level restrictions about giving surveys in the classroom. However, this was true for both the control and treatment groups so any impact should be consistent among groups.

Wilson (2013) said that ‘It is nearly impossible to isolate the effects of PD on student learning’. We took a holistic approach at measuring the entire programme before we began investigating specific aspects of it. Next steps include a closer look at daily activities in the PD and also the longevity of effect. We did not observe teachers’ teaching, which limits our interpretation of the behavioural results. We are planning teacher observation protocols for a subsequent study. School-based case studies and ethnographic research could help look at the systemic issues related to PD, such as the impact of whole school reform. This study could be categorised an *efficacy* trial (Wayne et al., 2008) in that our results and model have not yet been replicated outside of the museum. The museum is currently running a pilot project to help other museums start their own PD initiatives. Further research with those sites could help with generalisability of the programme model.

Conclusion

A goal of the study was to generate ecologically valid and generalisable results. We employed an experimental, cross-sectional design that covered two scientific domains with a large and diverse population of teachers and students. Our instruments consisting of multiple-choice and open-ended items students and teachers encounter in the everyday classroom. We found positive gains in student and teacher subject content knowledge along with teachers reporting that they conducted more student-centered classroom activities. However, we found no gains in attitudes towards science for students and limited attitude gains with teachers. The subject content knowledge gains all occurred in the first year of the programme, challenging assumptions that PD takes years to show impact. We also show that attitude gains do not necessarily need to precede change in achievement or behaviour suggesting that the process of PD impact could be less linear than commonly considered.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

C. Aaron Price  <http://orcid.org/0000-0002-8954-9758>

References

- Adams St Pierre, E., & Roulston, K. (2006). The state of qualitative inquiry: A contested science. *International Journal of Qualitative Studies in Education*, 19, 673–684.
- Akerson, V. L., & Hanuscin, D. L. (2007). Teaching the nature of science through inquiry: Results of a 3-year professional development programme. *Journal of Research in Science Teaching*, 44, 653–80.
- Akiba, M., & Liang, G. (2016). Effects of teacher professional learning activities on student achievement growth. *The Journal of Educational Research*, 109(1), 99–110.
- Anderson, D., Lawson, B., & Mayer-Smith, J. (2006). Investigating the impact of a practicum experience in an aquarium on pre-service teachers. *Teaching Education*, 17, 341–353.
- Asghar, A., Ellington, R., Rice, E., Johnson, F., & Prime, G. M. (2012). Supporting STEM education in secondary science contexts. *Interdisciplinary Journal of Problem-Based Learning*, 6, 4.

- Astor-Jack, T., McCallie, E., & Balcerzak, P. (2006). Professional development and the historical tradition of informal science institutions: Views of four providers. *Canadian Journal of Math, Science & Technology Education*, 6, 67–81.
- Avalos, B. (2011). Teacher professional development in teaching and teacher education over ten years. *Teaching and Teacher Education*, 27, 10–20.
- Barmby, Patrick, Kind, Per M., & Jones, K. (2008). Examining changing attitudes in secondary school science. *International Journal of Science Education*, 30(8), 1075–1093. <https://doi.org/10.1080/09500690701344966>
- Bevan, B., Dillon, J., Hein, G. E., Macdonald, M., Michalchik, V., Miller, D., ... Yoon, S. (2010). *Making science matter: Collaborations between informal science education organizations and schools*. Washington, DC: Centre for Advancement of Informal Science Education.
- Blank, R. K., de las Alas, N., & Smith, C. (2008). *Does teacher professional development have effects on teaching and learning?: Analysis of evaluation findings from programmes for mathematics and science teachers in 14 states*. Washington, DC: Council of Chief State School Officers. Retrieved from: http://www.ccsso.org/documents/2008/does_teacher_professional_development_2008.pdf
- Blonder, R., Benny, N., & Jones, M. G. (2014). Teaching self-efficacy of science teachers. In R. H. Evans, C. Czerniak, & J. Luft (Eds.), *The role of science teachers' beliefs in international classrooms: From teacher actions to student learning* (pp. 3–15). Rotterdam: Sense Publishers.
- Brewer, M.B., & Crano, W. D. (2000). Research design and issues of validity. In Harry T. Reis & Charles M. Juss (Eds.), *Handbook of research methods in social and personality psychology*. (2nd ed., pp. 3–16). Cambridge, UK: Cambridge University Press.
- Buczynski, S., & Hansen, C. B. (2010). Impact of professional development on teacher practice: Uncovering connections. *Teaching and Teacher Education*, 26, 599–607.
- Capps, D. K., & Crawford, B. A. (2013). Inquiry-based professional development: What does it take to support teachers in learning about inquiry and nature of science? *International Journal of Science Education*, 35, 1947–1978.
- Chatterji, M. (2005). Evidence on 'what works': An argument for extended-term mixed-method (ETMM) evaluation designs. *Educational Researcher*, 34, 14–24.
- Chen, J., Gotwals, A. W., Anderson, C., & Reckase, M. (2016). The influence of item formats when locating a student on a learning progression in science. *International Journal of Assessment Tools in Education*, 3(2), 101–122.
- Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, 45, 283–292.
- Chiu, A., Price, C. A., & Ovrachim, E. (2015). *Supporting elementary and middle school STEM education at the whole-school level: A review of the literature*. Chicago, IL: Museum of Science and Industry.
- Çil, E., Maccario, N., & Yanmaz, D. (2016). Design, implementation and evaluation of innovative science teaching strategies for non-formal learning in a natural history museum. *Research in Science & Technological Education*, 34, 325–341.
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, D. K., & Ball, D. L. (1999). *Instruction, capacity and improvement* (Report No. CPRE-RR-43). Philadelphia, PA: CPRE.
- Cooper, M. M. (2015). Why ask why? *Journal of Chemical Education*, 92, 1273–1279.
- Cox, C. A., & Carpenter, J. R. (1989). Improving attitudes toward teaching science and reducing science anxiety through increasing confidence in science ability in inservice elementary school teachers. *Journal of Elementary Science Education*, 1, 14–34.
- Cunningham, C. M. (2009). Engineering is elementary. *The Bridge*, 30, 11–17.
- Darling-Hammond, L. (2017). Teacher education around the world: What can we learn from international practice? *European Journal of Teacher Education*, 40, 1–19.
- Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, 38, 181–199.

- Desimone, L., Smith, T. M., & Phillips, K. (2013). Linking student achievement growth to professional development participation and changes in instruction: A longitudinal study of elementary students and teachers in title I schools. *Teachers College Record*, 115, 1–46.
- Duran, E., Ballone-Duran, L., Haney, J., & Beltyukova, S. (2009). The impact of a professional development programme integrating informal science education on early childhood teachers' self-efficacy and beliefs about inquiry-based science teaching. *Journal of Elementary Science Education*, 21, 53–70.
- Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement*, 35, 137–154.
- Grenier, R. S. (2010). 'Now this is what I call learning!' A case study of museum-initiated professional development for teachers. *Adult Education Quarterly*, 60, 499–516.
- Gupta, P., & Adams, J. D. (2012). Museum–university partnerships for preservice science education. In *Second international handbook of science education* (pp. 1147–1162). New York: Springer.
- Gupta, P., Adams, J., Kisiel, J., & Dewitt, J. (2010). Examining the complexities of school-museum partnerships. *Cultural Studies of Science Education*, 5, 685–699.
- Guskey, T. R. (2002). Professional development and teacher change. *Teachers and Teaching*, 8, 381–391.
- Guskey, T. R., & Yoon, K. S. (2009). What works in professional development? *Phi Delta Kappan*, 90, 495–500.
- Hammer, D., & Berland, L. K. (2014). Confusing claims for data: A critique of common practices for presenting qualitative research on learning. *Journal of the Learning Sciences*, 23(1), 37–46. <https://doi.org/10.1080/10508406.2013.802652>
- Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, 95(7–8), 798–812. <https://doi.org/10.1016/j.jpubeco.2010.11.009>
- Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M., & Miratrix, L. W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching*, 49, 333–362.
- Heredia, S., & Yu, J. (2015). *Exploratorium teacher institute induction programme: Results and retention*. San Francisco, CA: Exploratorium. Retrieved from <http://www.exploratorium.edu/sites/default/files/pdfs/ExploratoriumTipReport.pdf>
- Hill, H. C., Beisiegel, M., & Jacob, R. (2013). Professional development research: Consensus, crossroads and challenges. *Educational Researcher*, 42, 476–487.
- Holliday, G. M., Lederman, J. S., & Lederman, N. G. (2014). 'Wow! look at that!': Discourse as a means to improve teachers' science content learning in informal science institutions. *Journal of Science Teacher Education*, 25, 935–952.
- Huber, S. G. (2011). The impact of professional development: A theoretical model for empirical research, evaluation, planning and conducting training and development programmes. *Professional Development in Education*, 37, 837–853.
- Human Rights Commission. (2016, October 26). *Collecting transgender inclusive gender data in workplace and other surveys*. Retrieved from <http://www.hrc.org/resources/collecting-transgender-inclusive-gender-data-in-workplace-and-other-surveys>
- James, M., & McCormick, R. (2009). Teachers learning how to learn. *Teaching and Teacher Education*, 25, 973–982.
- Johnson, C. C., & Fargo, J. D. (2010). Urban school reform enabled by transformative professional development: Impact on teacher change and student learning of science. *Urban Education*, 45, 4–29.
- Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research*, 86(4), 945–980. <https://doi.org/10.3102/0034654315626800>
- Kim, S., Walker, M. E., & McHale, F. (2010). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement*, 47(1), 36–53.

- Korur, F., Vargas, R. V., & Serrano, N. T. (2016). Attitude toward science teaching of Spanish and Turkish in-service elementary teachers: Multi-group confirmatory factor analysis. *Eurasia Journal of Mathematics, Science & Technology Education*, 12, 303–320.
- Kyriakides, L., Christoforidou, M., Panayiotou, A., & Creemers, B. P. M. (2017). The impact of a three-year teacher professional development course on quality of teaching: Strengths and limitations of the dynamic approach. *European Journal of Teacher Education*, 40, 1–22.
- Lee, H. S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education*, 24, 115–136.
- Lesage, E., Valcke, M., & Sabbe, E. (2013). Scoring methods for multiple choice assessment in higher education – is it still a matter of number right scoring or negative marking? *Studies in Educational Evaluation*, 39, 188–193.
- Luft, J. A., & Hewson, P. W. (2014). Research on teacher professional development programmes in science. *Handbook of Research in Science Education*, 2, 889–909.
- Martin, W., Strother, S., Beglau, M., Bates, L., Reitzes, T., & McMillan Culp, K. (2010). Connecting instructional technology professional development to teacher and student outcomes. *Journal of Research on Technology in Education*, 43(1), 53–74.
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282.
- Melber, L. M., & Cox-Petersen, A. M. (2005). Teacher professional development and informal learning environments: Investigating partnerships and possibilities. *Journal of Science Teacher Education*, 16, 103–120.
- Musset, P. (2010). *Initial teacher education and continuing training policies in a comparative perspective* (OECD Education Working Papers, 48, 0–1).
- Nadeau, P. A., Ebel, D. S., Harlow, G. E., Landman, N. H., Pagnotta, A., Sessa, J., ... Shumer, M. (2013, December). Putting teachers-to-be in the field and the lab: Hands-on research at the American Museum of Natural History. In *AGU Fall meeting abstracts*.
- Nadelson, L. S., Seifert, A., Moll, A. J., & Coats, B. (2012). i-STEM summer institute: An integrated approach to teacher professional development in STEM. *Journal of STEM Education: Innovations and Research*, 13, 69.
- National Academies of Sciences, Engineering and Medicine. (2015). *Science teachers' learning: Enhancing opportunities, creating supportive contexts*. Washington, DC: National Academies Press. Retrieved from <http://nap.edu/21836>
- National Research Council. (2010). *Preparing teachers: Building evidence for sound policy*. Washington, DC: National Academies Press.
- Osborne, J., Simon, S., Christodoulou, A., Howell-Richardson, C., & Richardson, K. (2013). Learning to argue: A study of four schools and their attempt to develop the use of argumentation as a common instructional practice and its impact on students. *Journal of Research in Science Teaching*, 50, 315–347.
- Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, 25, 1049–1079.
- Penuel, W. R., Gallagher, L. P., & Moorthy, S. (2011). Preparing teachers to design sequences of instruction in earth systems science: A comparison of three professional development programmes. *American Educational Research Journal*, 48, 996–1025.
- Phillips, M., Finkelstein, D., & Wever-Frerichs, S. (2007). School site to museum floor: How informal science institutions work with schools. *International Journal of Science Education*, 29, 1489–1507.
- Postholm, M. B. (2012). Teachers' professional development: A theoretical review. *Educational Research*, 54, 405–429.
- Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher*, 29, 4–15.
- Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, 52, 354–379.

- Riegle-Crumb, C., Morton, K., Moore, C., Chimonidou, A., Labrake, C., & Kopp, S. (2015). Do inquiring minds have positive attitudes? The science education of preservice elementary teachers. *Science Education*, 99, 819–836.
- Ross, J., & Bruce, C. (2007). Professional development effects on teacher efficacy: Results of randomized field trial. *The Journal of Educational Research*, 101, 50–60.
- Ross, J. A., Bruce, C. D., & Hogaboam-Gray, A. (2006). The impact of a professional development program on student achievement in grade 6 mathematics. *Journal of Mathematics Teacher Education*, 9, 551–577.
- Rouweler, M. (2016). *Equipping pre-service teachers to improve science education at primary schools* (Master's thesis). Retrieved from <http://essay.utwente.nl/70952/1/Thesis%20M.J.M.%20Rouweler%202016%20.pdf>
- Sandholtz, J. H., & Ringstaff, C. (2014). Inspiring instructional change in elementary school science: The relationship between enhanced self-efficacy and teacher practices. *Journal of Science Teacher Education*, 25, 729–751.
- Schmidt, W. H., & Cogan, L. S. (2014). *An Investigation of the Museum of Science and Industry, Chicago's 2012-2013 Get Re-Energized Module* (Working Paper #40). East Lansing, MI: Education Policy Centre, Michigan State University.
- Setioko, W., & Irving, K. W. (2017, March 16–17). The changing role of museums in advancing science education. In *Conference proceedings, new perspectives in science education* (p. 666). Florence.
- Stokes, D., Evans, P., & Craig, C. (2017). *Developing STEM teachers through both informal and formal learning experiences*. Ediciones Universidad de Salamanca. Retrieved from <https://repositorio.grial.eu/handle/grial/919>
- Summers, R., & Abd-El-Khalick, F. (2017). Development and validation of an instrument to assess student attitudes toward science across grades 5 through 10. *Journal of Research in Science Teaching*, 55(2), 172–205.
- Tierney, R. D. (2013). Fairness in classroom assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 125–144). Thousand Oaks, CA: Sage.
- Traphagen, K., & Traill, S. (2014). *How cross-sector collaborations are advancing STEM learning*. Los Altos, CA: Noyce Foundation.
- van Aalderen-Smeets, S., & Walma van der Molen, J. (2013). Measuring primary teachers' attitudes toward teaching science: Development of the dimensions of attitude toward science (DAS) instrument. *International Journal of Science Education*, 35(4), 577–600. <https://doi.org/10.1080/09500693.2012.755576>
- van Aalderen-Smeets, S. I., Walma van der Molen, J. H., van Hest, E. G. C., & Poortman, C. (2017). Primary teachers conducting inquiry projects: Effects on attitudes towards teaching science and conducting inquiry. *International Journal of Science Education*, 39, 238–256.
- Van Driel, J. H., Meirink, J. A., van Veen, K., & Zwart, R. C. (2012). Current trends and missing links in studies on teacher professional development in science education: A review of design features and quality of research. *Studies in Science Education*, 48, 129–160.
- Vangrieken, K., Meredith, C., Packer, T., & Kyndt, E. (2017). Teacher communities as a context for professional development: A systematic review. *Teaching and Teacher Education*, 61, 47–59.
- Villegas-Reimers, E. (2003). *Teacher professional development: An international review of the literature*. Paris: International Institute for Educational Planning.
- Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: Motives and methods. *Educational Researcher*, 37, 469–479.
- Whitcomb, J., Borko, H., & Liston, D. (2009). Growing talent: Promising professional development models and practices. *Journal of Teacher Education*, 60, 207–212.
- Wilson, S. M. (2013). Professional development for science teachers. *Science*, 340, 310–313.
- Wunar, B. & Kowrach, N. (2017, March 16–17). The changing role of museums in advancing science education. In *Conference proceedings, new perspectives in science education* (p. 422). Florence.
- Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Report No. REL 2007–No.

033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Centre for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.

Yu, J. C., & Yang, H. J. (2010). Incorporating museum experience into an in-service programme for science and technology teachers in Taiwan. *International Journal of Technology and Design Education*, 20, 417–431.