Ha *et al. Evo Edu Outreach (2015) 8:11*
DOI 10.1186/s12052-015-0040-9

Evolution: Education and Outreach
a SpringerOpen Journal

RESEARCH ARTICLE

Open Access

The Long-Term Impacts of Short-Term Professional Development: Science Teachers and Evolution

Minsu Ha[1], Brian C Baldwin[2] and Ross H Nehm[3,4*]

Abstract

Background: Although a large body of work in science education has established the pervasive problem of science teachers' alternative conceptions about evolution, knowledge deficits, and anti-evolutionary attitudes, only a handful of interventions have explored the mitigation of these issues using professional development (PD) workshops, and not a single study to our knowledge has investigated if positive outcomes are sustained long after program completion. The central aim of our study was to investigate the long-term consequences of an intensive, short-term professional development program on teachers' knowledge of evolution, acceptance of evolution, and knowledge of the nature of science (NOS).

Methods: Program efficacy was examined using a pre-post, delayed post-test design linked to quantitative measures of teacher knowledge, performance (explanatory competence), and acceptance using published instruments shown to generate reliable and valid inferences.

Results: Our study is the first to report sustained large effect sizes for both knowledge of evolution, NOS, and acceptance change ~1.5 years after program completion. Concordant with other measures, teacher self-reports indicated that the PD program had lasting effects.

Conclusions: Our study suggests that short-term PD built using specific research-based principles can have lasting impacts on teachers' evolutionary knowledge and acceptance. Because evidence of sustained knowledge and belief change is prerequisite to downstream classroom studies (e.g., impacts on student learning), retention of evolutionary knowledge improvements and acceptance change emerge as central, but previously unstudied, components of teacher evolution PD.

Keywords: In-service teachers, Professional development, Evolution, Assessment, Knowledge and belief retention

Background

Evolution is universally recognized as a "core idea" in the life sciences by scientific and educational organizations (e.g., NSTA, NARST, AAAS), policy and standards documents (e.g., AAAS 1999, NRC 2012), and most US state standards. Although many policy and standards documents have attempted to bolster support for the teaching of evolution in schools, such efforts have little value if educators do not follow them; that is, if teachers refuse to teach evolution, project antipathy or ambivalence towards evolution to their students, or directly teach non-scientific alternatives to evolution (Moore 2002; Sickel and Friedrichsen 2013). Further diminishing the potential impacts of these important policy documents, many teachers who embrace state or national standards nevertheless have a weak grasp of evolutionary concepts, are partial to common alternative conceptions about the nature of science (NOS), natural selection, and/or earth science, or are not positioned to suppress community and cultural pressures to teach non-scientific alternatives to evolution (Berkman and Plutzer 2011; Brem et al. 2003; Gregory 2009; Nehm and Schonfeld 2007; Sickel

*Correspondence: ross.nehm@stonybrook.edu
[3] Center for Science and Mathematics Education, Stony Brook University (SUNY), 092 Life Sciences Building, Stony Brook, NY 11794, USA
Full list of author information is available at the end of the article

and Friedrichsen 2013). Teacher education and professional development about evolution are widely recognized as the critical link between the goals enumerated in numerous policy documents and standards on the one hand, and persistent public confusion and disbelief about this core scientific concept on the other (Berkman and Plutzer 2011; Nehm and Schonfeld 2007; Sickel and Friedrichsen 2013).

Evolution Professional Development

Although a very large body of work has focused on science teachers' knowledge, alternative conceptions, acceptance of evolution, religious conflict with teaching evolution, and the relationships among these variables (see Smith 2010; Sickel and Friedrichsen 2013, for excellent reviews), only a small number of intervention studies attempting to address these issues occur in the peer-reviewed literature (see Table 1). Scharmann and Harris (1992), for example, documented several important findings in a study of a 3-week teacher PD program: improved understanding of NOS, evolutionary principles, and acceptance of evolution; and reduced anxiety about teaching evolution in secondary science classrooms. In a follow-up study (Scharmann and Harris 1992) with a subset of participants (less than half of the original sample), no significant changes were found in these outcomes. Scharmann (1994) also demonstrated that a 2-week summer institute focusing on evolution helped science teachers to improve their acceptance of evolution and understanding of NOS. As in his previous study with Harris, Scharmann (1994) found that the summer institute reduced participants' self-reported anxieties about teaching evolution. While the results of these studies from nearly 20 years ago were very promising, and provide evidence for the efficacy of short-term interventions, the measures used to substantiate learning gains and acceptance levels have not been widely used in evolution education, making it difficult to align these results with most work in evolution education (see Table 1). Second, it is not clear if the small subset of participants (n = 9) who chose to complete the follow-up study were representative of the entire sample. Finally, the intervention did not appear to directly measure teachers' knowledge of evolution.

Firenze (1997) studied the effects of a 2-week intervention on New York science teachers' knowledge and perceptions of evolution. Firenze's intervention included content lectures, pedagogical practice, group work, laboratory work, field trips, and multi-media presentations. He used interviews, questionnaires, observations, and statistical methods to examine the efficacy of the program. However, Firenze used self-report measures (rather than objective measures) of teachers' knowledge and teaching ability. The self-report data revealed that teachers felt that they had more knowledge, more self-confidence, and more enthusiasm for evolution after completing the 2-week program. In addition, the program appeared to help the teachers find ways to use evolutionary theory as an overarching theme in their classes.

Crawford et al. (2005) examined the efficacy of technology-enhanced instruction on improving prospective science teachers' knowledge of evolution and the nature of science. The qualitative data from their research demonstrated that prospective teachers possessed increased understanding of evolution and NOS despite initial alternative conceptions. No follow-up studies were conducted.

Nehm and Schonfeld (2007), studying teachers from New York, reported that their semester-long intervention with >40 science teachers also produced significant and meaningful gains in teachers' knowledge of evolutionary concepts, reductions in common alternative conceptions, and more informed views of NOS. The authors found that, while the intervention had a significant impact upon knowledge and a reduction in alternative conceptions, there was only a modest improvement in overall evolutionary acceptance levels. At the end of the study, many teachers who improved their knowledge remained partial to teaching both evolution and creationism. No follow-up studies were conducted.

More recently, Nadelson and Sinatra (2010) conducted a very short-term experimental intervention (lasting <2 h) to examine whether improved understanding of situations of chance and NOS could produce positive learning outcomes. The experimental and control group received an online tutorial about common alternative conceptions of evolution and NOS; the experimental group also received an online tutorial about uncertainty in the context of evolution; and the control group received an online tutorial about the life and travels of Charles Darwin. Overall, Nadelson and Sinatra found that their intervention failed to improve teachers' knowledge of evolution, understanding of NOS, and situations of uncertainty relating to evolution, but it did improve participants' acceptance of evolution. However, there was not a significant difference in improving acceptance of evolution between the experimental and control groups. The durability of these changes were not examined.

Southerland and Nadelson (2012) recently examined teacher learning of evolution, including a focus on macroevolution. The 15-week course focused on affective factors (e.g., religious, emotional, and political dimensions (weeks 1–5), understanding both micro- and macroevolution (weeks 6–12), and connections between the two (weeks 13–15). Learners were provided various types

Table 1 Selected intervention studies with science teacher participants from the literature

Study	Location	n	Intervention duration	Measurement			Replicated
				Evolution knowledge	Evolution acceptance	NOS understanding	
Scharmann and Harris (1992)	Midwestern USA	19	90 h	10-item statement, 5-point Likert-type	5-item statement, 5-point Likert-type (Johnson and Peeples 1987)	29-item statement, 3-point Likert-type (Kimball 1967); 20-item statement, 5-point Likert-type (Johnson and Peeples 1987)	YES (Scharmann 1994)
Scharmann (1994)	Midwestern USA	24	84 h	10-item statement, 5-point Likert-type	5-item, 5-point Likert-type (Johnson and Peeples 1987)	29-item, 3-point Likert-type (Kimball 1967); 20-item, 5-point Likert-type (Johnson and Peeples 1987)	NO
Firenze (1997)	Northeastern USA	14	64 h	1-item, self-report Likert-type	n/a	n/a	NO
Crawford et al. (2005)	Northeastern USA	21	12 h	Open-response instrument (Bishop and Anderson 1990)	n/a	VNOS (Lederman et al. 2002)	NO
Nehm and Schonfeld (2007)	Northeastern USA	44	14 weeks	Open-response instrument (Bishop and Anderson 1990) and evolution content knowledge test (Nehm and Schonfeld 2007)	1 multiple-choice item	9-item, Likert-type (Nehm and Schonfeld 2007)	YES (unpublished)
Nadelson and Sinatra (2010)	n/a	89[a]	~1 h[b]	Concept inventory of natural selection (Anderson et al. 2002)	Measure of acceptance of the theory of evolution (Rutledge and Warden 1999)	Scientific attitude inventory II (Moore and Foy 1997)	NO
Southerland and Nadelson (2012)	n/a	14	15 weeks	CINS MUM	MATE I-SEA	VNOS	NO

[a] 44 for experimental group and 45 for control group.

[b] 1.5 h consisted of both instruction and pre-/post-testing.

of measures related to evolution, such as the MATE, VNOS, CINS, I-SEA, and MUM to self-evaluate their own achievements throughout the course (that is, the assessments were part of the course instruction). The results from their intervention (with 14 pre- and in-service biology teachers) showed that 85.7% improved their understanding of microevolution, 78.6% improved their understanding of macroevolution, 64.3% improved their acceptance of evolution, 71.4% improved their acceptance of human evolution, and 92.9% acquired a more informed understanding of NOS. As with many of the previous studies, no delayed post-tests were administered.

Building upon this small body of work on teacher evolution education is challenging for several reasons. First, in cases where the same general topics were conducted (e.g., NOS), different instruments were used, and the descriptions of the instruments suggest that they may not measure the same facets of the construct of interest (e.g., the VNOS vs. the ENOS vs. the Kimball test vs. the Scientific Attitude Survey II; Table 1). Second, some of the studies used recall-based assessments to measure aspects of teacher understanding (e.g., Crawford et al. 2005) whereas other studies used recognition-based multiple-choice assessments (e.g., Nadelson and Sinatra 2010); these two types of measures capture different aspects of knowledge (Opfer et al. 2012). Third, the duration (e.g., less than 2 h vs. one semester) and teacher sample sizes (e.g., 14 vs. 89) varied greatly across studies. Fourth, some of the measures that were used in the studies indirectly measured impact (e.g., *self-perceptions* of change: Firenze 1997; *perceptions* of NOS: Nadelson and Sinatra 2010) whereas other studies directly measured change (e.g., NOS: Scharmann 1994). Fifth, different types of research approaches characterized this body of work (e.g., mixed-methods vs. quantitative, comparison group vs. pre-post only). In sum, although the empirical work on teacher education related to evolution makes generalizations difficult because of different: (1) instructional foci; (2) measures of efficacy; (3) durations; (4) sample sizes and characteristics (pre-service vs. in-service); and (5) intervention types and study designs, many promising findings have emerged.

In addition to methodological differences among the evolution intervention studies that have been published, consistent outcomes (in cases where the same topics were targeted) were not found. For example, some studies reported that instructional treatments did not meaningfully affect teacher evolution acceptance levels, even in cases in which large learning gains were found (e.g., Nehm and Schonfeld 2007), whereas others have demonstrated significant changes in acceptance but not knowledge (e.g., Nadelson and Sinatra 2010). Thus, it remains to be established: (1) what outcomes—knowledge change, acceptance change, both, or neither—should be anticipated or considered attainable in PD interventions; (2) how much time [e.g., ~1 h in Nadelson and Sinatra (2010) vs. ~84 h in Scharmann and Harris (1992)] is realistically needed to achieve meaningful change (i.e., large effect sizes); and (3) whether any interventions have effects that persist beyond the last day of the intervention program (i.e., durability).

In summary, the small body of work on teacher education interventions relating to evolution is very heterogeneous in almost every aspect, and, perhaps as a result of such heterogeneity, the only consistent finding is that statistically significant changes can be observed immediately after some instructional targets (e.g., NOS, acceptance). Simply put, we do not know if evolution-focused teacher education classes or PD will typically produce significant changes in knowledge or acceptance, by how much, or why. More importantly, we have no robust empirical evidence that any of the studies we reviewed had a lasting impact (>1 year) on teacher knowledge or acceptance levels (notably, however, Scharmann and Harris did report sustained change in a small subset of 9 teachers from their original sample). The lack of a robust evidence base on knowledge and belief retention in particular motivated our current work on teacher professional development relating to evolution; if positive impacts are temporary, they are unlikely to impact student learning in the long run.

Knowledge and Belief Retention

It is well established that human memory is limited, and declarative and procedural knowledge tends to degrade over time (Custers 2010; Ebbinghaus 1966; Semb and Ellis 1994). For this reason, it is to be expected that educational learning gains and belief changes will not be retained indefinitely. This issue becomes particularly relevant for teacher education, as stakeholders need to not only know whether science teachers' content knowledge, beliefs, and pedagogical practices are acceptable at the time of graduation or certification, but also the relative durability of this knowledge over the course of professional practice. Much like declarative knowledge decay over time, Franke et al. (2001) noted that teachers' practices could also decline over time. Hewson (2007, p. 1187) likewise pointed out the lack of durability of teachers' actions during the transition from PD to classroom. Empirical work with NOS in particular has revealed that initial teacher learning gains may not be sustained. Akerson et al. (2006) specifically found that some pre-service teachers regressed to their former (naïve) ideas of NOS even though they initially improved their NOS understanding. Clearly, it is unrealistic to assume that gains achieved through PD will be infinitely sustained. An important but unanswered question is: How durable are knowledge and belief changes achieved in evolution PD?

Other than Scharmann and Harris's (1992) study on PD retention effects, no work to our knowledge has examined the durability of evolution knowledge and belief changes achieved through PD. Nevertheless, there is a fairly large and diverse literature about science learning retention rates in other participant populations that merits attention. Fig. 1 summarizes empirical data on knowledge retention rates for teachers, medical students, and college students over time. Several patterns are apparent: (1) a paucity of quantitative research on teachers' knowledge retention (compared to the number of studies on medical student and undergraduate science learning); (2) occasional decreases of knowledge retention over time for most groups; (3) several studies on medical students illustrating very high knowledge retention rates; and (4) knowledge retention rate varies by domain (e.g., general science vs. biochemistry).

While empirical research has shown that learners often gain knowledge after instruction but sometimes forget what they learned (Fig. 1), the empirical literature on teachers' belief retention is much more ambiguous (see Jones and Carter 2007, Sickel and Friedrichsen 2013). Many important theoretical arguments, however, have been advanced for the importance of "worldviews" on knowledge and belief (see, for example, Cobern 1996; El-Hani and Mortimer 2007; Smith and Siegel 2004 for reviews and discussion of worldviews). In line with theoretical arguments, many empirical studies have shown that acceptance change may be more difficult to achieve than knowledge gain. For example, Nehm and Schonfeld (2007) reported that the vast majority of science teachers who participated in a 14-week intervention on evolution significantly increased knowledge of evolution and NOS but still preferred teaching some amount of creationism. In contrast, Scharmann and Harris (1992) and Scharmann (1994) reported that a short-term teacher PD program on evolution and NOS improved teachers' acceptance of evolution. Despite the lack of a large research base on intervention effects on evolutionary acceptance, research has shown that beliefs are an

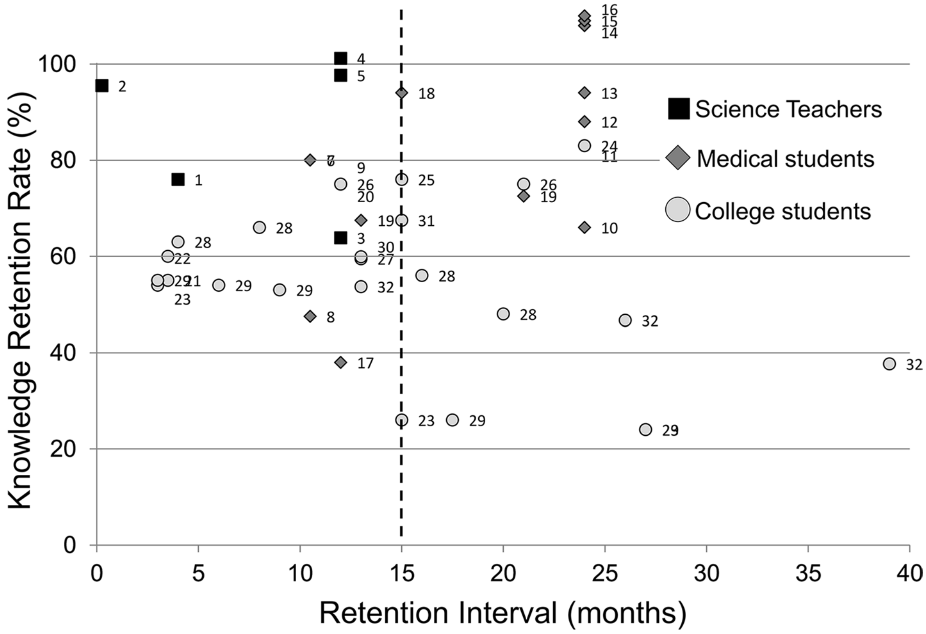

Fig. 1 Review of knowledge retention patterns in the literature (citations from Custers 2010; Semb and Ellis 1994, NCOSP, personal communication; full references in this legend can be found in the Custers (2010) and Semb and Ellis (1994)). The dashed line represents the time point at which our study examined teacher knowledge and belief retention. Knowledge Retention Rates (KRR) were calculated using the equation KRR = Delayed post-test score/post-test score *100. *1* Educational psychology (McDougall 1958), *2* Education (Zimmer 1985), *3* Physical science (D. Hanley, personal communication), *4* Life science (NCOSP, personal communication), *5* Earth science (NCOSP, personal communication), *6* Physiology (D'Eon 2006), *7* Immunology (D'Eon 2006), *8* Neuroanatomy (D'Eon 2006), *9* Behavioral Sciences (D'Eon 2006), *10* Biochemistry (Kennedy et al. 1981), *11* Anatomy (Kennedy et al. 1981), *12* Microbiology (Kennedy et al. 1981), *13* Physiology (Kennedy et al. 1981), *14* Bahavioral sciences (Kennedy et al. 1981), *15* Pathology (Kennedy et al. 1981), *16* Pharmacology (Kennedy et al. 1981), *17* Neurosciences (Levine and Forman 1973), *18* Science (Swanson et al. 1996), *19* Biology (Blunt and Blizard 1975), *20* Zoology (Cederstrom 1930), *21* Zoology (Greene 1931), *22* Psychology (Greene 1931), *23* Botany (Johnson 1930), *24* Biology (Kastrinos 1965), *25* Zoology (Tyler 1933), *26* Anatomy (Blunt and Blizard 1975), *27* Zoology (Cederstrom1930), *28* Zoology (Greene 1931), *29* Botany (Johnson 1930), *30* Science (Landauer and Ainslie 1975), *31* Zoology (Tyler 1933), *32* Zoology (Wert 1937).

important predictor of instructional emphasis on evolution (e.g., Berkman and Plutzer 2011; Moore 2007, 2008; Nehm and Schonfeld 2007). In sum, durability of knowledge and belief change is a centrally important—but under-researched—aspect of research on science teachers and evolution.

Research Design and Questions

Building upon findings from the intervention studies reviewed above, and employing the most widely used and robust measurement instruments available, we developed a research-based, short-term teacher professional development program about evolution and the nature of science for New Jersey (USA) teachers.

We investigated four overarching research questions:

1. What magnitudes (i.e., effect sizes) of evolution knowledge and acceptance change can be achieved in a short-term teacher professional development (PD) program?
2. How closely associated are learning gains for evolution content knowledge and learning gains for the nature of science (NOS)? Are these learning gains associated with acceptance change?
3. How durable are knowledge and belief changes 15 months after the PD program?
4. What effect did teachers think the PD program had on them?

Science Teacher Professional Development

While many different models of teacher professional development have been proposed in the literature (e.g., Bell and Gilbert 1996; Borko et al. 2008; Desimone 2009; Franke et al. 2001; Joyce and Showers 1988; Loucks-Horsley et al. 1998; Supovitz and Turner 2000), we situate our study within Desimone's (2009) conceptual framework (Fig. 2). In our view, a major strength of this model is that it provides a parsimonious, broadly applicable, birds-eye view of PD, encompassing the core features most salient to the education of teachers and their students. In brief, Desimone's model encompasses "…interactive, non-recursive relationships between the critical features of professional development, teacher knowledge and beliefs, classroom practice, and student outcomes." Desimone's (2009, p. 184–185) 'theory of action' for professional development maps putative causal pathways linking four central features (Fig. 2): that '(1) teachers experience effective professional development; (2) professional development increases teachers' knowledge, skills, attitudes, and beliefs; (3) teachers use their new knowledge, skills, attitudes, and beliefs to improve instruction and/or their approach to pedagogy; and (4) instructional changes foster increased student learning.' While this model is domain-general and has not been tested with causal research designs, it represents a consensus model built on a large body of empirical literature and comparative study findings and may provide a framework for synthesizing a broad array of educational research studies relating to PD (Desimone 2009, p. 185).

Our PD intervention included the five critical features of professional development in Desimone's (2009) model (Fig. 2): content focus (cross-cutting focus on evolution and NOS); active learning (inquiry-based activities); coherence (alignment with NJ State Standards and district goals); duration (the equivalent of one graduate class); and collective participation (student-centered instruction and collaborative learning) (Table 2). These features are in close alignment with what has been

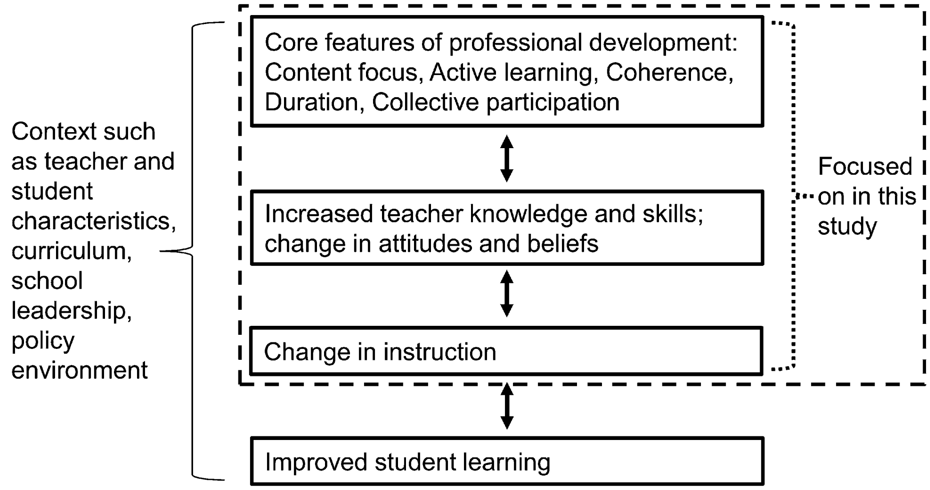
Context such as teacher and student characteristics, curriculum, school leadership, policy environment
Core features of professional development: Content focus, Active learning, Coherence, Duration, Collective participation
Increased teacher knowledge and skills; change in attitudes and beliefs
Change in instruction
Focused on in this study
Improved student learning

Fig. 2 Desimone's (2009) core conceptual framework for professional development. See text for details. Modified from Desimone (2009, p. 185).

Table 2 Overview of the NJ State Standards, intervention topics, activities, and readings used in the teacher professional development workshop

Day	Major topic and NJ Standards	Topics	Activities	Readings and reflections
Day 1	Science practices (Standard 5.1)	Student learning and assessment	Video: Private Universe	McComas (1996)
		Prior knowledge and alternative conceptions	Nature of Science Interviews	Abd-El-Khalick and Lederman (2000)
		Introduction to the nature of science (observation, inference, theory, law, etc.).	Developing NOS formative assessments	NOS alternative conceptions reflection essay
Day 2	Science practices (Standard 5.1)	Nature of science (continued)	Black box activity	Collins and Pinch (1998)
		Scientific models and modeling	Model testing activity	
		Types of experiments in science	Textbook analysis of NOS concepts	
Day 3	Heredity and reproduction (5.3.D)	Causes of variation; heredity patterns	DNA from the beginning (Cold Spring Harbor online activities)	Driver et al. (1994)
		Sexual reproduction and genetics	Genetics formative assessment development	
Day 4	Heredity and reproduction (5.3.D)	DNA, transcription, translation	Linking NOS and genetics lessons	Driver et al. (1994)
		Mendel's laws of heredity and NOS	Case study of the discovery of Huntington's disease	Genetics alternative conceptions reflection essay
		From observable phenotypic patterns to unobservable genetic causes	Genetics formative assessment development (revisions)	
Day 5	Heredity and reproduction (5.3.D)	Genetics and reproduction activities	Piloting genetics formative assessments	Gregory (2009)
		Natural selection key concepts	Pollination and Lily sexual reproduction	
	Evolution and diversity (5.3.E)	Typological thinking and essentialism	Mendels law's in corn cobs	
Day 6	Evolution and diversity (5.3.E)	Natural selection key concepts	M&M natural selection activity	Gregory (2009)
		History of corn domestication	Natural selection formative assessment development	Natural selection alternative conceptions reflective essay
			Snail scales of phenotypic variation	
Day 7	Interdependence (5.3.C)	Plants, animals, energy, and matter	Carnations and water flow	Driver et al. (1994)
		Photosynthesis and respiration	Peace Lily leaves	Photosynthesis alternative conceptions reflective essay
		Transpiration and water relations	Sun and shade leaf patterns	
			Transpiration patterns	
Day 8	Interdependence (5.3.C)	Review and synthesis	Question and discussions about content	Formative assessment revisions

written regarding the design of effective professional development opportunities for science teachers (e.g., Duschl et al. 2007; Loucks-Horsley et al. 1998; Supovitz and Turner 2000) and evolution education (reviewed in Nehm and Schonfeld 2007; Sickel and Friedrichsen 2013).

For in-service teachers, summer professional development programs have been commonly used to augment teachers' knowledge and improve their pedagogical practices. Our PD project differed from such models in that it consisted of a combination of a content-based summer workshop and academic-year, in-school curriculum topic study workshops. The State-funded PD also mandated the participation of many teachers from a few districts—echoing strategies for successful professional development through local systemic change (Banilower et al. 2007).

We developed and implemented a professional development model with a structured emphasis on strengthening NJ science teachers' understanding of the revised State content standards relating to the nature of science and evolution, in line with the school districts' efforts to improve student learning in these content areas. This approach is consistent with Desimone's (2009) notion of "coherence" (see above). The implementation plan also paid careful attention to well-documented strategies for successful PD implementation (Loucks-Horsley et al. 1998): providing ample opportunities for participant teachers to shape the content of their professional development, and creating a forum for teachers to reflect upon their experiences and pedagogical and curricular practices.

Our PD design specifically focused on using structured opportunities for improving teachers' content knowledge and pedagogical content knowledge through the use of workshops. The implementation strategy of using workshops and seminars in professional development has certain assumptions regarding its usefulness: first, external knowledge is viewed as valuable; second, learning outside of the work environment allows in-depth study and practice needed for success; and third, one size can fit all (Loucks-Horsley et al. 1998). In order to translate the broad topics of content, active learning, coherence, and collective participation, pedagogical and content experts (education and biology faculty) from both the home institution and another university designed 2 weeks of all-day workshops that focused on the development of the following core content areas: Science practices/the nature of science; Organization and development; Interdependence; Heredity and reproduction; and Evolution and diversity. Evolution and the nature of science were central, cross-cutting themes throughout the entire workshop.

Finally, the PD program was carefully designed using lessons learned and key strategies suggested in the science education literature, including: (1) a strong content focus (Desimone 2009; Supovitz and Turner 2000; Sickel and Friedrichsen 2013); (2) explicit discussions of the nature of science (NOS) and associated NOS alternative conceptions (Scharmann and Harris 1992; Sickel and Friedrichsen 2013); (3) reflective consideration of the relationships between science and religion (Smith and Scharmann 2008); (4) inquiry-based instruction in evolutionary concepts (Desantis 2009); and (5) explicit engagement with common alternative conceptions about natural selection and evolution using inquiry activities (Nehm and Schonfeld 2007).

It is important to mention that other professional development models (such as Loucks-Horsley et al. 1998) have similar frameworks in terms of common goals and characteristics (such as: shared understanding of effective classroom pedagogy; opportunities for teachers to learn content in a similar manner as their students; and, continuous reflection and assessment of professional development). Our model of PD incorporates these common principles of successful development, and so it is likely that our design could also be successful within one of the other well-accepted paradigms of science professional development.

Participants and Methods

Sample

Participants in the study voluntarily enrolled in a 10-day intensive (9 am–5 pm) teacher PD program in New Jersey (8 days involved interactive instruction, and 2 days involved collaborative project and assignment work). Participants totaled 28 teachers, including eighteen elementary school teachers (i.e., K-6) and 10 secondary science teachers (i.e., 7–12). The majority of participants (24) were female. They all reported that they took science courses in college, and that they taught science-related courses in the State of New Jersey. All participants remained in the teaching profession during the 1.5 years of our study.

We were concerned that elementary and secondary science teachers might differ in core background variables given differences in, for example, pre-service teacher preparation. However, we did not find significant or meaningful differences in pre-test scores for these two groups ($p > 0.05$ for pre-CINS, pre-misconception scores, pre-MATE scores and $p > 0.01$ for pre-KC scores). Given similar levels of prior knowledge and acceptance of evolution, we combined both elementary and secondary teachers into one sample in our analyses.

Study Design

We used a pre-post-delayed post-test, mixed-methods research design to investigate the impact of the professional development program on in-service science teachers' content knowledge and acceptance of evolution. Given the nature of the State-sponsored Math Science Partnership (MSP) summer professional development program, we were not able to establish a comparison group or randomize subjects to treatment and control conditions. Consequently, our study design prohibits making causal claims but does provide rich information on associations between the intervention and knowledge and acceptance change. In order to measure putative knowledge and attitude change, we employed several instruments that have been shown to generate reliable and valid inferences in comparable populations (see below). The instruments were administered at the beginning and end of the professional development program. Nearly all teachers completed the instruments (but completion rates differed slightly among instruments).

Evolution Knowledge Measures: Multiple-Choice

As we noted above, the majority of studies in the science education literature have used paper and pencil instruments to measure students' and teachers' knowledge of evolution. In order to provide robust measures of teacher knowledge, our study used two different instruments, with different formats. Importantly, both instruments have been subjected to reliability and validity evaluation. The first instrument we used is the multiple-choice Conceptual Inventory of Natural Selection (CINS). The CINS was developed to measure 10 evolutionary concepts using 20 items (Anderson et al. 2002). Despite displaying some psychometric problems at a fine-grained level (Battisti et al. 2010; Nehm and Schonfeld 2010), the CINS is generally recognized as a tool that generates valid inferences about overall evolutionary knowledge. Each item of the CINS has one correct response and several alternative conception distractors. The total score of the CINS instrument therefore ranges from 0 to 20. We employed all 20 CINS items in the pre- and post-test while we employed a part of the CINS test (8 items) for the delayed post-test. This was done in order to minimize the length of the test and maximize participation completion rates.

The original CINS paper suggests that the instrument was designed to measure knowledge of natural selection, but additional concepts are present, such as speciation. Many authors consider speciation to be a macroevolutionary concept (Futuyma 2009). For this reason, we consider the CINS to test both microevolutionary and macroevolutionary knowledge. For our sample, the reliability of the CINS (measured using Cronbach's alpha) was (pre-test) 0.84, (post-test): 0.88, and (delayed post-test): 0.61. The lower values for the delayed post-test are not surprising given that fewer items were used.

The twenty items in CINS consist of 10 natural selection concepts (biotic potential, natural resources, population stability, change in a population, limited survival, origin of variation, variation inheritable, origin of species, variation within a population, and differential survival) using three different contexts (the evolution of finches, guppies, and lizards). The three different contextual parts of CINS each cover a different number of concepts (finches: 8 concepts, guppies: 5 concepts, and lizards: 7 concepts). In addition, because the 'lizards' part consists of both 'origin of variation' and 'variation within a population', which are both part of the concept of variation, the actual number of concepts that the "lizard" section covers is six. Therefore, the construct coverage of the Finches section (8 different concepts) is the most complete section of the three-part CINS. To test the reliability of the abbreviated CINS (the 'finches' section: items 1–8 of the CINS), we administered the CINS to 46 pre-service teachers and advanced majors and found very strong correlation coefficients between the score of 'finches' part (1–8) and the score of 'guppies + lizards' part ($r = 0.851$, 95% CI 0.745–0.915). The correlation coefficient between the 'finches' section and the total CINS was very robust (0.932).

In addition to the reliability of the abbreviated CINS, we also examined the reliability of CINS for pre-service science teachers. Several empirical studies used the CINS to measure pre-service and in-service K-12 teachers' knowledge of natural selection (Ha et al. 2012; Nadelson and Sinatra 2009, 2010). Ha et al. (2012) used the CINS to measure 124 biology pre-service teachers' natural selection knowledge and found an acceptable reliability (Cronbach's alpha: 0.737). In addition, this study showed that the CINS was able to discriminate biology pre-service teachers' academic levels (i.e., years in the program; $F = 3.228$, $p < 0.025$, partial eta squared = 0.075). Nadelson and Sinatra (2010) also used the CINS to measure elementary, middle school, and secondary pre-service teachers. This study also reported acceptable Cronbach's alpha values (0.63) for the CINS. Nadelson and Sinatra (2009) also collected CINS data from educational professionals, 53.7% of whom reported science-teaching experience. The study showed adequate reliability (KR20 = 0.64). Thus, the CINS appears to have generated reliable inferences, and, based on its relationship to academic level, it is seen to have a degree of predictive validity in the context of measuring teachers' knowledge of natural selection concepts.

Evolution Knowledge Measures

The second instrument that we used to measure teachers' knowledge of evolution was the constructed-response ACORNS (Assessment of COntextual Reasoning about Natural Selection) instrument (Nehm et al. 2012). The ACORNS is a modified and expanded version of Bishop and Anderson's (1990) widely used instrument. Detailed studies of validity and reliability inferences were reported in the original studies (Nehm et al. 2012). We used four isomorphic ACORNS items differing in surface features (e.g., animal vs. plant, trait gain vs. trait loss) to assess teacher knowledge in the pre- and post-test. For the delayed post-test, we used two ACORNS items (one trait gain, one trait loss).

The format of the ACORNS items was: "A species of X [plants or animals] lacks Y. How would biologists explain how a species of X [with or without] Y evolved from an ancestral X species [without or with] Y?" (X = Snail/Rose/Penguin/Elm and Y = poison/thorns/flight/winged seeds). The ACORNS is considered to be a test of both microevolutionary and macroevolutionary knowledge because it asks students to explain the mechanisms that caused between-species (i.e., macroevolutionary) trait change. ACORNS responses were scored using the rubrics of Nehm et al. (2010). The rubrics include detailed scoring information for seven key concepts (accurate ideas) and six alternative conceptions or naïve ideas. Key Concept (KC) scores for each item ranged from 0 to 7, and alternative conception scores ranged from 0 to 6. Thus, the maximum KC score across all four items was 28, and the maximum alternative conception score was 24.

Key concept scores and alternative conception scores were also used to determine teachers' knowledge coherence, which refers to the consistency of explanatory concept use across items (in our case, the ACORNS items). Coherence of knowledge represents the stability of a concept across contexts (Kampourakis and Zogza 2009). KC and alternative conception scores were also used to categorize teachers' reasoning patterns as: explicitly scientific explanations, mixed explanations (naïve + scientific), explicitly naïve, or no clearly articulated or comprehensible idea (e.g., the teacher answered by rephrasing the question).

Two raters independently scored the ACORNS responses; one was a Ph.D. student in biology education and the other was a biologist. Kappa values for inter-rater agreement were >0.8 for all KCs and alternative conceptions. Consensus scores were established in all cases of disagreement. Reliabilities for the ACORNS were measured using Cronbach's alpha. Strong reliabilities were found for KCs (0.865 pre-test and 0.853 post-test) but weaker reliabilities were found for alternative conceptions (0.557 pre-test and 0.383 post-test). The reliabilities for the delayed post-test were 0.72 for KCs and 0.76 for alternative conceptions. Other studies have also noted that alternative conceptions are very context dependent (Nehm and Ridgway 2011; Nehm and Ha 2011), which explains the low internal reliability measures for alternative conceptions in our sample; students and teachers use different alternative conceptions depending on the item features, and so reliabilities of items designed to elicit different types of alternative conceptions in the same individual are not expected to be high. Given the reliability information in this study, our analyses used the sum of the key concept scores across the four items but analyzed the four alternative conception scores individually.

Measures of Evolution Acceptance

We measured teachers' acceptance of evolutionary theory using the MATE (Measure of the Acceptance of the Theory of Evolution) (Rutledge and Warden 1999). The MATE consists of 20 items covering six concepts, which include the process of evolution, the scientific validity of evolutionary theory, and the nature of science. The MATE instrument items are on a five-point Likert scale (strongly agree to strongly disagree). MATE scores were transformed into a 100-point scale so that we could compare our results with previously published studies. Prior studies have used the MATE on elementary, middle school, and secondary science teachers. Nadelson and Sinatra (2010), for example, used the MATE on a mixed sample of elementary, middle, and high school teachers (about evenly split between early- and late-grade levels). Using this sample, Nadelson and Sinatra reported that the MATE was capable of measuring acceptance change in response to treatment (a form of predictive validity evidence) and had acceptable reliabilities (Cronbach's alpha >0.6). We used the MATE given that prior validity and reliability evidence had been established using a sample very similar to our own. Specifically, we used Cronbach's alpha to measure internal reliability for the MATE and found values of 0.865 (pre) and 0.928 (post) for the original version and values of 0.825 (pre) and 0.923 (post) for the "scientist" version. We used the original version of the MATE for the delayed post-test. The reliability of the delayed post-test was 0.92. These values are closely aligned with what Rutledge and Sadler (2007) reported in their 2007 study of the MATE ($\alpha = 0.94$), and even stronger than the results of Nadelson and Sinatra (2010). In addition, as is reported below, the measure was sensitive to treatment.

Measures of Targeted NOS Knowledge

We measured teacher knowledge using open-response items from the VNOS-C (Views of Nature of Science form C, Lederman et al. 2002). In particular, we selected VNOS-C items 1, 3, 5, and 6 for the pre- and post-test, and items 3 and 5 for the delayed post-test. We used these particular items because they were most closely aligned with the content that was addressed in the workshop (See the Supplementary Materials for the exact wording of these VNOS items and scoring). Therefore, it must be noted that our targeted NOS knowledge measure (hereafter TA-NOS) does not necessarily reflect participant understanding of the *entire* construct of NOS (see Lederman et al. 2002). Two raters independently scored the four VNOS-C responses; one was a Ph.D. student in science education and the other was a science educator. Kappa values for inter-rater agreement were >0.8 for all four VNOS items. Consensus scores were established in all cases of disagreement. Reliabilities for the four VNOS items were measured using Cronbach's alpha. We found weak internal consistencies of the four VNOS items in both pre- and post-test (0.42 pre-test and 0.54 post-test). The reliability of the delayed post-test, however, was a more robust 0.88.

Self-perceptions of change measures

In the delayed post-test, 15 months after the PD workshop, we asked teachers to self-assess the degree to which their knowledge and acceptance changed, and the impact of the PD on their classroom practices. The participants were given five Thurston-scale items: (1) How much of the biology/evolution content covered in the summer workshop do you still remember? (2) How much of the targeted aspects of the nature of science (NOS) content covered in the summer workshop do you still remember? (3) To what extent did the summer workshop influence how you teach about biology/evolution? (4) To what extent did the summer workshop influence how you teach about the targeted aspects of nature of science (NOS)? (5) To what extent did the summer workshop influence your current attitudes toward evolution? The scales had answer options of "completely," "very much," "somewhat," "a little," and "not at all." The internal reliability (Cronbach's alpha) of the five self-perception items was 0.804 and these measures generally aligned with external measures of PD effects (see below).

Measures of Learning Gain Scores

Many studies in science education have begun to use learning gain scores to quantify the magnitude of change in instructional interventions (e.g., Hake 1998). We converted pre- and post-test scores into learning gain scores. We used both 'absolute learning gain scores' and 'normalized learning gain scores' because we were concerned about the impact of ceiling effects in normalized learning gain score calculations (e.g., pre-test scores constrained possible gains; see also Bao 2006). Absolute learning gain scores were converted by subtracting raw pre-test scores from raw post-test scores. Normalized learning gain scores, on the other hand, were converted by dividing the actual gains (post-test percentage scores − pre-test percentage scores) by the potential gains (100% − pre-test percentage scores). We compared teachers' learning gain scores among measured knowledge and acceptance variables (see above).

Imputing Missing Data and Statistical Analyses

Prior to performing our statistical analyses, item-level missing data was filled in with imputed values using multiple linear regressions on the cases without missing data. This involved imputations for 4.8% of the 336 values of variables across the 28 participants in the pre- and post-test design time points. Given that all six variables at each time point are highly correlated to each other, the multiple linear regression method to predict missing data using present data should be highly effective (Allison 2002). There was no item-level missing data in the delayed post-test.

Statistical tests to examine the efficacy of the PD program were performed on both the individual response measures (CINS, ACORNS-KC, ACORNS-MIS, MATE-P, MATE-S, TA-VNOS) and on a global measure using dimensionality reduction techniques. For the statistical tests to measure the change of individual variables between pre- and post-test, and pre-, post-, and delayed post-test, either repeated measures ANOVAs (using CINS, MATE-P, and MATE-S) or non-parametric Wilcoxon Signed Rank tests (ACORNS-KC, ACORNS-MIS, and TA-VNOS) were performed in accordance with the nature of the normality of the data. The effect size of each test was calculated using partial eta squared and Cohen's d (Lomax 2007).

To examine the overall change of all variables through the intervention, we used Categorical Principal Components Analysis (CATPCA) that allows for reduction of a set of variables (including both quantitative/continuous and categorical/ordinal variables) and provides component scores that can then be used in standard linear models (Linting et al. 2007). We first combined the key concept scores and alternative conception scores relating to the written explanation evolutionary *trait gain* items and the two written explanation *trait loss* items. Thus, the pre- and post-test data included three quantitative variables (CINS, MATE-P, and MATE-S) and eight categorical variables (ACORNS-KC-gain, ACORNS-KC-loss, ACORNS-MIS-gain, ACORNS-MIS-loss, and the four scores of TA-NOS items); the delayed post-test data also contained two quantitative variables (CINS and MATE-P) and six categorical variables (two key concept scores, two alternative conception or "misconception" [MIS] scores and two TA-NOS scores).

We used CATPCA to produce the loading scores based on the pre-PD variables only and then applied those same scores to the post-PD variables; this allows us to examine changes in the same construct from the pre- to post-PD situations. This technique also avoids creating a bias toward finding significance in the efficacy analysis. These steps were repeated with the pre-/post-/delayed post-test data set (n = 20), which was analyzed separately from the pre-versus post-data set (n = 28). In particular, loading scores from the 11 initial pre-PD variables in data set one and the eight initial variables for data set two were applied to standardized versions of the variables at the other time points to create component scores that were used in the final analyses. These analyses of the overall efficacy of the PD program compared the change of component scores between pre- and post-test, and among pre-, post-, and delayed post-tests using repeated-measure MANOVA models. In pairwise tests, Bonferroni corrections were made to account for multiple tests.

The Cronbach's alpha for the two-dimensional CATPCA model for the eleven variables in data set one (n = 28) was 0.941 (dimension 1: 0.874 and dimension 2: 0.561). The variance accounted for by the two-dimensional model was 62.9% (dimension 1: 44.3% and dimension 2: 18.6%). The pre-, post-, and delayed post-test data (n = 20) contained eight variables (e.g., CINS, MATE-P, two KC score, two MIS score, and two VNOS). The eight variables in this data set were loaded into the two-dimensional model of CATPCA and produced a Cronbach's alpha of 0.930 (dimension 1: 0.805 and dimension 2: 0.569). The variance accounted for by this two-dimensional model was 67.2% (dimension 1: 42.3% and dimension 2: 24.9%). Note that, unlike traditional PCA, with categorical variables in the mix in CATPCA, it is not possible to reach 100% of the variance explained. In sum, the CATPCA allowed us to consider all of the measures we had in a single analysis; this produces the strongest statistical power.

Finally, Spearman's rho was used to compute the magnitude of association among knowledge and acceptance of evolution and understanding of NOS. All of the statistical analyses were performed in SPSS$^{\text{TM}}$ version 19.

Results

Overall Program Effects

Our first test of program efficacy was performed using a repeated-measure MANOVA. As addressed in the methods section above, the two-dimensional transformed component scores were used as dependent variables. The independent variable of this test was the change between the pre- and post- time points. The repeated-measure MANOVA indicated that the intervention program significantly impacted teachers' knowledge of evolution, acceptance of evolution, and understanding of NOS ($F_{2,\,25} = 42.0$, $p < 0.001$, partial eta squared = 0.77, power = 1.00).

Knowledge and Alternative Conceptions of Evolution

We used the CINS and ACORNS instruments to assess science teachers' evolutionary knowledge. The repeated measures ANOVAs (pre- vs. post-intervention) demonstrated that overall CINS scores increased significantly whereas the ACORNS key concept scores did not change significantly (CINS: pre-test $M = 13.08$, $SD = 4.66$, post-test $M = 16.43$, $SD = 4.08$, $F_{1,\,27} = 25.8$, $p < 0.001$, partial eta squared = 0.49, power = 1.00; ACORNS-KC: pre-test $M = 6.32$, $SD = 4.63$, post-test $M = 7.28$, $SD = 3.77$, $F_{1,\,27} = 2.7$, $\text{p} > 0.05$, partial eta squared = 0.09, power = 0.35). The alternative conception scores for two items (e.g., snail, penguin) decreased significantly between the pre- and post-test (Snail: $z = -2.24$, $p = 0.025$; Penguin: $z = -2.56$, $p = 0.010$). Although we did not find significant changes in total ACORNS key concept scores, we did find significant improvements for two *individual* key concepts: *variability* and *phenotypic/genotypic distribution change* (Variability: $F_{1,\,27} = 14.3$, $p < 0.001$, partial eta squared = 0.35, power = 0.95; Phenotypic/genotypic distribution change: $F_{1,\,27} = 12.2$, $p < 0.01$, partial eta squared = 0.31, power = 0.92). We also found, rather unexpectedly, that the intervention was associated with significant *decreases* in the use of the concept of *limited resources* ($F_{1,\,27} = 25.6$, $p < 0.001$, partial eta squared = 0.49, power = 1.00).

We analyzed teachers' knowledge coherence and structure pre- and post-intervention. In the pre-test, only 14.3% of teachers consistently employed the KC *variation* in their explanations of evolutionary change across the four problem types; in contrast, 66.7% of teachers did so in the post-test. Although the percentage of teachers consistently using the KC *heredity* across contexts was low in the pre-test and in the post-test, twice as many teachers employed heredity in the post-test as in the pre-test (3.6 vs. 7.4%, respectively). The consistent application of the KC *differential survival* did not display sizable differences between the pre- and post-test (21.4 vs. 22.2%, respectively). In terms of participants' knowledge structures, half (50.9%) of the sample used exclusively scientific ideas in their evolutionary explanations in the pre-test, while 83.3% of participants' responses did so in the post-test. The frequencies of teachers' mixed models, explicitly naïve models, and no discernible models, all decreased in the post-test. However, it is important to note that 12.0% of teachers' responses in the post-test still included models comprised of both scientific and naïve ideas.

Some examples of teachers' ACORNS responses (i.e., their explanations of evolutionary change) are provided in the Additional file 1: Table S1. Although some responses did contain alternative conceptions, teachers used different magnitudes of key concepts to build their evolutionary explanations. Although fewer alternative conceptions were present in the post-test, many remained in teachers' responses in the form of "mixed" or "synthetic" models (cf. Nehm and Ha 2011; Vosniadou 2008). For example, participant 23 responded to the 'snail' item of the post-test in the following manner: "The poisonous snail species evolved from his ancestors because of mutation. This new species evolved in order to protect itself and to adapt to its environment to survive". Although participant 23 utilized the important scientific concept of mutation in her response, she also appeared to possess the goal-driven idea of evolution ('in order to'). It is important to note that prior empirical work has shown that in most cases teleological language in ACORNS responses is associated with problematic teleological reasoning (Rector et al. 2013; see Kampourakis 2014 for several chapters debating "acceptable" and "unacceptable" forms of teleological language). Thus, despite learning gains, this teacher appeared to have left the intervention with a mixed mental model of evolutionary change. Participant 9 responded to the 'rose' item in the post-test in the following manner: "If thorns protected roses in an environment that lacked predators, the thorns became less needed over time for the roses to be successful. Some flowers with characteristics (greater sun flower size, etc.) increased in proportion to the original thorny species". Although participant 9 utilized concepts such as limited resources (e.g., predators) and change in a population (e.g., increased in proportion), she still used teleological language in her explanation. As noted, the most prominent alternative conception identified in teachers' responses (both pre- and post-test) was problematic teleological language (i.e. need or goal-driven evolutionary change not associated with selection). Interestingly, teachers' teleological language was often associated with the use of the key concepts *differential survival* or *limited resources* (e.g. 'in order to survive better' or 'in order to protect against predators').

Acceptance of Evolution

We employed two versions of the MATE (teachers' personal acceptance [version P] and teachers' perceptions of scientists' acceptance [version S]) to quantify science teachers' acceptance of evolution before and after the intervention. The repeated measure ANOVA demonstrated that the personal acceptance MATE scores (MATE-P) were significantly less than teachers' perceptions of scientists' acceptance (MATE-S) in the pre-test (MATE-P $M = 79.81$, $SD = 11.98$, MATE-S $M = 84.91$, $SD = 11.71$, $F_{1,\,27} = 10.8$, $p < 0.01$, partial eta squared = 0.29, power = 0.89) whereas they did not differ significantly in the post-test (MATE-P $M = 85.30$, $SD = 12.51$, MATE-S $M = 87.75$, $SD = 11.81$, $F_{1,\,27} = 3.1$, p > 0.05, partial eta squared = 0.10, power = 0.40). Second, the repeated measure ANOVA demonstrated that personal acceptance MATE scores increased significantly post-test, whereas teachers' perceptions of scientists' acceptance did not change (MATE-P: $F_{1,\,27} = 33.4$, $p < 0.001$, partial eta squared = 0.55, power = 1.00; MATE-S: $F_{1,\,27} = 4.7$, $p = 0.04$, partial eta squared = 0.15, power = 0.55).

Targeted Aspects of the Nature of Science

Written examples of teachers' changing notions of TA-NOS for the four categories that we investigated are provided in the Additional file 1: Table S2. Dramatic changes in teachers' understanding of the targeted aspects of nature of science are apparent in the written responses; differences were also detected using statistical analysis (e.g., Wilcoxon Signed Ranks Test). Specifically, teachers' scores (naïve = 0, partially informed = 1, informed = 2) for all four TA-NOS items displayed significant increases from pre- to post-test (TA-NOS 1: z = 2.68, $p < 0.01$; TA-NOS 3: z = 4.05, $p < 0.001$; TA-NOS 5: z = 4.37, $p < 0.001$; TA-NOS 6: z = 3.04, $p < 0.01$). Teachers showed the greatest improvement in TA-NOS item 5 (the difference between scientific theories and laws).

Delayed Post-Test Scores

Figure 3 provides the results of comparisons among pre-, post-, and delayed post-test scores. Two issues are important to keep in mind when viewing these results. First, recall that we used selected item sets from the CINS, ACORNS-KC, and TA-NOS instruments to limit the length of the delayed post-test. Second, only 20 of the 28 teachers participated in the voluntary delayed post-test. Because of these two considerations, we separately compared scores from the pre-and post-test results using the whole sample (see Fig. 3, black line) and the pre-and post-test results for the portion of the sample that participated in the delayed post-test (see Fig. 3, gray line). Figure 3 shows that the patterns of pre- and post-test results, and the delayed post-test results, are almost identical with patterns from the whole sample.

We performed a repeated-measure MANOVA using the component scores of the CATPCA analysis. The results indicated that the intervention program was associated with significantly and meaningful improvements in teachers' knowledge of evolution, acceptance of evolution, and understanding of NOS ($F_{4,\,16} = 16.8$, $p < 0.001$, partial eta squared = 0.81, power = 1.00). The tests of within-subjects contrasts of the repeated measures MANOVA revealed that no significant decreases in the component scores were noted between the post- and delayed post-test (Dimension 1: $F = 0.01$, $p = 1.000$, partial eta squared = 0.00, power = 0.05; Dimension 2: $F = 2.75$, $p = 0.114$, partial eta squared = 0.13, power = 0.35); in contrast, strongly significant increases were noted between pre- and post-test responses (Dimension 1: $F = 54.13$, $p = 0.000$, partial eta squared = 0.74, power = 1.00; Dimension 2: $F = 38.44$, $p = 0.00$, partial eta squared = 0.70, power = 1.00). The pairwise comparisons using a Bonferroni correction showed the same results.

We also performed repeated measures ANOVA for individual quantitative variables (e.g., CINS, ACORNS-KC, MATE-P). The results for the CINS test revealed a significant increase between pre- and post-test scores, but no significant difference between the post- and delayed post-test scores (pre-post: $F_{1,\,19} = 14.5$, $p = 0.001$, partial eta squared = 0.43, power = 0.95; post-delayed post: $F_{1,\,19} = 0.2$, $p = 0.695$, partial eta squared = 0.01, power = 0.07). Tests of the MATE scores also illustrated a significant increase between the pre- and post-test but no significant difference between the post- and delayed post-test (pre-post:$F_{1,\,19} = 20.0$, $p < 0.001$, partial eta squared = 0.51, power = 0.99; post-delayed post: $F_{1,\,19} = 1.0$, $p = 0.331$, partial eta squared = 0.05, power = 0.16). (The pairwise comparisons using a Bonferroni method revealed the same results). The results of KC scores from the ACORNS items did not display significant differences among the pre-, post-, and delayed post-tests; but the mean values did show an increasing trend. Likewise, Friedman tests and multiple Wilcoxon Signed Ranks tests for differences in ACORNS MIS

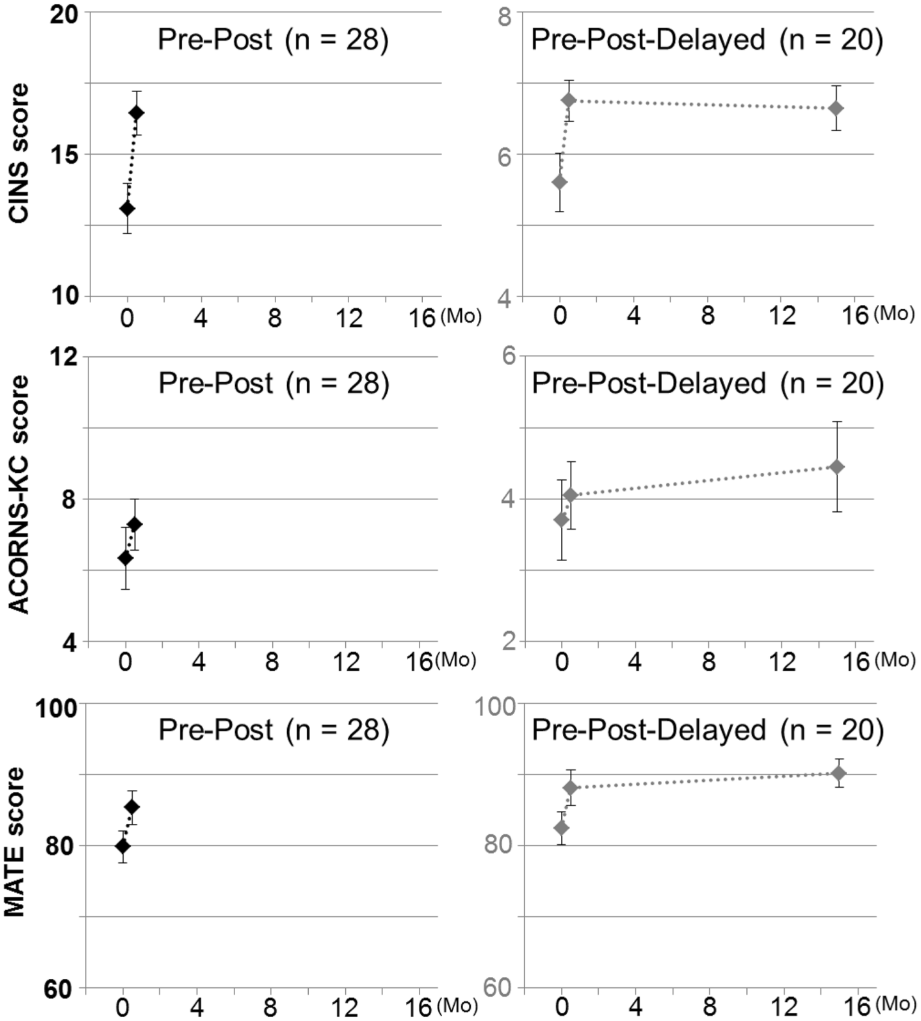

Fig. 3 Retention of knowledge and acceptance change 15 months after the PD intervention. *Left column* (results from complete MATE, complete CINS, and two ACORNS essay items): pre- and post-test mean scores, n = 28 (100% participation). *Right column* (results from complete MATE, partial CINS, and one ACORNS item): pre-, post- and delayed post-test, n = 20 (>70% participation); *Error bars* represent standard errors of the mean. A shorter version of the post-test was used to increase the likelihood that participants would complete the survey. Note that the pre-post results for the full and partial instruments show similar patterns. See text for instrument details and scoring.

scores (e.g. snail and rose items) across pre-, post- and delayed posttest illustrated MIS scores of Rose item for delayed post-test were significantly lower than the scores on the pre-test (Figure S1, Wilcoxon Signed Ranks test for pre- and delayed post-test: z = 2.65, $p = 0.008$).

Friedman tests and Wilcoxon Signed Ranks tests were used to examine differences in TA-NOS. Items 3 and 5 showed a significant increase between the pre- and post-test, and a non-significant decrease between the post- and delayed post-test (Item 3: Chi Square = 26.8, $p = 0.000$; Wilcoxon Signed Ranks test: pre-post, $z = 3.85$, $p = 0.000$; post-delayed post: $z = 1.82$, $p = 0.068$; Item 5: Chi Square = 21.6, $p = 0.000$; Wilcoxon Signed Ranks test, pre-post: $z = 3.72$, $p = 0.000$; post-delayed post: $z = 1.83$, $p = 0.068$).

Self-Perceptions of Change

Fifteen months after the PD workshop, participants were asked to self-report their perceptions of its impact (Fig. 4). Twenty of the twenty-eight participants voluntarily completed the delayed post-test. Among these participants, nearly half (45% for evolution and TA-NOS) reported that they remembered most of the biology/evolution content and the nature of science (TA-NOS) content. In addition, half of participants (45% for evolution and 60% for TA-NOS) reported that the workshop influenced

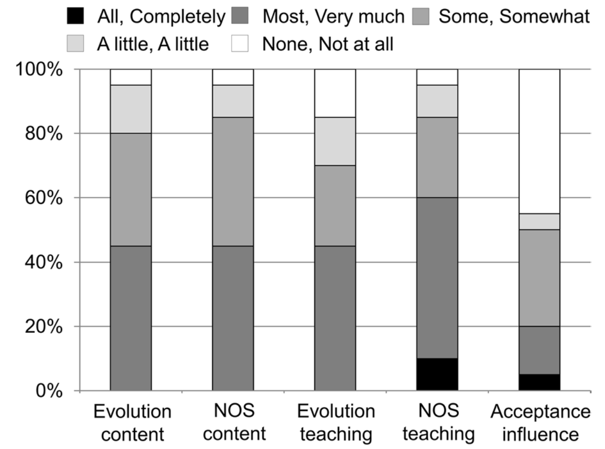

Fig. 4 Teachers' self-perceptions of knowledge/acceptance retention 15 months after the PD intervention. See text for full item descriptions. The Y-axis represents the percentage of participants in each of the Thurstone-scale categories.

how they taught about biology/evolution and TA-NOS (Fig. 4). Despite that the fact that the majority of delayed post-test participants perceived that their knowledge was retained and teaching was influenced by the PD program, 50% of participants thought their attitudes toward evolution were influenced a little or not at all. In contrast, 20% of participants reported that the summer workshop influenced their current attitudes toward evolution. The 50% of participants who thought their attitudes toward evolution were influenced a little or were not influenced displayed similar scores on the MATE instrument across the pre-, post- and delayed post-test. The repeated-measure ANOVA indicated significant increases in MATE scores between the pre- and post-test ($F_{1,7} = 6.6$, $p = 0.037$, partial eta squared = 0.49, power = 0.60) and no significant decreases in MATE scores between the post- and delayed post-test ($F_{1,7} = 0.1$, $p = 0.719$, partial eta squared = 0.02, power = 0.06). Interestingly, we found no significant correlations between self-perceptions of change and objective measures of knowledge and acceptance change (Spearman, $p > 0.05$). Thus, while some participants perceived that their attitudes did not change, the empirical measures contradicted this result to some degree.

Relationships Among Variables

Table 3 illustrates the Spearman's rho correlation coefficients among knowledge and acceptance variables (both pre- and post-intervention; ACORNS-MIS variable was excluded). In general, we found similar associations among variables pre- and post-intervention. First, the correlations between knowledge and acceptance in the pre-tests were higher than the correlations in the post-test (Pre CINS vs. Pre MATE-P: $r_{28} = 0.836$, $p < 0.001$; Pre ACORNS-KC vs. Pre MATE-P: $r_{28} = 0.830$, $p < 0.001$; Post CINS vs. Post MATE-P: $r_{28} = 0.714$, $p < 0.001$; Post ACORNS-KC vs. Post MATE-P: $r_{28} = 0.743$, $p < 0.001$). Second, the correlation between personal acceptance and scientists' acceptance of evolution in the post-test was higher than that the correlation in the pre-test (Pre MATE-P vs. Pre MATE-S: $r_{28} = 0.682$, $p < 0.001$; Post MATE-P vs. Post MATE-S: $r_{28} = 0.781$, $p < 0.001$). Overall, one of the most significant findings to emerge from the correlation analyses was that evolutionary knowledge (measured using the CINS and the ACORNS) and acceptance of evolution (measured using the MATE) were significantly associated, with surprisingly strong *r* values (e.g., 0.836 and 0.830).

Learning Gains

We examined the correlations among absolute learning gain scores and normalized learning gain scores for quantitative measured variables (e.g. CINS, ACORNS-KC, MATE-P, and MATE-S). We found that correlations among absolute learning gain scores for all of the variables (e.g. CINS, ACORNS, and MATE) were not significant ($p > 0.05$). Likewise, correlations among normalized learning gain scores for all of the variables were not significant ($p > 0.05$). Overall, gain scores for most measured variables were not significantly related.

Discussion

The central aim of this study was to investigate the long-term impacts of an intensive, short-term professional development program on teachers' knowledge of evolution, acceptance of evolution, and knowledge of the nature of science (NOS). While numerous studies over the past 50 years have documented different facets of teacher ambivalence or antipathy towards evolution (reviewed in Kim and Nehm 2011), a comparatively smaller body of empirical work in science education has involved interventions attempting to mitigate this core challenge (Table 1). Our PD workshop attempted to move beyond prior published work in the following ways: First, unlike all past studies, our intervention was built upon an explicit theoretical model of PD (Desimone 2009) and was clearly linked to specific approaches associated with successful PD in the literature, including a strong content focus, active learning (student-centered inquiry-based activities); coherence (alignment with NJ State Standards and district goals); duration (the equivalent of one graduate class); and collective participation (collaborative learning in school-based context) (Table 2). Second, the program utilized curricular and pedagogical strategies identified in the science education literature as effective for evolution instruction, such as a focus on the

Table 3 Spearman's rho correlations among instrument scores pre- to post-intervention

	CINS	ACORNS-KC	MATE-P	MATE-S	TA-NOS 1	TA-NOS 3	TA-NOS 5	TA-NOS 6
Pre-survey								
CINS	1							
ACORNS-KC	0.775**	1						
MATE-P	0.836**	0.830**	1					
MATE-S	0.550**	0.425*	0.682**	1				
TA-NOS 1	0.537**	0.439*	0.501**	0.335	1			
TA-NOS 3	0.247	0.163	0.212	0.263	0.173	1		
TA-NOS 5	0.454*	0.379*	0.494**	0.282	0.15	0.283	1	
TA-NOS 6	0.286	0.334	0.411*	0.239	0.085	0.086	0.314	1
Post-survey								
CINS	1							
ACORNS-KC	0.617**	1						
MATE-P	0.714**	0.743**	1					
MATE-S	0.443*	0.598**	0.781**	1				
TA-NOS 1	0.368	0.237	0.394*	0.361	1			
TA-NOS 3	0.638**	0.372	0.406*	0.292	−0.002	1		
TA-NOS 5	0.314	0.297	0.180	0.092	0.356	0.276	1	
TA-NOS 6	0.594**	0.347	0.312	0.268	0.059	0.372	0.080	1

CINS natural selection knowledge multiple-choice test, *ACORNS-KC* open-response evolution key concept knowledge test, *MATE* acceptance of evolution test, *MATE-P* Personal acceptance test, *MATE S* Scientists' acceptance test, *TA-NOS* the nature of science knowledge test.

* $p < 0.05$, ** $p < 0.01$.

relationships between science and religion (content) and student ideas/alternative conceptions (cognition). Third, the program focused on three core areas (evolution knowledge, evolution acceptance, and NOS) that have only sometime been united in prior teacher intervention studies (see Table 1). Fourth, unlike some past studies that have relied upon multiple-choice tests as measures of PD efficacy, we supplemented such tools with scientific practice tasks (particularly explaining evolutionary change). Fifth, our study is the first to determine whether short-term gains have any staying power (>1 year later) using empirical (including practice) measures and self-perception data. We now discuss some specific findings relevant to future studies.

Science Teachers' Evolution Knowledge

We used two assessments—one multiple choice and the other constructed-response—to measure the impact of the intervention on teachers' knowledge of evolution. While covering the same general content (evolutionary change and its causes), measures derived from the two instruments revealed different learning gain patterns. Teachers showed statistically significant improvement in the number of correct answers chosen on the multiple-choice assessment but did not show statistically significant improvement on the use of accurate key concepts in the constructed-response assessment. We did, however, find a significant and meaningful reduction in the number of alternative conceptions in teachers' explanations on the constructed response assessment. In particular, significant decreases were noted in teachers' use of *teleology* in their explanations of evolutionary change. Because the reduction in *teleology* is associated with increases in the use of the scientifically accurate concepts of '*variability*' and '*population change*', the intervention may help teachers adopt 'population thinking' and overcome the well-documented cognitive bias of essentialism (Sinatra et al. 2008). Methodologically, these findings indicate that measuring teachers' accurate knowledge using MC assessments alone may provide a limited picture of conceptual improvements in understanding (Nehm and Schonfeld 2008).

In addition, the open-response assessment depicted a change in knowledge coherence patterns and the frequencies of mixed models (that is, explanations comprised of both scientific and naïve elements). Although more than half of the participants employed *variability* across all four open-response items, the majority of teachers inconsistently employed *heritability* and *differential survival/reproduction* in the post-test. In addition, after the intervention, 12.0% of teachers displayed mixed models. Although the dramatic increase of explicitly scientific ideas, and the decrease in the frequency of mixed models and explicitly naïve models provide support for the effectiveness of the PD program, several teachers failed to meet a minimum competency benchmark (i.e., being able to explain evolutionary change free of alternative conceptions).

We consider the scores on the MC test to be representative of teachers' knowledge gains, specifically their ability to recognize accurate scientific information. We view the absence of significant changes in teachers' performance on the constructed-response assessment as indicative of their difficulties in using, applying, and communicating that knowledge. Teachers' communication and explanation skills are best measured using constructed-response tests, and yet comparable types of assessments were lacking in most prior teacher survey studies, association studies, and intervention studies (e.g., Table 1; however, see Crawford et al. 2005). Thus, while knowledge increase and alternative conception decrease were notable outcomes of our intervention, and many others in the literature, teachers' abilities to communicate robust explanations of evolutionary change were modest. Future teacher PD programs should therefore (1) employ measures of efficacy more closely aligned to pedagogical practice; and (2) provide opportunities for teachers to explain evolutionary change orally and in writing. Such changes may have greater connection to important classroom competencies, such as teaching evolution effectively.

Science Teachers' NOS Knowledge

The nature of science (NOS) has long been associated with evolutionary theory in the science education literature (Kim and Nehm 2011; Nehm et al. 2009; Scharmann 1994; Rutledge and Warden 2000; Trani 2004) and is widely considered to be a necessary prerequisite to understanding evolution (e.g., Kennedy et al. 1998). As noted above, our program covered core NOS concepts (observation vs. inference; theory vs. law; etc.) during the first 2 days of the 10-day program. Moreover, TA-NOS concepts were highlighted repeatedly when teaching about other topics (e.g., Mendelian ratios; fossils; photosynthesis and transpiration; etc.). Although TA-NOS was not the primary focus of the intervention, teachers displayed the largest levels of improvement in their understanding of TA-NOS relative to other measured variables (see Additional file 1: Figure S2). This finding supports the theoretical rationale in the science education literature, namely that explicit and active learning about TA-NOS is helpful to the learning of evolution. Indeed, our results may be added to three of the four other intervention studies that we reviewed that documented similar findings (i.e., Nehm and Schonfeld 2007; Scharmann and Harris 1992; Scharmann 1994).

The magnitudes of TA-NOS knowledge change documented in our study are in alignment with those achieved in interventions focusing on TA-NOS alone. Akerson et al. (2000), for example, taught reflective, explicit, activity-based instruction on TA-NOS for graduate students in elementary education for one semester (weekly in 3-h blocks) and found that 24% of participants developed adequate understanding of empiricism and 56% of participants developed adequate understanding of the differences between theories and laws. Our study found larger gains, and illustrated that 71% of participants developed adequate understanding of empiricism and 86% of participants developed adequate understanding of the differences between theories and laws. In terms of knowledge retention, Akerson et al. (2006) investigated pre-service elementary teachers who completed a science methods course focusing on explicit-reflective instruction in TA-NOS. The authors reported that 5 months after the intervention, 41% of participants retained their ideas of scientific empiricism and 94% of participants retained understanding of the differences between theories and laws. Our retention findings were not as impressive (53 and 70%, respectively), but they were measured much longer after the intervention than in Akerson et al. (2006) study.

The intervention studies that we reviewed used different pedagogical and curricular approaches to teach evolution and NOS than described in our study (e.g., Smith and Scharmann 2008), supporting conventional wisdom that different approaches to teaching subject matter can nevertheless generate comparable learning gains. Because previous studies reported z values from Wilcoxon signed-ranks tests, and not traditionally used partial eta squared scores, we also generated *z* values so that we could compare our results with those from previous results. The *z* values for teachers' TA-NOS knowledge gain in our study were TA-NOS 1 (2.68, $p < 0.01$), TA-NOS 3 (4.05, $p < 0.001$), TA-NOS 5 (4.37, $p < 0.001$), and TA-NOS 6 (3.04, $p < 0.01$), compared to 2.74 ($p < 0.01$) in Scharmann and Harris's (1992) study, 3.06 ($p < 0.01$) in Scharmann's (1994) study, and 4.20 ($p < 0.01$) in Nehm and Schonfeld's (2007) study. All of the studies, which combined both evolution and TA-NOS, demonstrated highly significant z values. This suggests, but does not causally demonstrate (because of the study designs), that teaching both evolution and TA-NOS together tends to generate improvements in TA-NOS understanding. This is an emerging theme in evolution education research.

Science Teachers' Acceptance of Evolution

We administered the MATE instrument, one of the most widely used measures of evolution acceptance, to the teachers in our program (Rutledge and Sadler 2007).

Notably, we administered two versions of the MATE instrument: personal acceptance (MATE-P) and perceptions of scientists' acceptance (MATE-S) (see "Participants and Methods"). Moore (2007, 2008) indicated that many biology teachers who are very religious are conflicted about teaching evolution because of their religious affiliations. They may nevertheless teach evolution because they recognize that scientists accept it, even if they do not personally accept it. Therefore, it is important to administer both versions of the MATE instrument in order to precisely measure teachers' acceptance of evolution, and changes in acceptance levels, relative to their perceptions of scientists' views.

Regardless of measurement type (MATE-P vs. S), teachers in our program displayed significant increases in their acceptance of evolution, although to begin with they had relatively high levels of acceptance. Teachers' acceptance levels after the intervention (88/100) were very close to a 'very high level' of MATE scores (89–100). Prior to the intervention, in contrast, teachers' acceptance levels were lower (80/100). Rutledge and Warden (1999, 2000) reported "high acceptance" levels (77.6/100) for Indiana biology teachers as did Korte (2003) for Ohio biology teachers (87.5/100). Trani (2004) likewise reported MATE scores of 85.9 for Oregonian biology teachers. Although research on teacher's acceptance of evolution using the MATE has not been conducted on a national level, most previous research showed that American biology teachers in general possess a "high level" of evolutionary acceptance. Our results indicate that our teacher sample began our intervention with average scores similar to other teachers from the Midwestern United States.

In contrast to studies of in-service teachers, samples of pre-service teachers and undergraduate science students tend to display slightly lower acceptance levels. Rutledge and Sadler (2007), for example, reported that non-biology college students at Middle Tennessee State displayed low levels of acceptance as measured by MATE scores (55.9/100). Interestingly, the MATE scores from students at The Ohio State University displayed somewhat different results. Kim and Nehm (2011) reported that non-major and biology major students at Ohio State displayed relatively high MATE scores (77.9 and 80.8, respectively). In a different socio-cultural context, Turkish pre-service biology teachers were found to have MATE scores of 63.7 (Deniz et al. 2008). In Korea, Ha et al. (2012) reported that Protestant pre-service biology teachers displayed MATE scores of 65.5 whereas non-religious pre-service biology teachers displayed MATE score of 76.0. Overall, it is clear that even when the same instrument is used, similar populations of students (undergraduates, pre-service teachers) display variable acceptance levels depending on geography (Tennessee vs. Ohio) and culture (Korea vs. Turkey). Nevertheless, greater educational attainment is associated with greater acceptance of evolution.

It has been reported that changing belief is far more difficult than changing knowledge of evolution (Nehm and Schonfeld 2007; see also Jones and Carter 2007). In our study, however, we found that learning gains (effect sizes) for the multiple-choice knowledge measure were slightly smaller than the effect size of the acceptance of evolution measure (partial eta squared for CINS = 0.49; partial eta squared for MATE-P = 0.55). We also a noted a very small effect size for constructed-response knowledge measures (partial eta squared for ACORNS-KC = 0.09). Our study suggests that knowledge increase was smaller than acceptance increase.

Teachers' Knowledge and Belief Retention

While the participants showed a significant increase in knowledge of evolution and the nature of science (e.g. CINS, ACORNS, and TA-NOS scores) and acceptance of evolution (e.g., MATE scores) in the post-test, it is important to know whether these changes were sustained long after the PD experience ended. No studies to our knowledge have investigated whether changes associated with evolution PD have long-lasting effects (>1 year). The delayed post-tests, conducted 15 months after the intervention, indicated that participants retained knowledge gains of both evolution and TA-NOS, and retained acceptance of evolution gains. Given that the delayed post-tests were conducted when the participants were classroom teachers, the significant knowledge and belief retention suggests that the PD program may have influenced teaching practices. In addition, the results of teachers' self-evaluation in the delayed post-test also indicated that the PD program retained its influence. The retention of teachers' acceptance changes is a particularly encouraging finding given that this goal has been out of reach for many past studies (e.g., Nehm and Schonfeld 2007). Since teachers' evolutionary beliefs are known to be an important factor relating to the instructional time devoted to evolution (Moore 2007, 2008), sustaining belief change is a particularly salient feature of evaluating PD impacts. Using this criterion, our PD program was successful.

Associations Among Intervention Variables vs. Learning Gains

The results of correlation analyses revealed that the relationships among knowledge and acceptance of evolution, and understanding of TA-NOS, were robust ($r > 0.6$). We discuss the relationship between knowledge and acceptance of evolution first. As we noted in the literature review section, there have been several studies reporting the relationship between teachers' knowledge and acceptance of evolution (Deniz et al. 2008; Ha et al. 2012; Korte 2003; Rutledge and Warden 2000). Although it is difficult to generalize from these studies because of inconsistent findings, more educated participants (e.g. in-service teachers) tend to display higher correlation coefficients between knowledge and acceptance of evolution than less educated participants (e.g. pre-service teachers). Our sample showed strong correlation coefficients between acceptance of evolution for both CINS and ACORNS scores (pre-test: $r = 0.86$ for CINS, $r = 0.79$ for ACORNS; post-test $r = 0.68$ for CINS, $r = 0.66$ for ACORNS). Thus, the higher values documented in our study relative to prior work is in line with general patterns in the literature.

The second relationship that we discuss is between acceptance of evolution and understanding of NOS. Lombrozo et al. (2008) reported that NOS scores were significantly and moderately correlated with evolution acceptance scores ($r = 0.40$, non-biology major college students; although see criticisms of their instrument by Neumann et al. (2011). Johnson and Peeples (1987) also reported acceptance and understanding associations for biology majors and college students. The correlation values that they documented were also moderate (but statistically significant) for both biology majors and college students ($r = 0.45$ and 0.45, respectively). Rutledge and Warden (2000), who recruited a similar sample to our study, also reported strong correlations between comparable measures similar to those documented in our study ($r = 0.7$, ~0.6, respectively).

Initially, we considered the strong and significant correlations between knowledge and acceptance of evolution and understanding of NOS as sufficient to support a pedagogical strategy of teaching evolution and NOS together. We assumed, in particular, that a better understanding of evolution may facilitate a change in acceptance of evolution, and an improved understanding of NOS may lead to a greater acceptance of evolution. Our assumption here was based on previous studies concerning the relationships among knowledge and acceptance of evolution and understanding of NOS (Deniz et al. 2008; Johnson and Peeples 1987; Korte 2003; Lombrozo et al. 2008; Rutledge and Warden 2000). However, it is very important to emphasize that the correlations among learning gain scores for nearly all variables (CINS, ACORNS-KC, ACORNS-MIS, MATE-P, MATE-S, and TA-NOS) in the present study were not significant. Although our initial assumption was rejected, this result is important because it may provide evidence in support of the idea that the relationships among levels of knowledge and acceptance of evolution and understanding of TA-NOS are not cause-effect relationships.

If knowledge and acceptance of evolution and understanding of NOS are in fact causally associated with one another, then learning one factor would affect learning of another factor. Our findings for learning gains, in contrast, do not support a cause-effect relationship between these variables. It is possible that other latent variables explain gains across all variables, but we lack evidence to support or reject such a view. Further work, using causal study designs, is needed to determine if in fact teaching these subjects together produces a synergistic effect. Indeed, science educators (including us) have too readily accepted associations among knowledge of evolution and acceptance of evolution, and knowledge of TA-NOS and understanding evolution, as suggestive of causal connection. While reasonable, there is at present no evidence to support such a view, and our data relating to learning gains suggest otherwise.

Differences Between Delayed Post-Test Scores and Self-Perception Scores

In the delayed post-test, we employed two types of measures of teachers' evolutionary knowledge and acceptance: self-perception measures and empirical measures. As noted above, the results from these two types of measures were not in alignment. Similar patterns have been found in medical education research (see Lai and Teng 2011; McCormack et al. 2004). Specifically, Mabe and West (1982) reported that self-perception accuracy may relate most strongly to participants' intelligence, achievement status, and internal locus of control rather than objective correspondence to actual magnitudes of change. Indeed, many exogenous variables likely mediate self-perception measures and lead to weak alignment with objective measures. It could be that self-perceived, self-reported measures (such as those used in Firenze 1997) do not robustly capture true magnitudes of change. Without a third, independent measure, however, we cannot determine which of the two measures generates more valid inferences about change. Nevertheless, the finding raises questions about how best to measure the impact of PD programs on science teachers.

Overall Efficacy and Implications for Policy and Practice

Many studies have reported that evolution is difficult to learn, and changing acceptance levels is difficult to achieve. Moreover, sufficient understanding of TA-NOS is lacking in many science teachers. The overall results of our multivariate analyses indicated that our program was highly effective in addressing all of these issues; our intervention did improve teachers' knowledge of evolution (partial eta squared for CINS [0.49] and for ACORNS-KC [0.09]), acceptance of evolution (partial eta squared for MATE-P [0.55], for MATE-S [0.15]), and understanding of TA-NOS (Cohen's d of VNOS 1, 3, 5, and 6 were respectively 0.61, 1.37, 2.27, and 0.86). The overall efficacy of this program, measured using partial eta squared, indicated that 77% (pre-/post-tests) and 81% (pre-/post-/delayed post-tests) of participants benefited from the program. Such successful findings across all measures may be attributed in part to our overall teaching strategy: (1) explicit teaching about TA-NOS; (2) discussions of the relationships between science and religion; and (3) inquiry-based exploration of evolutionary concepts.

Study Limitations and Future Research Directions

The greatest limitation of our study is that it did not investigate participant teachers' classroom actions with respect to evolution, or if they changed as a consequence of the intervention. Even though the intervention was judged to be successful in many respects, it is by no means clear whether the knowledge gains and acceptance changes that we documented actually impacted classroom instruction in any meaningful sense. It is possible that the benefits of the PD program produced no downstream effects and had little impact on K-12 students' learning experiences. This is a serious limitation of all of the intervention studies that we reviewed in the literature (see Table 1) and should be addressed in future work. This limitation complicates our interpretation that the intervention was truly "successful." Nevertheless, large percentages of teachers self-reported, 15 months after the PD intervention, that it impacted their classroom practices. This is an encouraging finding, but objective evidence is needed to substantiate this claim.

A second limitation is that the study did not explore two important aspects of teacher evolution education that have significant downstream effects on student learning: (1) teachers' preferences for teaching evolution (e.g., Nehm and Schonfeld 2007; Griffith and Brem 2004) and (2) teachers' pedagogical content knowledge (e.g., Asghar et al. 2007; Großschedl et al. 2014). As Sickel and Friedrichsen (2013) have noted, handling controversy and being able to effectively teach evolutionary ideas are central features of effective teacher education programs. Future PD programs should consider these important elements in their design and execution.

Our study focuses on teachers from New Jersey, a State with relatively strong evolution standards and a well-educated citizenry (see subnormalnumbers.blogspot; http://subnormalnumbers.blogspot.kr/2010/04/acceptance-of-evolution-by-state.html). New Jersey is also characterized by above-average scores on the National Assessment of Educational Progress (NAEP) science test. While there is no direct measurement of comparison in science achievement between states, NAEP scores indicate that New Jersey students scored in the top third of all states in the 2011 science assessment (Heinz, personal communication 2015).

The northeastern United States is a region characterized by high levels of evolution acceptance (see http://subnormalnumbers.blogspot.kr). These factors limit our ability to generalize our finding that short term PD can have meaningful and lasting effects on teachers. For example, a sample of teachers from another region of the country with lower evolution acceptance levels might not have responded to our PD program in the same way as our NJ sample. Future work is clearly needed to explore the effects of regional differences on PD efficacy and durability. Given that our study is the first to explore retention patterns, it provides a benchmark for future work.

A final limitation of our study is that it did not explore many variables salient to effective PD, including sample size (How many teacher participants make for an effective PD experience?), participant homogeneity (Would single-grade samples be more or less effective?), duration (Would 1 week or 3 weeks generate more or less change?), methodology (Would qualitative explorations of teacher knowledge and belief change produce comparable findings?), and theoretical framing (Would other PD models, such as the 5-E approach, be more or less effective?). Clearly, much important work remains to be done, and our study only scratches the surface of a large and complex challenge in evolution education.

Conclusion

Many politicians, scientists, and science educators are concerned about American science teachers' low levels of knowledge and acceptance of evolution, particularly because of their crucial role in connecting the knowledge of scientists with the literacy of future generations. While the production of educational studies whose aim is to document problems with science teacher knowledge and acceptance levels continues unabated, remarkably less work has been devoted to addressing the problem, or rigorously documenting the features of effective intervention programs. Our findings provide clear and convincing evidence that well-designed teacher PD programs can achieve significant, meaningful, and sustained impacts upon both teachers' knowledge and acceptance of evolution, and their understanding of NOS. More efforts in the science education community and by funding agencies should be directed at implementing evidence-based, short-term intervention programs to target science teachers rather than conducting and funding additional studies of teacher ambivalence towards evolutionary science.

Additional File

Additional file 1:Figure S1. ACORNS individual pre-post item scores. **Figure S2.** VNOS pre-post test scores. **Table S1.** Examples of teacher responses to the ACORNS. **Table S2.** VNOS items and teacher responses.

Authors' Contributions

MH analyzed the data, reviewed literature, and contributed to the writing of the paper; BB recruited teachers, oversaw the PD program, collected the data, and reviewed the paper; RN designed the study, interpreted results, and contributed to the writing of the paper. All authors read and approved the final manuscript.

Author details

[1] Division of Science Education College of Education, Kangwon National University, Hyoja-dong, Chuncheon-si, Gangwon-do 200-701, South Korea. [2] New Jersey Center for Science, Technology, and Mathematics, Kean University, 1000 Morris Ave., Union, NJ 07083, USA. [3] Center for Science and Mathematics Education, Stony Brook University (SUNY), 092 Life Sciences Building, Stony Brook, NY 11794, USA. [4] Department of Ecology and Evolution, Stony Brook University (SUNY), Stony Brook, NY 11794, USA.

Acknowledgements

This material is based upon work supported by the National Science Foundation REESE 0909999 and assessments developed as part of TUES-1322872. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Compliance with Ethical Guidelines

Competing Interests

The authors declare that they have no competing interests.

Received: 21 March 2015 Accepted: 13 July 2015
Published online: 25 July 2015

References

Abd-El-Khalick, F., & Lederman, N. G. (2000). Improving science teachers' conceptions of nature of science: a critical review of the literature. *International Journal of Science Education, 22*(7), 665–701.

Akerson, V. L., Abd-El-Khalick, F., & Lederman, N. G. (2000). Influence of a reflective explicit activity-based approach on elementary teachers' conceptions of nature of science. *Journal of Research in Science Teaching, 37*(4), 295–317.

Akerson, V. L., Morrison, J. A., & McDuffie, A. R. (2006). One course is not enough: preservice elementary teachers' retention of improved views of nature of science. *Journal of Research in Science Teaching, 43*(2), 194–213.

Allison, P. D. (2002). Missing data: quantitative applications in the social sciences. *British Journal of Mathematical and Statistical Psychology, 55*(1), 193–196.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching, 39*(10), 952–978.

Asghar, A., Wiles, J. R., & Alters, B. (2007). Canadian pre-service elementary teachers' conceptions of biological evolution and evolution education. *McGill Journal of Education, 42*(2), 189–210.

Banilower, E. R., Heck, D. J., & Weiss, I. R. (2007). Can professional development make the vision of the standards a reality? The impact of the national science foundation's local systemic change through teacher enhancement initiative. *Journal of Research in Science Teaching, 44*(3), 375–395.

Bao, L. (2006). Theoretical comparisons of average normalized gain calculations. *American Journal of Physics, 74*, 917–922.

Battisti, B. T., Hanegan, N., Sudweeks, R., & Cates, R. (2010). Using item response theory to conduct a distracter analysis on conceptual inventory of natural selection. *International Journal of Science and Mathematics Education, 8*, 845–868.

Bell, B., & Gilbert, J. (1996). *Teacher development: a model from science education*. London/Washington: Falmer Press.

Berkman, M. B., & Plutzer, E. (2011). Defeating creationism in the courtroom, but not in the classroom. *Science, 331*, 404–405.

Bishop, B. A., & Anderson, C. W. (1990). Student conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching, 27*(5), 415–427.

Borko, H., Jacobs, J., Eiteljorg, E., & Pittman, M. E. (2008). Video as a tool for fostering productive discussions in mathematics professional development. *Teaching and Teacher Education, 24*(2), 417–436.

Brem, S. K., Ranney, M., & Schindel, J. (2003). Perceived consequences of evolution: college students perceive negative personal and social impact in evolutionary theory. *Science Education, 87*(2), 181–206.

Cobern, W. W. (1996). Constructivism and non-western science education research. *International Journal of Science Education, 18*(3), 295–310.

Collins, H. M., & Pinch, T. J. (1998). *The golem at large: what you should know about science* (2nd ed.). Cambridge: Cambridge University Press.

Crawford, B. A., Zembal-Saul, C., Munford, D., & Friedrichsen, P. (2005). Confronting prospective teachers' ideas of evolution and scientific inquiry using technology and inquiry-based tasks. *Journal of Research in Science Teaching, 42*(6), 613–637.

Custers, E. J. F. M. (2010). Long-term retention of basic science knowledge: a review study. *Advances in Health Sciences Education, 15*(1), 109–128.

Deniz, H., Donnelly, L. A., & Yilmaz, I. (2008). Exploring the factors related to acceptance of evolutionary theory among Turkish preservice biology teachers: toward a more informative conceptual ecology for biological evolution. *Journal of Research in Science Teaching, 45*(4), 420–443.

Desantis, L. R. G. (2009). Teaching evolution through inquiry-based lessons of uncontroversial science. *The American Biology Teacher, 71*(2), 106–111.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181–199.

Driver, R., Squires, A., Rushworth, P., & Wood-Robinson, V. (1994). *Making sense of secondary science: research into children's ideas*. New York: Routledge.

Duschl, R., Schweingruber, H. A., & Shouse, A. W. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academy Press.

Ebbinghaus, H. (1966). *U¨ber das Geda¨chnis. Untersuchungen zur Experimentellen Psychologie. Nachdruk der Ausgabe Leipzig* 1885. Amsterdam, Netherlands: E. J. Bonset. [English text available at http://psychclassics.yorku.ca/Ebbinghaus/index.htm].

El-Hani, C. N., & Mortimer, E. F. (2007). Multicultural education, pragmatism, and the goals of science teaching. *Cultural Studies of Science Education, 2*(3), 657–702.

Firenze, R. F. (1997). The identification, assessment, and amelioration of perceived and actual barriers to teachers' incorporation of evolutionary theory as a central theme in life science classes through a two-week institute and follow-up studies. Unpublished doctoral dissertation. Binghamton University State University of New York, Vestal, NY.

Franke, M. L., Carpenter, T. P., Levi, L., & Fennema, E. (2001). Capturing teachers' generative change: a follow-up study of professional development in mathematics. *American Educational Research Journal, 38*(3), 653.

Futuyma, D. (2009). *Evolution* (2nd ed.). Sunderland: Sinauer Associates.

Gregory, T. R. (2009). Understanding natural selection: essential concepts and common alternative conceptions. *Evolution: Education and Outreach, 2*, 156–175.

Griffith, J. A., & Brem, S. K. (2004). Teaching evolutionary biology: Pressures, stress, and coping. *Journal of Research in Science Teaching, 41*(8), 791–809.

Großschedl, J., Konnemann, C., & Basel, N. (2014). Pre-service biology teachers' acceptance of evolutionary theory and their preference for its teaching. *Evolution: Education and Outreach, 7*(1), 1–16.

Ha, M., Haury, D. L., & Nehm, R. H. (2012). Feeling of certainty: Uncovering a missing link between knowledge and acceptance of evolution. *Journal of Research in Science Teaching, 49*(1), 95–121.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*, 64–74.

Hewson, P. W. (2007). Teacher professional development in science. In S. K. Abell & N. S. Lederman (Eds.), *The handbook of research on science teaching* (pp. 1179–1203). Mahwah: Lawrence Erlbaum.

Johnson, R. L., & Peeples, E. E. (1987). The role of scientific understanding in college: student acceptance of evolution. *The American Biology Teacher, 49*, 93–98.

Jones, M. G., & Carter, G. (2007). Science teacher attitudes and beliefs. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 1067–1104). Mahwah: Lawrence Erlbaum Associates Inc.

Joyce, B., & Showers, B. (1988). *Student achievement through staff development*. New York: Longman.

Kampourakis, K. (2014). *Understanding evolution*. Cambridge: Cambridge University Press.

Kampourakis, K., & Zogza, V. (2009). Preliminary evolutionary explanations: a basic framework for conceptual change and explanatory coherence in evolution. *Science & Education, 18*(10), 1313–1340.

Kennedy, D., Moore, J., Alberts, B., Scott, E., Ezell, D., Singer, M., et al. (1998). *Teaching about evolution and the nature of science*. Washington, DC: National Academy Press.

Kimball, M. E. (1967). Understanding the nature of science: A comparison of scientists and science teachers. *Journal of Research in Science Teaching, 5*(2), 110–120.

Kim, S. Y., & Nehm, R. H. (2011). A cross-cultural comparison of korean and american science teachers' views of evolution and the nature of science. *International Journal of Science Education, 33*(2), 197–227.

Korte, S. E. (2003). The acceptance and understanding of evolutionary theory among Ohio secondary life science teacher. Unpublished master thesis, Ohio University.

Lai, N. M., & Teng, C. L. (2011). Self-perceived competence correlates poorly with objectively measured competence in evidence based medicine among medical students. *BMC Medical Education, 11*(25), 1–8.

Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. S. (2002). Views of nature of science questionnaire: toward valid and meaningful assessment of learners' conceptions of nature of science. *Journal of Research in Science Teaching, 39*(6), 497–521.

Linting, M., Meulman, J. J., Groenen, P. J. F., & van der Koojj, A. J. (2007). Nonlinear principal components analysis: introduction and application. *Psychological Methods, 12*(3), 336–358.

Lomax, R. G. (2007). *An introduction to statistical concepts* (2nd ed.). Mahwah: Lawrence Erlbaum Associates.

Lombrozo, T., Thanukos, A., & Weisberg, M. (2008). The importance of understanding the nature of science for accepting evolution. *Evolution: Education and Outreach, 1*, 290–298.

Loucks-Horsley, S., Hewson, P. W., Love, N., & Stiles, K. E. (1998). *Designing professional development for teachers of science and mathematics*. Thousand Oaks: Corwin Press.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: a review and meta-analysis. *Journal of Applied Psychology, 67*(3), 280–296.

McComas, W. F. (1996). Ten myths of science: reexamining what we think we know about the nature of science. *School Science and Mathematics, 96*(1), 10–16.

McCormack, G., Giles-Corti, B., Lange, A., Smith, T., Martin, K., & Pikora, T. (2004). An update of recent evidence of the relationship between objective and self-report measures of the physical environment and physical activity behaviours. *Journal of Science and Medicine in Sport, 7*(1), 81–92.

Moore, R. (2002). Teaching evolution: do state standards matter? *Bioscience, 52*(4), 378–381.

Moore, R. (2007). The differing perceptions of teachers and students regarding teachers' emphasis on evolution in high school biology classrooms. *The American Biology Teacher, 69*(5), 68–271.

Moore, R. (2008). Creationism in the biology classroom: what do teachers teach and how do they teach it? *The American Biology Teacher, 70*(2), 79–84.

Moore, R. W., & Foy, R. L. H. (1997). The scientific attitude inventory: a revision (SAI II). *Journal of Research in Science Teaching, 34*(4), 327–336.

Nadelson, L. S., & Sinatra, G. M. (2009). Educational professionals' knowledge and acceptance of evolution. *Evolutionary Psychology, 7*, 490–516.

Nadelson, L. S., & Sinatra, G. M. (2010). Shifting acceptance of the understanding evolution website. *The Researcher, 23*, 13–29.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core Ideas*. Washington, DC: The National Academies Press.

Nehm, R. H., Beggrow, E. P., Opfer, J. E., & Ha, M. (2012). Reasoning about natural selection: diagnosing contextual competency using the ACORNS Instrument. *The American Biology Teacher, 74*(2), 92–98.

Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching, 48*(3), 237–256.

Nehm, R.H., Ha, M., Rector, M., Opfer, J., Perrin, L., Ridgway, J., Mollohan, K. (2010). Scoring guide for the open response instrument (ORI) and evolutionary gain and loss test (EGALT). Technical Report of National Science Foundation REESE Project 0909999. 40 p.

Nehm, R. H., & Ridgway, J. (2011). What do experts and novices "see" in evolutionary problems? *Evolution: Education and Outreach, 4*(4), 666–679.

Nehm, R. H., & Schonfeld, I. S. (2007). Does increasing biology teacher knowledge of evolution and the nature of science lead to greater preference for the teaching of evolution in schools? *Journal of Science Teacher Education, 18*, 699–723.

Nehm, R. H., & Schonfeld, I. S. (2008). Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. *Journal of Research in Science Teaching, 45*(10), 1131–1160.

Nehm, R. H., & Schonfeld, I. S. (2010). The future of natural selection knowledge measurement: a reply to Anderson et al. (2010). *Journal of Research in Science Teaching, 47*(3), 358–362.

Nehm, R. H., Kim, S. Y., & Sheppard, K. (2009). Academic preparation in biology and advocacy for teaching evolution: biology versus non-biology teachers. *Science Education, 93*(6), 1122–1146.

Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: rasch-based analyses of a nature of science test. *International Journal of Science Education, 33*(10), 1373–1405.

Opfer, J. E., Nehm, R. H., & Ha, M. (2012). Cognitive foundations for science assessment design: knowing what students know about evolution. *Journal of Research in Science Teaching, 49*(6), 744–777.

Rector, M. A., Nehm, R. H., & Pearl, D. (2013). Learning the language of evolution: lexical ambiguity and word meaning in student explanations. *Research in Science Education, 43*(3), 1107–1133.

Rutledge, M. L., & Sadler, K. C. (2007). Reliability of the measure of acceptance of the theory of evolution (MATE) instrument with university students. *The American Biology Teacher, 51*, 275–280.

Rutledge, M. L., & Warden, M. A. (1999). Development and validation of the measure of acceptance of the theory of evolution instrument. *School Science and Mathematics, 99*, 13–18.

Rutledge, M. L., & Warden, M. A. (2000). Evolutionary theory, the nature of science and high school biology teachers: critical relationships. *The American Biology Teacher, 62*(1), 23–31.

Scharmann, L. C. (1994). Teaching evolution: the influence of peer teachers' instructional modeling. *Journal of Science Teacher Education, 5*, 66–76.

Scharmann, L. C., & Harris, W. M. (1992). Teaching evolution: understanding and applying the nature of science. *Journal of Research in Science Teaching, 29*(4), 375–388.

Semb, G. B., & Ellis, J. A. (1994). Knowledge taught in school: what is remembered? *Review of Educational Research, 64*(2), 253–286.

Sickel, A. J., & Friedrichsen, P. (2013). Examining the evolution education literature with a focus on teachers: major findings, goals for teacher preparation, and directions for future research. *Evolution: Education and Outreach, 6*(1), 23.

Sinatra, G. M., Brem, S. K., & Evans, E. M. (2008). Changing minds? Implications of conceptual change for teaching and learning about biological evolution. *Evolution: Education and Outreach, 1*, 189–195.

Smith, M. U. (2010). Current status of research in teaching and learning evolution: II. Pedagogical issues. *Science & Education, 19*(6), 539–571.

Smith, M. U., & Scharmann, L. (2008). A multi-year program developing an explicit reflective pedagogy for teaching pre-service teachers the nature of science by ostention. *Science & Education, 17*(2), 219–248.

Smith, M. U., & Siegel, H. (2004). Knowing, believing, and understanding: What goals for science education? *Science & Education, 13*(6), 553–582.

Southerland, S. A., & Nadelson, L. S. (2012). An intentional approach to teaching evolution: Making students aware of the factors influencing learning of microevolution and macroevolution. In K. S. Rosengren, S. Brem, E. M. Evans, & G. Sinatra (Eds.), *Evolution challenges: integrating research and practice in teaching and learning about evolution* (pp. 348–374). Cambridge: Oxford University Press.

Supovitz, J. A., & Turner, H. M. (2000). The effects of professional development on science teaching practices and classroom culture. *Journal of Research in Science Teaching, 37*(9), 963–980.

Trani, R. (2004). I won't teach evolution; it's against my religion. And now for the rest of the story. *The American Biology Teacher, 66*, 419–427.

Vosniadou, S. (2008). The framework theory approach to the problem of conceptual change. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 3–34). New York: Routledge.