J Sci Teacher Educ (2014) 25:807–821
DOI 10.1007/s10972-014-9401-4

Effective, Sustained Inquiry-Based Instruction Promotes Higher Science Proficiency Among All Groups: A 5-Year Analysis

Jeff C. Marshall · Daniel M. Alston

Published online: 17 October 2014

Abstract Student's performance in science classrooms has continued to languish throughout the USA. Even though proficiency rates on national tests such as National Assessment of Educational Progress are higher for Caucasian students than African-Americans and Hispanics, all groups lack achieving desired proficiency rates. Further, the Next Generation Science Standards detail a new higher benchmark for all students. This study analyzes a professional development (PD) project, entitled Inquiry in Motion, designed to (a) facilitate teacher transformation toward greater quantity and quality of inquiry-based instruction, (b) improve student achievement in science practices and science concepts, and (c) begin to narrow the achievement gap among various groups. This 5-year PD study included 11 schools, 74 middle school teachers, and 9,981 students from diverse, high minority populations. Findings from the quasi-experimental study show statistically significant gains for all student groups (aggregate, males, females, Caucasians, African-Americans, and Hispanics) on all three science Measure of Academic Progress tests (composite, science practices, and science concepts) when compared to students of non-participating teachers. In addition to an increase in overall performance for all groups, a narrowing of the achievement gap of minority students relative to Caucasian students was seen. When combined with other studies, this study affirms that, when facilitated effectively, inquiry-based instruction may benefit all students, for all demographic groups measured.

J. C. Marshall (✉) · D. M. Alston
Eugene T. Moore School of Education, Clemson University, Clemson, SC 29634-0705, USA
e-mail: marsha9@clemson.edu

Present Address:
J. C. Marshall
Eugene T. Moore School of Education, 404A Tillman Hall, Clemson University, Clemson, SC 29634-0705, USA

Keywords Achievement gap · Assessment · Inquiry-based instruction · Middle school science · Student achievement · Professional development

Introduction

According to several indicators, American students have continued to underperform relative to the standards we have set in science education (Lauko, Grigg, & Brockway, 2006; Martin, Mullis, Foy, & Stanco, 2012; Schmidt, McNight, & Raizen, 2002; US Department of Education, Institute of Education Sciences, & National Center for Education Statistics, 2011). Among these reports, data show that during the past decade, that performance has generally stagnated or increased just slightly overall relative to the national benchmarks; plus, large achievement gaps continue to persist (Hartney & Flavin, 2014; Lauko et al., 2006; Rowen, Hall, & Haycock, 2010; US Department of Education et al., 2011). Specifically, the gap for the eighth-grade science students on the 2011 National Assessment of Education Progress (NAEP) showed 43 % of Caucasian students and 42 % of Asian students were proficient versus only 10 % of African-American and 16 % of Hispanic students (US Department of Education et al., 2011). Even though great strides have been taken toward improving minority performance over the past 70 years, much work is still needed to narrow gaps among ethnic and gender groups (Span & Rivers, 2012).

Within the current condition of languishing performance, a new benchmark for learning, Next Generation Science Standards (NGSS), has been introduced that effectively raises the performance expectations for what all students in K-12 science classes should know and be able to do (Achieve, 2013; National Research Council, 2012). High-stakes assessments are typically the real driver for change, and while they are not yet in place, it is clear that the assessments for the new performance expectations will need to challenge students to demonstrate mastery of higher-order thinking skills and knowledge (Krathwohl, 2002). Admittedly, these new assessments will be fundamentally different from most of the past standardized science tests that assessed mostly factual lower-level skills and knowledge. NAEP, also referred to as the Nation's Report Card, provides an exception to the previous rule of testing mostly lower-order thinking skills (e.g., recall and define) because their assessments measure both science concepts and science practices and challenge student thinking in higher-order skills (e.g., design, evaluate, and analyze).

The implementation of the NGSS represents a fundamental shift from the past standards for two significant reasons: (a) The degree of higher-order skills that all students are expected to master has increased dramatically, and (b) the level to which scientific practices and scientific content are integrated has dramatically increased. To clarify these differences further, the rise in expectations is clearly evidenced by the following two illustrations. First, in State X, which was lauded as having one of the strongest science standards according to the Fordham Institute (Gross et al., 2005), we found that 82 % of the previous high school life science standards were written for lower-order thinking (remembering and understanding) versus *NGSS* life science performance expectations which have 6 % at lower order (remembering and understanding), 29 % at middle order (applying and analyzing), and 65 % at higher order (evaluating and creating) (Krathwohl, 2002; Marshall, 2014). Second, prior state and national standards (National Research Council, 1996) placed science practices (inquiry) separate from science content and concepts. NGSS is written, so that the practices are integrated with the content being learned (Achieve, 2013). Clearly, these differences necessitate a reshaping of past curriculum and instruction because teaching a student to remember, recall, and define is very different from teaching, so that students can model complex ideas, plan and conduct an investigation, or provide evidence-based arguments.

For decades, science education has espoused, even though it has not been broadly adopted the importance of inquiry-based instruction (American Association for the Advancement of Science, 1993; Bybee et al., 2006; Marshall, 2013; National Research Council, 2000). While inquiry-based instruction is not the only instructional strategy geared toward encouraging higher-order thinking, it provides a vehicle by which teachers can engage their students in experiences that go beyond lower-level thinking (Marshall, 2013). Aspects crucial to inquiry-based instruction (e.g., active engagement with evidence, challenging curriculum geared to develop meaningful understandings, formative assessments utilized to inform teaching practices, allowing students the opportunities to investigate and construct their own knowledge, and engaging in classroom discourse) have been found to be key characteristics of effective science teachers (Rennie, Goodrum, & Hackling, 2001; Tyler, 2003). Further, effective teachers frequently excel with all students at levels far exceeding lower performing teachers. Although the metrics continue to be debated, highly effective teachers move students between 25 and 50 % more than would be expected for normal growth (Aaronson, Barrow, & Sander, 2007; Gordon, Kane, & Staiger, 2006; Hanushek, 1992, 2011). With this in mind, it is not surprising that inquiry-based strategies have been shown to reduce the achievement gap of minority students (Geier et al., 2008).

Social constructivism posits that it is through our interaction with others and the world around us that we make, understand, and construct knowledge (Vygotsky, 1978). Since none of our experiences are the same, guided inquiry-based instruction provides important opportunities for students to explore concepts and ideas within groups before formal explanations (sense-making) can occur (Marshall, 2013). Further, there is no relationship between student achievement and the time spent lecturing (Van Klaveren, 2011), and when teachers engage students in shared experiences which allow them to collaboratively explore conceptual ideas before explanation (student or teacher), which inquiry-based instruction promotes, the amount of time necessary for teacher explanation decreases (Marshall, Smart, Lotter, & Sirbu, 2011). NAEP 2009 and 2011 science results support the value of these shared experiences along with learning that formally engages students in science (US Department of Education et al., 2011). Specifically, eighth-grade students who are engaged in hands-on experiences almost every day versus only once or twice a month, according to teacher self-report, scored on average nine points higher in 2009 and eight points higher in 2011 on the NAEP science test.

Numerous other studies confirm the importance of effective forms of inquiry instruction, particularly when learning and retention for all is the emphasis. Specifically, knowledge, reasoning, and argumentation were greater for students involved in inquiry-based instruction versus more commonplace instruction (Johnson, 2009; Wilson, Taylor, Kowalski, & Carlson, 2009). Research that synthesized findings from 138 studies also indicates a positive effect relative to inquiry-based instruction (Minner, Levy, & Century, 2009).

Despite our knowing that knowledge and understanding are often co-created in social settings (Vygotsky, 1978) and despite us knowing that inquiry-based instruction helps to challenge and encourage critical thinking, it is evident that effective inquiry-based instruction is far from the norm in most classrooms (Marshall, Horton, Igo, & Switzer, 2009a). Moreover, considerable time is needed to transform practice from a classroom focused on inquiry-based instruction versus one where teacher transmission of knowledge predominates. Research suggests that transformation guided by professional development (PD) initiatives need to be sustained over significant periods of time (Garet, Porter, Desimone, Birman, & Yoon, 2001; Loucks-Horsley, Stiles, Mundry, Love, & Hewson, 2010; Supovitz & Turner, 2000; Yoon, Duncan, Lee, Scarloss, & Shapley, 2007).

Our research builds from the need to improve student achievement for all groups of learners in science classrooms. The PD that this research focuses on sought to transform teacher practice relative to inquiry-based instruction with the expectation that student achievement would increase for all groups of students. Specifically, three specific research questions were addressed in this study: (a) Do student proficiency levels increase for those engaged in effective inquiry-based instruction? (b) Do classrooms that utilize inquiry-based instruction demonstrate a narrowing of the achievement gap for minority students? (c) Do proficiency rates for both males and females increase in classrooms where teachers utilize inquiry-based instruction?

Methods

Intervention

This quasi-experimental design (with four groups) occurred over a 5-year time period. Specifically, the study led by the faculty associated with the Inquiry in Motion Institute provided PD for science teachers in 11 different middle schools (grades 6–8) located in five different school districts. The goals of the PD were threefold: (a) increase the quantity and quality of inquiry-based instruction and learning within classrooms, (b) increase student achievement in science, and (c) begin to narrow the achievement gap. This study focuses primarily on the second and third goal. While increases in student performance were expected with regards to science practices (as identified by NGSS), we were not sure whether achievement on science concepts (as identified by NGSS) would also follow. Although the measurement of the first goal is not the focus of this study, specific improvements in teacher practice that resulted due to the PD have been well documented in previous work (Marshall et al., 2011).

At least 60 % of the science teachers were required to participate from a given school before a partnership was established. Commitment included approximately 80 h per year of involvement, which included: (a) two full weeks of summer interactions, (b) four group follow-up meetings during the academic year, (c) four individual classroom observations, and (d) individual meetings with teachers to address personal issues surrounding challenges and transformation. Year 1 involvement focused more heavily on improving individual instructional practice. Year 2 and 3 participants continued to focus on improving individual practice, but participants were also expected to design and implement a school or district focused initiative that emphasized furthering inquiry-based instruction in their setting. PD experiences were modeled after well-regarded research concerning transforming teaching practices (Banilower, Boyd, Pasley, & Weiss, 2006; Darling-Hammond, 2000; Loucks-Horsley et al., 2010; Supovitz & Turner, 2000). Specifically, during the summer, teachers learned inquiry-based instructional approaches through personal experience with the 4E $\times$ 2 Instructional Model (Marshall, Horton, & Smart, 2009b); they created inquiry-based lesson exemplars in teams, used electronic quality of inquiry protocol (EQUIP) (see http://iim-web.clemson.edu/?page_id=166) to make planning more intentional; and they focused on improving content deficiencies (Marshall, Smart, & Horton, 2010, Marshall et al., 2011). In the follow-up sessions during the academic year, participants discussed successes and challenges relative to implementing lesson exemplars, targeted areas of improvement based on EQUIP indicators, and modified previous exemplars.

Participants and Schools

Five school districts, 11 schools, 74 teachers, and 9,981 students were involved in this study. Table 1 provides detailed information regarding the four teacher groups (non-participants, first year participants, second year participants, and third year participants) and school information. The student composition was 50.3 % male and 49.7 % female. Further, the self-reported student ethnicity was as follows: 49.7 % Caucasian, 30.4 % African-American, 7.4 % Hispanic, and 12.4 % other or unknown. Of the 11 schools, eight would be classified as suburban and three would be rural. All were within the same region of the state. High-needs/low-performing schools were the primary targets, and seven of the 11 met these criteria as defined by the state.

Table 1 Breakdown of schools and teachers included in the study

	Total	Non-participant[a]	1st year participant	2nd year participant	3rd year participant
Schools	11	7	11	8	6
Teachers	74	32	47	19	11

Second and third years are included in the count of first years; therefore, the total reflects the actual numbers of teachers and schools and not their times involved in the PD

[a] Non-participants = year 0

Instrument: MAP Science Test

The Measure of Academic Progress (MAP) science test published by Northwest Evaluation Association, NWEA, was used to measure student achievement and to determine whether or not students attained proficiency. MAP provides a reliable and valid assessment of student knowledge associated with science concepts and science practices and is used by schools in 48 states (Northwest Evaluation Association, 2004). As an adaptive test, MAP provides either more or less challenging items, depending on students' success or failure on previous questions.

MAP has several inherent strengths. Because test items are aligned to science standards, it has high predictive validity for state assessments (Cronin, Kingsbury, Dahlin, Adkins, & Bowe, 2007; Northwest Evaluation Association, 2005). Second, because it is adaptive, MAP provides a broader, more robust sample of the entire domain than a fixed-form test does (Northwest Evaluation Association, 2003). Third, teachers know the standards being tested but do not know the individual items; this eliminates issues such as test familiarity and teaching to specific test items.

Students took the MAP test in both the fall and spring. The fall score was used to establish a baseline metric and to ensure that the students from the various groups (non-participants, first, second, and third year) were similar from the beginning. The spring score was used to determine student growth and to determine which students had earned a proficiency rating by year's end.

MAP tests all domains of science (life, physical, earth, and space), so even though teachers do not teach all of these domains each year, the test actually shows a combination of growth in the domains taught and retention in the domain not taught. All grade level teachers taught the same core content, but the content taught by sixth-, seventh-, and eighth-grade teachers taught differed. Specifically, sixth-grade teachers taught weather and climate, energy, diversity of life; seventh-grade teachers studied matter and its interactions, living systems, heredity, and ecosystems; and eighth-grade teachers learned forces and motion, waves, Earth's place in the universe, Earth's system, and Earth's history.

Data Collection and Analysis

MAP science scores were collected from NWEA for each project year and then were consolidated into a single SPSS data file for later analysis. Proficiency can be determined in numerous ways (e.g., aligning to state standard proficiency or by a set cut score). This analysis utilizes both these examples. Math and reading MAP tests have regularly been aligned to the high-stakes state test, and in each case, a score aligned to the 40 % percentile relative to the MAP norm was found to align with proficiency for the state test. As such, we have adopted a 40 % percentile score as the proxy for proficiency for each portion of the MAP test used for this study (science composite, science concepts, science practices). Since not all students take the state high-stakes science test (half take science and half take social studies), we felt that using the alignment to math and reading was a reasonable proxy. Proficiency is individually determined for each student since the norm for each grade and each year is unique. It is important to note that although the Next Generation Science Standards are now being enacted in the classrooms of many states across the country, no national standardized test is available at this time. So although MAP is used as a proxy to reflect the new direction that the standards are moving, it is fair to say that the concepts and practices that are measured for this analysis would be closer to the alignment of the National Science Education Standards (NSES) (concepts and processes). Both NSES and NGSS seek to have the practices of science integrated in the learning of the content, but NGSS makes this more explicit.

An analysis of variance (ANOVA) was performed to determine whether the starting means were significantly different for the various intervention years. If significant differences were noted, then a post hoc was performed to determine where the differences occurred.

Next, a series of ANOVAs were performed on the spring data set (final student performance). The first examined whether the percentage of students earning proficient on the MAP science composite was a function of the intervention year of the teacher (non, first, second, and third). The second analysis considered the proficiency for students' of participants on the MAP science concepts only portion of the test with the third analyzing proficiency for only the MAP science practices. Further, ANOVAs were completed to study whether there was an effect of the intervention year of the teacher on the resulting student proficiency for Caucasians, African-Americans, and Hispanics. The same approach was used to see whether there were gender differences.

Levene's test to challenge the assumption of equal variances was performed for each ANOVA. If Levene's test was significant, Games–Howell post hoc was run for any significant ANOVAs to determine which pairwise comparisons were significant. Tukey's HSD was run in cases when Levene's test was not significant.

Finally, one-sample *t* tests were performed to see whether any of the minority groups were able to significantly narrow the achievement gap based on the intervention year.

Results

Intervention Groupings

In an effort to make sure that teachers associated with each intervention year (non, first, second, and third) were beginning with nearly equivalent groups of students, an ANOVA was conducted, and a significant difference was found between the starting scores of students in the fall on MAP science composite [$F(3, 19{,}822) = 12.528$, $p < .001$], MAP science concepts [$F(3, 9{,}837) = 7.188$, $p < .001$], and MAP science practices [$F(3, 9{,}981) = 6.046$, $p < .001$].

Since the pattern is the same for each of the three tests (science composite, science concepts, and science practices), the data that follow illustrate the mean differences for the fall MAP science composite scores. Using Games–Howell post hoc comparisons (Levene's test was significant), no statistical difference was noted

($p > .9$) between student starting scores (mean, SD) of non-participants (208.44, 11.61), first year participants (208.43, 11.33), and second year participants (208.87, 11.39). However, a significant difference was found between the starting scores of students of third year participants (210.09, 11.87) and non-participants ($p < .001$), first year participants ($p < .001$), and second year participants ($p < .01$). This indicates that all students except those from the classrooms of third year participants were starting from a similar place with the exception of third year participants' students who began significantly higher. Thus, the assumption is made that at the time of the initial fall test, students of non-participants, first year participants, and second year participants were nearly equivalent (based on starting mean scores) and that differences in the spring proficiency scores can be attributed at least in part to the first and second year teachers being involved in the intervention.

Student Proficiency

An ANOVA was performed to examine whether the percentage of students earning proficient on the MAP science composite is a function of the intervention year of the teacher (non, first, second, and third). Table 2 shows the means and standard deviations for each intervention year. The one-way ANOVA (see Table 3) of percentage of students earning proficient revealed a statistically significant main effect [$F(3, 19{,}822) = 28.147$, $p < .001$].

The test for homogeneity of variance was significant [Levene $F(3, 19{,}822) = 121.546$, $p < .05$], leading to Games–Howell post hoc comparisons being conducted to determine which of the pairings had significant mean differences. These results (see Table 4) indicate that students of teachers who were involved in first year, second year, and third year of the intervention, more often earned proficient scores when compared to those who were not part of the intervention. Moreover, the differences continued to be significant between groups in all cases except for between second year and third year participants of the intervention, which showed no statistical significance.

Table 2 Means and standard deviations for percentage of students earning proficient

Method	*n*	Mean	SD
Non-participant	5,547	68.2	46.57
First year	8,378	70.6	45.57
Second year	3,904	76.1	42.68
Third year	1,997	75.0	43.33
Total group	19,826	71.4	45.17

Table 3 Analysis of variance for students earning proficient

Source	SS	*df*	MS	*F*	*p*
Between	171,601.6	3	57,200.5	28.147	.000
Within	40,282,886.9	19,822	2,032.2		
Total	40,454,488.6	19,825			

Table 4 Games–Howell post hoc results and effect size of percentage of proficient by year of intervention

Group	Mean	Mean differences $(\bar{X}_i - \bar{X}_j)$ Effect size in parentheses			
		0	1	2	3
0. Non-participant	68.2	0			
1. First year participant	70.6	−2.36* (.052)	0		
2. Second year participant	76.1	−7.83*** (.18)	−5.47*** (.12)	0	
3. Third year participant	75.0	−6.75*** (.15)	−.4.38*** (.10)	1.09	0

* $p < .05$; *** $p < .001$

The MAP science composite test considers the aggregate of science concepts and science practices. This composite view of student achievement provides a summary of learning that unites concepts and practices which is the target for effective inquiry-based learning (Marshall, 2013). When the MAP science test is disaggregated into science concepts and science practices, a similar overall pattern is observed. Specifically, the one-way ANOVA of percentage of students earning proficient revealed a statistically significant main effect for the MAP science concepts test [$F(3, 9{,}837) = 12.059$, $p < .001$] and the MAP science practices test [$F(3, 9{,}981) = 15.713$, $p < .001$]. Further, the test for homogeneity of variances was significant for the MAP science concepts test [Levene $F(3, 9{,}837) = 50.712$, $p < .05$] and the MAP science practices test [Levene $F(3, 9{,}981) = 70.160$, $p < .05$]. Once again Games–Howell post hoc was used to determine whether individual pairings between intervention years were significant. Figure 1 provides a summary of student proficiency for all three MAP tests for each year of the intervention.

Proficiency Rates Based on Gender and Ethnicity

When considering gender and ethnicity proficiency rates, the third year participants were omitted from the analysis since these teachers' students began from a significantly higher starting point (based on fall MAP score). When studying gender, an ANOVA showed no significant difference between gender and proficiency rates for students of non-participating teachers [$F(1, 5{,}545) = 1.286$, $p = .257$], first year participants [$F(1, 8{,}376) = .875$, $p = .350$], and second year participants [$F(1, 3{,}902) = .568$, $p = 451$].

When studying ethnicity, an ANOVA showed significant differences among the three largest demographic student groups (Caucasian, African-American, and Hispanic) for non-participants [$F(2, 4{,}604) = 226.6$, $p < .001$], first year participants [$F(2, 7{,}283) = 344.2$, $p < .001$], and second year participants [$F(2, 3{,}688) = 153.8$, $p < .001$] (see Fig. 2). Significant growth was also noted for the proficiency rates of Caucasian, African-American, and Hispanic students between non-participants, first year, and second year participants. The only exception was a significant drop was observed for Hispanic students between non-participant and first year intervention. Note that one likely reason for the drop seen in year 1 with

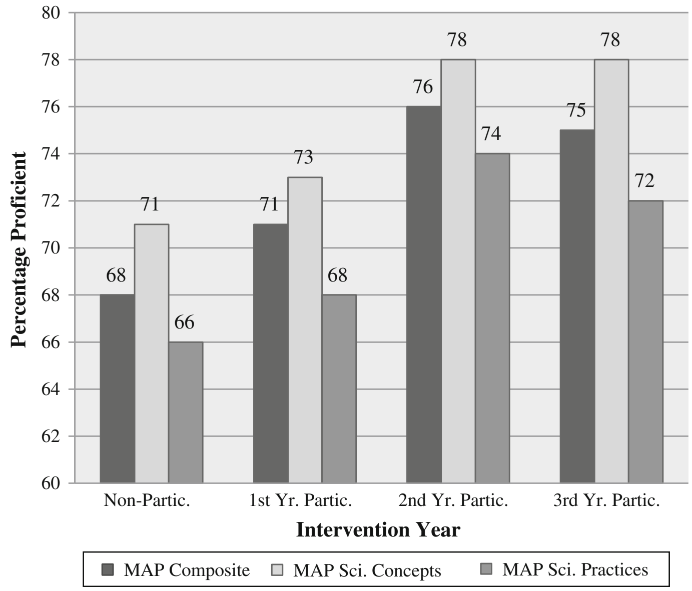
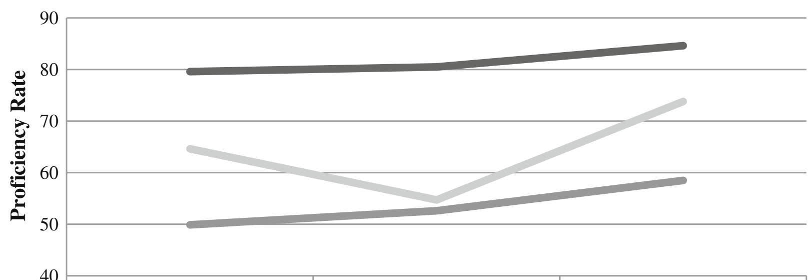


Fig. 1 Percentage scoring proficient on MAP versus intervention year of students' teachers



	Non-participant	1st Yr. Partic.	2nd Yr. Partic.
Caucasian	79.6	80.5	84.6
Hispanic	64.6	54.7	73.8
Afr. Amer.	49.9	52.6	58.5

Fig. 2 Percentage scoring proficient on MAP versus intervention year of teachers

Hispanic students was that one Title I school had a very high Hispanic population that began at a significantly lower starting mean. Thus, the final proficiency rates would be expected to likewise be lower. However, the second year intervention data show a significant improvement that exceeded the original non-participant starting point for Hispanic students. The net proficiency increase from non-participant to second year intervention for African-American students was 8.6 %, for Hispanic students was 9.2 %, for Caucasian students was 5.0, and 7.9 % when all ethnicities were compiled together.

Despite the increase in proficiency rates for all three major ethnic groups studied, with non-participants, African-American students began 29.7 % behind Caucasians, while Hispanic students began 15.0 % behind Caucasians. By the second year, the gap between African-American students' and Caucasian students narrowed to 26.2 % (3.5 % narrowing). Though African-American students grew more than Caucasian students, the 58.5 % (SD = 49.30) proficiency rate was not statistically higher than the expected 55.9 % expected if there were no change at all in the gap, $t(1, 135) = 1.78$, $p = .07$. Further, Hispanic student proficiency achievement gap to Caucasian students was narrowed to 10.8 % (4.2 % narrowing). Likewise, this 73.8 % (SD = 44.04) proficiency was not statistically different than the expected 69.6 %, $t(274) = 1.58$, $p = .11$.

Discussion

The project goal was to substantially improve the quality of inquiry-based instruction being led in middle school classrooms with the expectation being that this would lead to significant increases in student achievement for all demographic groups studied. This is one of few studies that we were aware of that actually link the PD to student achievement outcomes associated with inquiry-based instruction. While we expected learning to measurably increase when science practices were measured, we were not sure the degree to which science concepts would also be affected. The results show that the highlighted intervention was successful on many levels in terms of improving student proficiency. First, when comparing the proficiency rates of students of non-participating teachers with students of second year participants, all groups (collective, male, female, Caucasian, African-American, and Hispanic) showed significant growth (see Figs. 1, 2). Second, these increases were noted for all three MAP tests (science composite, science practices, and science concepts) (see Fig. 1). This suggests that when inquiry-based instruction effectively links learning with key conceptual ideas, which was an explicit goal of the PD intervention, learning gains can be seen in both the science practices (e.g., interpreting graphs and analyzing data) and the science concepts (e.g., understanding concepts such as energy, chemical interactions, or genetics). This aligns with the NAEP data that suggest that when inquiry-based instructional practices are frequently used in the classroom learning for all student groups results (Lauko et al., 2006; US Department of Education et al., 2011).

In the past, the explicit separation and wording of the standards encouraged teachers to teach scientific practices (or process skills) separate from science content (National Research Council, 1996). While this disjunctive treatment of instruction and learning may have been sufficient at the time, NGSS does not afford teachers the option of using such an approach while facilitating student mastery relative to the performance expectations (Achieve, 2013). For instance, students cannot discuss and plan a scientific study without being immersed in the content of the discipline. Inquiry-based instruction provides a solid means to achieve proficiency relative to the performance expectations set forth by NGSS (Achieve, 2013; National Research Council, 2012), and other state standards that emphasize having students engage in the learning of science through experiences that require, among other things, students to model complex ideas, plan scientific investigations to test ideas, communicate and justify ideas, and think critically and deeply about concepts.

As debate continues regarding how to narrow the achievement gap, a report from The Educational Trust suggests that perhaps the typical definition (closing the percentage among) is too simplistic (Rowen et al., 2010). Rather, a more comprehensive view includes considering four perspectives: (a) Gap between groups decreases over time, (b) all groups gain over time, (c) magnitude of current gap is minimized, and (d) group comparison in other districts, states, etc., shows improvement. Our study clearly demonstrates that all groups are gaining in proficiency over time (b), that the magnitude of the gap is decreasing (c), and that gains far exceed the non-participants (d). The gap is getting smaller (a), though not significantly so, but the greatest accomplishment is that all groups: male, female, Caucasian, African-American, and Hispanics grew significantly. Further, all of these groups grew in terms of both science practices and science concepts.

Because this study seeks to show that student achievement (specifically proficiency) increases when middle school teachers facilitate effective guided inquiry-based instruction, there are numerous limitations as well. First, as one moves further from the intervention (the PD experience), more variables potentially intervene in the final outcome (student achievement). Further, many of the components of the PD intervention have been detailed in *Succeeding with Inquiry in Science and Math Classrooms*, but this by itself will not ensure that the same results will occur with all teachers and their students. Future case studies may be helpful in determining which teachers were most successful and why as well as who struggled and why. MAP provides a solid assessment of student achievement in science in the middle grades, but it, nor any other national assessment, has fully been implemented to measure these new science practices and concepts.

Implications

We have known for quite some time that learning needs to engage the student, work through alternative conceptions, build on prior knowledge, and provide opportunities for students to explore conceptual ideas before the formal explanation occurs (National Research Council, 2000, 2012; Vygotsky, 1978). However, teachers have often struggled with how to move from a more commonplace, transmission of knowledge, framework to one that effectively engages the learner (Anderson, 2002). Further, the performance expectations from NGSS demand strategies such as inquiry-based instruction to allow students to model concepts, provide evidence to support claims, and design investigations to study complex phenomena.

The highlighted study provides one example to illustrate how we can begin to move our students to higher proficiency when proficient inquiry-based instruction becomes central to the classroom-learning experience. Though many questions still exist about how to improve teacher effectiveness and student achievement, this intervention showed that all student groups can be moved toward higher proficiency in middle school science classrooms. Specifically, this study shows that teachers begin to excel when PD is focused on sustained interactions/interventions that help to scaffold teacher practices from prior typically confirmatory activities or lecture to more intentional interactions that challenge students to model concepts, explain and justify thinking, and question ideas and the world around them. This intentionality of practice is predicated on getting teachers to facilitate a paradigm of having students explore (data, ideas, and concepts) before formal explanations occur. From this intervention, we have learned that it is not sufficient to just share and model the desired behavior. In order to change instructional practice, teachers must truly believe there is a need for change; they must have multiple opportunities to see and practice with a new instructional approach; and possibly most important of all, they must be provided with clear feedback and support as they scaffold from their own older approaches to newer more effective research-based approaches.

While there are admittedly many factors not discussed in this paper (e.g., beliefs, school, and cultural), that are also crucial for students to succeed, it was our goal to determine whether inquiry-based instruction could make a difference regardless of demographic group. It is our hope that these findings, when combined with the unstudied factors (e.g., beliefs, school, and cultural), will pave the way for educators to design better environments and strategies that can lead to maximized learning for all students. Perhaps using Hierarchical Linear Modeling will also help in the future to study the nested nature of students with different teachers.

Though this study focuses on in-service teachers, we have begun to transfer the learnings of this PD to our pre-service teacher preparation. Specifically, we have learned that pre-service teachers, like in-service teachers, need numerous semesters to become proficient and comfortable implementing effective guided inquiry-based instruction. As a result, our pre-service teachers now work with facilitating effective inquiry in four different courses as opposed to one in the past. The results of this change are still being studied.

Acknowledgments This material is based in part upon the work supported by the National Science Foundation under Grant No. DRL-0952160. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*, 95–135.

Achieve. (2013). *Next generation science standards*. Retrieved from http://www.nextgenscience.org/

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

Anderson, R. D. (2002). Reforming science teaching: What research says about inquiry. *Journal of Science Teacher Education, 13*, 1–12.

Banilower, E. R., Boyd, S. E., Pasley, J. D., & Weiss, I. R. (2006). *Lessons from a decade of mathematics and science reform: A capstone report for the local systemic change through teacher enhancement initiative*. Chapel Hill, NC: Horizon Research.

Bybee, R. W., Taylor, J. A., Gardner, A., Scotter, P. V., Powell, J. C., Westbrook, A., & Landes, N. (2006). *The BSCS 5E instructional model: Origins, effectiveness, and applications* (p. 49). Colorado Springs: BSCS.

Cronin, J., Kingsbury, G. G., Dahlin, M., Adkins, D., & Bowe, B. (2007). *Alternate methodologies for estimating state standards on a widely-used computerized adaptive test*. Paper presented at the national council on measurement in education, Chicago, IL.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Journal of Education Policy Analysis, 8*, 1–40.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*, 915–945.

Geier, R., Blumenfeld, P. C., Marx, R. W., Krajcik, J. S., Fishman, B., Soloway, E., & Clay-Chamber, J. (2008). Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. *Journal of Research in Science Teaching, 45*, 922–939.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job: The Hamilton Project White Paper 2006-11*. Washington, DC: The Brookings Institute.

Gross, P., Goodenough, U., Lerner, L., Haack, S., Schwartz, M., & Schwartz, R. (2005). *The state of the state science standards*. http://www.edexcellence.net

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy, 100*, 84–117. doi:10.1086/261808

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30*, 466–479.

Hartney, M. T., & Flavin, P. (2014). The political foundations of the black-white education achievement gap. *American Politics Research, 42*, 3–33.

Johnson, C. C. (2009). An examination of effective practice: Moving toward elimination of achievement gaps in science. *Journal of Science Teacher Education, 20*, 287–306.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*, 212–218. doi:10.1207/S15430421tip4104_2

Lauko, M. A., Grigg, W. S., & Brockway, D. (2006). *The nation's report card: Science 2005 (NSES 2006-466)*. U.S. Government Printing Office. Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main2005/2006466.pdf

Loucks-Horsley, S., Stiles, K. E., Mundry, S., Love, N., & Hewson, P. W. (2010). *Designing professional development for teachers of science and mathematics* (3rd ed.). Thousand Oaks, CA: Corwin Press.

Marshall, J. C. (2013). *Succeeding with inquiry in science and math classrooms*. Alexandria, VA: ASCD & NSTA.

Marshall, J. C. (2014). ASTE: A paradigm shift is underway—are you and your students ready? Paper presented at the NSTA, Boston, MA.

Marshall, J. C., Horton, B., Igo, B. L., & Switzer, D. M. (2009a). K-12 science and mathematics teachers' beliefs about and use of inquiry in the classroom. *International Journal of Science and Mathematics Education, 7*, 575–596. doi:10.1007/S10763-007-9122-7

Marshall, J. C., Horton, B., & Smart, J. (2009b). 4E x 2 instructional model: Uniting three learning constructs to improve praxis in science and mathematics classrooms. *Journal of Science Teacher Education, 20*, 501–516.

Marshall, J. C., Smart, J., & Horton, R. M. (2010). The design and validation of EQUIP: An instrument to assess inquiry-based instruction. *International Journal of Science and Mathematics Education, 8*, 299–321.

Marshall, J. C., Smart, J., Lotter, C., & Sirbu, C. (2011). Comparative analysis of two inquiry observational protocols: Striving to better understand the quality of teacher facilitated inquiry-based instruction. *School Science and Mathematics, 111*, 306–315.

Martin, M. O., Mullis, I. V. S., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 international results in science*. Chestnut Hill, MA: International Association for the Evaluation of Educational Achievement.

Minner, D. D., Levy, A. J., & Century, J. (2009). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching, 47*(4), 1–24.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academies Press.

National Research Council. (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. Washington, DC: National Academies Press.

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

Northwest Evaluation Association. (2003). *Technical manual*. Lake Oswego, OR: Author.

Northwest Evaluation Association. (2004). *Reliability and validity estimates: NWEA achievement level tests and measure of academic progress*. http://www.nwea.org

Northwest Evaluation Association. (2005). *NWEA reliability and validity estimates: Achievement level tests and measures of academic progress*. Lake Oswego, OR: Author.

Rennie, L. J., Goodrum, D., & Hackling, M. (2001). Science teaching and learning in Australian schools: Results of a national study. *Research in Science Education, 31*, 455–498.

Rowen, A. H., Hall, D., & Haycock, K. (2010). *Gauging the gaps: A deeper look at student achievement*. Washington, DC: The Education Trust.

Schmidt, W. H., McNight, C. C., & Raizen, S. A. (2002). *A splintered vision: An investigation of U.S. science and mathematics education*, from http://imc.lisd.k12.mi.us/MSC1/Timms.html

Span, C. M., & Rivers, I. D. (2012). Reassessing the acheivement gap: An intergenerational comparison of African American student achievement before and after compensatory education and the elementary and secondary education act. *Teachers College Record, 114*(6), 1–17.

Supovitz, J. A., & Turner, H. (2000). The effects of professional development on science teaching practices and classroom culture. *Journal of Research in Science Teaching, 37*, 963–980. doi:10.1002/1098-2736(200011)37:9<963::Aid-Tea6>3.0.Co;2-0

Tyler, R. (2003). A window for a purpose: Developing a framework for describing effective science teaching and learning. *Research in Science Education, 33*, 273–298.

U.S. Department of Education, Institute of Education Sciences, & National Center for Education Statistics. (2011). *The nation's report card, 2009–2011 science assessments*. Retrieved from http://nationsreportcard.gov/science_2011/

Van Klaveren, C. (2011). Lecturing style teaching and student performance. *Economics of Education Review, 30*, 729–739.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge: Harvard University Press.

Wilson, C. D., Taylor, J. A., Kowalski, S. M., & Carlson, J. (2009). The relative effects and equity of inquiry-based and commonplace science teaching on students' knowledge, reasoning, and argumentation. *Journal of Research in Science Teaching, 47*, 276–301.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement (issues and answers report, REL 2007 No. 033)*. Washington, DC: U.S. Department of Education, Regional Educational Laboratory Southwest.