Res Sci Educ
DOI 10.1007/s11165-011-9218-5

Project Clarion: Three Years of Science Instruction in Title I Schools among K-Third Grade Students

Kyung Hee Kim · Joyce VanTassel-Baska · Bruce A. Bracken · Annie Feng · Tamra Stambaugh · Lori Bland

Abstract The purpose of the study was to measure the effects of higher level, inquiry-based science curricula on students at primary level in Title I schools. Approximately 3,300 K-3 students from six schools were assigned to experimental or control classes (N=115 total) on a random basis according to class. Experimental students were exposed to concept-based science curriculum that emphasized 'deep learning' though concept mastery and investigation, whereas control classes learned science from traditional school-based curricula. Two ability measures, the Bracken Basic Concept Scale-Revised (BBCS-R, Bracken 1998) and the Naglieri Nonverbal Intelligence Test (NNAT, Naglieri 1991), were used for baseline information. Additionally, a standardized measure of student achievement in science (the MAT-8 science subtest), a standardized measure of critical thinking, and a measure for observing teachers' classroom behaviors were used to assess learning outcomes. Results indicated that all ability groups of students benefited from the science inquiry-based approach to learning that emphasized science concepts, and that there was a positive achievement effect for low socio-economic young children who were exposed to such a curriculum.

Keywords Science curriculum · Low income learners · Gifted · Inquiry-based learning · Concept development · Concept mapping · Primary age/level or early childhood

The continued issuance of national reports and studies highlights the central role of concept development and reasoning skills in learning. A recent national report underscored the importance of teaching science and math for deep conceptual understanding at the precollegiate level to ensure opportunities for advanced learning (National Research Council [NRC] 2002). The report acknowledged the central role of concept development in the process of learning science deeply, noting the principles of learning with understanding, using metacognitive strategies, building on prior knowledge, creating a community of

K. H. Kim (✉) · J. VanTassel-Baska · B. A. Bracken · A. Feng · T. Stambaugh · L. Bland
School of Education, The College of William and Mary, Room 3122, 301 Monticello Avenue, PO Box 8795, Williamsburg, VA 23187, USA
e-mail: kkim@wm.edu

Published online: 06 May 2011

learners to sustain modeled behavior, and differentiating between learners and their beliefs about learning. Current national standards also call for student understanding of general as well as specific science concepts (NRC 1996).

National and international assessment data also suggest that the nature of science learning in American classrooms is inadequate. The most recent National Assessment of Educational Progress report indicated that although 97% of the 17-year-olds understood some basic scientific principles, only 41% of these students demonstrated detailed scientific knowledge and could evaluate the appropriateness of scientific procedures (National Center for Education Statistics [NCES] 2001). The TIMSS-R study in 1999 showed no improvements in U.S. eighth grade achievement, which provided further evidence that U. S. student performance is low relative to international peers entering high school (NCES 2000).

Project Clarion, a 5-year scale-up Javits project in its fourth year of operation, was one response to these problems. The purpose of the project was to target low income, high ability learners and measure the effects of higher level, inquiry-based science curricula. Specific project objectives include: 1) implement instrumentation sensitive to low socio-economic learners for purposes of enhanced identification and assessment of learning, 2) write, implement, refine and extend research-based concept curriculum units of study in grades Pre-Kindergarten, one, two, and three, 3) develop and implement professional training models for teachers, administrators, and broader school communities, and 4) conduct research on short term and longitudinal student learning gains, as well as investigate the mechanisms that promote institutionalization of innovation through curriculum scaling up.

Literature Review

Research in cognitive science suggests that sound principles of teaching and learning may result in heightened student achievement. For example, analogical reasoning, metacognition, concept mapping, collaborative learning, and articulation of thinking all appear individually, and collectively, to contribute to enhanced science learning (e.g., Boyer et al. 2001; Kwon and Lawson 2000; Novak, 1998; Ziegler, 1995). Consistent with the assumptions of cognitive theory, much of the research on concept development has been domain-specific, with emphasis on mathematical conceptual understanding (e.g., Rittle-Johnson and Alibali 1999) and the teaching of science concepts (e.g., Kwon and Lawson 2000).

Research on teaching science concepts has continued to highlight the importance of learners' previous experiences as an important feature in conceptual understanding. For example, studies have shown that procedural knowledge structures (reasoning patterns) in science predict readiness for instruction in descriptive and theoretical concepts (Johnson and Lawson 1998; Kwon and Lawson 2000). Individual concepts also exist within complex conceptual systems such that knowledge and understanding of these concepts deepens by learning related concepts (Mintzes et al. 1998), a process that facilitates transfer of conceptual understanding from one domain to another. Wardekker (1998) suggested that scientific concepts are best taught through reflexive dialogue and in contexts where concepts can be demonstrated through and related to multiple applications. Research also shows that concept maps and analogies are useful instructional tools in helping students develop science concepts (Krajcik 1991; Novak, 1998; Pankratius 1990).

Studies on conceptual development and reasoning in science and mathematics have yielded important understanding about processes of learning in these domains, processes

that underpin this project. These include the insights that: perception precedes conceptualization, and conceptual understanding is built upon perceptual categorical representations in a foundational and continuous way (Quinn and Eimas 1997); the process used by good learners is a powerful resource in designing educational interventions (Boyer et al. 2001); self-explanation improves conceptual understanding (Chi et al. 1989; Pine and Messer 2000; Zeigler 1995); and conceptual change is grounded by conceptualizations, and engagement in generalized integration of information (Linn and Songer 1991). These research-based insights were employed to build an enhanced mathematical component to the project, one that integrates core science and math concepts in the Bracken curriculum (Bracken 1986) with the higher order concept of systems found in the William and Mary Problem Based Learning (PBL).

A review of early childhood projects also reveals the learning efficacy of intensive, high quality, and pervasive interventions with children between the ages of 4–8 (Ramey and Ramey 1998). Unlike programs that are ill-defined and primarily provide funding streams such as Title I or Head Start (Office of Child Development 1965), projects like the Perry Preschool Program (Schweinhart and Weikart 1983) and the Abecedarian project (Campbell and Ramey 1995) have utilized developmental timing, project intensity, direct provision of learning experiences, program breadth and flexibility, differentiation, and on-going use as principles of effectiveness in implementation. Success for All, a more contemporary program, has also shown reading and math gains when schools regrouped students by instructional level and provided follow-up tutoring services differentially (Borman and Hewes 2002).

Purpose of the Study

The purpose of this study was to assess the learning gains of the participants on relevant instruments. Specifically, the questions of interest were:

> Is there a difference between experimental and control group students in longitudinal effects on scores on the Metropolitan Achievement Test (8th ed., MAT-8, Harcourt Brace Educational Measurement 2000) and in their performance on the Test of Critical Thinking (TCT, Bracken et al. 2003)?

a. Are there differences between the two groups in longitudinal effects on the MAT or in their performance on the TCT scores by students' ethnicity, gender, and/or school?
b. If there is a difference in longitudinal effects on the MAT or in performance on the TCT scores among schools, do teachers' classroom behaviors, as measured by a science behavior teacher observation scale, affect students' performance?

Methods

Participants

As of the 2008–2009 school year, Year IV of Project Clarion, 3,307 students across three Title I school districts in the state of Virginia have participated in the project from a Rural (n=957 [29.0%]), an Exurban (n=1,413 [42.7%]), and a Suburban school district (n=937 [28.3%]). Within each school district, two Title I designated schools participated. A Title I school is a school in which low-income students make up the majority of student

participants. Teachers were randomly assigned to either the control group or the experimental group. In the rural district, experimental and control students were designated by school rather than classrooms. Within each building, classrooms of Kindergarten, first, second, and third graders participated in the study. In the exurban and suburban school districts, experimental and control groups were randomly assigned within schools.

Within the total group, there were similar numbers of male (n=1,642; 49.7%) and female (n=1,644; 49.7%) participants. By ethnicity, the student participants were: Caucasian (36.7%), Hispanic (24.4%), Asian (16.0%), African American (15.4%), Native American (0.5%), and Other (1.8%). Further, 1.5% of the total group was Limited English Proficient (LEP) students, and 2.8% students of the total group were in a special education program.

The number of participating districts increased in Year III of the project. In Year II, two school districts, an Exurban and a Rural school district, participated in the project. In Year III, a Suburban school district also joined the project.

Instrumentation

Five different instruments, used as part of the study, are reported on here. These include two ability measures (one verbal and one nonverbal), the Bracken Basic Concept Scale-Revised (BBCS-R, Bracken 1998) and the Naglieri Nonverbal Intelligence Test (NNAT, Naglieri 1991) for baseline information, a standardized measure of student achievement in science (the MAT-8 science subtest), a standardized measure of critical thinking (the TCT), and a measure for observing teachers' classroom behaviors. Additional performance-based instruments were also employed to assess specific curriculum outcomes in content, scientific process, and concept areas. However, results from these measures are not reported on in this paper.

The Bracken Basic Concept Scale-Revised (BBCS-R) The BBCS-R is designed to assess the basic concept development of children in the age of 2 years 6 months through 7 years 11 months. It is used to measure students' understanding of 308 basic language concepts distributed across 11 conceptual categories: Colors, Letters, Numbers/Counting, Sizes, Comparisons, Shapes, Direction/Position, Self-/Social Awareness, Texture/Material, Quantity, and Time/Sequence. It is individually administered, and the concepts are presented orally within the context of complete sentences and visually in a multiple-choice format (Bracken 1998). The BBCS-R is a revised instrument and has been reported to have high reliability with the internal consistency reliability of .98 (see Bracken 1998) and validity (see Bracken 1998; Bracken and Crawford 2006). The BBCS-R requires 30 min per student in individual administrations.

The Naglieri Nonverbal Intelligence Test (NNAT) The NNAT is a 38-item matrix analogy test with nonverbal content of spatial reasoning analogies. It has a 30 min-administration time and minimal verbal directions. It is available in multiple grade-based forms and administered in either a group or individual format. It has been reported to have adequate reliability (with the internal consistency reliability for raw scores of between .81 and .89.) and validity for screening and research purposes, and it has been widely used for the identification of gifted students (see Naglieri 1991).

The Metropolitan Achievement Test (8th ed., MAT-8) The MAT-8 is a well-known and widely used and validated group administered achievement test. Appropriate science portions of the MAT-8 were used for pre and post assessments of science understanding.

Students at grades 1 to 3 were administered the MAT-8, taking about an hour for each class. The KR-20 and KR-21 coefficients were greater than .80.

The Test of Critical Thinking (TCT) The TCT assesses children's critical thinking within seven life domains (i.e., social, affect, academic, competence, family, physical, and spiritual) using Paul's model of critical thinking (Bracken et al. 2003). Only 3rd graders were administered the TCT for graduating out of the program, which took approximately 45 min.

Teacher Classroom Observation Scale-Revised (COS-R) A researcher-designed adapted scale was used to assess teachers' classroom instructional efficacy in six dimensions and corresponding student responses. It was adapted from the Classroom Observation Scale-Revised (COS-R), (VanTassel-Baska et al., 2005). Scale indicators were aligned with the science curriculum in order to judge fidelity of the implementation in experimental teacher classrooms. The scale consisted of items related to teaching the content topics of the units, the scientific process, and the concept of change. For the scale, content validity was .98; internal consistency was .92; and inter-rater reliability was .89. The means for each item were calculated for the observed teachers and were categorized into three levels (below the 25th percentile, between the 25th percentile and the median, and above the median) based on the entire teachers' mean scores.

Identification The NNAT and the BBCS-R were administered to collect baseline data for identification and diagnostic purposes and information about the sample at varying points throughout the study. The NNAT was administered only once to all participating students, typically at the point of entry into the project. The NNAT produces standard scores with a mean of 100 (SD=15). Students with ability levels above 130 were labeled as high ability and those students who scored between 115 and 129 were considered promising learners. Those who scored between 100 and 114 were classified as typical learners and those who scored between 85 and 99 were classified as low end learners. Students who scored below 85 were classified as atypical. The percentage of students in the experimental and control group by category is presented in Table 1. For Year I, the experimental group (n=658, M=102.33, SD=20.79) and control group (n=496, M=101.17, SD=19.93) did not differ at baseline on the NNAT, $F(1, 1152)=0.915$, $p=.339$, *Partial* $\eta^2=0.001$.

However, assigning students to the same experimental or control group for Years II and III was not allowed due to school circumstances. In fact, the experimental group had statistically significantly lower mean scores than control group at baseline on the NNAT for both Year II, $F(1, 810)=4.753$, $p=.030$, *Partial* $\eta^2=0.006$, and Year III, $F(1, 835)=5.233$, $p=.022$, *Partial* $\eta^2=0.006$.

Table 1 Percentage of students by ability level group based on the NNAT scores

Group	NNAT	Control	Experimental
High ability	>130	5%	4%
Promising learner	115–129	14%	14%
Typical learners	100–114	30%	22%
Low end learners	85–99	27%	31%
Atypical	<85	24%	29%

The BBCS-R was individually administered at the beginning of each school year to randomly selected K-first graders to collect baseline data on basic concept attainment in order to diagnose needs for intervention. Due to teacher and principal objections in using instructional time for these assessments and the lack of project personnel to handle all the individual administrations of this instrument, the sampling plan was employed. For Year I, the experimental group (n=495) and control group (n=400) did not differ at baseline on the BBCS-R, $F(1, 893)=0.014$, $p=.906$, *Partial* $\eta^2=0.000$. The BBCS-R produces standard scores with a mean of 100 (SD=15). Students whose standard scores were below 100 were identified as students who had attained fewer basic concepts. These students were flagged to receive additional help in attaining basic concepts, central to benefitting from the Clarion curricula, which is designed for high ability learners. In order to facilitate this opportunity, a center for basic concepts was developed by project staff for each classroom.

The BBCS-R was also individually administered to K-first graders for Years II and III. Similar to the results at baseline on the NNAT for Years II and III, the experimental group had statistically significantly lower mean scores than control group at baseline on the BBCS-R for both Year II, $F(1, 344)=6.235$, $p=.013$, *Partial* $\eta^2=0.018$, and Year III, $F(1, 94)=9.837$, $p=.002$, *Partial* $\eta^2=0.095$.

Procedures

This project incorporated a quasi-experimental research design to measure the effects of inquiry-based science curriculum units on student achievement and critical thinking with a focus on developing the science talents of economically disadvantaged students in Title I schools. Year I of Project Clarion consisted of curriculum writing and securing school district participation. During Year II, the intervention curriculum was piloted in classrooms. Feedback on unit implementation was solicited, and revisions to the curriculum were made based on data from reviewers and teachers from Title I schools. Pre/post and baseline assessments were administered in control and experimental classrooms, and data were analyzed. Ongoing professional development also occurred formally for experimental teachers at least twice per year as well as informally through visitations to schools and job-embedded training based on teacher needs. Classroom observations were conducted for implementation fidelity and to guide professional development. Years III and IV continued the implementation of the curriculum intervention, ongoing professional development, and pre/post assessment data collection and analysis. The three year data collection shared in this paper was based on Years 2–4 of the project.

Intervention

All units designed for this intervention incorporated the Integrated Curriculum Model (ICM, VanTassel-Baska 1986) as the curriculum framework and empirical findings on effective instructional strategies for teaching science from the NRC (2002). Each unit includes an inquiry-based approach to learning science that focuses on an overarching concept of either change or systems with an emphasis on advanced processes that help students “do” science, which aims especially for early elementary school students and economically disadvantaged students. In each of the units, students take on the role of a scientist by learning the scientific process in order to answer a question or solve a real-world problem, which is also targeted towards younger students. All units integrate critical thinking and metacognition by emphasizing higher-level questions, science reflection

journals and prompts, and teacher-student discussion. The NRC (2005) has emphasized that the use of deliberate scaffolds to aid instruction are found to be especially effective for younger children.

The units have undergone multiple revisions. In Years I and II, 11 units were written and piloted in a variety of PreK, first, second, and third grade classrooms in Title I schools. Teacher feedback was solicited and major revisions were made that resulted in combining or omitting different units. At the end of Year III, eight units were significantly revised and underwent external reviews by content experts as well as solicitation of feedback from experimental teachers through focus groups and teacher journals. Changes to each unit were made, based on reviews from all sources.

During the development of the units, each was aligned to national and state standards. Recently, units were also aligned to the Virginia state science assessment (third grade Standards of Learning) as well as one of the standardized assessments administered: the MAT-8, science subtest.

Professional Development

Professional development, both on and off site, has occurred at least twice each year for the experimental teachers and administrators. The Project Clarion manager and co-principal investigators also made visits to each district to provide an overview of Project Clarion and hold informational meetings regarding implementation.

One unique feature of the Project Clarion professional development model was the use of project staff as facilitators for implementation in buildings. Project staffs, dubbed "ambassadors," were assigned to each school district to serve as a liaison between the school district and the grantee. The role of the ambassador was to assist teachers with Project Clarion unit implementation. Ambassadors provided support to teachers in a variety of ways that differentiated for each teacher and building, based on described or perceived needs. Ambassadors conducted informal inservices, showcased model lessons and strategies, or provided assistance, suggestions or guidance to teachers. Classroom observations were also conducted to judge intervention fidelity and to guide professional development throughout the year.

Results

Students' Performance in Science Understanding

The means and standard deviations for the pre and post MAT at Years I, II, and III for both the experimental group and the control group are presented in Table 2. For Year I, $F(1, 811)=0.78$, $p=.378$, and for Year III, $F(1, 267)=0.17$, $p=.683$, the results showed no statistically significant difference in the MAT posttest scores between the experimental and control groups. However, for Year II, the experimental group had a statistically significant higher mean score than the control group did, $F(1, 1097)=14.56$, $p<.001$, *Partial* $\eta^2=0.013$(a medium effect).

Students' Performance in Critical Thinking

The means and standard deviations for the TCT scores for both the experimental group and the control group are presented in Table 4. The results showed no statistically significant

Table 2 Means and standard deviations of the scores on the MAT-8 for control and experimental groups for years I, II, and III (N=2,182)

Year	Group test	Control (n=958)		Experimental (n=1,224)	
		$M(SD)$	n	$M(SD)$	n
I	Pre	19.7(4.5)	356	20.5(4.5)	458
	Post	22.1(4.7)		22.9(4.8)	
II	Pre	18.8(4.8)	470	19.7 (4.6)	630
	Post	20.6(4.8)		22.0(4.6)	
III	Pre	19.7(5.4)	132	20.8(5.5)	136
	Post	22.3(5.5)		23.4(5.6)	

difference in the TCT scores between the experimental and control groups for Year I, $F(1, 114)=0.24$, $p=.627$. However, the experimental group had a statistically significant higher mean score than the control group did for both Year II, $F(1, 276)=8.09$, $p=.005$, *Partial* $\eta^2=0.028$ (a medium effect), and Year III, $F(1, 401)=12.34$, $p<.001$, *Partial* $\eta^2=0.030$ (a medium effect).

Analyses Using Hierarchical Linear Modeling

We attempted to use a hierarchical linear modeling (HLM) analysis in order to account for the nesting of students in classrooms, in addition to the earlier analyses, because the hierarchical model takes into account the dependence among students within classrooms. A 3-level HLM (students: level 1; classrooms: level 2; schools; level 3) was not used due to there being only six schools. Moreover, the two schools in the suburban district joined Project Clarion later than the original four schools in the rural and exurban school districts. Thus, school and district information was included in level 2, which is the classroom level. A series of three 2-level multilevel models was conducted using HLM version 6.04 (Raudenbush et al. 2004) to investigate the extent of variation at Level 1 (students) and Level 2 (classroom) for Years I, II, and III post-MAT and Years II and III TCT scores, respectively.

The results of the HLM analyses indicated that Intervention was not a statistically significant predictor of post-MAT scores except for Year II, which are similar to the results of the ANCOVA analyses. However, the results of the HLM analyses indicated that Intervention was not a statistically significant predictor of TCT scores, which are different from the results of the ANOVA analyses. This may be because statistical power was much lower for the multilevel analysis than for the GLM analyses in which students were the unit of analysis. Further, because there were no data collected on characteristics of teachers, schools, or school districts for the level 2, no other meaningful results were obtained. The data sets for this study were not well set up for conducting a HLM analysis.

Students' Performance in Science Understanding Longitudinally

Assigning students to the same experimental or control group for Years II and III was not allowed due to the school circumstances as stated earlier. Thus, some of the experimental group students from Year I became members of the control group for Year II and/or Year III, and some of the control group students from Year I became members of the

experimental group for Year II and/or Year III. Therefore, the Intervention was divided into two groups for the purpose of the analysis: The experimental condition was for students who had been in an experimental group for 3 years, and the control condition was for students who had never been in the experimental group or had been in the experimental group only for 1 year. The zero to 1 year versus 3 year categorization was based on the conclusion that the use of higher-level reform of education programs tends to take at least 2 years of intensive teacher training in order to demonstrate the intended student outcomes of the reform (Borko et al. 1993).

Due to the late participation of the Suburban school district, students' longitudinal gains on the MAT-8 science test scores across 3 years reflect the results only of the Exurban and the Rural school districts. The means and standard deviations for the pre and post MAT scores at Year I, Year II, and Year III for both the experimental group and the control group are presented in Table 3.

In order to assess differences between the Interventions in students' longitudinal gains on the MAT-8 science test from Year I to Year III, a repeated measures analysis of variance (ANOVA) of pre and post MAT scores at Year I, Year II, and Year III by Intervention was conducted. The multivariate test results showed that the main effect for time is significant, Wilks's $\Lambda=0.30$, $F(5, 109)=50.70$, $p<.001$, *partial* $\eta^2=0.699$ (a large effect), indicating that the MAT scores were increased significantly over time.

The interaction of time*Intervention, Wilks's $\Lambda=0.87$, $F(5, 109)=3.27$, $p=.009$, *partial* $\eta^2=0.130$ (a medium effect), was significant, indicating that increasing MAT scores over time were influenced by time *Intervention. The experimental group students started with higher scores than the control group students on the pre test for Year I (even though there were no differences between the two groups on their baseline BBCS-R and NNAT scores). Both groups of students increased their scores on the post test for Year I. Then both groups of students decreased their scores on the pre test for Year II. The experimental group scores increased more than the control group scores on the post test for Year II. In Year III the experimental group students started with pre test scores similar to their Year II post test scores while control group students started with increased scores. Finally, both of the groups increased their scores on the post test for Year III.

Comparison by Intervention and ethnicity In order to assess differences between the Interventions and ethnicity in students' longitudinal gains on the MAT from Year I to Year III, a repeated measures ANOVA of pre and post MAT scores at Year I, Year II, and Year III by both Intervention and ethnicity was conducted. Neither of the interactions of time*ethnicity, Wilks's $\Lambda=0.82$, $F(20, 336)=1.01$, $p=.452$, nor time*Intervention*ethnic-

Table 3 Means and standard deviations of the scores on the MAT-8 for control (never or 1 year Clarion) and experimental (3 years clarion) groups for Years I, II, and III ($N=115$)

Year	Group test	Control ($n=63$) *M* (*SD*)	Experimental ($n=52$) *M* (*SD*)
I	Pre	19.3 (3.8)	21.6 (3.6)
	Post	21.4 (3.7)	23.4 (3.0)
II	Pre	18.7 (4.4)	19.7 (3.9)
	Post	20.1 (4.2)	23.3 (4.3)
III	Pre	20.8 (5.8)	21.7 (4.7)
	Post	23.2 (5.4)	24.2 (5.1)

ity, Wilks's Λ=0.80, F(20, 336)=1.20, p=.254, was significant, indicating that increasing MAT scores over time was not influenced by ethnicity or Intervention*ethnicity.

Comparison by Intervention and Gender The multivariate test results showed that neither of the interactions of time*gender, Wilks's Λ=0.97, F(5, 107)=0.77, p=.574, nor time*Intervention*gender, Wilks's Λ=0.94, F(5, 107)=1.32, p=.261, was significant, indicating that increasing MAT scores over time was influenced neither by the gender effect nor by the Intervention*gender effect.

Comparison by Intervention and School The multivariate test results showed that the interaction of time*school was not significant, Wilks's Λ=0.88, F(15, 290)=0.92, p=.0539, indicating that increasing MAT scores over time was not influenced by school. The interaction of time*Intervention*school was significant, Wilks's Λ=0.87, F(5, 105)=3.10, p=.012, *partial* η^2=0.129 (a medium effect), indicating that increasing MAT scores over time was influenced by interaction of Intervention*school.

Comparison by Intervention and teacher The results of a repeated measures ANOVA of MAT scores by Intervention and teacher showed a significant teacher effect, F(2, 53)=5.33, p=.008, *Partial* η^2=0.168 (a medium effect), but showed a non-significant Intervention*-teacher effect, F(2, 53)=1.25, p=.296. As Fig. 1 shows, when looking at the differences within each level of the classroom observation scores, both the teachers with scores above the median as well as the teachers with scores between the median and the 25th percentile had students who performed well in the experimental group. Among experimental teachers

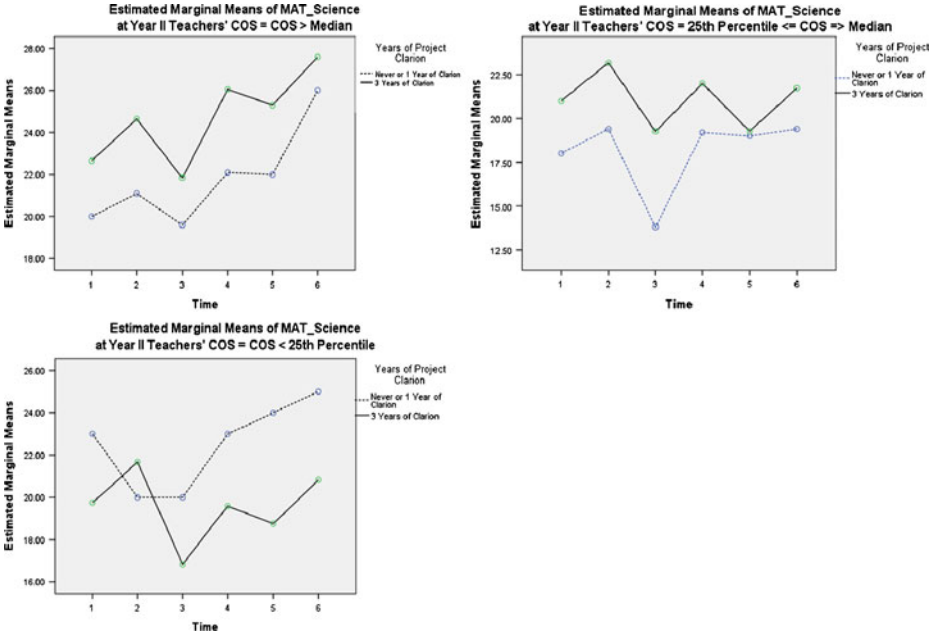

Fig. 1 Longitudinal growth of the scores on the MAT-8 for control (never or 1 year Clarion) and experimental (3 years Clarion) groups for students who had teachers with Classroom Observation Scale-Revised (COS-R) scores above the median, students who had teachers with COS-R between the median and the 25th percentile, and students who had teachers with COS-R below the 25th percentile (N=59)

with scores below the 25th percentile, however, the students in their classrooms did not perform well and actually did worse than the control group students.

The multivariate test results showed that the interaction of time*teacher was not significant, Wilks's $\Lambda=0.74$, $F(10, 98)=1.57$, $p=.126$, indicating that increasing MAT scores over time was not significantly influenced by the teacher.

The interaction of time*Intervention*teacher was not significant, Wilks's $\Lambda=0.80$, $F(10, 98)=1.18$, $p=.313$, indicating that increasing MAT scores over time was not influenced by the interaction of Intervention*teacher.

Students' Performance in Critical Thinking

Because no students took the TCT twice and participants consisted of three groups of third graders—one for each year—the scores were assembled into one variable with no year distinction and analyzed. Maintaining consistency for analyses using the MAT scores and TCT scores, the Intervention was divided into two groups: Experimental group students were those who had been in an experimental group for 3 years, and control group students were those who had never been in the experimental group or had been in the experimental group only for a year. An ANOVA was conducted with the TCT scores serving as a dependent measure and the Intervention serving as an independent variable. The means and standard deviations for the test scores for both the experimental group and the control group are presented in Table 4. The experimental group had a significantly higher mean score than the control group, $F(1, 192)=7.60$, $p=.006$, *Partial* $\eta^2=0.038$ (a medium effect).

Comparison by Intervention and ethnicity An ANOVA was conducted to examine any differences in scores on the TCT between experimental and control group students as well as among students' ethnicity. The results showed that neither the ethnicity effect, $F(4, 184)=2.22$, $p=.069$, nor the Intervention*ethnicity effect, $F(4, 184)=0.95$, $p=.439$, was significant, indicating that students' TCT scores were not influenced by either students' ethnicity or the interaction of Intervention*ethnicity.

Comparison by Intervention and Gender An ANOVA was conducted to examine any differences in scores on the TCT between experimental and control group students as well as between students' gender. The results showed that neither the gender effect, $F(1, 190)=0.28$,

Table 4 Means and standard deviations of the scores on the TCT for control and experimental groups for years I, II, and III ($N=797$) and for control (never or 1 year Clarion) and experimental (3 years Clarion) groups ($N=194$)

Group year	Control ($n=369$)		Experimental ($n=428$)	
	M (*SD*)	*n*	*M* (*SD*)	*N*
I	20.5 (6.3)	53	20.0 (5.6)	63
II	14.9 (5.1)	108	16.8 (5.8)	170
III	13.7 (5.5)	208	15.6 (5.4)	195
Combined*	14.3 (6.1)	106	16.7 (5.6)	88

Combined * ($N=194$) consisted of students who had no or 1 year Clarion (control group) and students who had 3 years Clarion (experimental group)

p=.600, nor the Intervention* gender effect, F(1, 190)=0.54, p=.465, was significant, indicating that students' TCT scores were not influenced by either students' gender or the interaction of Intervention*gender.

Comparison by Intervention and School An ANOVA was conducted to examine any differences in scores on the TCT between experimental and control group students as well as among students' schools. The results showed that the school effect was significant, F(5, 186)=3.98, p=.002, *partial* η^2=0.097(a large effect), indicating that students' TCT scores were significantly influenced by the school that students attended. Post hoc tests showed that the Rural-A school students performed significantly better than both of the Exurban-A and Exurban-B school students on the TCT. The Intervention* school effect was also significant, F(1, 186)=8.78, p=.003, *partial* η^2=0.045 (a medium effect), indicating that students' TCT scores were significantly influenced by the interaction of Intervention*-school: the Exurban-B school students performed significantly better in the experimental group than in the control group on the TCT.

Comparison by Intervention and teacher An ANOVA was conducted to examine any differences in scores on the TCT between experimental and control group students as well as among students' teachers. The results showed that neither the teacher effect, F(2, 95)= 1.91, p=.154, nor the Intervention*teacher effect, F(2, 95)=0.33, p=.722, was statistically significant, indicating that students' TCT scores were not significantly influenced by either the teacher effect or the Intervention*teacher effect. However, the students whose teachers' observation scores were above the median performed the best; the students whose teachers' observation scores were between the median and 25th percentile performed next best; and the students whose teachers' observation scores were below the 25th percentile performed the worst although the resultant effect was not statistically significant.

Discussion

Students' Performance in Science Understanding

The results of both ANCOVA and HLM analyses for Year I and Year III were not significantly different in the MAT-8 Science score gains between the experimental and control groups, which might be because Year I was a pilot study. It might be also because the experimental group had statistically significant lower mean scores at baseline on the BBCS-R and NNAT than the control group for Year III. Or, it might be because the use of multiple curriculum reform emphases tends to take at least 2 years to demonstrate the intended outcome of the reform (Borko et al. 1993), something that could not be well tested, given the conditions of this study. It is noteworthy that the experimental group had a significantly higher mean score than the control group did for Year II even though the experimental group had significantly lower mean scores at baseline on the BBCS-R and NNAT than the control group for Year II.

Students' Performance in Critical Thinking

For Year I, the results showed no significant difference in the TCT scores between the experimental and control groups, which might be because Year I was a pilot study or

because the use of multiple curriculum reform emphases tends to take at least 2 years to demonstrate the intended outcome of the reform (Borko et al. 1993), similar to the results of the MAT scores. The experimental group had significantly higher mean scores than the control group did for both Year II and Year III. Although the results of HLM analyses did not show a statistical significance, the results of the ANOVA analyses are noteworthy because the experimental group had significantly lower mean scores at baseline on the BBCS-R and NNAT than the control group for Years II and III.

Students' Performance in Science Understanding Longitudinally

Regarding longitudinal growth on the MAT, although there were no differences between the two groups on longitudinal growth on the MAT for Year I, the students who had Clarion for 3 years tended to increase scores significantly more than the students who never had Clarion or had Clarion for a year. This is especially noteworthy because the experimental group had statistically significant lower mean scores at baseline on the BBCS-R and NNAT than the control group for Years II and III. Students' ethnicity, gender, school, and teacher did not affect longitudinal growth on the MAT scores, except for the Intervention. However, among the students who had Clarion for 3 years, only the Exurban-A school did not increase their MAT scores, compared to the students who never had Clarion or had Clarion for a year.

On average for each pre and post MAT score set, the students who had Clarion for 3 years tended to perform better than students who never had Clarion or had Clarion for a year, regardless of their ethnicity. Although among the entire group of participants, Caucasian students performed better than Hispanic students while male students performed better than female students. However, among the students who had Clarion for 3 years, female students performed as well as male students on the MAT. Thus, Clarion participants equalized performance by ethnicity and gender.

On average, for each pre and post MAT score set, both the Rural-A and Rural-B school students performed better than the Exurban-A school students. In the Exurban-A school, students who had never had Clarion or had Clarion for a year performed better than the students who had Clarion for 3 years. Because of this inconsistent result with other findings, teachers' classroom teaching behaviors were examined, using the classroom observation scores. The students who had teachers whose teaching strategies were reported to be effective by the Clarion ambassadors tended to perform better than the students who had teachers whose teaching strategies were reported to be not as effective. Moreover, among the students who had teachers whose teaching strategies were reported to be ineffective, the students who had Clarion for 3 years performed not even as well as the students who never had Clarion or who had Clarion for a year. Thus, even when targeted interventions are implemented in schools, if the teachers in the schools are not effective, then it seems to be difficult to achieve the intended outcomes of the intervention. This is consistent with previous literature in that teacher effectiveness has been reported to be the primary determinant for students' progress (Sanders and Horn 1998). Sanders and Rivers (1996) have reported that students who had ineffective teachers for 3 years decreased their achievement on mathematics up to 54% regardless of their ability. Therefore, classroom instruction must be closely monitored in order to achieve the intended outcomes of an educational program. The ambassadors for Clarion assessed teachers' classroom behaviors and used the results of the evaluation for professional development. However, the frequency of the professional development may have been insufficient. Moreover, Kimball (2002) has reported that few teachers changed their instructional practice as a result of

evaluation, and furthermore, most teachers did not consider the evaluation process as an incentive to get professional development opportunities.

Glynn and Winter (2004) concluded that conditions that support the implementation of new teaching and learning strategies include teachers' collaborative interaction with students, high level activities in the lesson, and more importantly, teachers' sound classroom management techniques. Because several Clarion teachers were unable, in some classrooms, to implement the Bracken basic concept centers or to group students effectively based on pre-assessment results, students' growth may have been affected.

Comparison of Experimental and Control Group Students' Performance in Critical Thinking

Students who had Clarion for 3 years tended to perform better than students who never had Clarion or had Clarion for a year on their performance on the TCT, regardless of their ethnicity and gender. In terms of school, the Rural-A school students performed better than both the Exurban-A and Exurban-B school students on the TCT. However, the Exurban-B school students performed well particularly when they had Clarion for 3 years although the students did not perform as well as Rural-A school students due to the low performance of the students who never had Clarion or who had Clarion for a year. When students had Clarion for 3 years, the Exurban-A school students performed the lowest on the MAT, whereas the Exurban-B school students performed the highest on the TCT. The reason for low performance on the MAT among the students who had Clarion for 3 years in the Exurban-A school might be related to their teachers' lack of effectiveness. Similar results were also found in the students' performance on the TCT although the effect was not statistically significant. When looking at the average TCT score differences, students who had teachers whose teaching strategies were reported to be effective by the Clarion ambassadors tended to perform better than the students who had teachers whose teaching strategies were reported to not be as effective. Similar to the performance on the MAT, the students who had teachers whose teaching strategies were reported to be ineffective and who had Clarion for 3 years showed lower performance compared to the students who never had Clarion or had Clarion for a year.

Conclusion

The results of the Project Clarion intervention suggest that students in general benefitted from the science inquiry-based approach to learning that emphasized science concepts, and that there was a positive achievement effect for low socio-economic young children exposed to such a high-powered inquiry-based science curriculum. It would seem to suggest that templates for curriculum development might want to emphasize the role of concepts in developing science understandings and skills. Moreover, the use of high-powered curriculum with all students in low income schools is supported by this study. The results of the students' performance on the TCT would suggest that teaching science in this way could impact positively on critical thinking, even among primary age students.

Implications

The strongest implication of this study rests on the difficulty for universities in carrying out collaborative research in school districts. As with many other reform projects that have tried

to collect systematic data on student and teacher performance, this project suffered from a lack of follow through on the part of teachers, schools, and districts in recognizing the importance of implementing a predetermined design to which both parties had agreed. The inability and/or unwillingness of districts to keep experimental and control groups intact across 3 years in the targeted schools caused great difficulty in the analysis stage of the project. Moreover, the number and nature of assessments that needed to be collected and analyzed put a strain on the project staff as well. Individual assessments, for example, expended multiple days of graduate student time that had to be carefully calibrated to when the schools would allow students to be assessed and the general timeline of the project for implementation. Often, these pre and post assessments were done "just in time".

It is also fair to suggest that these mega Javits projects are trying to do all things for everyone, and in the process may not be targeting resources efficiently. Schools appear to need greater support in buildings for implementation than what can be realistically provided by project staff. Yet instructional coaches in buildings are already assigned to provide other types of support to teachers, and thus are often not available to supplement teacher needs in a special project such as Clarion. Because Title I schools in general underidentify their gifted population, the focus on curriculum designed for this group is often viewed with suspicion and seen as not essential for all learners. Because many schools do not cluster or group gifted or promising students in heterogeneous classrooms, the effects of a specially designed curriculum may be diminished for the very population of interest.

Moreover, because teachers at primary level often limit their science teaching time, the implementation of a project targeted in this content area and grade levels may expect to experience problems, both because of lack of teacher expertise and practice with the subject area and the attitude that it is less important than what is assessed on the state test. Although science became a tested subject in the third year of the project at third grade level, the impact on teachers of that reality was slow to take, especially for those teaching at the K-2 levels. It also would appear from the results of this project that random assignment of teachers, while empirically sound, does not result in a fair test of a complex innovation geared at top learners.

The professional development component of these projects also appears to be problematic from inception. Since we are using regular classroom teachers, randomly assigned, as the group to elevate gifted students' learning, there is an assumption that these teachers would have some knowledge and skill in working with these learners in school. Yet none of them in Project Clarion had the basic 12 h preparation for working with the gifted in their backgrounds. To think that these teachers could learn to work with this population effectively through differentiated opportunities in a subject area they also have limited background in as a result of 2 days of initial training and follow-up during implementation stretches credulity. If serious positive change is to be effected in these schools, then researchers need to have greater degrees of freedom to control implementation variables and to insist on school administrator engagement with monitoring classroom implementation.

Higher level thinking is not easy for the majority of the adult population. Some studies have suggested that not all teachers are capable of employing it on demand. Yet the gifted community is suggesting that all learners can progress to such higher levels of thinking, often in the absence of direct instruction, well-designed materials, an effective teacher, and sufficient practice in the most challenged schools in our nation. Lessons learned from this project suggest that the significant implementation problems that abound with such large scale studies ultimately compromise results in unacceptable ways.

Our modest results from this study would suggest that designing curriculum, training teachers, and implementing innovative instructional practices may not be worth the effort in schools where certain ground rules on implementation practices cannot be valued enough to be followed and taken seriously. Leadership at the building level must exert instructional leadership by holding teachers accountable for both what and how they are teaching, especially in a federally funded project for which they are receiving funding for implementation in addition to free professional development and materials. Otherwise, the rhetoric of using research-based approaches is empty. Educational writers who tout a no-group policy as research-based further muddies the waters for administrators who are responsible for the implementation of flexible grouping.

Future directions for research would suggest several caveats in designing studies. One major issue is the choice of instrumentation that will be sensitive to the nature of reform-based curriculum. While standardized achievement tests have excellent psychometric properties, they rarely align well with curriculum designed to address higher level skills and concepts, causing an underestimation of student learning in critical areas. Another issue for researchers to tackle is examining the affective response to innovative science curriculum by probing young children's enhanced interest in science as a result of participation in such a program and their enhanced attitudes toward doing science in school. Finally, researchers on similar types of studies should recognize the "slippery slope" of school-based practices and enter the arena fully prepared to confront implementation problems.

References

Borko, H., Mayfield, V., Marion, S., Flexer, R., & Cumbo, K. (1993). Teachers developing ideas and practices about mathematics performance assessment: successes, stumbling blocks, and implications for professional development. *Teaching and Teacher Education, 13*, 259–278.

Borman, G. D., & Hewes, G. M. (2002). The long-term effects and cost-effectiveness of success for all. *Educational Evaluation and Policy Analysis, 24*, 243–266.

Boyer, P., Bedoin, N., & Honore, S. (2001). Relative contributions of kind- and domain-level concepts to expectations concerning unfamiliar exemplars: developmental change and domain differences. *Cognitive Development, 15*, 457–479.

Bracken, B. A. (1986). *Bracken concept development program*. San Antonio: The Psychological Corporation.

Bracken, B. A. (1998). *Bracken basic concept scale-revised*. San Antonio: Harcourt Assessments.

Bracken, B. A., & Crawford, E. (2006, June). *Project Clarion: A concept-based science curriculum*. Paper presented at the National Association for the Education of Young People 15th National Institute for Early Childhood Professional Development, San Antonio, TX.

Bracken, B. A., Bai, W., Fithian, E., Lamprecht, S., Little, C., & Quek, C. (2003). *Test of critical thinking*. Williamsburg: The College of William and Mary, Center for Gifted Education.

Campbell, F. A., & Ramey, C. T. (1995). Cognitive and school outcomes for high-risk African-American students at middle adolescence: positive effects of early intervention. *American Educational Research Journal, 32*, 743–772.

Chi, M. T. H., Hutchinson, J. E., & Robins, A. F. (1989). How inferences about novel domain-related concepts can be constrained by structural knowledge. *Merrill-Palmer Quarterly, 35*, 27–62.

Glynn, S. M., & Winter, L. K. (2004). Contextual teaching and learning of science in elementary schools. *Journal of Elementary Science Education, 16*(2), 51–63.

Harcourt Brace Educational Measurement. (2000). *Metropolitan achievement test* (8th ed.). San Antonio: Harcourt Assessment.

Johnson, M. A., & Lawson, A. E. (1998). What are the relative effects of reasoning ability and prior knowledge on biology achievement in expository and inquiry classes? *Journal of Research in Science Teaching, 35*, 89–103.

Kimball, S. M. (2002). *Analysis of feedback, enabling conditions, and fairness perceptions*. University of Wisconsin, Wisconsin Research in Education.

Krajcik, J. S. (1991). Developing students' understanding of chemical concepts. In S. M. Glynn, R. H. Yeany, & B. K. Britton (Eds.), *The psychology of learning science: International perspective on the psychological foundations of technology-based learning environments* (pp. 117–145). Hillsdale: Erlbaum.

Kwon, Y., & Lawson, A. E. (2000). Linking brain growth with the development of scientific reasoning ability and conceptual change during adolescence. *Journal of Research in Science Teaching, 37*, 44–62.

Linn, M. C., & Songer, N. B. (1991). Cognitive and conceptual change in adolescence. *American Journal of Education, 99*, 379–417.

Mintzes, J. J., Wandersee, J. H., & Novak, J. D. (1998). *Teaching science for understanding*. San Diego: Academic.

Naglieri, J. A. (1991). *Naglieri nonverbal ability test*. San Antonio: Harcourt Assessments.

National Center for Education Statistics. (2001). *The nation's report card: Science 2000*. Washington: United States Department of Education, Office of Educational Research and Improvement.

National Center for Education Statistics. (2000). *Pursuing excellence: Comparisons of international eighth-grade mathematics and science achievement from a U.S. perspective, 1995 and 1999*. Washington: United States Department of Education.

National Research Council. (1996). *National science education standards*. Washington: National Academy Press.

National Research Council. (2002). *Learning and understanding: Improving advanced study of mathematics and science in U.S. high schools*. Washington: National Academy Press.

National Research Council. (2005). *How students learn: History, mathematics, and science in classroom*. Washington: National Academy Press.

Novack, J. D. (1998). *Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations*. Mahwah: Erlbaum.

Office of Child Development. (1965). *Recommendations for a head start program by a panel of experts*. Washington: U.S. Department of Health, Education, and Welfare.

Pankratius, W. J. (1990). Building an organized knowledge base: concept mapping and achievement in secondary school physics. *Journal of Research in Science Teaching, 27*, 315–333.

Pine, K. J., & Messer, D. J. (2000). The effect of explaining another's actions on children's implicit theories of balance. *Cognition & Instruction, 18*, 35–51.

Quinn, P. C., & Eimas, P. D. (1997). A reexamination of the perceptual-to-conceptual shift in mental representations. *Review of General Psychology, 1*, 271–287.

Ramey, C. T., & Ramey, S. L. (1998). Early intervention and early experience. *American Psychologist, 53*, 109–120.

Raudenbush, S., Bryk, A., Cheong, Y. E., Congdon, R., & du Toit, M. (2004). *HLM6: Hierarchical linear and non-linear modeling*. Lincolnwood: Scientific Software International.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural understanding: does one lead to the other? *Journal of Educational Psychology, 91*, 175–189.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee value-added assessment system (TVASS) database: implications for educational research and evaluation. *Journal of Personnel Evaluation in Education, 12*, 247–256.

Sanders, W. I., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Schweinhart, L. J., & Weikart, D. P. (1983). *The effects of the Perry preschool program on youths through age 15: A summary*. In *Consortium for longitudinal studies, as the twig is bent—lasting effects of preschool programs*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

VanTassel-Baska, J. (1986). Effective curriculum and instructional models for talented students. *Gifted Child Quarterly, 30*, 164–169.

VanTassel-Baska, J., Quek, C., & Feng, A. (2005). *The classroom observation scale-revised*. Williamsburg: The College of William and Mary, Center for Gifted Education.

Wardekker, W. L. (1998). Scientific concepts and reflection. *Mind, Culture, and Activity, 5*, 143–153.

Zeigler, E. F. (1995). Competency in critical thinking: a requirement for the "Allied Professional". *Quest, 47*, 196–211.