## Abstract Title Page

***Not included in page count.***

**Title:**

The Relative Effects of Inquiry-Based and Commonplace Science Teaching on Students' Knowledge, Reasoning and Argumentation about Sleep Concepts: A Randomized Control Trial

**Author(s):**

Christopher D. Wilson, Joseph A. Taylor, Susan M. Kowalski, and Janet Carlson

BSCS Center for Research and Evaluation

## Abstract Body

### Background/context:

From Dewey to the Standards, inquiry has been an increasingly prominent theme in multiple science education reform movements, yet the transition from theory and advocacy to practice and policy has been disappointing. The paradox of educational reform without change is not exclusive to the sciences (Cuban, 1988; Woodbury & Gess-Newsome, 2002), but it is nevertheless surprising that such a sustained and largely consistent drive for reform has had such little impact on teacher practice. Two recent large scale studies from Horizon Research Inc. (Weiss et al., 2003, and Hudson, McMahon & Overstreet, 2002) highlight the uncommonness of inquiry-based teaching in the United States. From classroom observations and interviews with 364 science and mathematics teachers, Weiss et al. (2003) found that inquiry was a focus of only 2% of science lessons in grades 9-12. This finding mirrors those in a survey of 5278 teachers (Hudson, McMahon & Overstreet, 2002), in which teaching practices and student objectives characteristic of inquiry consistently occurred with less frequency and emphasis than traditional teaching methods and learning goals.

Many barriers to implementing inquiry in a manner consistent with the vision of the National Science Education Standards have been described in the literature (Anderson, 2002), among the most the recent of which is the *No Child Left Behind* (NCLB) legislation (US Department of Education, 2002). The NCLB act and the associated accountability movement has led to an increased emphasis on standardized testing to measure teacher and school effectiveness, which in turn, some have argued (see Blanchard et al. 2008), has resulted in teaching practices that are at odds with those advocated in the national science education reform documents (e.g. AAAS, 1993, 2000; NRC, 1996, 2000). NCLB and the current climate therefore present one further obstacle to inquiry's role in reform, in that accountability and inquiry-based teaching can appear incompatible to teachers (Blanchard et al. 2008).

While No Child Left Behind and the accountability movement have resulted in a shift in the way we assess teacher and school effectiveness, they have also resulted in a shift in the expectations for evidence in education research. Federal policies have begun to advocate *evidence-based reform* – in which the adoption of programs or practices is based on rigorous research conducted with methods derived from the medical and pure sciences; particularly randomized experiments (Slavin, 2008). We are therefore met with a challenge. If, within the climate of accountability and evidence-based reform, the cumulative vision of a century of science education reform is to be revealed in the transition of inquiry-based teaching from theory and advocacy to practice and policy, the question inevitably becomes: *what is the evidence that demonstrates the effectiveness of inquiry-based teaching, and what is the nature of that evidence?*

While there is a growing body of research which suggests that student understanding is enhanced by inquiry-based teaching, only recently have studies begun to use experimental designs. From the perspective of the accountability movement, the evidence for the effectiveness of inquiry-based teaching can therefore only be seen as inconclusive. There have also recently been challenges to inquiry-based instruction (Chen and Klahr, 1999; Klahr and Nigam, 2004; Kirshner, Sweller & Clark, 2006) that have received much attention .In this study, we address

this ambiguity by employing the methods of “scientifically-based research” (Slavin, 2008, Shavelson & Towne, 2002) to examine differences between the learning gains of students who receive inquiry-based instruction, and students who receive instruction on the same content, but in an instructional unit designed around commonplace teaching practices. Since significant achievement gaps by gender, race/ethnicity, and socioeconomic status remain (Clewell & Campbell, 2002) despite the long-standing call for science for all Americans, we disaggregate data by various student variables to examine if inquiry-based instruction can provide equitable science education. We also measure three different goals of science education: scientific knowledge; reasoning with scientific models; and construction/critique of scientific explanations. As described below, we use the Horizon Research Inc. survey and interview data (Weiss et al., 2003 and Hudson, McMahon & Overstreet, 2002) to define “commonplace teaching”, and use The Biological Sciences Curriculum Study (BSCS) 5E instructional model, or the “5Es” (Bybee, 1997) to organize the inquiry-based unit.

**Purpose/objective/research question/focus of study:**

1. To what extent can differences in student learning between the inquiry-based and commonplace groups be attributed to randomized group assignment?

2. What differences in achievement by treatment group exist specific to the learning goals of knowledge, reasoning, and argumentation?

3. Does student race/ethnicity, gender, or socio-economic status account for variation in posttest scores above and beyond variation accounted for by pretest scores and group assignment?

**Setting:**

Both groups of students were taught during the summer in a controlled classroom setting by the same teacher. The students were unaware of the purpose of the study, their group assignment, and as much as was possible, the existence of the other treatment group. To remove the possibly confounding effects of multiple teachers, both units were taught by the same teacher in a controlled classroom setting. The teacher selected for this study had 27 years of experience teaching in public schools, a Ph.D. in curriculum and instruction, and experience teaching with a wide range of traditional and inquiry-based materials.

**Population/Participants/Subjects:**

Sixty students were recruited and randomly assigned to a group that would receive inquiry-based instruction organized around the 5Es, or a group that would receive instruction on the same content but organized around commonplace teaching practice (as defined by the Horizon data). With respect to gender, race, age, and free and reduced lunch status, no significant differences were present in the make up of each of two treatment groups *(Insert Table 1 Here)*. The study participants came from 24 schools from seven districts from across a range of urban, suburban, and rural areas; five of the students attended private schools and two were home-schooled.

**Intervention/Program/Practice:**
Both units were built around the *Sleep, Sleep Disorders, and Biological Rhythms* unit from the NIH Curriculum Supplement Series (BSCS, 2003). All students participated in 14 hours of instruction and testing over the course of two weeks in the summer.

*The Commonplace Unit.* The two research documents from Horizon Research, Inc. described previously (Weiss et al., 2003; Hudson et al., 2002) were used to help establish commonplace teaching practice. Each of the lessons in the NIH sleep unit was modified to reflect the frequency of teaching practices illustrated by patterns in the data from the Horizon Research, Inc. studies. To reflect commonplace practices, changes were also made to the order of the lessons, as well as to the connections between the lessons. Rather than merely focusing on didactic approaches to teaching, the commonplace unit included strategies and activities such as group work and experiments in the same frequency as the survey and interview data.

*The Inquiry-Based Unit.* Despite the original NIH sleep unit being organized around the BSCS 5Es, the unit was reviewed to insure consistency with teaching science as inquiry within the BSCS 5E model. A small number of changes were made to more fully represent the BSCS 5E Instructional Model and the processes of scientific inquiry.

Both sets of materials were reviewed and revised by expert curriculum developers to insure that while the instructional approaches differed, the learning goals remained the same.

**Research Design:**
Since one goal of this study was to make causal inferences about the effectiveness of inquiry-based teaching, an experimental design (randomized control trial) was used.

**Data Collection and Analysis:**

**Data sources: *1. Pretest and Posttest***
The pretest was identical to the posttest, and contained 4 multiple-choice items, 8 true or false items, and 5 constructed response items. The objective items were designed to focus on ‘facts’ and vocabulary about sleep, while the constructed response items required students to apply scientific models of sleep behavior to reasoning about data presented in new contexts. The mean item difficulty for the dichotomous items was 0.789, with the total test having a reliability index (Cronbach’s alpha) of 0.695. All items were included in the analysis since each had a positive discrimination index. To score the constructed response items we needed a set of levels representing increasingly sophisticated ways of reasoning with scientific models of sleep behavior. The process of developing these levels was modeled after Chen, Mohan and Anderson (2008), as were the initial notions of the levels themselves. Table 2 shows the 5 levels that were used to score the constructed response items. Since four raters each scored one fourth of the total set of pretests and posttests, inter-rater reliability needed to be calculated to ensure consistency in scoring between raters. A sample of >10% (n=12) of the tests were scored by all four raters, and inter-rater reliability was calculated using the intra-class correlation coefficient ($\alpha = .78$).

We used hierarchical regression (ordinary least squares) to address the questions posed in this study. Since students came to the clinical setting of the study from a variety of classes and schools in the area, nesting (multicollinearity) at the class or school level was not a factor and

multi-level modeling was not necessary. The first factor in the analysis was a student's pre-test score. Based on theoretical considerations and research (Muller, Stage, & Kinzie, 2001), we chose to test each factor in turn in the following order: pre-test score, FRL status, race/ethnicity, and gender.

**Data sources:** ***2. Clinical Interviews***
A thirty-minute standardized open-ended interview protocol was developed around the topics of sleep behavior, circadian rhythms and the biological clock. During these interviews, students were presented with sleep data in the form of actograms – representations that were not used during either instructional unit. Based on the data in the actograms, students were guided through the construction of explanations that included environmental and physiological explanations for the observed data; asked for alternative explanations for their observations; and asked to critique given explanations for the patterns in the data. Each interview was recorded on video. As a framework for scoring the interviews, we began with the modification of Toulmin's argumentation model presented by McNeill et al. (2006), in which students' explanations are scored according to the quality of their claim ("an assertion or conclusions that answers the original question"), evidence ("scientific data that supports the claim") and reasoning ("a justification that shows why the data count as evidence to support the claim"). A sample of students' interviews was scored by the same four raters as the pretests and posttests, and the extent of scoring agreement between raters was evaluated and discussed. Minor changes were made to the rubric based on these discussions, and rules for scoring certain types of responses were developed.

**Data sources:** ***3. CLES and RTOP***
Each class session was observed by two external evaluators, who completed the Reformed Teaching Observation Protocol (RTOP) for each unit. The master teacher also took extensive notes after each lesson, recording his pedagogical moves and differences between his teaching in the two units. Each class session was recorded on video. At the end of the unit, all students completed a survey containing a subset of 17 items from the Constructivist Learning Environment Survey (CLES). The CLES scores for the Inquiry-based Unit were significantly higher than the scores for the Commonplace Unit [$t(55) = 3.195$, $p < 0.01$], as were the scores for each of the RTOP subscales (for each $p < 0.01$).

**Findings/Results:**

***Total Test Scores***
Students in the Inquiry group had significantly higher posttest scores than students in the Commonplace group [$F(1,55) = 4.662$, $p < 0.05$], controlling for variance in the students' pretest scores. The effect size (Cohen's d) for this difference was 0.27. Figure 1 shows the effects of the two instructional units on student learning.

***Level 5 Understanding***
Of the five levels used to score the constructed response items, Level 5 (*Model-based accounts connected across scales*) represents the type of reasoning that is a desirable goal of secondary science education (Chen, Mohan & Anderson, 2008). As such, we next examined how the achievement of Level 5 reasoning differed between the students in the Commonplace and

Inquiry-based groups. Students in the Inquiry-based group gave a significantly higher fraction of responses at Level 5 than students in the Commonplace group, [$F(1,56) = 2.746$, $p < 0.05$], controlling for variance in the students' pretest scores via analysis of covariance. The effect size (Cohen's d) for this difference was 0.42. Figure 2 shows the effects of the two instructional units on the frequency of Level 5 accounts.

***Examining Gaps in Student Achievement by Race, Gender and Socio-Economic Status***

In the calculation of F-change statistics for the hierarchical regression, only group assignment contributed to the model above and beyond pre-test score. FRL status, race/ethnicity, and gender did not account for variation in post-test scores above and beyond other factors. Students in the inquiry group demonstrated significantly better learning from pretest to posttest and the posttest scores were *not dependent* on a student's membership in any specific demographic group. We further examined scores on the pretest and posttest via one-way ANOVA. The only significant difference in scores between white and non-white students was on the posttest for the students in the commonplace unit [$F(1,27) = 5.530$, $p = 0.026$] indicating that the commonplace science teaching led to a significant achievement gap by race, whereas the inquiry-based instruction did not.

***Interviews and Argumentation***

Analysis of the argumentation scores from the standardized interviews showed that students in the Inquiry group had significantly higher scores for Claims [$F(1,54) = 4.253$, $p < 0.05$], Evidence [$F(1,54) = 9.794$, $p < 0.01$] and Reasoning [$F(1,54) = 5.051$, $p < 0.05$] than students in the Commonplace group, controlling for variance in the students' pretest scores via analysis of covariance. The effect sizes (Cohen's d) for each difference were 0.48, 0.61 and 0.48 respectively.

**Conclusions:**

Using scientifically-based research methods that meet the standards required by the evidence-based reform movement to establish causality, this study found that students in an inquiry-based classroom reached significantly higher levels of achievement than students experiencing commonplace teaching. The superior effectiveness of the inquiry-based instruction was consistent across a range of learning goals (knowledge, scientific reasoning, and argumentation) and types of measures (dichotomous items, open-response items, and clinical interviews). This study therefore contributes to the growing body of evidence demonstrating the effectiveness of inquiry-based teaching; supports the claims about inquiry in national science education reform documents (e.g. AAAS, 1993, 2000; NRC, 1996, 2000); and refutes the claims made by Kirshner, Sweller & Clark (2006) in response to the findings by Klahr and colleagues (Chen & Klahr, 1999; Klahr & Nigam, 2004).

Since students in the inquiry-based group outperformed students receiving commonplace instruction on each of the knowledge, scientific reasoning, and argumentation measures, this study provides evidence that teachers need not compromise the quality of their teaching to see increases in student achievement in an age of accountability.

# Appendixes

## Appendix A. References

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

American Association for the Advancement of Science (AAAS). (2000). Designs for science literacy. New York: Oxford University Press.

Blanchard, M. R., Annetta, L. A., & Southerland, S. A. (2008). Investigating the effectiveness of inquiry-based versus traditional science teaching methods in middle and high school laboratory settings. Paper presented at the annual conference of the National Association for Research in Science Teaching, Baltimore, MA.

BSCS. (2003). Sleep, sleep disorders, and biological rhythms. Bethesda, MA: National Institutes of Health (NIH) (NIH Publication No. 04-4989).

Bybee, R. W. (1997). Achieving scientific literacy: From purposes to practices. Portsmouth, NH: Heinemann Educational Books.

Chen, Z., & Klahr, D. (1999). All other things being equal: Children's acquisition of the control of variables strategy. Child Development, 70(5), 1098–1120.

Chen, J., Mohan, L., & Anderson, C. W. (2008). Developing a K-12 learning progression for carbon cycling in socio-ecological systems. Paper presented at the 2008 annual meeting of the National Association for Research in Science Teaching, Baltimore MD.

Clewell, B. C., & Campbell, P. B. (2002). Taking stock: Where we've been, where we are, where we're going. Journal of Women and Minorities in Science and Engineering, 8, 255–284.

Cuban, L. (1988). Constancy and change in schools (1880s to the present). In P. W. Jackson (Ed.), Contributing to educational change (pp. 85–105). Berkley, CA: McCutchan.

Hudson, S. B., McMahon, K. C., & Overstreet, C. M. (2002). The 2000 national survey of science and mathematics education: Compendium of tables. Chapel Hill, NC: Horizon Research.

Kirschner, P.A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry based teaching. Educational Psychologist, 41, 75–86.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. Psychological Science, 15, 661–667.

McNeill, K. L., Lizotte, D. J, Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. The Journal of the Learning Sciences, 15(2), 153–191.

Muller, P. A., Stage, F. K., & Kinzie, J., (2001). Science achievement growth trajectories: Understanding factors related to gender and racial-ethnic differences in precollege science achievement. American Educational Research Journal, 38, 981–1012.

National Research Council (NRC). (1996). National science education standards. Washington, DC: National Academic Press.

National Research Council (NRC). (2000). Inquiry and the national science education standards. Washington, DC: National Academic Press.

Shavelson, R. J., & Towne, L. (Eds.). (2002). Scientific research in education. Committee on Scientific Principles for Education Research. Division on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. Educational Researcher, 37, 5–14.
U.S. Department of Education. (2002). No Child Left Behind: A desktop reference. Washington, DC: Available from http://www.ed.gov/offices/OESE/reference.
Weiss, I. R., Pasley, J. D., Smith, P. S., Banilower, E. R., & Heck, D. J. (2003). Looking inside the classroom: A study of K-12 mathematics and science education in the U.S. Chapel Hill, NC: Horizon Research.
Woodbury, S., & Gess-Newsome, J. (2002). Overcoming the paradox of change without difference: A model of change in the arena of fundamental school reform. Educational Policy, 16(5), 763–782.

**Appendix B. Tables and Figures**

Table 1. Summary of student demographic data. *n = 26, two home-schooled students in the commonplace group did not answer this question.

| | Commonplace Unit (n=28) | Inquiry-Based Unit (n=30) |
|---|---|---|
| Gender | 61% male, 39% female | 47% male, 53% female |
| Race (% non-white) | 21% | 23% |
| Age (mean) | 15.1 | 14.9 |
| Free and Reduced Lunch | 12%* | 10% |

Table 2. Description of the levels used to score the pretest and posttest constructed response items.

| Level | Description |
|---|---|
| 5 | Model-based accounts connected across scales |
| 4 | Appropriate but superficial connections between organismal and physiological systems |
| 3 | Alludes to hidden physiological mechanisms |
| 2 | Accounts restricted to the organismal level |
| 1 | Stories at the organismal level based on personal experience / cultural models |
| 0 | No response / unintelligible / negligible |

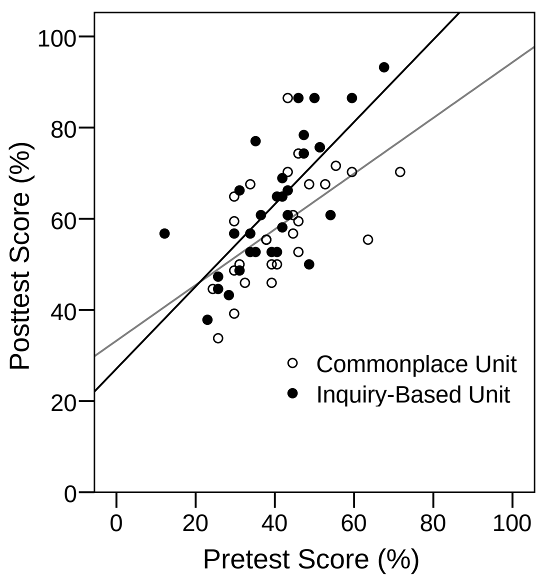


*Figure 1*. Pretest-Posttest bivariate distribution for the students receiving instruction from the commonplace and inquiry-based units. The slopes of the regression lines are significantly different [$F(1,55) = 4.662$, $p < 0.05$].

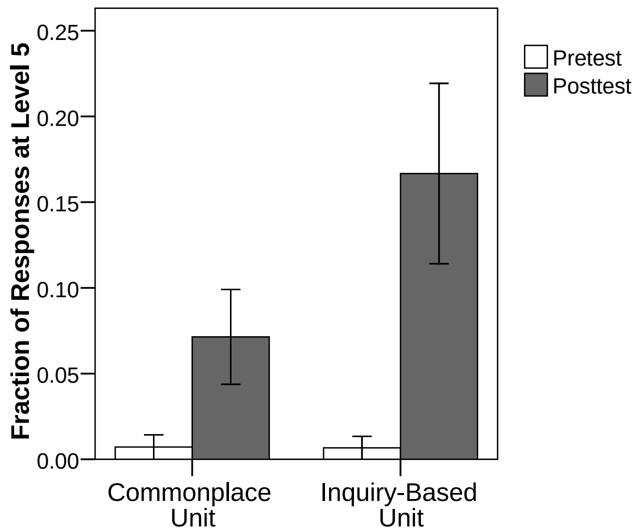


*Figure 2*. Significant differences [$F(1,56) = 2.746$, $p < 0.05$] in the frequency of posttest Level 5 accounts between the Commonplace and Inquiry-based groups. Error bars = +/- 1SE.