Appendix A2
Study Methods: Classroom Observation

A2.1 Introduction

The classroom observation study was designed to document classroom practices and interactions so that we could characterize IBL and non-IBL teaching practices, including their variation, and link these practices to student outcomes. We sought to address the following questions

- How are classrooms designated "IBL" alike or different from comparative classrooms?
- What practices and features are commonly applied in IBL courses, and what are the variations among them?
- What classroom features are seen in course sections where there are good student outcomes? How do these compare with classroom features of sections where outcomes are less positive?

We sought to address these questions with high-quality, internally consistent data yet without extensive investment of both observation and analysis time. Thus we designed a protocol that could be largely completed during class in real time, by trained but non-expert observers. The protocol documented classroom activities with simple, quantifiable indicators of student and faculty behaviors, rather than requiring subtle judgments of pedagogical effectiveness as used in instruments such as the Reformed Teaching Observation Protocol (RTOP) (Pilburn et al., 2000). To augment the quantitative data, we asked observers to provide notes and comments to help us interpret the data and to capture their own evidence-based judgments of less tangible variables such as classroom atmosphere. Moreover, we did not videotape or transcribe course sessions, because of the cost involved of collecting and, especially, analyzing such data. These choices sacrifice some detail but enabled us to gather a large volume of data across many sections.

A2.2 Study sample

Three campuses participated in the observation study. Altogether, 52 course sections were observed for multiple class sessions by trained observers, 36 sections in Year 1 and 16 in Year 2. Seven of the Year 2 sets were omitted from the analysis because there was a high incidence of missing data or too few hours of class time were observed. Another set of observations was procedurally valid but was omitted from the data analysis as representing an "experimental" hybrid course that was not easily classified as IBL or non-IBL.

For two large lecture courses that also included a recitation session, we sought to ensure that the observation data to represent a student's overall experience, not just the lecture. The lecture session was observed for multiple periods, and separately, a sample of recitation periods was observed (e.g. 6 recitations taught by different TAs). Observation data from lecture and recitation were combined in a 3:1 weighted average to reflect the three hours of lecture and one hour of recitation that any student would experience in a week.

With these adjustments, the total observation sample included 42 course sections: 31 IBL and 11 non-IBL sections of 18 different courses on 3 campuses. All the non-IBL sections observed were chosen from courses that also offered IBL sections, but fewer non-IBL sections were observed because (1) a comparable non-IBL section was not available for every IBL course we wished to observe; (2) some non-IBL instructors declined to participate; and (3) the non-IBL courses emphasized lecture. Since these were more homogeneous in style, a smaller sample seemed to be representative. This last assumption is borne out by the data: non-IBL courses exhibit much narrower distributions around the mean for nearly every practice-oriented variable.

A2.3 Observation protocol

The observation protocol was based on a study by Gutwill-Wise (2001; see also ModularCHEM Consortium, 1998, 1999) comparing active-learning and more traditional versions of a reformed undergraduate chemistry courses at two institutions. The content of the protocol was adapted using information from a preliminary study of five IBL Mathematics Centers, which included sites visits and observation of two to six IBL class sessions at each campus. Additional information was drawn from focus groups with students and graduate teaching assistants, and interviews with campus leaders and faculty. The protocol included three main components:

A. A summary sheet, where observers recorded basic data about the time and date of the class, the instructor, and the number, gender and apparent ethnicity of students attending. Observers also provided, in writing, their overall impressions of features such as classroom interactions, mood, morale, and any special context features (e.g., class held the day before an exam). This sheet is included as Exhibit E2.1.

B. An observer survey, where observers estimated the proportion of students who participated in class and rated the frequency of 14 student and instructor behaviors on a 5-point Likert-like scale (1=never to 5=very often). Behaviors included offering ideas, asking questions, working with others, listening to others, setting the pace or direction, giving feedback, and were chosen as indicators of the classroom atmosphere and interactions of class members. The observer survey is included as Exhibit E2.2.

C. A classroom log, where observers tracked class activities, leadership roles, and question-asking behaviors, as detailed below. The classroom log is included as Exhibit E2.3.

The classroom log included three different classification schemes, each based on a set of simple letter codes to record categories of behavior: class activities, leadership roles, and question-asking. The scheme for class activities (Scheme 1) was adapted from Gutwill-Wise (2001) to incorporate all activity types observed or reported in the preliminary site visits. Coding classroom activities enables us to determine differences in practice between IBL and non-IBL courses, and to gather a rough measure of the level of inquiry actually implemented in an IBL course. Scheme 2 was added to document the active roles of instructors, TAs, and students. Scheme 3 addressed the nature of questions asked by instructors and students.

Starting at the beginning of class, observers recorded the clock time and categorized the main activity and the lead role, using Schemes 1 and 2 respectively.

Scheme 1: Main activity

- B Addressing class business, procedural activity (e.g. returning papers)
- L Professor lecturing—presenting *previously prepared* material. This may include response to student questions during a lecture, but the L code is retained if the question does not turn into a multi-student discussion.
- E Extended explaining, *in response to* question or difficulty (instructor or student). This is an extended discourse or mini-lecture, different from L because it is not pre-planned, but responsive to an issue that arises on the spot.
- G Working on a problem or an example in groups, in seats or at the board (informal—class working in groups on problems, instructors circulate among groups)
- P Students presenting a solution or proof (individuals or groups).
- D Class discussing or critiquing a solution that has been presented. Usually whole-class.
- C Students working at computers—on a problem, modeling task, visualization, etc.
- O Other (describe)

Scheme 2: Lead role (these codes also used to identify anyone asking a question)

- f Faculty or lead instructor
- t Teaching assistant—any graduate or undergraduate student assisting the lead instructor
- s Single student
- ns New student (first time to participate today)
- rs Repeat student (has already participated today)
- g Group of students
- c Whole class/multiple roles

Observers were asked to make a real-time judgment of when the activity or lead role changed, and to record the time and the new letter code when either changed. They could also write comments to clarify activities or transitions. In practice, this required some judgment, since many activities changed spontaneously and observers had to decide, for example, at what point a dialogue between one student and a professor became a whole-class discussion involving multiple students. Leadership by a "new" or "repeat" student was recorded so that we could distinguish participation by a few, very active students, or a broader group.

In addition to the activity and role codes, observers recorded and coded each question that was asked by an instructor or student. We focused on question-asking because questions are central to "inquiry." Summaries of the literature indicates that questions are related to students' cognitive activity (Gall, 1970; Wilen, 1982; Edwards & Bowman, 1996). Teachers tend to ask mostly low-level questions requiring recall rather than higher-order reflection, but the cognitive level of teachers' questions correlates positively to the cognitive level of students' reply. Thus

the type of questions asked by instructors should relate to students' thinking and learning. Student question-asking has been studied less than instructor questioning, but Edwards and Bowman (1996) suggest that a shift in the incidence of student-asked questions may reflect a shift in the view of the teacher from sole classroom authority to one who guides student-teacher interaction. Kawanaka and Stigler (1999) argue that not only the number but the type of student questions is related to learning. Thus we documented the frequency and type of student questions, as a further indicator of the cognitive demand and inquiry nature of a course.

Observers used Scheme 2 to indicate the question-asker and Scheme 3 to indicate the question type. Scheme 3 is a simplified version of Bloom's taxonomy (1956) adapted from Gall (1970) to include questions with functions other than cognitive ones.

Scheme 3: Question type

- R Recall or factual. Closed-ended; there is a right answer. Lowest in cognitive demand.
- E Explanatory/descriptive—seeks to draw out or build toward an explanation of how or why something is done. More cognitively demanding than recall questions.
- C Critiquing—asking for evaluation or judgment of an idea. High in cognitive demand.
- S Stretching/linking—asking for expansion, connection, creativity. These are high on cognitive demand.
- M Monitoring/involving. Includes instructor monitoring (Did you understand this? Are we ready to move on?) and student self-monitoring, checking, clarifying (Am I sure I understood this?). Also questions intended to generate metacognition or to involve others (Sally, what do you think?). These are process-oriented more than cognitive.
- B Business/procedural, accomplishing course business (Did everyone get their paper back? Who wants to present next? When is our homework due? Will that be on the test?)
- X Other, unknown, or unclear in type

Question types were linked to the activity episode during which each question occurred, based on findings by Edwards and Bowman (1996) that student questions are influenced by the instructional format in which the question occurred. Thus, for example, we can analyze the number and types of questions that occurred during lecture or during discussion. Observers listened but did not track questions during small group work, because they could not attend to more than one group at a time and we did not have a way to choose a group randomly. In practice, observers found categorizing question types (Scheme 3) to be more difficult than categorizing the activity or lead role (Schemes 1 and 2).

A2.4 Data Collection

Because the research team members could not be present on all three campuses for the length of time needed to observe multiple class sessions, we recruited classroom observers on each campus. We provided a job description to the campus leader or internal evaluator, who then recruited or circulated the description. Different pools of people were available on different

campuses, but all the observers had degrees in mathematics and interest in teaching, including graduate students in mathematics and mathematics education and a doctorally trained mathematics education researcher.

We trained the initial group of observers in a 2-3 hour session conducted in person at their own campus. Observers read advance materials about the study design and observation protocols, and in the training session we reviewed the materials and practiced applying the protocol to video-recordings of two different IBL classrooms. Every observer signed a confidentiality agreement to keep the data and their own opinions confidential. On two campuses, we retained our observers from Year 1 to Year 2. On the third campus, the campus project evaluator, who had previously participated in the training, trained a new graduate student, and we held a conference call to discuss the protocol and answer his questions. At the beginning of Year 2, we also sent new copies of the protocol (with a few minor revisions called out) and a reminder of responsibilities to the returning observers.

To ensure a representative sampling of actual classroom processes, each class was visited multiple times at least two different points in the academic term. The aim was to capture 8-12 hours of class time for each course section. In practice, observers' adherence to this schedule varied, but we were nonetheless able to document several class sessions for every section observed. On average, 6.9 hours of class time were observed for every section included in the analysis, for a total of 298 hours of observation. During this time, over 2200 distinct episodes of instructional activity and over 10,300 questions were logged.

A2.5 Data analysis

Raw data were entered into a pre-formatted Excel spreadsheet by undergraduate assistants, who also assisted with compiling and tallying data for individual class sections, using Excel's conditional counting and summing functions. Text data (such as comments and notes) were transcribed for qualitative analysis.

For the observer survey, we computed mean ratings for each survey item across all observed sessions. For the classroom log data, most variables were analyzed on a cumulative or average basis over all class hours observed, rather than by class session, because class sessions varied in length from 45 to 90 minutes. Frequencies of specific behaviors or events are normalized to an hourly basis (e.g. episodes of discussion per hour of class time).However, variables that depend on student attendance (e.g. % of all students who ask a question) are necessarily based on class sessions, because attendance varies by session. For question-asking variables, we normalize to hours of non-group work, because questions were not tracked during small group work (e.g. questions asked by students per hour of non-group work).

Variables constructed for analysis included:

- Percentage of observed class time spent on each of the activity types in Scheme 1
- Mean frequency of each activity in Scheme 1, per hour of class observed

- Percentage of observed class time under leadership by each participant type in Scheme 2
- Mean frequency of each leadership role in Scheme 2, per hour of class observed
- Mean number of questions asked by each participant type in Scheme 2, per hour of non-group work
- Percentage of all questions asked by each participant type in Scheme 2
- Mean number of students, and percentage of all students in attendance, who asked a question, for a given class session
- Percentage of all questions of each question type in Scheme 3
- Mean number of questions per hour of each question type in Scheme 3
- Mean frequency of Scheme 3 question types for each Scheme 1 activity type (e.g. number of recall questions per hour during discussion or during lecture)

We computed basic descriptive statistics for IBL and non-IBL courses, and t-tests for statistical significance, using Microsoft Excel. Analyses of correlations among observation variables and relationships between observation and student survey data were conducted using SPSS.

A2.6 Data reliability and validity

We performed several types of checks on the data to ensure that they were as precise and accurate as possible. Observer coding omissions represent one type of uncertainty in the data. Observers omitted very few codes for class time by instructional activities and leadership role (Schemes 1 & 2). Counting and coding questions was more difficult, in part because questions sometimes came rapidly. Questions that were uncategorized by asker (Scheme 2) accounted for 1.2% of all questions, and those uncategorized by type (Scheme 3) accounted for 4.5% of all questions. Because Scheme 2 is straightforward to apply, we use that omission rate to estimate the rate of simple error (forgetting to circle a choice), i.e. 1.2%. Because Scheme 3 is more difficult to apply, the higher omission rate likely reflects real difficulty in categorizing questions.

Early in the study, some pairs of observers visited the same class session so that they could compare notes and discuss issues. We used these observations as a check on inter-rater reliability, though we did not repeat this test after observers were experienced and confident. Comparison of independent observations show that observers agreed to a very high extent in categorizing the nature of the activity (Scheme 1) and lead role (Scheme 2) but were less well aligned in recording question-asking behaviors. The total counts of questions were most variable, depending on how observers counted a single utterance with multiple embedded questions (e.g. recording “What do you think? Did everybody follow?” as 1 or 2 questions). Observers categorized questions with reasonable consistency, however: While their raw numbers of questions varied, the percentages of questions categorized both by asker (Scheme 2) and type (Scheme 3) were the same within 5 to 20%.

On the observer survey, observers did not differ by more than one point on any single item in their rating of the frequency of various student and instructor behaviors. Typically, they agreed exactly on a majority of the items, and the sum of their scores on all 14 ratings differed by less than 5% of the total possible rating points.

Given the high volume of data, data entry errors are another potential source of uncertainty. Because two codes were recorded for classroom time (Schemes 1 and 2) and two codes for question (Schemes 2 and 3), we could catch most data-entry errors by comparing the number of minutes or questions logged under each scheme, which should be identical. Where the totals did not match, we checked the spreadsheet data against the raw data and were able to identify and correct nearly all such errors. We estimate the proportion of undetected data entry errors as rather less than 0.5%.

A2.7 References cited

Bloom, B. S., ed. (1956). *Taxonomy of educational objectives: Handbook 1: Cognitive domain*. New York: David McKay.

Edwards, S., & Bowman, M. A. (1996). Promoting student learning through questioning: A study of classroom questions. *Journal on Excellence in College Teaching*, **7**(2), 3-24.

Gall, M. D. (1970). The use of questions in teaching. *Review of Educational Research,* **40**(5), 707-721.

Gutwill-Wise, J. (2001). The impact of active and context-based learning in introductory chemistry courses: An early evaluation of the modular approach. *Journal of Chemical Education,* **78**, 684-690.

Kawanaka, T., & Stigler, J. W. (1999). Teachers' use of questions in eighth-grade mathematics classrooms in Germany, Japan, and the United States. *Mathematical Thinking and Learning,* **1**(4), 255-278.

ModularCHEM Consortium (1998, January 23-24). MC^2 evaluation progress report: January 1998, in *Fourth Meeting of the ModularCHEM Consortium National Visiting Committee*. (Report to the National Visiting Committee) Berkeley, CA: ModularCHEM Consortium.

ModularCHEM Consortium (1999, January 29-30). ModularChem Consortium evaluation report to the National Visiting Committee, in *Fifth Meeting of the ModularCHEM Consortium National Visiting Committee*. (Report to the National Visiting Committee) Berkeley, CA: ModularCHEM Consortium.

Pilburn, M., Sawada, D., Falconer, K., Turley, J., Benford, R., & Bloom, I. (2000). *Reformed Teaching Observation Protocol (RTOP).* Tempe, AZ: Arizona Collaborative for Excellence in the Preparation of Teachers.

Wilen, W. W. (1982). *Questioning skills, for teachers. What research says to the teacher.* Washington, DC: National Education Association.

Exhibit E2.1: COVER SHEET and SUMMARY
Classroom Observation Protocol for Mathematics

Date: Class start time: Observer:

Class end time:

Course name/number: Instructor:

Description of student population counted or estimated (circle one)

# of students	White women	White men	Women of color	Men of color	TOTAL
Present at start					
Entering later					

Notes – class context, interactions, mood, morale. What was interesting about this class? Please record what you observed as well as how you interpret it.

Exhibit E2.2: Classroom Observation Post-Survey

Fill out this survey right after observing a class. It aims to capture some general features to complement the detailed log. Use your own assessment based on what you saw and experienced, using the scale (1 = Never to 5 = Very often) to indicate the extent to which the activity or feature occurred in the session you saw. Please add any explanatory comments.

A. What percentage of students (approximately) participated in class? (circle one)

0-25% 25-50% 50-75% 75-100%

Was participation generally representative by gender and ethnicity? Yes No
Please comment.

B. To what extent did (single or groups of) students …	**Never**				**Very often**
1. Offer their own ideas during class?	1	2	3	4	5
2. Ask instructor/TAs questions?	1	2	3	4	5
3. Review or challenge other students' work?	1	2	3	4	5
4. Work together with other students?	1	2	3	4	5
5. Set the pace or direction of class time?	1	2	3	4	5
6. Get help from other students?	1	2	3	4	5
7. Receive personal feedback on their work?	1	2	3	4	5

C. To what extent did the instructor and TAs …	**Never**				**Very often**
1. Listen to students' ideas/explanations?	1	2	3	4	5
2. Express their own ideas or solutions to problems?	1	2	3	4	5
3. Set the pace or direction of class time?	1	2	3	4	5
4. Give concrete feedback on students' work?	1	2	3	4	5
5. Offer help to students?	1	2	3	4	5
6. Establish an overall positive atmosphere?	1	2	3	4	5
7. Summarize or place class work in a broader context?	1	2	3	4	5

Exhibit E2.3: Classroom Observation Log – Classroom Activities & Questioning Behaviors

Date: Course: Observer: Page _____ of _____

I. Classroom Activities. Mark start time & general activity at each change.			**II. Question-Asking Behaviors.** Mark asker & categorize.	
Time	Main activity B L E G P D C O	Lead role f t s g c	Asked by	Question type (with notes)
			f t ns rs	R E C S M B X
			f t ns rs	R E C S M B X
			f t ns rs	R E C S M B X
			f t ns rs	R E C S M B X
			f t ns rs	R E C S M B X
			f t ns rs	R E C S M B X
			f t ns rs	R E C S M B X
			f t ns rs	R E C S M B X
			f t ns rs	R E C S M B X
			f t ns rs	R E C S M B X
			f t ns rs	R E C S M B X
			f t ns rs	R E C S M B X
			f t ns rs	R E C S M B X