Appendix A5: Study Methods for Mathematics Tests

A5.1 Introduction

In addition to students' own reports of their learning gains from a mathematics course (Ch. 3, Appendix A3), we gathered information from tests about possible changes in students' content knowledge and mathematical thinking during a college mathematics course. The first test, called *Learning Mathematics for Teaching* (LMT), offered us well-validated instruments for measuring pre-service teachers' cognitive gains from an IBL mathematics course. The other test, nicknamed the *Proof Test*, measured students' ability to evaluate a mathematical argument and determine its validity (see Exhibit E5.1). Using these two tests we gathered two mid-sized data sets on the development of students' mathematical knowledge and thinking. These studies addressed our research questions:

- How do students' mathematical knowledge and thinking change during an IBL college mathematics course?
- How do the changes differ by student groups, especially between IBL and non-IBL students?
- How do the changes align with results from other measures of students' learning gains?

We also used one more method to get comparative data on students' learning gains during a mathematics course. We refer to this as the *Instructor Ratings* instrument (see Exhibit E5.2). We designed a rubric that asked instructors to assess their students' learning from their course by rating the students on both their initial expertise in mathematics (as a check on instructors' perceptions) and overall learning gain in mathematics from the course. Because these data were mainly used to check the validity of our other measures of student mathematics learning and gains, results from this data set are reported in this Appendix.

A5.2 Learning Mathematics for Teaching (LMT) Tests

A5.2.1 LMT Tests

We used well-validated instruments, called *Learning Mathematics for Teaching* (LMT) tests, in studying pre-service teachers' gains in mathematical knowledge from an IBL mathematics course. The LMT instruments have been developed and validated by a team at the School of Education, University of Michigan, for assessing professional development courses for K-12 mathematics teachers (Hill, Schilling & Ball, 2004). Their project investigates the mathematical knowledge needed for teaching, and how such knowledge develops as a result of experience and professional learning. The LMT tests reflect both the mathematical content that teachers teach and the special knowledge they need to teach that content to students. The LMT measures are not designed to make statements about <u>individuals'</u> mathematical knowledge but rather to compare the mathematical knowledge of <u>groups of teachers</u> (such as those participating in particular courses) and how their knowledge develops over time.

The test items are designed to measure the development of mathematical knowledge needed for teaching: solving problems, using definitions, and identifying adequate explanations (Hill, Schilling & Ball, 2004). Each item and form has been piloted with over 600 elementary teachers, yielding information about scale reliability and item characteristics. Some examples of released LMT test items can be found at the project's website (Learning Mathematics for Teaching, n.d.). After participating in a training session provided by the LMT developers, we signed a terms of use contract for using the instruments, which helps to protect the utility and validity of the items by keeping them confidential except for research and evaluation purposes (for example, the items may not be used for teaching).

For this study, we chose a pre-test and related post-test on elementary *Number Concepts and Operations* in order to study changes in pre-service teachers' mathematical knowledge during an IBL course. The pre-test consisted of 24 and the post-test of 23 items. Standardized IRT (item response theory) scores provided by the developers were applied to match the results between pre-test and post-test. We use these scores in analyzing our data and reporting results.

We added a separate section at the end of a pre- and post-test with demographic questions about students':

- gender,
- class year,
- prior teaching experience (if any), in total number of years,
- grade level of prior mathematics teaching experience (if any),
- intended grade level for future mathematics teaching.

Items on teaching experience were included because some graduate students in the study might be completing teacher certification programs after some prior teaching experience.

We verified the course, section and instructor for each student. In addition, we asked students to establish an identifier that helped us to match students' pre- and post-test answers and also their survey responses.

A5.2.2 Study sample for the LMT Test

Data from the LMT measures were gathered from students taking targeted IBL courses for pre-service teachers at two campuses. Altogether, we got pre- and post-test data from pre-service teachers in several distinct two-course sequences preparing teachers for elementary and middle school, elementary, or secondary teaching. In all, six data sets from three course sequences were received during the two academic years 2008-2010. No comparative (non-IBL) sections of these courses were offered at any of the campuses. The results reported in Chapter 5 are based on data from the 109 pre-service teachers who took both the pre-test and post-test at two campuses. The sample is described in Table A5.1, showing demographic characteristics of students in the three main groups, divided by the target audience (elementary, elementary/middle, or secondary) of the courses in which the students participated.

Table A5.1: LMT Test Sample Teachers by Gender, Ethnicity, and Race

Indicator	**Group 1** Elementary		**Group 2** Elementary/middle		**Group 3** Secondary		**Total**	
	Count	%	Count	%	Count	%	Count	%
Gender								
Women	27	*100*	62	*97*	9	*50*	98	***90***
Men	0	*0*	2	*3*	9	*50*	11	***10***
	27	*100*	64	*100*	18	*100*	109	***100***
Ethnicity								
Hispanic or Latino	6	*23*	4	*7*	1	*6*	11	***11***
Not Hispanic or Latino	20	*77*	57	*94*	15	*94*	92	***89***
	26	*100*	61	*100*	16	*100*	103	***100***
Race								
Asian	4	*20*	3	*5*	1	*10*	8	***9***
Multiracial	2	*10*	5	*8*	1	*10*	8	***9***
White	14	*70*	53	*87*	8	*80*	75	***82***
	20	*100*	61	*100*	10	*100*	91	***100***

Table A5.1 indicates that most of the pre-service teachers who took both the pre- and post-test are women (90%). We had only 11 men in our sample—a typical distribution of gender among teachers. In particular, all the elementary teachers were women and only two in Group 2 (elementary/middle school) teachers were men. Half of the secondary school teachers were men, but their number was still low in our sample overall.

Table A5.1 also shows that only 11% (11) of the pre-service teachers in our sample overall were Hispanic or Latino, although their proportion was a bit larger among elementary teachers (23%). Most (82%) of the students were white and the number of students reporting other races was 16 (18%). Again, this shows the sample had little variation in students' demographic characteristics.

Table A5.2: LMT Test Sample by Academic Status and Major Subject

Indicator	Group 1		Group 2		Group 3		Total	
	Count	%	Count	%	Count	%	Count	%
Academic status								
Sophomore	1	*4*	2	*3*	2	*11*	5	***5***
Junior	8	*30*	53	*83*	3	*17*	64	***59***
Senior	18	*67*	6	*9*	13	*72*	37	***34***
Graduate, other	0	*0*	3	*5*	0	*0*	3	***3***
	27	*100*	64	*100*	18	*100*	109	***100***
Major subject								
Math or applied math	0	*0*	14	*23*	14	*88*	28	***27***
Science, engineering, computer science	3	*12*	8	*13*	1	*6*	12	***12***
Non-science	23	*89*	40	*65*	1	*6*	64	***62***
	26	*100*	62	*100*	16	*100*	104	***100***

Table A5.2 displays pre-service teachers' academic status and college majors. Typically, these students were well along in their academic careers. Only five were second-year students and all the others were juniors (59%) or seniors (34%). Slight variation appeared between the three groups. Most of the elementary and secondary school teachers were seniors (67%), whereas nearly all the elementary/middle school teachers (83%) were juniors.

The distribution of the elementary and middle school teachers' college majors was typical. Most of them had a non-science (education) as their major subject (89%, 65%), whereas nearly all of the secondary school teachers (88%) reported a major in Mathematics or Applied Mathematics.

A5.2.3 Methods for Administering and Analyzing the LMT Test

Both the pre- and post-test were administered as a paper-and-pencil test in class by the IBL project coordinators at two campuses. The coordinators were provided with written instructions for administering the tests and preserving students' confidentiality. The written tests were returned to researchers, coded, and the data entered into separate SPSS data files by trained student assistants. Finally, the data on students' responses to the test items were matched with their responses on our survey instruments (Ch. 3). A smaller sample set of students' LMT test results was matched with instructor ratings (see Section A5.4).

The raw scores on the pre- and post-test were converted to standardized IRT scores according to a scoring table provided by the developers of the tests. All the analyses were performed by using these standardized scores that also enabled matching of students' pre-test scores to their post-test scores. To report results, we also apply the IRT scoring table in illustrating average score gains in mathematical knowledge from actual LMT test scores. In addition to descriptive statistics, we applied correlational analysis and parametric tests (independent and pairwise t-tests, ANOVA) to analyze the data. Stepwise regression was used to study the extent to which LMT test score gains were explained by other measures of mathematical knowledge and learning.

A5.3 Proof Test

We used another test, which we refer to as the “Proof Test,” to measure students’ mathematical knowledge and thinking. This test studied students’ ability to evaluate a mathematical argument and determine its validity. Test data were gathered from interviews and paper-and-pencil tests.

A5.3.1 Proof Test instrument

The proof test was based on items on evaluating mathematical arguments that were designed by Weber (2009). We reformulated the original test into a paper-and-pencil test. Wording of some of the claims was clarified with a few additional words suggested by a mathematics professor. We also numbered the lines in the arguments so that students could reference specific lines in their comments. In order to obtain equal numbers of answers to each argument, we used two tests, forms A and B, which were identical except for the order of the arguments.

The one-hour test included nine of the ten original arguments from Weber (2009) on algebra, number theory and calculus (see Exhibit A5.1). Three of these arguments were valid and six arguments had some flaws for students to detect. Each argument was followed by structured questions to probe:

- Did students understand the argument?
- To what extent were students convinced by the argument?
- To what extent did students find it to have explanatory power?
- Did students consider the argument to be a mathematical proof?

Students answered the first three questions on a scale between 1 (strong disagreement) and 5 (strong agreement). On the fourth question, students assessed whether an argument was a proof (fully rigorous, not fully rigorous, not a proof, don’t understand). At the end of this question, we requested students’ explanations for their reasoning behind their decisions (see Exhibit A5.1).

A cover sheet for the proof test gathered demographic data on students, including ethnicity, race, gender, class year, academic major, and whether or not students planned to become a K-12 teacher. We also verified the course sections in which students were enrolled, the number of their college mathematics courses they had taken before and during or after the target course, and their expectations for their course grade (see Exhibit A5.1).

A5.3.2 Study sample for Proof Test

The first data set was gathered from one-on-one problem-solving interviews. Later, the test was revised into a paper-and-pencil form that was administered either in class to all students, or out of class to volunteers. In the interviews, students were asked to verbally explain the reasoning behind their answer about each argument, and these responses were recorded. On the paper-and-pencil test, students wrote down their reasoning about whether or not each argument was a mathematical proof. Both the interviews and the paper-and-pencil tests took an hour.

In all, we obtained tests from 42 IBL students (27 men, 15 women) and 35 non-IBL students (19 men, 16 women) at the end of a mathematics course. Of these, 24 students (14 IBL, 10 non-IBL) took an interview and 53 students (28 IBL, 25 non-IBL) a paper-and-pencil test. Most of the students were volunteers (63) who were paid a modest honorarium for participating. Only 14 students took an in-class post-test. In addition, we got pre/post-test data from one section (20 pre-, 14 post-tests). Table A5.3 displays demographics of the students, for IBL and non-IBL students separately.

Table A5.3: Proof (Post-) Test Sample by Gender, Ethnicity, Race, and Course Type.

Indicator	**IBL students**		**Non-IBL students**		**Total**	
	Count	%	Count	%	Count	%
Gender						
Women	15	*36*	16	*46*	31	***40***
Men	27	*64*	19	*54*	46	***60***
	42	*100*	35	*100*	77	***100***
Ethnicity						
Hispanic or Latino	3	*8*	2	*6*	5	*7*
Not Hispanic or Latino	37	*93*	33	*94*	70	***93***
	40	*100*	35	*100*	75	***100***
Race						
Asian	14	*34*	14	*44*	28	***38***
White	25	*61*	16	*50*	41	***56***
Other race	2	*5*	2	*6*	4	***6***
	41	*100*	32	*100*	73	***100***

The proportion of men among IBL students exceeded that among non-IBL students. Otherwise, the sample looked like our other student samples. Only five students' ethnicity was Hispanic or Latino, and only 4 students reported a race other than white or Asian. However, one third of IBL but 44% of non-IBL students were Asian.

Table A5.4: Proof (Post-) Test Sample by Course Type and Academic Status.

Indicator	IBL students		Non-IBL students		Total	
	Count	%	Count	%	Count	%
Academic status						
First-year	0	*0*	0	*0*	0	***0***
Sophomore	7	*18*	7	*21*	14	***19***
Junior	13	*33*	14	*41*	27	***37***
Senior	20	*50*	13	*38*	33	***45***
Other	0	*0*	0	*0*	0	***0***
	40	*100*	34	*100*	74	***100***
Major subject						
Mathematics	30	*83*	26	*87*	56	***85***
Natural science	1	*3*	1	*3*	2	***3***
Math/Natural Science	1	*3*	2	*7*	3	***5***
Non-science	0	*0*	0	*0*	0	***0***
Math/Non-science	4	*11*	1	*3*	5	***8***
	36	*100*	30	*100*	66	***100***

According to instructors, our target courses represented an introductory or mid-level proof-based course. But, in practice, we found that many students had substantial proving experience in prior courses. This is also reflected in Table A5.4: most of the students were seniors or juniors. Nearly all were also pursuing a math major.

A5.3.3 Methods for Proof Test

We used descriptive statistics and parametric (T-test) or non-parametric (Mann-Whitney) tests to examine differences between student groups in responses to the three first structured questions (see Exhibit A5.1). Table A5.5 displays averages of students' ratings on the three structured questions, for each argument separately.

Table A5.5: Average Student Ratings of Single Arguments

Argument	Course type					
	IBL			Non-IBL		
	Understanding	Conviction	Explanation	Understanding	Conviction	Explanation
Valid arguments						
Arg 1	4.8	4.7	4.3	4.7	4.7	4.5
Arg 3	4.6	4.4	4.3	4.5	4.3	4.2
Arg 4	4.4	4.2	4.4	4.2	4.0	4.1
Invalid arguments						
Arg 2	4.8	4.5	4.4	4.8	4.4	4.3
Arg 5	4.3	2.1	1.9	4.4	2.2	1.8
Arg 6	4.6	3.3	3.4	4.7	3.5	3.5
Arg 7	4.3	2.1	2.1	4.8	2.4	2.4
Arg 8	4.3	3.3	3.3	4.5	3.5	3.5
Arg 9	4.4	3.5	3.4	4.1	3.1	2.9

Scales (1-5):

Understanding: 1=not understand fundamental details to 5=understand completely.

Conviction: 1=not convinced at all to 5=completely convinced.

Explanation: 1=does not explain to 5=really illuminates why it is true.

Differences in frequency distributions between student groups in answers to the fourth question were compared by using a non-parametric test (Chi2).

In addition to the four structured questions, we analyzed students' written reasoning about each argument. These data came from 53 students (28 IBL, 25 non-IBL). These qualitative data were coded and analyzed qualitatively according to eight main themes related to the nature of students' criteria for assessing the arguments. The categories were derived from preliminary analysis of a subset of written comments and finalized using inductive content analysis (Miles & Huberman, 1994; Strauss & Corbin, 1990) of the complete set of written comments. Table A5.6 presents the eight main themes and the 29 sub-categories under the main themes, and the frequencies of student comments in each category.

Table A5.6: Frequency of Criteria Used in Written Comments about Reasoning

Main category	Subcategory	Number of students reporting		
	Description	once	2-3 times	≥4 times
Understanding	Some step(s) or the whole argument are/are not understandable	19	8	-
False statement	Makes false statement about an argument	18	2	-
Inadequate reasoning	Lacks justification	16	19	1
	Uses empirical/perceptual evidence	8	-	-
	Experience: have seen the argument/proof before	4	-	-
	An argument is complete	4	-	-
Use of steps as criterion	Steps are acceptable/not acceptable	13	7	-
	All steps are included/not included	15	1	-
	Proved step by step	5	1	-
Formalism	Theorems, formulas included/not included	15	5	1
	Written/not written formally	3	-	-
	External structure is correct/incorrect	16	6	1
	Mathematical rules, concepts, terms are included/not included	19	3	-
	Doesn't remember the concepts, terms, definitions, proof	1	1	-
Quality of explanation or reasoning	Requests (more) explanation, reasoning	20	16	5
	Level of explanation/reasoning is assessed	8	4	-
	Good explanation/reasoning in the argument	14	7	-
	Visual aid is used/not used	8	1	-
	The steps are clearly stated/explained	5	1	2
Rigor	Lack of/rigor of steps found in an argument	25	18	2
	Critical about a claim/presupposition	8	3	-
	Detects the flaw(s) in an argument	18	20	2
	Does not accept a picture/graph/equation as a rigorous way to prove	9	2	-
	Does not accept empirical evidence	30	2	-
	Suggests a more rigorous way to prove	10	1	-
Assessment of an argument as a whole	Beauty, appearance, ease of argument as a whole	11	6	-
	Style of (presentation or writing in) an argument	20	13	1
	Logic of argument as a whole	11	7	1
	Subject/mathematical level of an argument	11	2	-

A5.4 Instructor Ratings of Student Expertise

We conducted a small experiment using instructor ratings of students' mathematical expertise and learning, as one more method to get comparative data on IBL students' learning gains during a mathematics course. This is called the *Instructor Ratings* instrument.

A5.4.1 Instructor Ratings instrument

We designed a rubric that asked instructors to assess their students' learning from their course by rating the students on both their initial expertise in mathematics and their overall learning gain from the course. Instructors gave their answers in a prepared Excel spreadsheet (Exhibit 5A.2).

We asked instructors to give two ratings for each student: their overall level of expertise in mathematical knowledge and thinking at the *start* of the course, and overall *gain* or improvement in mathematical knowledge and thinking by the *end* of the course. That is, we asked instructors to distinguish students' incoming *ability* from their *learning* in their course. The instructors gave their ratings on a scale between from 5 to 1: very high, high, moderate, low, very poor or strongly lacking (see Exhibit E5.2).

A5.4.2 Study sample for Instructor Ratings

The sample on instructor ratings is based on data from four sections at one campus. We asked the campus coordinators to establish an identifier for each student that was later matched with the other data sets from these students. Students in two sections were math-track students and the two other sections represented courses for elementary/middle school pre-service teachers. Instructors rated students' mathematical expertise at the end of a course. In all, we matched instructor ratings from 27 math-track students and 37 pre-service teachers to the other data sets.

A5.4.3 Methods for Instructor Ratings

We compared data from the instructor ratings to data on the same students' learning gains from a mathematics course as measured by the SALG-M, using correlational analysis. Descriptive statistics and parametric (independent and paired T-test, ANOVA) or non-parametric tests (Chi2) were used to study subgroup differences in initial mathematical expertise and gains in mathematical expertise.

A5.4.4 Results from Instructor Ratings

Table A5.7 displays frequencies for the instructor ratings, for math-track students and pre-service teachers separately. The two IBL student groups differed from each other. On average, pre-service teachers' ($M=3.4$) rated initial mathematical expertise exceeded that of math-track students ($M=2.6$, $p<0.01$). But math-track students' ($M=3.4$) rated gains were higher than that of pre-service teachers ($M=2.3$, $p<0.001$). Also, comparisons between initial expertise and gain in expertise showed a difference between these two student groups. While there was a clear improvement in math-track students' rated mathematical expertise ($p<0.001$), pre-service teachers' gains in mathematical expertise were rated clearly lower ($p<0.001$) than their initial mathematical expertise. While 48% of math-track students had high or very high rated gain in mathematical expertise during an IBL course, only 5% of pre-service teachers' gain in mathematical expertise was rated at this level.

Table A5.7: Instructor Ratings by Course Type

Instructor rating	**Math-track students**		**Pre-service teachers**		**Total**	
	Count	%	Count	%	Count	%
Initial expertise	27	*100*	37	*100*		***100***
very high	-	-	5	*4*	5	***8***
high	2	*7*	15	*41*	17	***27***
moderate	11	*41*	10	*27*	21	***33***
low	14	*52*	5	*14*	16	***25***
very poor	-	-	5	*5*	5	***8***
Gain in expertise	27	*100*	37	*100*	64	***100***
very high	3	*11*	1	*3*	4	***6***
high	10	*37*	1	*3*	11	***17***
moderate	11	*41*	11	*30*	22	***34***
low	2	*7*	18	*49*	20	***31***
very poor	1	*4*	6	*16*	7	***11***

A5.5 Reliability and Validity of Mathematics Tests

A5.5.1 LMT Tests

The *Learning Mathematics for Teaching* (LMT) instruments are carefully developed and well-validated instruments (Hill, Schilling & Ball, 2004). The test items are designed to measure the development of mathematical knowledge needed for teaching: solving problems, using definitions, and identifying adequate explanations (Hill, Schilling & Ball, 2004). Each item and form has been piloted with over 600 elementary teachers, yielding information about scale reliability and item characteristics. Our sample from two campuses was also large enough to detect real gains and differences among students. However, because IBL methods were used in all sections of the courses targeted to pre-service teachers that were available for this study, we had no opportunity to compare student learning with that in a traditionally taught course.

At the start of the course, pre-service teachers reported their score (1-5) on the AP Calculus test (if they had taken it), their current estimated undergraduate GPA, and their expected grade in the present course. They also reported their expected course grade at the end of the two-term sequence. We checked to see how well these other academic measures correlated with students' LMT scores. Table A5.8 summarizes findings on these correlations.

Table A5.8: Correlations of Students' LMT Scores with Other Achievement Indicators

LMT score component	Expected course grade *at start*	Expected course grade *at end*	AP Calculus score *(1-5)* *at start*	Estimated GPA *at start*	Cognitive gains (SALG-M)	Affective gains (SALG-M)	Instructor rating: gain in math expertise
Pre-test	+ **	+ **	+		+*	+ **	+
Post-test	+ **	+ **	+ **			+ *	+
Score gain	+	+	+				

* p< 0.05, ** p< 0.01

Students' estimated undergraduate GPA did not correlate with the LMT test scores. This is understandable since these students' undergraduate studies were not usually focused on mathematics; their GPA represents a broad range of courses and not mathematical ability. However, all the three LMT test measures—pre-test, post-test, and test score gain—did correlate positively with students' AP test scores. This result shows that the LMT tests measured mathematical knowledge that was somewhat related to that measured by AP Calculus tests, even though the LMT test content addressed number and operations, not calculus. However, this finding is limited: we had AP test data from just 36 students, who came mostly from one university, and only the correlation between AP test score and LMT post-test score was statistically significant ($p<0.01$).

Students' expected grades reported at both the start and end of their course correlated positively with both LMT pre- and post-test IRT scores ($p<0.01$). But the positive correlations between expected grades and students' test score *gains* were weaker. In general, students who expected a higher grade in the course tended to earn higher LMT scores and to make greater LMT test score gains, but this latter relation was not statistically significant. In other words, students who thought they would get a good grade did get better test scores, but their expected grade was not well linked to their LMT test score growth.

We also examined students' LMT score gains in comparison with their self-reported cognitive, affective, and social gains from the SALG-M survey instrument (Ch. 3). Students' self-reported gains correlated positively with the LMT pre-test scores, while LMT post-test scores were positively related only to gains in confidence ($p<0.05$) and positive attitude ($p<0.01$). However, students' self-reported gains were not generally related to LMT test score gains. But among those students who started with the lowest initial LMT scores, LMT test score gains were positively (but not statistically significantly) related to cognitive, affective and social learning gains. That is, students' LMT score gains were consistently reflected only in the self-reports from students with low pre-test scores.

For a subset of students, we compared LMT test score gains with instructor ratings of students' gain in mathematical expertise during an IBL course. No direct correlation appeared between instructor ratings of students' gain in mathematical expertise and LMT score gains. However, instructor ratings of students' expertise in the beginning of a course were (statistically significantly) positively related to both LMT pre- and post-test scores, and slightly to their LMT-test gain scores. That is, the instructor's assessment of initial mathematical expertise was consistent with LMT pre-test and post-test scores; instructors could identify stronger or weaker students overall. But the instructor was a less successful judge of *learning* as measured by LMT score gains. This is a similar result to the relationship between instructors' assessment of gains in mathematical expertise and students' self-assessed grade, in that students and instructors both judge relative performance with some accuracy, but do not accurately predict learning.

A5.5.2 Proof Test

The proof test was intended to measure students' ability to assess mathematical arguments on algebra, number theory and calculus. It was based on items on evaluating mathematical arguments that were designed and previously tested by Weber (2009). After gathering and analyzing proof test data from individual student interviews, we reformulated the original test into a paper-and-pencil test that was further reviewed by a mathematics professor. Students were provided with an opportunity to offer written feedback or additional comments on the arguments. Both the student interviews and written test sheets indicated that students did not have difficulty in understanding the questions or goals of the proof test. However, the proof test sample was not the same as the samples from our surveys. Thus, we were unable to compare results from proof test with other indicators of student learning or gains.

Our ability to draw strong conclusions is limited by our sample of students. The students who volunteered to take the proof test were strong mathematics students, based on their self-reported grades and high numbers of prior mathematics courses taken. We surmise that differences in the responses and reasoning of IBL vs. non-IBL students are less easily detected among this group than among lower-achieving or less experienced mathematics students. Thus, while the test itself seems to be sensitive to differences in students' understanding, our sample is not optimized to detect group differences that might result from IBL instruction focused on proof processes. Moreover, this particular test is likely to be more sensitive in "introduction to proof" courses where enrollment is controlled or sequenced in such a way as to assure that most students have relatively little prior proof experience. In this sample, many students had proof experience already, and we cannot rule out that the test measured expertise developed in earlier courses.

A5.5.3 Instructor ratings

We used instructor ratings in order to get comparative data on IBL students' learning gains during a mathematics course. We checked to see how well initial expertise and gains in expertise correlated with each other and how these instructor ratings correlated with other indicators of mathematical knowledge or academic gains. Because the patterns in instructor ratings for math-track students and pre-service teachers differed (see section A5.4.4), we studied these relations within the two student groups separately. Table A5.9 summarizes findings from the nonparametric (Spearman) correlations between mathematical expertise and other indicators of student knowledge or learning.

Table A5.9: Correlation of Mathematical Expertise with Other Performance Indicators, by Student Group

Instructor rating (1-5)	Initial expertise	Expected course grade *at start*	Expected course grade *at end*	AP Calculus score *at start*	Estimated GPA *at start*
Math-track students					
Initial expertise		+	+ **	+	+
Gain in expertise	+	+	+	-	+
Pre-service teachers					
Initial expertise		+ **	+ **	+ *	+ *
Gain in expertise	-*			-	+

* p< 0.05, ** p< 0.01

Instructor ratings of initial mathematical expertise were somewhat consistent with students' mathematical performance level as indicated by their self-reported AP Calculus test score and GPA level at the start of a course. This applied more clearly to the ratings by the pre-service teachers' instructor than to those of the instructor of the math-track students. Moreover, ratings of initial mathematical expertise were consistent with both the grade expectations of both student groups at the start and end of a course. Students who expected a higher course grade were rated higher in mathematical expertise by their instructor, and the opposite was true for students who had lower course grade expectations.

However, in courses for pre-service teachers, the instructor's ratings of students' gains in mathematical expertise did not correlate with students' grade expectations at the end of a course . Similarly, in math-track courses, the correlation between the instructor-rated gains in expertise and students' own grade expectation at the end of a course was only slightly positive. That is, students' own expectations of their success in an IBL course did not match with their instructors' ratings of their learning gains. This may be an accurate assessment of the situation, if grades are seen by both parties to reflect achievement rather than learning.

Among pre-service teachers, initial mathematical expertise (as assessed by the instructor) was positively related to all other indictors of knowledge or gains. That is, instructor ratings were consistent with students' own expectations and the external performance indicators of AP test scores and estimated GPA at start of an IBL course. Moreover, the negative correlation between pre-service teachers' initial expertise and gain in expertise indicted that instructor ratings of gain tended to be higher for students who started with lower mathematical expertise. Further analysis by student groups showed that, on average, those pre-service teachers who started with high or very high initial expertise got lower ratings in gain in expertise than other pre-service teachers ($p<0.05$). Instructors saw these students as initially strong, therefore making lower gains. Also pre-service teachers' AP test score correlated slightly negatively to their instructor-rated gain in mathematical expertise. These results are consistent with other findings on pre-service teachers (see Section 5.2.3): weaker students at start of an IBL course may gain more than students with stronger mathematical background.

In contrast, among math-track students, instructor ratings of students' gains in mathematical expertise correlated slightly positively with their ratings of initial expertise. That is, students who started with stronger mathematical expertise also tended to gain more during an IBL course. But, unlike the result for pre-service teachers, the correlation was not statistically significant for math-track students. Overall, instructors seemed less successful in assessing learning during an IBL course than they were in assessing the initial level of students' mathematical expertise.

We also examined correlations between instructor-rated mathematical expertise and students' self-reported learning gains (SALG-M, Ch 3). Table A5.10 displays the results of the nonparametric (Spearman) correlations.

Table A5.10: Correlations of Gain in Expertise with Learning Gains (SALG-M).

Instructor rating (1-5)	Mathematical concepts & thinking	Application	Affective gains	Social gains
Initial expertise				
Gain in expertise	+	-	+	

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

IBL students' self-reported gains were not generally related to their initial mathematical expertise, as rated by their instructor. But students' instructor-rated gains in mathematical expertise were modestly related to their self-reported learning gains. Overall, students who had higher gains in mathematical expertise, as rated by their instructor, also tended to self-report higher gains in understanding concepts and mathematical thinking and problem-solving. They also tended to report higher affective learning gains. Students' gains in collaboration did not relate to their gain in mathematical expertise. Rather, higher instructor ratings in gain in mathematical expertise were slightly negatively related to students' self-reported gain in application of mathematical knowledge.

These results indicate that students' own assessments of learning were somewhat consistent with their instructor's ratings of their gains in mathematical expertise during an IBL course. However, the correlations were slight and mostly applied to students with initial poor to moderate mathematical expertise.

In sum, instructor ratings of gains in mathematical expertise varied between the student groups and courses. Moreover, they were mostly not consistent with other indicators of student learning. But students' self-reported learning gains were moderately in line with their instructor's ratings. Instructors may not be as successful in assessing their students' learning during as they are in assessing the initial level of their students' mathematical expertise. This applied especially to pre-service teachers and students who started with high mathematical expertise.

A5.6 References cited

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal, 105*, 11-30.

Learning Mathematics for Teaching (LMT) Project website (no date). Accessed 5/11/2011 at http://www.sitemaker.umich.edu/lmt

Miles, M. B., & Huberman, A. M. (1984). *Qualitative data analysis : a sourcebook of new methods*. Thousand Oaks, CA: Sage Publications.

Strauss, A., & Corbin, J. M. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Thousand Oaks, CA: Sage Publications.

Weber, K. (2009). Proving is not convincing. *Conference on Research in Undergraduate Mathematics Education,* Raleigh, NC, February 26-March 1, 2009.

Instructions: Assessing Mathematical Arguments

Thank you for participating in our study! This problem-solving test is part of a research study on how students learn to construct and evaluate mathematical proofs.

The problem-solving session will take about 1 hour of your time and will include 9 problems. For each problem, you will be asked to examine a mathematical argument and to decide whether or not you think it is a valid mathematical proof. You will then answer a few questions about each argument and write in some comments to help us understand your reasoning. Your comments do not need to be lengthy, but please show the thinking that led you to your answer. For convenience, the lines of each argument are numbered so that you can refer to them in your answer if you wish. Please work steadily, but do not rush. If you need more work space for any answer, please use the space on the last page and note the problem number.

Your participation is voluntary. You may skip questions or tasks that you do not wish to answer, or choose not to participate. Your answers are anonymous and will not be reported in any way that can identify you individually; they will be reported in groups with answers from other students from your course and other schools.

When you have finished the problems or are nearly out of time (whichever comes first), please complete question #10.

By taking this test, in part or whole, you agree that we may use this data to understand and improve the quality and effectiveness of college mathematics education. Thanks for your help!

Start time:

End time:

Argument 1

Claim: For all real numbers *a* and *b*: $(a + b)^2 = a^2 + 2ab + b^2$

Line:

1 $(a + b)^2 = (a + b)(a + b)$

2 $(a + b)(a + b) = a(a + b) + b(a + b)$

3 $a(a + b) = a^2 + ab$

4 $b(a + b) = ba + b^2$

5 So $(a + b)(a + b) = a^2 + ab + ba + b^2 = a^2 + 2ab + b^2$

For each question, circle the answer (1 to 5) that best fits your thinking about the argument.

A. Do you feel that you understood the argument that was presented?

1	2	3	4	5
There are fundamental details of the argument that I don't understand				I understand the argument completely

B. Are you convinced by this argument?

1	2	3	4	5
Not convinced at all				Completely convinced

C. Does this argument explain why the assertion is true?

1	2	3	4	5
No, the argument does not explain why the assertion is true				Yes, it really illuminates why the assertion is true

D. Would you consider this argument to be a mathematical proof?

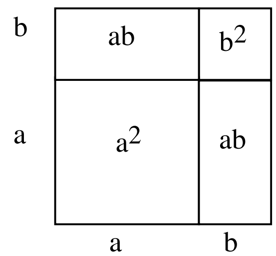
1. ______ Yes, I consider this argument to be a fully rigorous mathematical proof.
2. ______ Yes, I consider this argument to be a proof, although not fully rigorous.
3. ______ No, I think this argument does not meet the standards of a proof.
4. ______ Not sure, because I don't fully understand the argument.

Please explain your reasoning about **why** you think **this is** or **is not** a mathematical proof.

Argument 2

Claim: For all real numbers *a* and *b*: $(a + b)^2 = a^2 + 2ab + b^2$

Consider the diagram at right:



Line:

1 The length and width of the square are each (a+b), so the area of the diagram is
2 $(a+b)(a+b) = (a+b)^2$.
3 The area can also be found by adding the areas of the four cells of the square whose
4 areas are a^2, ab, ab, and b^2, which is $a^2 + 2ab + b^2$.
5 So $(a+b)^2 = a^2 + 2ab + b^2$.

For each question, circle the answer (1 to 5) that best fits your thinking about the argument.

A. Do you feel that you understood the argument that was presented?

1	2	3	4	5
There are fundamental details of the argument that I don't understand				I understand the argument completely

B. Are you convinced by this argument?

1	2	3	4	5
Not convinced at all				Completely convinced

C. Does this argument explain why the assertion is true?

1	2	3	4	5
No, the argument does not explain why the assertion is true				Yes, it really illuminates why the assertion is true

D. Would you consider this argument to be a mathematical proof?

1. _______ Yes, I consider this argument to be a fully rigorous mathematical proof.
2. _______ Yes, I consider this argument to be a proof, although not fully rigorous.
3. _______ No, I think this argument does not meet the standards of a proof.
4. _______ Not sure, because I don't fully understand the argument.

Please explain your reasoning about **why** you think **this is** or **is not** a mathematical proof.

Argument 3

Claim: For all natural numbers $n > 1$, $n^3 - n$ is divisible by 6.

Line:

1 $n^3 - n = n(n^2 - 1) = n(n+1)(n-1)$.
2 Either n is even or $n+1$ is even.
3 Since both numbers are factors of $n^3 - n$, $n^3 - n$ is even.
4 Because $n-1$, n, and $n+1$ are three consecutive numbers, one of them is divisible by 3.
5 So $n(n+1)(n-1)=n^3 - n$ is divisible by 3.
6 Since $n^3 - n$ is even and divisible by 3, $n^3 - n$ is divisible by 6.

For each question, circle the answer (1 to 5) that best fits your thinking about the argument.

A. Do you feel that you understood the argument that was presented?

1	2	3	4	5
There are fundamental details of the argument that I don't understand				I understand the argument completely

B. Are you convinced by this argument?

1	2	3	4	5
Not convinced at all				Completely convinced

C. Does this argument explain why the assertion is true?

1	2	3	4	5
No, the argument does not explain why the assertion is true				Yes, it really illuminates why the assertion is true

D. Would you consider this argument to be a mathematical proof?

1. ______ Yes, I consider this argument to be a fully rigorous mathematical proof.
2. ______ Yes, I consider this argument to be a proof, although not fully rigorous.
3. ______ No, I think this argument does not meet the standards of a proof.
4. ______ Not sure, because I don't fully understand the argument.

Please explain your reasoning about **why** you think **this is** or **is not** a mathematical proof.

Argument 4
Claim. There is no real number x which solves the equation $4x^3 - x^4 = 30$.

Line:
1 Consider the function, $f(x) = 4x^3 - x^4$. Because $f(x)$ is a polynomial of degree 4
2 and the coefficient of x^4 is negative, $f(x)$ is continuous and will approach $-\infty$ as x
3 approaches ∞ or $-\infty$. Hence, $f(x)$ must have a global maximum. The global maximum
4 will be a critical point. $f'(x) = 12x^2 - 4x^3$. If $f'(x) = 0$, then $x = 0$ or $x = 3$. $f(0) = 0$. $f(3) = 27$.
5 Since $f(3)$ is the greatest y-value of f's critical points, the global maximum of $f(x) = 27$.
6 Therefore $f(x) \neq 30$ for any real number x. $4x^3 - x^4 = 30$ has no real solutions.

For each question, circle the answer (1 to 5) that best fits your thinking about the argument.

A. Do you feel that you understood the argument that was presented?

1	2	3	4	5
There are fundamental details of the argument that I don't understand				I understand the argument completely

B. Are you convinced by this argument?

1	2	3	4	5
Not convinced at all				Completely convinced

C. Does this argument explain why the assertion is true?

1	2	3	4	5
No, the argument does not explain why the assertion is true				Yes, it really illuminates why the assertion is true

D. Would you consider this argument to be a mathematical proof?
1. _______ Yes, I consider this argument to be a fully rigorous mathematical proof.
2. _______ Yes, I consider this argument to be a proof, although not fully rigorous.
3. _______ No, I think this argument does not meet the standards of a proof.
4. _______ Not sure, because I don't fully understand the argument.

Please explain your reasoning about **why** you think **this is** or **is not** a mathematical proof.

Argument 5

Claim. Any even integer greater than two can be written as the sum of two prime numbers.

Consider the following table:

Even	Sum of two primes		
4	2+2		
6	3+3		
8	3+5		
10	3+7,	5+5	
12	5+7		
14	3+11,	7+7	
16	3+13,	5+11	
18	5+13,	7+11	
20	3+17,	7+13	
22	3+19,	5+17,	11+11
24	5+19,	7+17,	11+13
26	3+23,	7+19,	13+13

Line:

First, note that each even number between 4 and 26 can be written as the sum of two primes. Second, note that the number of pairs of primes that work appears to be increasing. For 4, 6, 8, and 12, there is only one prime pair whose sum is that number. For 22, 24, and 26, there are three prime pairs whose sum is that number. Every even number greater than 2 will have at least one prime pair whose sum is that number. For large even numbers, there will be many prime pairs that satisfy this property.

For each question, circle the answer (1 to 5) that best fits your thinking about the argument.

A. Do you feel that you understood the argument that was presented?

1	2	3	4	5
There are fundamental details of the argument that I don’t understand				I understand the argument completely

B. Are you convinced by this argument?

1	2	3	4	5
Not convinced at all				Completely convinced

C. Does this argument explain why the assertion is true?

1	2	3	4	5
No, the argument does not explain why the assertion is true				Yes, it really illuminates why the assertion is true

(continued, next page)

D. Would you consider this argument to be a mathematical proof?
1. _______ Yes, I consider this argument to be a fully rigorous mathematical proof.
2. _______ Yes, I consider this argument to be a proof, although not fully rigorous.
3. _______ No, I think this argument does not meet the standards of a proof.
4. _______ Not sure, because I don't fully understand the argument.

Please explain your reasoning about **why** you think **this is** or **is not** a mathematical proof.

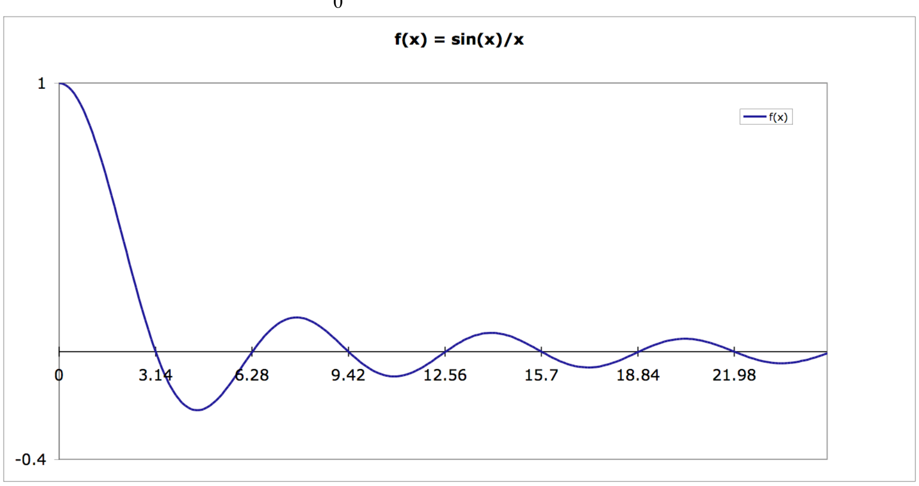
Argument 6

Claim. $\int_0^\infty \frac{1}{x}\sin x dx > 0$

The graph of $f(x) = \frac{1}{x}\sin x$ is given below.

Line:

1 $\int_0^\infty \frac{1}{x}\sin x dx > 0$ means that $f(x) = \frac{1}{x}\sin x$ has more area above the x-axis than below it.
2 To show this, note that it is clear from the graph that the first positive region—between 0
3 and π (about 3.14)—has more area than the first negative region—between π and 2π
4 (between 3.14 and 6.28). The second positive region has more area than the second
5 negative region. The third positive region has more area than the third negative region.
6 Since each positive region has a greater area than the negative region to the right of
7 it, the overall area of $\int_0^\infty \frac{1}{x}\sin x dx$ will be positive.



(continued, next page)

For each question, circle the answer (1 to 5) that best fits your thinking about the argument.

A. Do you feel that you understood the argument that was presented?

1	2	3	4	5
There are fundamental details of the argument that I don't understand				I understand the argument completely

B. Are you convinced by this argument?

1	2	3	4	5
Not convinced at all				Completely convinced

C. Does this argument explain why the assertion is true?

1	2	3	4	5
No, the argument does not explain why the assertion is true				Yes, it really illuminates why the assertion is true

D. Would you consider this argument to be a mathematical proof?

1. ______ Yes, I consider this argument to be a fully rigorous mathematical proof.
2. ______ Yes, I consider this argument to be a proof, although not fully rigorous.
3. ______ No, I think this argument does not meet the standards of a proof.
4. ______ Not sure, because I don't fully understand the argument.

Please explain your reasoning about **why** you think **this is** or **is not** a mathematical proof.

Argument 7

Claim: Let n be a natural number. If n^2 is divisible by 3, then n is divisible by 3.

Line:

1 We need to show that n is divisible by 3.
2 If n is divisible by 3, then there exists an integer k such that $n = 3k$.
3 $n^2 = (3k)^2 = 9k^2$.
4 So n^2 is divisible by 9.
5 All numbers divisible by 9 are also divisible by 3.
6 So if n^2 is divisible by 3, then n is divisible by 3.

For each question, circle the answer (1 to 5) that best fits your thinking about the argument.

A. Do you feel that you understood the argument that was presented?

1	2	3	4	5
There are fundamental details of the argument that I don't understand				I understand the argument completely

B. Are you convinced by this argument?

1	2	3	4	5
Not convinced at all				Completely convinced

C. Does this argument explain why the assertion is true?

1	2	3	4	5
No, the argument does not explain why the assertion is true				Yes, it really illuminates why the assertion is true

D. Would you consider this argument to be a mathematical proof?

1. ______ Yes, I consider this argument to be a fully rigorous mathematical proof.
2. ______ Yes, I consider this argument to be a proof, although not fully rigorous.
3. ______ No, I think this argument does not meet the standards of a proof.
4. ______ Not sure, because I don't fully understand the argument.

Please explain your reasoning about **why** you think **this is** or **is not** a mathematical proof.

Argument 8

Claim: Let *f(x)* be a real valued function and let *a* and *b* be real numbers such that *b*> *a*. Then $\int_a^b |f(x)|dx \geq \int_a^b f(x)dx$

Line: (Proof by cases).

1 Either $f(x) \geq 0$ or $f(x) < 0$.

2 Case 1: $f(x) \geq 0$.

3 If $f(x) \geq 0$, then $|f(x)| = f(x)$.

4 Thus, $\int_a^b |f(x)|dx = \int_a^b f(x)dx$.

5 Case 2: $f(x) < 0$.

6 If $f(x) < 0$, then $\int_a^b f(x)dx \leq 0$.

7 Since $|f(x)| > 0$, then $\int_a^b |f(x)|dx \geq 0$.

8 So $\int_a^b |f(x)|dx \geq 0 \geq \int_a^b f(x)dx$.

9 Thus, $\int_a^b |f(x)|dx \geq \int_a^b f(x)dx$.

For each question, circle the answer (1 to 5) that best fits your thinking about the argument.

A. Do you feel that you understood the argument that was presented?

1	2	3	4	5
There are fundamental details of the argument that I don't understand				I understand the argument completely

B. Are you convinced by this argument?

1	2	3	4	5
Not convinced at all				Completely convinced

C. Does this argument explain why the assertion is true?

1	2	3	4	5
No, the argument does not explain why the assertion is true				Yes, it really illuminates why the assertion is true

(continued, next page)

D. Would you consider this argument to be a mathematical proof?
1. ______ Yes, I consider this argument to be a fully rigorous mathematical proof.
2. ______ Yes, I consider this argument to be a proof, although not fully rigorous.
3. ______ No, I think this argument does not meet the standards of a proof.
4. ______ Not sure, because I don't fully understand the argument.

Please explain your reasoning about **why** you think **this is** or **is not** a mathematical proof.

Argument 9

Claim. Let $f(x) = \ln x$. Then $f(x) \to \infty$ as $x \to \infty$.

Line:

1 Let a and b be positive real numbers with $a > b$.
2 Dividing both sides by b gives:
3 $a/b > 1$ (since b is positive).
4 $\ln(a/b) > 0$ (since $\ln x > 0$ when $x > 1$)
5 $\ln(a) - \ln(b) > 0$ (by the rules of logarithms)
6 $\ln(a) > \ln(b)$
7 Hence, for positive reals a and b, if $a > b$, then $f(a) > f(b)$.
8 Therefore, $f(x) \to \infty$ as $x \to \infty$.

For each question, circle the answer (1 to 5) that best fits your thinking about the argument.

A. Do you feel that you understood the argument that was presented?

1	2	3	4	5
There are fundamental details of the argument that I don't understand				I understand the argument completely

B. Are you convinced by this argument?

1	2	3	4	5
Not convinced at all				Completely convinced

(continued, next page)

C. Does this argument explain why the assertion is true?

1	2	3	4	5
No, the argument does not explain why the assertion is true				Yes, it really illuminates why the assertion is true

D. Would you consider this argument to be a mathematical proof?

1. _______ Yes, I consider this argument to be a fully rigorous mathematical proof.
2. _______ Yes, I consider this argument to be a proof, although not fully rigorous.
3. _______ No, I think this argument does not meet the standards of a proof.
4. _______ Not sure, because I don't fully understand the argument.

Please explain your reasoning about **why** you think **this is** or **is not** a mathematical proof.

10. Please provide some information about your personal and math background. These data help us check that we are gathering answers from a diverse group of students. Please check the choice that fits you best.

a) Ethnicity (check one)
_____ Hispanic or Latino/a
_____ Not Hispanic or Latino/a

b) Race (check one or more)
_____ American Indian or Alaskan Native
_____ Asian
_____ Black or African American
_____ Native Hawaiian or other Pacific Islander
_____ White
_____ Other; please specify: ________________

c) Gender (check one)
_____ female
_____ male

d) Class year (check one)
_____ First-year
_____ Sophomore
_____ Junior
_____ Senior
_____ Other, please specify: _________

e) Academic major (check one or more)
_____ mathematics
_____ natural science; please specify: _________
_____ engineering; please specify: _________
_____ non-science; please specify: _________

f) Are you preparing to become a K-12 teacher? (circle one)

Yes, elementary Yes, secondary No Maybe

g) What college math courses have you taken **before** this course ? List course names or numbers.

h) What **other** college math courses are you taking **this fall**? List course names or numbers.

i) What grade do you expect to receive in this course? (check one)

A+ _________	B+ _________	C+ _________	D _________
A _________	B _________	C _________	F _________
A- _________	B- _________	C- _________	

Other (please explain):

Exhibit 5A2: Instructor Ratings

Dear IBL Instructor,
We seek your assessment of individual students' learning from your course. We will compare your views with students' self-assessment of their learning from this course, in order to understand how well students' self-judgments of their mathematical learning correlate with the assessment of experienced mathematicians familiar with their work.

We ask you to give two ratings for each student: his/her overall level of expertise in mathematical knowledge and thinking at the START of the course, and his/her overall GAIN or improvement in mathematical knowledge and thinking by the END of the term. We ask you to distinguish, as best you can, students' *incoming* ability (which may depend on their educational background, preparation, and talent) from their *learning* in your course. Both of these can be factors in students' final performance or grade, so we ask you to separate them as best you can based on your observations of their work for your course.

Please focus on students' overall mathematical knowledge and thinking, even though you may have observed other kinds of abilities, growth, or learning gains in your course. And feel free to add any comments on this sheet, if you wish.

Please rate your students in comparison with other students in this course or at this level in your program. Use this scale for both ratings:

5 = very high, 4 = high, 3 = moderate, 2 = low,
1 = very poor or strongly lacking

First two letters of first name	**INITIAL expertise in mathematics**	**Overall learning GAIN in mathematics from your course**