## GEOLOGICAL SCIENCES

**Annual Assessment Report, 2014-15**

## Executive Summary

To determine if our measures of student performance actually align with our expectations and learning objectives, the Department of Geological Sciences examined our existing summative assessments in core courses for undergraduate majors. We categorized every question from midterms and finals in our core courses based on their cognitive level according to Bloom’s Taxonomy. We observed variation between faculty, as well as a general trend towards higher level questions on final exams than on midterms. The results served as a springboard for an ongoing discussion about what our assessments should look like in the future – both exams and beyond.

## Background

In 2014-15, our data collection was student-driven. Students in GEOL 501 (Teaching and Learning Natural Science) are upper division undergraduates and graduate students. As part of their coursework, students complete a mini research project on either classes they take or teach (as teaching associates). These projects allow us to investigate what happens in our clasess from the perspective of student researchers. Some examples of students’ research questions this year:

What makes online courses effective?

How do students use PDF’s of Powerpoint slides in undergraduate courses?

Is there a correlation between exam scores and attendance in introductory courses?

Does class size impact student performance in introductory courses?

Are test questions with image interpretation easier or harder for students?

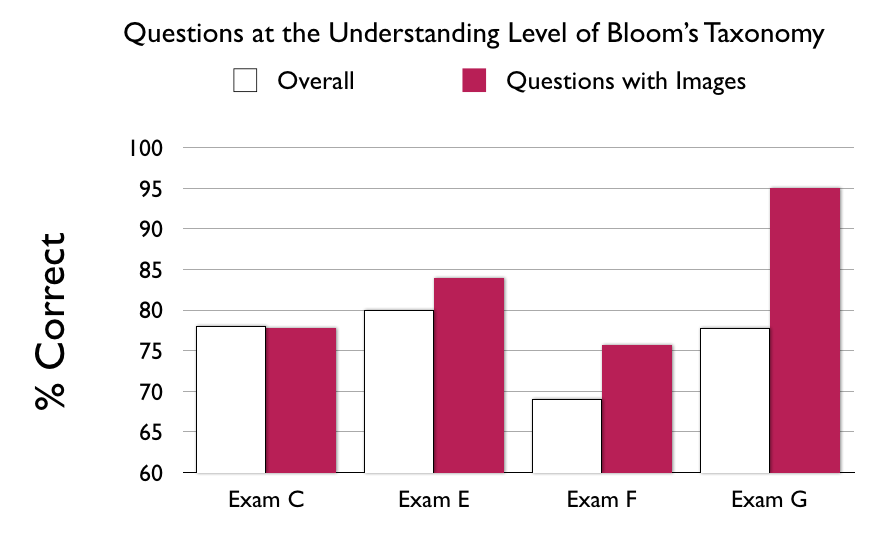
The emphasis was on large introductory courses rather than courses for our majors, perhaps because many of the research studies discussed in the class investigated such courses because of the large sample sizes available. Our core courses tend to be small and produce only qualitative data, a research approach that tends to be less familiar for our developing scientists.

The methods used by one study motivated a larger investigation into our core courses. Shawna Couplin, a graduating senior, wanted to know how students performed on test items with images. She recognized that questions using images required different skills than text-only questions, and therefore might be more challenging. She collected exams from 4 different instructors along with aggregate data about student performance on each item. As she examined the questions with and without images, she recognized that there were differences in the complexity or cognitive demand expected even within the questions with no images. For example, many of the multiple choice items were straight recall questions that required very little understanding. Other questions were more conceptual and asked students to apply their knowledge. Were any differences in student performance related to the images themselves, or the fact that questions with images were often higher order questions? Even within images questions, some questions were image-recognition while others required complex interpretations. She coded each question based on the level of Bloom’s taxonomy it required. She noted two obvious patterns in the data below:

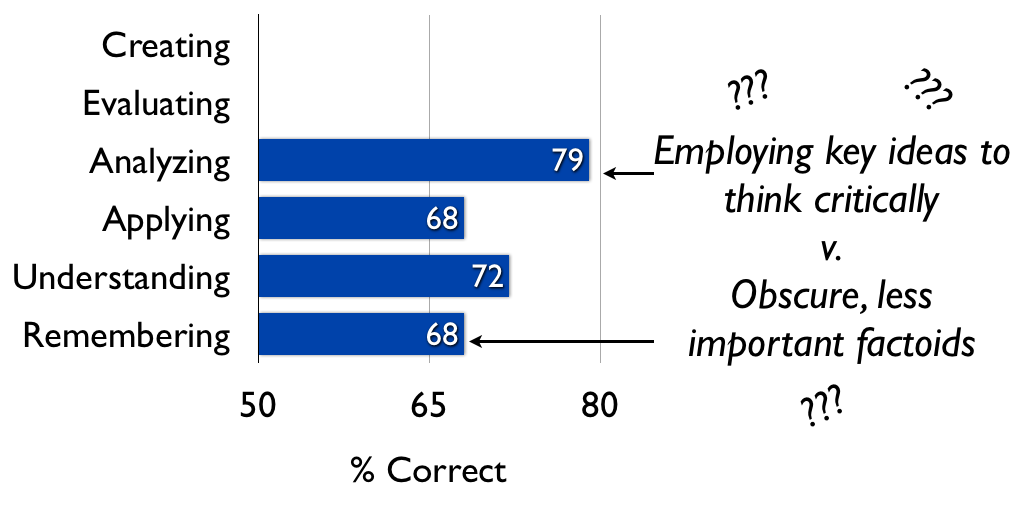
* No questions on these multiple choice exams got above the level of ‘Applying’
* There is a very large amount of variation between instructors, and even between different exams by the same instructor.



In terms of her original research question, there were too few questions using images at most cognitive levels and she was only able to make a direct comparison with questions at the “Understanding” level of Bloom’s taxonomy. She found that questions with images consistently produced higher scores than questions at the same level on Bloom’s taxonomy without images. Her main explanation for these results was that most professors rely heavily on image rich Powerpoint slides during lecture sessions. Having the assessment match the learning process could result in higher performance on the assessment. Alternatively, students may actually be wired to represent their understanding as internal visual images. When assessment items match this modality, students might perform better.



When our assessment team looked more closely at her data, we discovered a curious effect. Different levels of Bloom’s taxonomy were not necessarily easier or harder. In fact, analyzing two exams in more detail, we found that students scored higher on higher order questions than they did on lower order factual recall questions. This seemingly counterintuitive idea may be related to the fact that the remembering items were testing individual factoids that were often more obscure (because the test designer didn’t want to make the recall questions “too easy”). By contrast, the questions requiring analysis required students to think critically about the big ideas from the course material. Bloom’s taxonomy level and difficulty level are two independent scales; instructors can design challenging recall questions or easy application questions.



## Analysis of summative assessment in 2014-15 core courses

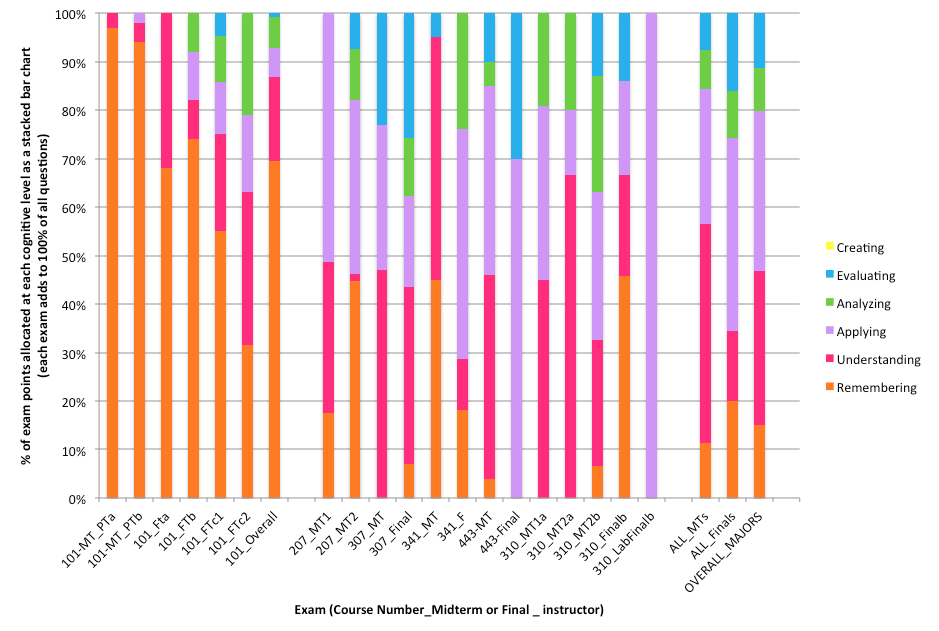
The extreme variation between instructors identified by Couplin’s research caused a great deal of discussion within our department. This is the first year we are implementing a new set of core courses for our majors. Part of the course proposals approved by EPC included a uniform strategy of embedded assessments, but this strategy was only vaguely designed during the course proposal stage. Now that we are faced with actually teaching the courses, serious thought must go into their design and implementation. The idea is to ensure that each course includes an “authentic assessment” where students produce work comparable to that produced by geologists in academic or industry workplaces. The instruction itself should be framed around achieving this performance expectation. Different classes in our majors have different culminating assessments, but the majority have midterms and final exams. We decided to investigate the question, “Across our curriculum, what have we been asking our students to do during summative assessments (exams)?” The awareness we gain about our prior practice can help inform our future assessment implementation and serve as a baseline for comparison. Note that our goal is not to evaluate student performance but instead to take stock of the expectations for learning expressed in our prior summative assessments.

### Findings

The graph below shows a comparison of different exams within our department. The group on the left includes a sampling of GEOL 100-level courses by different instructors as well as an aggregate for all 101 courses. The group in the middle shows the sequence of our core. The group on the right shows three aggregations for our core courses for majors (midterms, final exams, and overall). The stacked bars can be thought of like individual pie charts showing the relative proportion of different type classes. It should be noted that the items were weighted by point value, so a single item with a high point value receives a larger weight in our analysis.

#### Uncertainty in our findings

We should note that categorizing items based on Bloom’s taxonomy is subjective and requires knowing the context in which the questions were asked (i.e., when an instructor asks students to compare and contrast two ideas, this could be a very high level task if the situation is novel, or it could be a very low order task if an entire lecture session was devoted to comparing the two features and the student is only expected to summarize that session.). The questions were coded by an independent rater and not by the course instructor, so context was often assumed based on the amount of space provided for students to answer the question and assumptions about the typical content, textbook materials, and teaching of these materials. As a result, any given bar could easily be under or overestimated, perhaps by as much as 10-20%. Even with this subjectivity, some of the trends we observed are qualitatively robust. These trends are what we are relying on to engage in discussion – no faith is placed in the absolute quantities of any given item or midterm’s results.



#### Robust trends we observed

**We ask higher order questions of our majors than for our GE students**. On average, 70% of the GE questions were at the Remembering level while only 17% of questions for our majors were at that cognitive level. As with the general education courses, there was variation between classes and instructors, with one class asking absolutely no remembering questions on the final exam up to 45% remembering questions on a midterm in another class.

**We ask lower order questions on our midterms than our final exams**. On midterms, an average of 68% of questions are at the lowest levels of Remembering or Understanding. On final exams, that number is only 28%. With only one exception, every course’s final exam had higher order questions than midterms in the same course taught by the same instructor.

**Application questions are the most common in courses for our majors**. About 33% of all questions were application-based. We should note that this analysis does NOT include laboratory finals exams, which are administered as separate exams. About half of the lab classes have a final exam, while others have a final project.

**Truly creative endeavors are absent from all *exams* considered**. This is totally understandable given the limited timeframe of exams. The exam that came closest to a creative endeavor was a 24-hour take-home final exam (GEOL 443\_Final). Other summative assessments such as final projects are more likely to include such high order tasks, though none were examined in this year’s assessment analysis.

### Discussion

In previous years, we have assessed our students’ scientific thinking in 5 domains that relate primarily to SLO 3 within our program (“**demonstrate skills in** standard data-gathering and **data-analysis method**s in both lab and field settings”):

1. Hypothesis generation
2. Planning investigations
3. Analyzing data (*quantitative literacy*)
4. Evaluate Conclusions (*critical thinking*)
5. Reflect on Conclusions and Formulate Next Steps

Reports for 2012 and 2013 from our department document the room for growth in the Analyzing data and Evaluating Conclusions domains of scientific thinking. ‘Analyzing’ and ‘Evaluating’ are used in a similar manner to their usage in Bloom’s taxonomy. On average, our exams test at these levels only about 20% of the time, but a high fraction of the exams (especially midterms) have no more than 5% of the questions at the level of Analyze or above. These findings raise the question of whether or not we are giving students sufficient practice at these skills. If not, should we be? Or is an application focus for our questions most appropriate for the undergraduate level? The key word in our SLO is the ‘methods’ of data analysis, and we are clearly asking them to acquire and apply standard techniques.

While application focus questions are most common in our exams, a full 55% of our midterms and more than 45% of the items on our exams overall are at the Remembering and Understanding level. Should we consider eliminating these items in favor of additional time spent on applying, analyzing, or evaluating? Such assessment items require that students remember and understand the content.

**Missing pieces of assessment from this analysis**

The idea for this analysis came from one of our students and was implemented by the assessment liaison. As mentioned above, the faculty in our department were intrigued by this analysis, but really want to see our full picture of summative assessment that includes our laboratory exams and projects. These activities are inherently designed around tasks at different cognitive levels and were excluded from this report. Those items carry a larger weight in the minds of our faculty, so we clearly need to do a broader analysis of this missing piece. But this does raise an interesting question that if the lab items carry so much mental focus for our faculty, should we be doing something different with the weighting of our units or grades (our core lecture courses are 3 units with required 1 unit labs as corequisites). Some faculty tie the lecture and lab grades together while some report them separately. A lab grade tied together would decrease the overall weight of the lecture final exams and place the actual weight where it belongs on the lab components.

#### Useful comparisons

Below, we highlight a sampling of questions where questions on the left side are at a lower cognitive level than the questions on the right side that test a similar skill or disciplinary content idea.

|  |  |
| --- | --- |
| 1. Define the following:    1. Effective stress    2. pure shear    3. simple shear    4. …   *(Remembering/Understanding)* | 1. Compare and contrast the following terms—define what the terms mean and explain how they are similar, how they are different, how they compare and/or how they are related.   1. yield stress and viscosity 2. pure shear and simple shear 3. …   *(Evaluating)* |

|  |  |
| --- | --- |
| 1. Describe the bed-length balance test.   *(Understanding)*   1. Is this cross section balanced? (Hint: Use the bed-length balance test)   *(Applying)* | 1. In a geologic report that I read recently, I saw the following statement: “We used the bed-length balance-test to evaluate our cross sections. The four bed lengths that we measured were equivalent to within 1% of each other, demonstrating that these cross sections are accurate”.   In the space below explain, interpret, and critique this statement for a hypothetical student who is just learning to draw cross sections (in other words, explain, interpret, and critique this statement to one of next year’s structure students).  *(Evaluating)* |

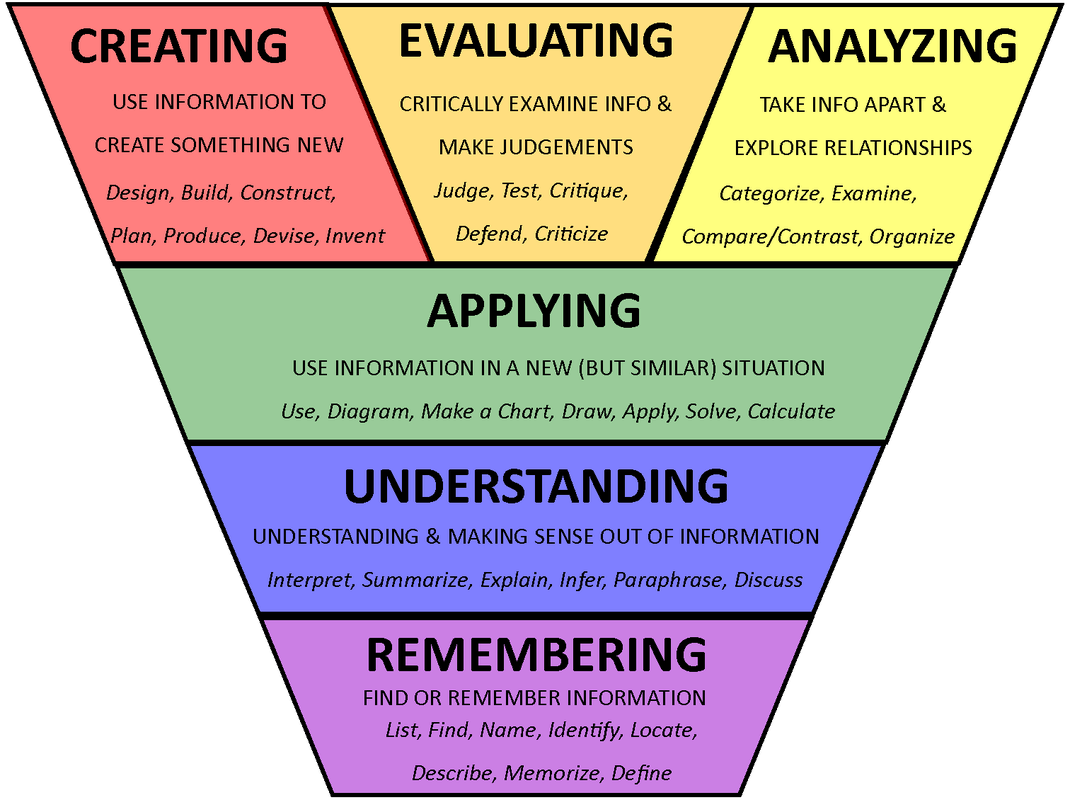
|  |  |
| --- | --- |
| You carefully study 4 clastic rocks and, based on their mineralogy, plot them on the ternary diagram as rocks *w*, *x*, *y* and *z* (see diagram at left). Which of these is the most mature? Explain your answer.  *(Analyzing)* | Rocks x and y have very similar compositions, but your labmate thinks that rock y is more mature. Give two reasons why your labmate might have come to this conclusion (i.e. what else might help you determine maturity?).  *(Evaluating)* |

|  |  |
| --- | --- |
| Find the equilibrium mineral assemblage in A and B.  *(Applying)* | Are the equilibrium mineral assemblages the same in A and B? If not please explain why not? What has changed?  *(Evaluating)* |

|  |
| --- |
| **An exemplary take home final exam:**  You are a geologist recently hired by the PH (Pima-Henry) Oil Company.  They have targeted a potential oilfield to be drilled, and have obtained data from 8 locations in the basin (shown below).  The main geologist on the project, Dr. Louie, was mysteriously abducted by a Hipparion, and the data from site E in addition to the basin analysis remains uncompleted.  Your task for your final exam is to interpret the stratigraphy at site E and complete the basin analysis to provide a detailed report on basin history.  You are given data from 8 locations within the basin that surround site E, including stratigraphic thicknesses for lithofacies observed, the relative time-of-deposition for each lithofacies (T1 is oldest, T5 is youngest), typical (but non-unique) photographs of the lithofacies, and information on paleocurrents.  You may also find the seismic line and subsidence data useful for your interpretation.  Please pay careful attention to the scale-indicators (ie. hammers or scales) in each photo.  GOALS:  1) Stratigraphic column from site E (15 pts) *(Applying)*  2) Assuming that each time interval "T" lasts 5 million years, create a deposition rate plot for Column E and show the deposition rates for your lithofacies (15 pts) *(Applying)*  3) A fence diagram that represents the basin lithofacies and shows your ability to correlate strata across the basin (20 pts).  Use at least 3 fences. *(Applying)*  4) Paleogeographic maps from times T2 and T4. (two separate maps, 20 pts total) *(Applying)*  5) 1-3 page write-up summarizing the basin history over time and any interpretations of basin formation.  You should use any references from class or elsewhere (30 pts). *(Analyzing/ Creating)* |

## Next steps

Our department is implementing our new curriculum for the first time this semester. Our assessment liaison will work with each faculty member one-on-one to discuss summative assessments. We will also develop a template for culminating assessments so that all faculty have a model from which to build (and potentially rubrics for grading them, if it seems like they can be applied universally).



by [**Maggie Hos-McGrane**](https://www.blogger.com/profile/06906722339671067160), <http://www.maggiehosmcgrane.com/2014/09/flipping-grade-4-and-flipping-blooms.html>