Proceedings of the Interdisciplinary STEM Teaching and Learning Conference

Volume 1 Article 5

5-2017

Assessing Student Progress and Performance across the Curriculum

Christopher G. Brown
Georgia Gwinnett College, cbrown37@ggc.edu

Jennell M. Talley
Georgia Gwinnett College, jtalley@ggc.edu

Bekah Ward Dr.
Georgia Gwinnett College, rward1@ggc.edu

Christopher I. Brandon Jr.
Georgia Gwinnett College, cbrandon@ggc.edu

Jill Penn
Georgia Gwinnett College, jpenn1@ggc.edu

See next page for additional authors

Assessing Student Progress and Performance across the Curriculum

Abstract
Evaluation of student learning is of paramount importance to the educational community and allows reflection on program successes and weaknesses; however, best practices are hotly debated. This project designed and implemented an assessment system in which an identical, mixed-format assessment was given to all levels of students in the Georgia Gwinnett College biology program at the start of the semester for academic years 2014-15, 2015-16 and Fall of 2016. The assessment contained multiple choice and free-response questions, and evaluated lab reports from core courses in the biology program. This system allows for longitudinal assessment of students, provides quick results for timely action, and can allow analysis of interesting demographic questions. We found student achievement on program goals was lower than previously assessed and student performance on multiple choice questions was higher than free-response questions. There was a modest, but temporary, gain in performance on the ability to effectively communicate science. Additionally, males outperformed their female counterparts and Hispanics underperformed their non-Hispanic peers.

Keywords
program goals, longitudinal comparison, higher education, formative and summative assessment

Authors
Christopher G. Brown, Jennell M. Talley, Bekah Ward Dr., Christopher I. Brandon Jr., Jill Penn, and Elisabeth Javazon

Assessing Student Progress and Performance across the Curriculum: A Tool to Evaluate Program Success Quickly

Christopher Brown, Georgia Gwinnett College
Jennell Talley, Georgia Gwinnett College
Rebekah Ward, Georgia Gwinnett College
Christopher Brandon, Georgia Gwinnett College
Jill Penn, Georgia Gwinnett College
Elisabeth Javazon, Georgia Gwinnett College

Abstract: Evaluation of student learning is of paramount importance to the educational community and allows reflection on program successes and weaknesses; however, best practices are hotly debated. This project designed and implemented an assessment system in which an identical, mixed-format assessment was given to all levels of students in the Georgia Gwinnett College biology program at the start of the semester for academic years 2014-15, 2015-16 and Fall of 2016. The assessment contained multiple choice and free-response questions, and evaluated lab reports from core courses in the biology program. This system allows for longitudinal assessment of students, provides quick results for timely action, and can allow analysis of interesting demographic questions. We found student achievement on program goals was lower than previously assessed and student performance on multiple choice questions was higher than free-response questions. There was a modest, but temporary, gain in performance on the ability to effectively communicate science. Additionally, males outperformed their female counterparts and Hispanics underperformed their non-Hispanic peers.

Keywords: program goals, longitudinal comparison, higher education, formative and summative assessment

Acknowledgements: This work was funded by a grant from the University System of Georgia STEM Initiative II Project. We would like to thank all the faculty in the biology discipline at Georgia Gwinnett College who volunteered their services and class time to make this project a success.

Introduction

Collegiate programs frequently determine a set of goals that reflect the required outcomes of the program. Evaluating student performance on program goals is of vital importance to determine progress through the program and identify targets for future remediation (Boyer 1990). In other words, faculty should know how well their students meet the goals set for them and adjust accordingly. Ideally, students become increasingly proficient in content knowledge and essential skills pertaining to their given field, i.e., seniors should display a higher mastery of outcomes than juniors, who are more capable than sophomores, and so on (Gardner et al. 1983). Graduates should possess the abilities expected of a budding professional and therefore be capable of success in a relevant field or post-graduate program. The ongoing process of improving assessment and evaluation began in earnest in 1918, has since experienced many significant changes in focus, including the Reagan administration report *A Nation At Risk*, and more recently has received new impetus from the Obama administration's "College Scorecard" initiative (Sims 1992). Furthermore, faculty should play a creative and consistent role in the development and implementation of any program assessment so they and their students can benefit (Emil and Cress 2014, Stohlman 2015).

Content knowledge, conceptual understanding and acquisition of essential skills often determine student progress over the duration of an academic program. Metrics used to determine proficiency on goals include longitudinal standardized exams, exit exams, portfolio building, and capstone or senior field experience analysis (Banta et al. 2009, Ruben 2016). Each metric has benefits and costs and proper assessment is often time-consuming. Exit exams are relatively quick and provide data comparable across students and campuses, but they often do not directly address a given institution's progress towards specific goals (Astin 2012). Exit exams also do not provide a baseline of performance or a sequence of progress; perhaps student performance at the end of a program is the same as it was at the beginning, or students make gains but fall short of a preset numerical goal (Tucker 2006). Students may also refrain from putting forth their best effort on exams not linked directly to success in a course. Alternatively, questions relevant to course goals can be embedded in exams given in diverse courses (Astin 2012), but the data then reflect progress in courses, not in programs, and must be aggregated, thus losing specificity. This type of assessment can also introduce bias from professors with expectations for individual students

DOI: 10.20429/stem.2017.010105

with whom they are familiar (Imrie et al. 2014). Neither exit exams nor exam-embedded questions provide true longitudinal data, which can allow educators to pinpoint areas of weakness in the curriculum. Here, we describe a comprehensive method of program evaluation that provides detailed longitudinal assessment of program goals while minimizing time constraints and mediating potential biases.

The described changes in evaluation methods took place within the biology discipline in the school of Science and Technology at Georgia Gwinnett College (GGC). GGC is a relatively new public, four-year, access institution that has grown rapidly in the last decade (from fewer than 100 students to almost 13,000 since being established in 2005). However, classes remain relatively small; biology classes are usually 24 students. Additionally, the college is highly diverse, 73% non-white, and has many students (36%) who are the first in their family to attend college. Importantly, biology majors at GGC are demographically representative of the entire school.

Biology majors are expected to show proficiency in content and laboratory skills as determined by seven program goals designed by the discipline's faculty (Table 1). Previously, program goals were assessed by measuring course goals using exam-embedded questions given to every student. These multiple-choice questions were on the final exam of every section of every course and graded by the corresponding professor. If an evaluation tool is to serve as formative for the faculty to modify the program, it must reveal a chain of causality; meaning instruction provides (or does not provide) increasing content knowledge for students (Hawthorne 1989). Because the previous method lacked consistent, unbiased, longitudinal assessment of any of the program goals, we were unable to robustly assess content achievement in any program goals.

Table 1. Program Goals for the biology program at Georgia Gwinnett College.

1. Communicate in oral and written form the ability to locate, critique, and utilize scholarly resources.
2. Demonstrate proficiency in basic lab skills and experimental design.
3. Apply basic chemistry and math to the study of the life sciences.
4. Know the structures and functions of cells.
5. Know the structures and functions of biomolecules (nucleic acids, pro teins, lipids, carbohydrates).
6. Explain the sources of genetic variation and determine patterns of inheritance. Describe the role of evolutionary mechanisms in biological diversity.

For many years, students at all ranks had consistently reached our 'satisfactory rate' of 70% or more for every goal. These high levels of success could either have been due to genuine measurement of knowledge and skills or insufficient rigor on questions. As a result, faculty confidence in assessment was low. Surveys conducted at the discipline level suggested faculty did not agree with the following assertions: (1) our current assessment methods significantly help to inform teaching, (2) accurately reflect our majors' knowledge and skills, or (3) the amount of time and energy we spend on assessment is appropriate. These results broadly match faculty opinion of assessment elsewhere (Emil and Cress 2014).

Therefore a new measurement tool and evaluation method was designed by faculty in the biology program in order to improve our ability to discern areas in need of remediation. The measurement tool consisted of a single comprehensive exam given to randomized subsets of students from all ranks in the core courses required for completion of a biology degree. The exam included open-ended questions (free response) in addition to multiple-choice questions to evaluate application, rather than simple retrieval. The importance of free response questions in the assessment of higher order learning is well established (e.g., Birenbaum and Kikumi Tatsuoka 1987, Becker and Johnston 1999, Nichols and Sugrue 1999, Resnick and Zurawsky 2007, Heyborne et al. 2011). Essay and short answer questions can allow students to cover a wider range of content than a multiple choice or matching question, they more easily assess the integrative and/or applied levels on Bloom's taxonomy as students are typically asked to "apply" or "explain", and allow students to express their reasoning for a given answer, providing important information for formative assessment. The exam was administered to a random selection of courses at the *beginning* of the semester, thus uncoupling student performance with professor evaluation, 'teaching to the test,' or confusing student knowledge with 'cramming' for a final. Identical tests were given to students at all levels (freshmen, sophomores, juniors, and seniors), providing consistency and facilitating an instantaneous, longitudinal comparison of student performance before completion of the program. In addition to the standardized tests, we gathered a sample of lab reports from classes common to all students in the program. Student-written lab reports facilitated assessment of goals 1 and 2 which pertain to scientific communication and experimental design. Tests, as well as lab reports, were scored simultaneously by a panel of biology faculty from diverse sub-disciplines.

DOI: 10.20429/stem.2017.010105

Grading was blind; graders had no knowledge of student identity, rank, or course.

We hold this assessment method reduces subjectivity, while providing detailed analysis of student progress through a program in time to affect change. It is efficient and provides faculty full control over program assessment while not being overburdensome. Here, we provide a description of the method with brief examples of the data it provides. Our method is not specific to biology or STEM programs and easily could be applied to other curricula at other institutions.

Methods

Design and Administration of the Assessment Exam

To assess the program, we designed and administered a standardized test to a sample of students at each level of the biology program. The exam consisted of twenty to twenty-five multiple-choice questions and one or two open-ended, short answer questions (free response). At first, questions were created by faculty in the discipline, but recent versions of the exam consisted of vetted questions derived from open-source concept maps (e.g., American Association for the Advancement of Science: http://assessment.aaas.org/topics/ and San Diego State University Division of Undergraduate Studies: http://go.sdsu.edu/dus/ctl/cabs.aspx). Each question was directly linked to a program goal.

The exam was administered to a randomized, representative set of core biology major courses during the first week of the class; if the course had a lab then the test was administered during lab. Half of the common courses were assessed in the fall and half in the spring. Therefore, all courses common to the core were evaluated each academic year. The program goals evaluated are shown in Table 1. A total of 558 biology majors from fall of 2014 through fall of 2016 were evaluated. Students were required to take the exam, but were asked for informed consent to allow use of their responses in publication. Only data from students who gave consent are presented in this paper. Exam questions were optimized over the duration of the project, thus no questions were used on more than one exam during the duration of the study. Tests were given by a designated test administrator (i.e., a faculty member not associated with the course) at the beginning of each semester. Test administrators read from a standardized script which provided reasoning for the exam. At first, students provided an anonymous identification code, but more recently, students provided their

student ID in order to analyze additional information collected by the Office of Academic Assessment, such as college admission test scores and grade point averages (GPA) and true academic rank. The anonymous identification code or student ID were on both the multiple choice and free response sections of the exam to ensure easy tracking of individual student performance.

Grading Exams

Multiple-choice questions were scored with Scantron ScoreIT software. For each free response question, a panel of full-time faculty graders worked collaboratively to create a rubric before grading (see Stevens and Levi 2013 for information about rubrics). Graders were blind to the identity, current course, and rank of the students being assessed. This was done in an attempt to remove potential biases that can arise when grading the work of students with knowledge of expected performance. Additionally, faculty were given ten control questions used for standardization of the free response to attempt to discern grading bias. For the first two years, faculty who volunteered for grading were awarded a modest stipend for their day's work. More recently, administration and grading of the exam fell to the program goals committee.

Collection of Demographics

At the time of the exam, a separate survey was given to students to assess demographic data such as gender, age, and race. To avoid influencing performance by drawing attention to cultural groups, i.e., stereotype threat (Steele, et al. 2002), this survey was given only after completion of the content sections of the exam. The demographics survey also gathered data about major, career plans, enrollment status (full-time or part-time), and workload.

Determining Rank

Unfortunately, determining level in the program (e.g., freshmen, sophomore, junior, or senior) is difficult since students often do not correctly report their rank, or have deferred taking courses in the program for several semesters resulting in their rank by hours not equivalent to progress in the major courses. Therefore, we determined rank in the program by combining information on each student's self-reported rank, the class in which the test was administered, and which classes they report having passed. Together, these responses provide a more accurate measurement of each student's rank with

DOI: 10.20429/stem.2017.010105

respect to their degree. When 'rank' is mentioned throughout the paper, it is their rank using the above described 'algorithm'. Our ranks include freshmen, sophomore, junior, and senior

Grading of Lab Reports

Our first program goal addresses students' ability to communicate scientifically and perform scholarship. To assess this goal, we asked professors from core biology courses to submit lab reports assigned during the semester. Lab reports were stripped of class and student identification when graded, but unfortunately, they could not be completely anonymized because the subject matter of the class dictated the subject of the lab report. Thus faculty who have taught the course were potentially able to surmise the course of origin. Demographic data was not taken for students submitting lab reports. Therefore, lab reports were analyzed by comparing classes, which roughly corresponded to rank. To control for professor grading differences, faculty scored lab reports together in the same room, used the same grading rubric and tried to standardize grading using a 'practice' lab report. Additionally, faculty were unknowingly given five of the same lab reports to allow for detection of significant differences in grading. Faculty who volunteered for grading were awarded a modest stipend for their day's work.

Statistical Analysis

Difference between means were tested with Student's t-tests and ANOVA. Significant differences among groups were compared using Tukey-Kramer post hoc tests. Comparison of scores on multiple-choice and free response questions was performed with paired t-tests with individual students as replicates. Sample sizes vary depending on the comparisons being tested and whether or not the particular exam required students to provide the relevant information. Analysis was performed using JMP 13 statistical software from SAS.

Results

Demographics

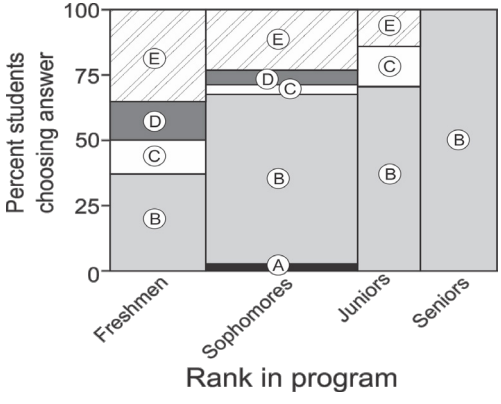
Race and ethnicity data, which were only measured in the fall of 2014, matched those of the biology department and school as a whole, indicating our sample was representative. 63% of students tested were female, 57% of students

were traditional college age (18-22 y), and 82% maintained a full-time college schedule (12 credit-hours or more). 82% of students had a job outside of school, with 18% of the total working more than 30 hours a week. A third of students indicated English was not the primary language spoken in their home. 47% of students surveyed intended to enter medical school after completing their bachelor's degree, 29% planned on attending graduate school, and the remaining 24% were split between careers in other health professions, education, or an unlisted field. Career plans did not differ noticeably across ranks.

Item Analysis

Using ScoreIt, individual questions were analyzed to evaluate student performance across ranks in the program and to identify moments in the academic experience where key student misconceptions were addressed. An example of the data available by question is shown in Figure 1; it shows the percentage of students who chose each answer (A-E) for each class rank. More students progressively chose the correct answer B, while E was progressively chosen less frequently. Answer A appears to be a distractor, while answer D is eliminated as a plausible choice by students by their junior year. Using this information, we can analyze each question to determine how students progress through the program at a conceptual level.

Figure 1. Diagram of student choices on an example multiple-choice question across rank. Each multiple-choice question linked to a core concept was analyzed individually using ScoreIt and JMP. The relative proportion of answers was broken down by rank within the program. Column width corresponds to sample size.



DOI: 10.20429/stem.2017.010105

In addition, ScoreIt provides point-biserial correlation analysis, which correlates (1) the likelihood each question is answered correctly with (2) the students' overall grades on the exam (Varma 2008). A question with a low point-biserial value is one more likely to be answered correctly by students who did poorly on the exam overall than students who did well overall. Such questions should be evaluated for confusing wording or for not matching the style or content of the rest of the exam.

Overall Scores

Mean scores of all tests combined was 48 +/- 17%, well below the historical goal of 70% set by the program's faculty. There was a significant interaction between student rank within the program and the semester the test was given (Fig. 2), indicating differences in the questions on the exam across semesters. Full factorial ANOVA analysis confirmed both rank and exam are significant determinants of overall score (Table 2). However, most tests showed a significant jump only between incoming freshmen to first-semester sophomores. After the freshman year, there were no differences among the top three ranks, excepting the fall 2016 exam, when seniors scored significantly higher than their lower-ranked peers.

Influence of Sex and Ethnicity

Effects of demographic differences were also assessed and we report a few intriguing findings here. Across exams, males performed significantly better than females (male score = 52 +/- 18, female score = 45 +/- 16, $t = 4.3$, $df = 446$, $p < 0.0001$). Additionally, students self-reporting as non-Hispanic performed significantly better than Hispanics/Latinos (non-Hispanic score = 49 +/- 18, Hispanic/Latino score = 44 +/- 15, $t = 3.3$, $df = 217$, $p = 0.001$). This reduction seems to only apply to students from Hispanic backgrounds in which English is not the primary language spoken in the home and did not hold for other ethnicities with English as a second language (Fig. 3). Indeed, comparison of overall score on the exam suggests that among Hispanic students, the language spoken at home is correlated with content acquisition. This is not the case for other ethnicities.

Figure 2. Mean score on multiple-choice per semester per rank. The dotted line represents the traditional passing score of 70%. Means +/- s.e. shown. Means sharing the same letter do not differ significantly at the 95% confidence level

based on the Tukey mean comparison method.

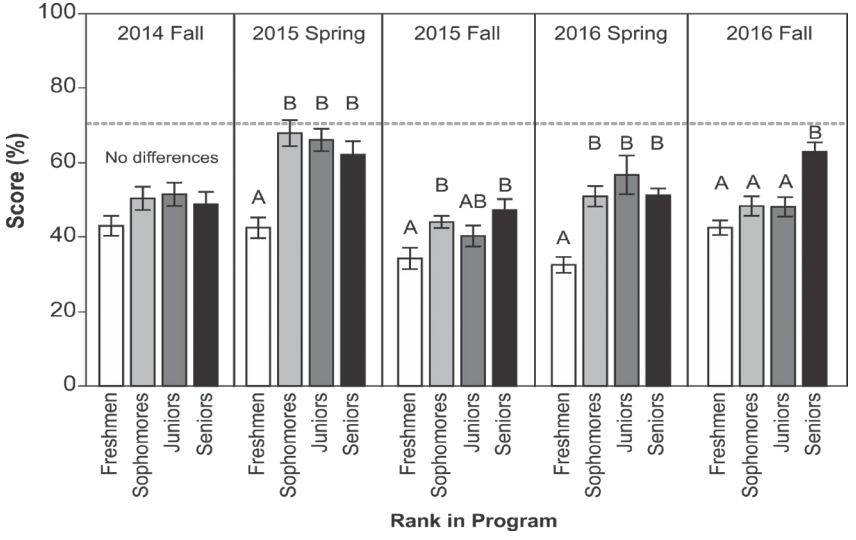
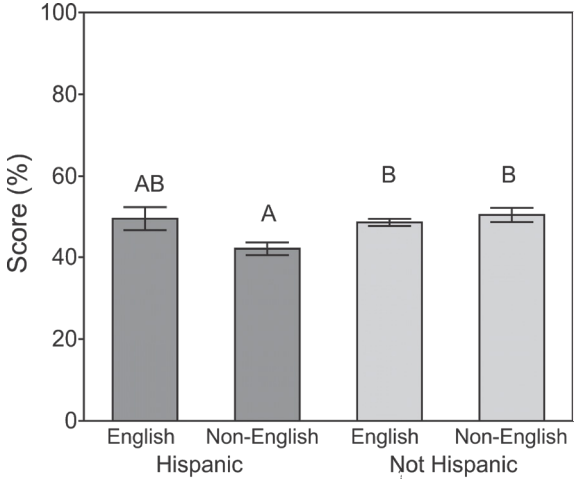


Table 2. Full factorial ANOVA showing significant differences among ranks within program, but also a significant effect of the semester in which the test was given.

Source of variation	Degrees of freedom	Sum of squares	Mean squares	F	P
Semester	4	14527.18	3631.80	16.13	**<0.0001**
Rank in program	3	18942.27	6314.09	28.03	**<0.0001**
Semester x Rank	12	8196.08	683.01	3.03	**0.0004**
Error	593	133539.84	2535.58		
Total	612	181696.86			

DOI: 10.20429/stem.2017.010105

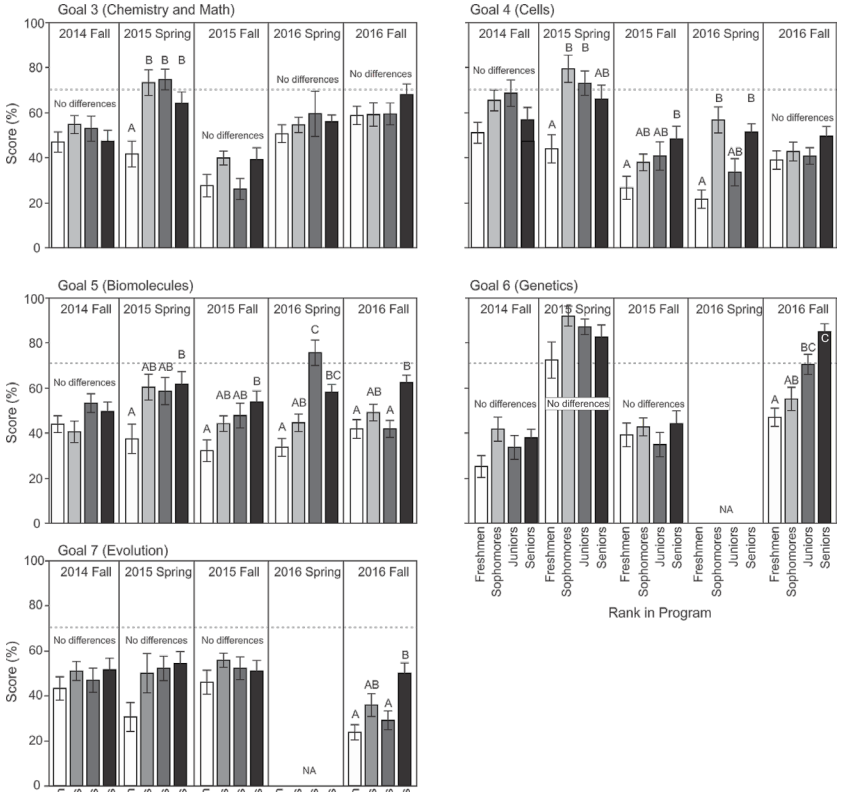
Figure 3. Mean score of Hispanic and Non-Hispanic students from homes that either speak English as the predominant language versus another language. Means +/- s.e. shown. Means sharing the same letter do not differ significantly at the 95% confidence level based on the Tukey mean comparison method.



Scores across Goals

In addition to being able to determine overall progression through the program and evaluating the effects of specific demographics, using our method, we were also able to assess if any patterns existed for each goal. Figure 4 shows performance varied across goals and differently between exams. The only mildly consistent trend is freshman tend to do worse on the goals compared to all other rank of student. One major exception was goal 7. Students across all ranks in the program consistently scored lower on questions pertaining to evolution, regardless of the exam administered.

Figure 4. Mean score in multiple-choice for each goal per semester per rank in program. Progress on each goal for each exam given (semester). Goal 6 and 7 were not assessed in the spring of 2016. The most common effect is a difference in performance between freshmen and the other ranks. Means +/- s.e. shown. Means within semesters sharing the same letter do not differ significantly at the 95% confidence level based on the Tukey mean comparison method.



Overall Scores on Free Response Questions

The free response, short answer questions targeted all of the goals over the course of this project. Because the biology discipline has recently been interested in gaining insight into student's understanding of goal 4, it was assessed most often during this study. The average of all the free response scores, broken down by rank and goal is shown in Figure 5. Similar to the multiple choice section of the exam, freshman often underperformed their higher ranking peers. Again goal 7 showed the lowest gains overall, whereas goals 3 and 5 showed some of the highest gains.

Interestingly, for most goals, students scored significantly higher on the multiple-choice versions of assessment than the free response, excepting goal 3 (Chemistry and Math), which showed the opposite result (Table 3).

Figure 5. Mean score on free response questions for each goal per rank in

DOI: 10.20429/stem.2017.010105

program per semester. Means +/- s.e. shown. Means within semesters sharing the same letter do not differ significantly at the 95% confidence level based on the Tukey mean comparison method.

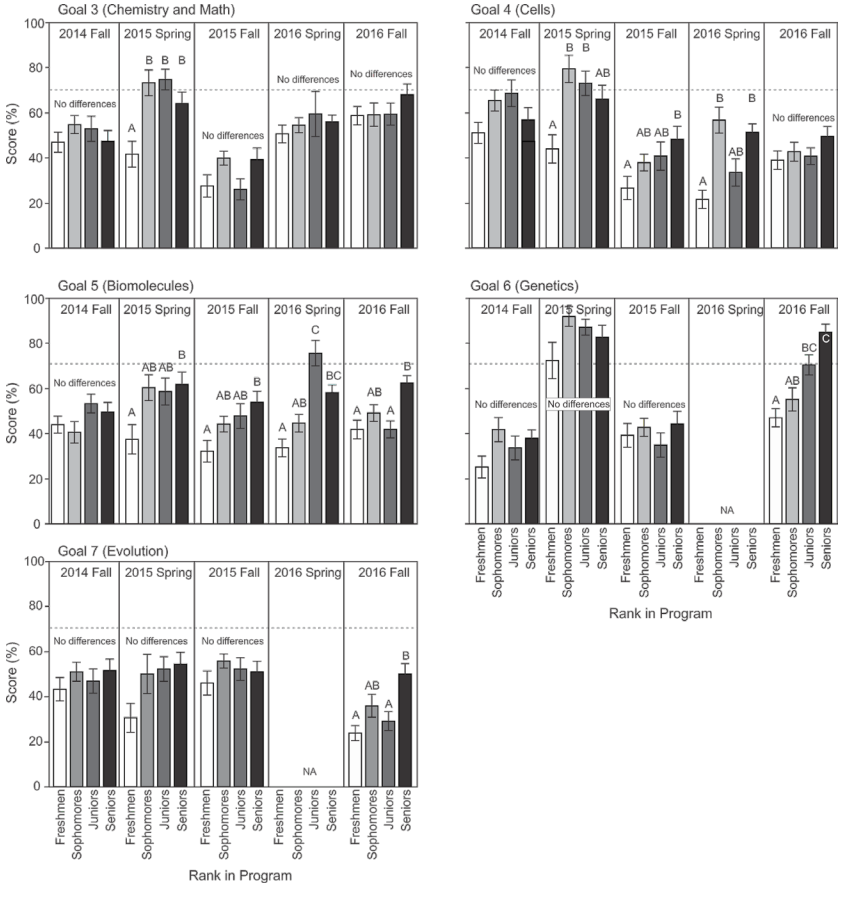


Table 3. Comparison of performance on multiple-choice versus free response questions. Analysis consisted of paired t-tests which control for among-student differences.

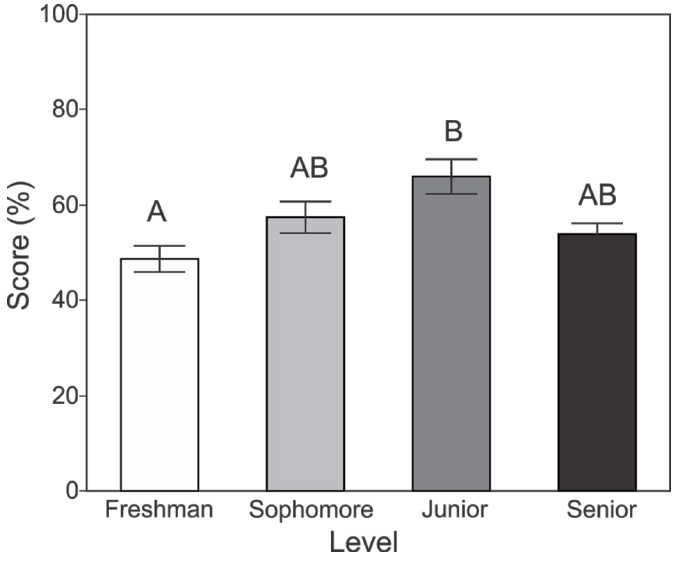
Goal	Semester	N	Multiple-Choice (mean %)	Free-response (mean %)	Mean difference	*t*	*p*
Goal 3 (Chemistry and Math)	Fall 2014	69	51.44	60.20	-8.75	-2.33	**0.0228**
	Fall 2015	58	32.75	74.38	-41.62	-10.36	**<0.0001**
Goal 4 (Cells)	Fall 2014	61	63.01	51.68	11.33	3.98	**0.0002**
	Spring 2015	38	71.05	41.56	29.49	7.91	**<0.0001**
	Fall 2016	21	52.38	50.95	1.45	0.578	0.56
Goal 5 (Biomolecules)	Fall 2014	39	49.25	44.82	4.432	2.87	**0.0054**
	Spring 2015	37	56.76	60.95	-4.19	-0.86	0.3921
Goal 6 (Genetics)*							
Goal 7 (Evolution)	Fall 2015	31	53.76	33.07	20.69	5.10	**<0.0001**

Analysis of Lab Reports

Lab reports were collected from each of the core classes that has a corresponding lab and graded by a committee of volunteer faculty in the fall of 2014, and spring of 2015 and 2016. Faculty worked together on a specified day to complete grading of the lab reports using a standard lab report rubric and started grading after first standardizing to one lab report. There was a difference between freshman and junior level courses, though this did not persist into the more senior-level course (biochemistry) (Fig. 6).

DOI: 10.20429/stem.2017.010105

Figure 6. Mean score on lab reports across course level. Lab reports were submitted by professors of core courses and did not come with student demographic data. Therefore, lab reports are divided by approximate level of the course. Means +/- s.e. shown. Means marked with different letters are significantly different.



Discussion

Our proposed method is efficient, informative, and effective. Our pilot program shows the assessment tool provides actionable information in the first weeks of a semester with minimum impact on student, professor, or class time and it has already provided novel data, unavailable using our previous method, which suggests areas of targeted remediation. For instance, our data indicate previous assessment methods overestimated performance, as scores differed greatly from the typical 70-80% scores (Fig. 2). Unfortunately, the progressive acquisition of core biology content goals was not found significant in these data, neither in the exam nor our evaluation of lab reports, although there are suggestions of improvement over the duration of the program, particularly after freshman year (Figs. 2, 4-6). These results are somewhat disconcerting, but provide useful information to begin addressing the issues. For instance, the spring of 2014 showed no gains overall in any goal; which could be due to spring

2014 being the first semester we designed and implemented the assessment. Afterward, deliberate effort was made to validate the questions used for assessment.

By analyzing student choices on individual questions across ranks (as in Figure 1), we provide a longitudinal measurement tool and a potentially powerful way to identify which courses address specific student misconceptions. Or alternatively, these data can reveal times in a student's academic career when a misconception is not appropriately dispelled or potentially created. This is even more impactful as the data become validated historically. Obviously, this requires a reusable measurement tool which is currently still under development in our institution.

Our method allows easy evaluation of each program goal individually, and we did find adequate gains, as well as higher overall scores, for some goals, suggesting satisfactory performance of our program in these areas. Other goals, however, are in need of immediate focus, for example, goal 7, evolution (Figs. 4 & 5). Student understanding of evolution is often lagging, especially in the United States where nearly 40% of Americans profess denial of the theory (Miller et al. 2006). One possible use of these data would be to identify and assess a key misconception or alternative conceptions, such as how natural selection works, a major tenet of evolutionary theory. We can examine the misconception via the granularity of item analysis (Fig. 1) and by designing a module could remediate the issue. Afterward, the same assessment question could be given to all students who took the class in which this module was tested, but at the start of the next semester. Remediation, or lack thereof, would have strong support. Such evidence would provide an argument to disseminate the use of the module, or to ask for funds for additional supplies to further address it. This type of immediate action planning is quite possible using our method.

Performance on lab reports also shows improvement through the ranks, however with important caveats. The lab reports we assessed were provided voluntarily by professors teaching core courses in the curriculum. Therefore, they cannot be associated with individual students and lack demographic data. Instead, we approximate student rank using the course in which the lab report was assigned. Unfortunately, different courses have different requirements for their lab reports and lab reports are based on vastly different experimental styles. Therefore, we cannot guarantee graders are not influenced by their expectations of the course. Remarkably, there is a decline in progress at the senior level

DOI: 10.20429/stem.2017.010105

(Fig. 6). Lab reports at the senior level came exclusively from biochemistry courses, which are taught by both biology and chemistry faculty, who often have different visions of the style desired in a lab report. Again, we cannot account for differences in professor requirements, but we suggest there may be discipline-specific differences as well. One hindrance in students' ability to properly communicate science may be related to varying expectations and standards for lab reports or other written projects from different subdisciplines.

Free response questions often provide more thorough assessment of student skills and knowledge and can address concepts higher on Bloom's taxonomy (Biranbaum and Tatsuka 1987), although the data on this is mixed (Hogan 1981). However, when used for formative assessment, the choice of question type can also influence future student achievement (Heyborne et al. 2011). We found students typically performed better on multiple-choice questions than free response for the same goal in the same semester. One notable exception to these findings is goal 3 (Chemistry and Math) (Table 3). Perhaps students are more accustomed to word problems in chemistry and math or are more likely to work through a problem rather than guess, when choices are not provided. These results may also relate to the level of Bloom's required for MCQ vs. free response questions.

The inconsistencies of exam questions, demonstrated in Figure 2, do warrant further investigation into the style of questioning. Perhaps these differences are because of the classes students have taken or are due to differences in question difficulty. However, because exams were given during different semesters, the student body itself may have changed. This is especially likely given the rapid growth of GGC. In the future, exams could be cycled to more directly compare progress over time. We are currently investigating using vetted questions from published sources to better standardize our exam.

It is important to note this type of longitudinal approach is not without its critics within the field of assessment and evaluation (Yorke and Zaitseva 2013). Astin (2012) argues measurement tools similar to ours are not informative because there are too many confounding factors to determine causality. Was it, for example, passing a genetics course that shifted aggregate junior's answers on question 7 to C? Is it true that upperclassmen are always a representative sample of the cohort of freshmen they were several years ago? Sometimes factors as basic as retention complicate the data. Additionally, because the multiple choice and free response sections of our assessment are likely considered low-stakes

by students, there is concern students do not take the test seriously. Motivating students can increase their performance on low-stakes assessments (Hawthorne et al. 2015). To encourage earnest participation and reduce student anxiety, we gave our assessment the first week of class by a faculty member not directly linked to the class. Additionally, a standardized script was read emphasizing the importance of their participation and how it will benefit their program and thus their education in our program. This was done in an attempt to increase their intrinsic motivation for doing their best on the exam.

One of our ongoing attempts to address some of the above concerns is the use of traceable identifiers for students who take the exam. This will allow us to compare scores with Grade Point Average as well as entry exam scores. Additionally, because the exam will continue to be given each year, students will likely take the assessment more than once in the course of their time at GGC. This allows us to examine a cohort (albeit quite small) for whom we can say with more confidence our program affected. This pool could be expanded by intentionally choosing classes with students already tested.

Despite caveats in longitudinal assessment, the data regarding Hispanic and male students do not rely on those same assumptions about progress and are therefore possible sources of insight into our institution, if not all higher education. Students of Hispanic origin scored lower than non-Hispanic students (Fig. 3). This is most likely due to students using English as a second language (ESL). Many schools provide resources to aid ESL students (Kim et al. 2015) and GGC is no exception. It is informative to know our data identified the difficulties dual language students face and point to further differences based on student origins (see Hambleton et al 2004 for more).

Of note are the consistent trends that self-identified males perform better in aggregate than females. Although similar results have been reported elsewhere, are far from novel (e.g., Hill et al. 2010), GGC may provide an atypical example given that most biology majors are female (62% in 2014-2015, Runck 2015). Despite being the majority, females appear to have lower educational content acquisition. This suggests an avenue for possible programmatic remediation and could relate to the lower rate of employment of females in STEM careers (Beede et al. 2011). Further investigation is required.

Because we were interested in how students were progressing through our biology program, the portfolio or exit exam approach is not an appropriate tool for our purposes. We want to understand which areas of our program are

DOI: 10.20429/stem.2017.010105

doing well and which may need more attention, thus at GGC we more closely approximate a value-added approach to assessment, with respect to both our stakeholders (students, administration, state education officials) and the concerns of our faculty. This approach is largely due to the nature of our institution and the associated mission. However, this does not preclude the use of salient data for formative assessment of our work as educators. Specifically, the use of traceable identifiers may allow us to measure specific modules for the effect on remediation of key misconceptions. The granularity of the measurement tool we have created allows for a potentially powerful lens to examine the effect of specific changes in course content or emphasis. Overall, we find this method generally easy to use and unique in its ability to provide an abundance of diverse and useful information related to our students' progression through our program.

References

"AAAS Science Assessment Beta." AAAS Science Assessment ~ Topics. N.p., n.d. Web. 07 Mar. 2017. <http://assessment.aaas.org/topics/>.

Astin, A. W. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. Rowman & Littlefield Publishers.

Banta, T. W., Jones, E. A., & Black, K. E. (2009). *Designing Effective Assessment: Principles and Profiles of Good Practice*. San Francisco, CA: JosseyBass.

Becker, W. E., & Johnston, C. (1999). The relationship between multiple choice and essay response questions in assessing economics understanding. *Economic Record*, 75, 348–357.

Beede D., Julian T., Langdon D., McKittrick G., Khan B., & Doms M. (2011). Women in STEM: A Gender Gap to Innovation, Economics and Statistics Administration Issue Brief 04–11, Washington, DC: U.S. Department of Commerce.

Birenbaum M. & Tatsuoka, K. K. (1987). Open-Ended Versus Multi ple-Choice Response Formats—It Does Make a Difference for Diagnos tic Purposes. *Applied Psychological Measurement*, 11(4), 385 - 395.

Boyer, E. (1990). *Scholarship Reconsidered: Priorities of the Professoriate*. Carn egie Foundation for the Advancement of Teaching.

"Conceptual Assessments in Biology | SDSU." Conceptual Assessments in Biolo gy | SDSU. N.p., n.d. Web. 07 Mar. 2017. <http://go.sdsu.edu/dus/ctl/cabs.aspx>.

Emil, S., & Cress, C. (2014). Faculty perspectives on programme curricular as sessment: individual and institutional characteristics that influence participation engagement. *Assessment & Evaluation In Higher Education*, 39(5), 531-552.

Gardner, D. P., et. al. (1983). *A Nation At Risk: The Imperative For Educational Reform*. Report to the Secretary of Education. Available at http://www.ed.gov/pubs/NatAtRisk/letter.html

Hambleton, R. K., Merenda P. F., & Spielberger., C. D. (2004). *Adapting educa tional and psychological tests for cross-cultural assessment*. Psychology Press.

Hawthorne, E. M. (1989). E*valuating employee training programs: A re search-based guide for human resource managers*. Quorum

DOI: 10.20429/stem.2017.010105

Books. https://ssl.gstatic.com/s2/profiles/images/silhouette96.png

Hawthorne, K. A., Bol, L., Pribesh, S., & Suh, Y. (2015). Effects of Motivational Prompts on Motivation, Effort, and Performance on a Low-Stakes Standardized Test. *Research and Practice in Assessment*, 10, 30-38.

Heyborne, W. H., Clarke, J. A., & Perrett, J. J. (2011). A Comparison of Two Forms of Assessment in an Introductory Biology Laboratory Course. *Journal Of College Science Teaching*, 40(5), 28-31.

Hill, C., Corbett, C., & St Rose, A. (2010). *Why so few? Women in Science, Technology, Engineering, and Mathematics. American Association of University Women*. 1111 Sixteenth Street NW, Washington, DC 20036.

Hogan, T.P. (1981). *Relationship between free response and choice-type tests of achievement*. Washington, DC: U.S. Department of Education, National Assessment of Educational Progress.

Imrie, B. W., et al. (2014). *Student assessment in higher education*: a handbook for assessing performance. Routledge.

Kim, Y. K., Hutchison, L. A., & Winsler, A. (2015). Bilingual education in the United States: an historical overview and examination of two-way immersion. *Educational Review*, 67(2), 236-252.

Miller, J.D., Scott, E.C. & Okamoto, S. (2006). Public acceptance of evolution. Science, 313, 765–766.

Nichols, P., & Sugrue, B. (1999). The lack of fidelity between cognitively complex constructs and conventional test development practice. *Educational Measurement: Issues and Practice*, 18, 18–29

Resnick, L. B., & Zurawsky, C. (2007). Science education that makes sense. American Educational Research Association Research Points, 5, 1.

Ruben, B. D. (2016). *Excellence in Higher Education Guide : A Framework for the Design, Assessment, and Continuing Improvement of Institutions, Departments, and Programs*. Sterling, Virginia: Stylus Publishing.

Runck, C. (2015). *Biology 2014 Annual Program Review*. Georgia Gwinnett College, Lawrenceville, GA.

Sims, S. J. (1992). *Student outcomes assessment*: A historical review and guide to program development. Westport, CT: Greenwood Publishing Group.

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. Zanna

(Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 379 – 440). New York, NY: Academic Press.

Stevens, D. D., & Levi, A. J. (2013). I*ntroduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning*. Stylus Publishing, LLC.

Stohlman, T. (2015). The Road to Redemption:Reclaiming the Value in Assess ment Retention Exams. *Journal of the Scholarship of Teaching and Learning*, 15(5), 64-71.

Tucker, D. E. (2006). Direct measures: Examinations. In W. G. Christ, *Assessing media education: A resource handbook for educators and administra tors* (pp. 373-396). Mahwah, NJ: LEA.

Varma, S. (2008). *Preliminary item statistics using point-biserial correlation and p-values*. Available at:https://jcesom.marshall.edu/media/24104/Item-Stats-Point-Biserial.pdf (Accessed 14 January 2017).

Yorke, M., & Zaitseva, E. (2013). Do cross-sectional student assessment data make a reasonable proxy for longitudinal data? *Assessment & Evalua tion In Higher Education*, 38(8), 957-967.

DOI: 10.20429/stem.2017.010105