SOCIAL SCIENCE

Evaluating the extent of a large-scale transformation in gateway science courses

Rebecca L. Matz[1]*, Cori L. Fata-Hartley[2], Lynmarie A. Posey[3], James T. Laverty[4], Sonia M. Underwood[5], Justin H. Carmel[5], Deborah G. Herrington[6], Ryan L. Stowe[3], Marcos D. Caballero[7], Diane Ebert-May[8], Melanie M. Cooper[3]

We evaluate the impact of an institutional effort to transform undergraduate science courses using an approach based on course assessments. The approach is guided by *A Framework for K-12 Science Education* and focuses on scientific and engineering practices, crosscutting concepts, and core ideas, together called three-dimensional learning. To evaluate the extent of change, we applied the Three-dimensional Learning Assessment Protocol to 4 years of chemistry, physics, and biology course exams. Changes in exams differed by discipline and even by course, apparently depending on an interplay between departmental culture, course organization, and perceived course ownership, demonstrating the complex nature of transformation in higher education. We conclude that while transformation must be supported at all organizational levels, ultimately, change is controlled by factors at the course and departmental levels.

INTRODUCTION

The need to transform undergraduate science, technology, engineering, and mathematics (STEM) education is not new (*1*, *2*), but the sense of urgency has accelerated because of a number of recent, influential reports (*3*–*5*). For example, the 2012 President's Council of Advisors on Science and Technology report (*5*) highlighted the need for increased numbers of STEM graduates and focused on improving recruitment and retention strategies for STEM fields. The report specifically targeted the first 2 years of postsecondary education because of the gatekeeper role that they tend to play for a wide range of students. Goals for transforming STEM education include supporting a web of diverse pathways to STEM degrees and careers and incorporating evidence-based teaching practices in courses, such as increasing student engagement through active learning. Although incorporating active learning into courses improves student outcomes in most instances and disciplines, and particularly for underrepresented students (*6*), didactic instruction still dominates most STEM disciplines (*7*).

In 2013, Michigan State University (MSU) was awarded a grant from the Association of American Universities as part of their Undergraduate STEM Education Initiative (STEM Initiative) (*8*). The goal of the STEM Initiative is to change the culture in higher education so that successful teaching and learning in STEM are recognized and rewarded as a vital and important aspect of faculty work at research-intensive universities. The MSU project focused on gateway courses in chemistry, physics, and biology, all of which play a substantial role in the curricula for STEM degree programs. While many course transformation efforts focus on faculty incorporating active learning methods, the MSU project is situated in an approach described by *A Framework for K-12 Science Education* (*9*). The *Framework* was developed for K-12 education, but we argue that the approach is transferable to the undergraduate level as well (*10*).

Three-dimensional learning

The *Framework* provides a vision for science education that emphasizes three dimensions: scientific and engineering practices (herein, scientific practices) in the context of crosscutting concepts and core disciplinary ideas. Scientific practices are the ways in which scientists and engineers use their knowledge and can be considered the disaggregated components of inquiry (*11*). Crosscutting concepts are ideas that transcend disciplines and provide ways to connect ideas and phenomena across disciplines. Core ideas are concepts that are central to a discipline, can be used to explain many phenomena, and can help students (and scientists) understand new phenomena. This integrated vision for teaching and learning in science is known as the three-dimensional learning (3DL) approach and provides the common theme and goal for our change efforts. This approach differs from the active learning paradigm in that it requires a reframing of curriculum that leads to changes in both the instructional activities and concomitant assessment tasks used to evaluate student knowledge and abilities. Further, active learning does not inherently engage students in scientific practices, but we contend that scientific practices are inherently active.

Over 3 years, many faculty teaching gateway chemistry, physics, and biology courses at MSU took part in one or more of three key activities, all centered on 3DL (table S1): (i) discussion groups within the disciplines aimed at implementing the 3DL approach, (ii) a 2-year interdisciplinary professional development program about 3DL, and (iii) periodic campus workshops and seminars.

i) To promote participation, ownership, and development of a shared vision (*12*–*14*), the disciplinary discussions initially focused on building faculty consensus around a set of core ideas for each discipline (*15*–*17*). We began with disciplinary core ideas to tap into the inherent interest that faculty have in their subject matter and followed this with discussions about what students should be able to do with their knowledge—that is, doing science requires that students use scientific practices.

[1]Hub for Innovation in Learning and Technology, Michigan State University, East Lansing, MI 48824, USA. [2]College of Natural Science, Michigan State University, East Lansing, MI 48824, USA. [3]Department of Chemistry, Michigan State University, East Lansing, MI 48824, USA. [4]Department of Physics, Kansas State University, Manhattan, KS 66506, USA. [5]Department of Chemistry and Biochemistry; Florida International University, Miami, FL 33199, USA. [6]Department of Chemistry, Grand Valley State University, Allendale, MI 49401, USA. [7]Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, USA. [8]Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA.
*Corresponding author. Email: matz@msu.edu

ii) The professional development program, called the STEM Gateway Fellowship, was instituted to provide a forum and support for faculty working to transform teaching in large-enrollment, introductory STEM courses. Fellows develop 3DL assessments and related instructional materials throughout the program, and each cohort is purposefully interdisciplinary to encourage communication and collaboration across disciplines.

iii) Members of the research team led professional development workshops and seminars focused on the 3DL approach through various MSU programs such as the STEM Teaching Essentials workshops and triannual STEM Alliance seminars.

Among these activities, we could practically control for balanced representation of faculty across the disciplines only in the Fellowship program, although numerous faculty from chemistry, physics, and biology attended both the disciplinary discussions and workshops and seminars (table S1). Together, in broadly engaging faculty with the 3DL approach through these activities, we hypothesized that changes in both instructional practices and the nature of assessments might occur in all three disciplines over time (*10*). Here, we report our evaluation of this transformation effort and discuss why these results differed across departmental and course-level domains.

Measuring impact

There are numerous approaches to characterizing the extent of transformation. Some approaches focus on measuring the degree to which active learning and transformed pedagogies are implemented (*18*), and many observation protocols have been developed and used to identify the types of interactions between faculty and students in the classroom (*19*–*22*). These protocols have been productive for investigating changes in how instructors teach. However, these approaches generally provide little evidence about what students are expected to know and do in a course. Faculty surveys about teaching provide self-reported data that identify pedagogical and cultural change. While it is true that awareness of pedagogical strategies is a prerequisite to engaging in these approaches to teaching, it does not guarantee that they are implemented with fidelity or that student outcomes will improve (*23*–*27*).

Other approaches to measuring change have used course exam scores and grades as evidence of effective instructional reform, but again, it is rare that details are provided about what students are asked to know and do (*6*). Pre- and post-assessments of student understanding are often reported on the basis of selected response tests, such as concept inventories that are designed to identify student difficulties with particular ideas, but these tests do not reveal student abilities to connect and use these knowledge fragments, and in general, multiple-choice tests may overestimate student understanding (*28*, *29*). Multiple methods should be used to provide compelling evidence of change, particularly with respect to what students are expected to know and be able to do upon completion of a course and when they graduate.

Here, therefore, we use an approach that characterizes change in course assessments, as they send strong signals to students about what instructors value (*30*–*34*), in addition to more traditional measures of transformation, such as the types and frequencies of classroom activities and average course grades. Our hypothesis is, if the nature of a course has changed to incorporate 3DL, then the concomitant assessments (specifically, course exams) also should have changed. If the curriculum changes while all usual course assessments are retained, it is unlikely that students will engage in 3DL, negating the intent of any transformation effort. Toward this end, we applied the Three-dimensional Learning Assessment Protocol (3D-LAP) (*16*) to course exams over a 4-year transformation period. The 3D-LAP facilitates characterizing whether assessment items have the potential to elicit evidence of 3DL, that is, items that require students to use a core idea in combination with a scientific practice and a crosscutting concept.

We applied the 3D-LAP to representative exams from chemistry, physics, and biology courses and determined the proportion of exam points that reflected the 3DL approach. Here, we address the following primary research questions: (RQ1) How did the use of 3DL assessment items change over time in the different disciplines, and how did corresponding student outcome measures change over time? Given our contention that students are inherently doing active learning when they are engaging with scientific practices, our secondary research question is as follows: (RQ2) What is the relationship between the amount of time instructors spend lecturing and the use of 3DL assessment items?

METHODS

This study was performed at MSU, a large, public university with very high research activity. The work was determined exempt from Institutional Review Board review and in part predicated on an agreement that we would not publish longitudinal information about individual instructors, even if deidentified. Data were collected over 4 years from eight large-enrollment science courses that constitute the core introductory sequences for a substantial number of MSU STEM degree programs. The courses of interest were General Chemistry I and II (Chem I and II), Algebra-based Physics I and II (Phy-A I and II), Calculus-based Physics I and II (Phy-C I and II), and Introductory Biology I and II (Bio I and II). While faculty who teach these courses are generally housed in the College of Natural Science, expectations about teaching and assessments are communicated primarily by department and discipline. We collected three types of data from these courses: (i) exams, (ii) video recordings of class meetings, and (iii) course grades and D-grade, F-grade, and withdrawal (DFW) rate information about the students enrolled in each course. Exams and video recordings were collected with instructor consent. We report the data by discipline for chemistry and physics and by course for biology because the two biology courses have different organizational structures and share fewer overlapping instructors than do the chemistry and physics courses.

Exam data

Representative exams from each course section were collected from instructors beginning 1 year before implementation of the transformation project (year 0) and continued for three additional years (years 1 to 3). We note that in some cases, particularly for the chemistry and physics courses, multiple sections of a single course in a given semester used common exams; these exams were counted only once. Further, when instructors used the same exams verbatim from year to year, which was rare, we coded each unique exam only once. In total, we collected and analyzed 4020 questions from 134 unique exams, fully representing all 185 course sections of the eight courses that were offered during the 4-year span (Table 1). Identifying information about the instructor(s), course, section, and term offered was removed from each exam so that the exams were identifiable only by the researcher who organized the data. Each exam and every question on each exam were tagged with a unique identification number, and multipart questions were identified as clusters (e.g., if a question had a part

Table 1. Summary of exam data collected in years 0 to 3 from introductory chemistry, physics, and biology courses. A detailed summary by course is provided in table S2.

Discipline	Course sections offered	Unique instructors	Unique exams coded	Questions coded*		
				Individual questions	Clusters	Total
Chem I and II	49	21	32	718	93	811
Phy-A and Phy-C I and II	78	35	34	479	230	709
Bio I	34	20	40	1705	184	1899
Bio II	24	9	28	528	83	611
Total	185	87	134	3430	590	4020

*When the same question was used on different exams (by the same or a different instructor), it was coded every time.

"a" and "b," then these parts were coded together as a single cluster) (*16*). We also recorded the number of points associated with each question.

Interrater reliability on practice items was achieved using a subset of the exams, as discussed in Laverty *et al.* (*16*), and then all items within each disciplinary area were coded with the 3D-LAP by team members with expertise in that discipline [we refer readers to the S1 Exemplars Supporting Information in Laverty *et al.* for a substantial set of example 3DL assessment items, and to Underwood *et al.* (*35*) and Laverty and Caballero (*36*) for comparisons between traditional and 3DL assessment items]. In alignment with the intent of the *Framework* (*9*) that the three dimensions be integrated, we report the results in a binary way, that is, an assessment item either met the criteria for all three dimensions or it did not. After all items were coded, we determined the fraction of points on each exam that was associated with questions coded as 3D. In this way, we are able to compare different sections of the same course within a discipline and compare courses across different disciplines.

Table 2. Summary of video data collected in years 1 and 3 from introductory chemistry, physics, and biology courses. A detailed summary by course is provided in table S3.

Discipline	Unique course sections offered*	Unique course sections recorded	Videos recorded/ videos attempted (%)
Chem I and II	20	18	54/54 (100%)
Phy-A and Phy-C I and II	31	27	74/81 (91%)
Bio I	18	11	30/33 (91%)
Bio II	12	9	26/27 (96%)
Total	81	65	184/195 (94%)

*When the same instructor(s) taught multiple sections of the same course during the same term, we assumed that the sections were taught similarly and recorded only one unique section.

Video data

Video recordings of class meetings were also collected from the same eight courses during years 1 and 3. The recordings focused on the instructors and any materials they were presenting such as board work and slide presentations, not on the students. We recorded 65 (80%) of the 81 unique course sections offered (Table 2); that is, when an instructor taught multiple sections of the same course in the same semester, we attempted to record only one of their sections. Three recordings of class meetings were collected from each unique course section approximately 1, 2, and 3 months into the semester. We note that, for 9 of the 65 course sections, we were able to collect only two recordings and, for 1 section, we could collect only one recording. These missing recordings occurred for various unsystematic reasons including equipment failure and last-minute schedule changes, and they are spread roughly evenly across the biology and physics courses, whereas the chemistry video data are complete.

The recordings were coded using an observation protocol developed for this project (*37*) that attends to the teaching activities in each class meeting (descriptions of the teaching activities and how the videos were processed are provided in the supplementary materials). Two-thirds of the year 1 recordings were coded initially, and the remaining third was mixed and coded with the year 3 recordings (in this second set, coders did not know the year of the recording). Although students and instructors engaged in many different activities during any particular class meeting, here, we report only the percent of time spent lecturing, defined as instructor-directed presentation of content-related information. Most of the other teaching activities refer to an active learning activity, such as the use of audience response systems (e.g., clicker questions) and in-class tasks (e.g., think-pair-share questions).

Institutional data

Course grade and summative DFW rate data were collected from the appropriate student information systems for all students who were enrolled in the eight courses of interest anytime during years 0 and 3. The DFW rate indicates the proportion of students who either withdrew from the course or earned a grade of 1.5 or lower on a 4.0 scale.

RESULTS AND DISCUSSION

RQ1: Change in use of 3DL assessment items and student outcomes over time

The results from coding the exam data with the 3D-LAP (Fig. 1) show that the proportion of points allocated to 3DL questions increased

over the 4 years of the project within each discipline; however, this relationship is statistically significant only for Chem I and II ($r_s = 0.72$, $P < 0.001$) and Bio I ($r_s = 0.37$, $P < 0.05$). There are stark differences by discipline and even by courses within a discipline. These differences suggest that although the 3DL approach to change was common across the disciplines, the culture and environment in which change occurs are important and warrant further investigation.

In addition, we find statistically significant correlations between an increase in the proportion of 3DL points and both an increase in mean final course grade and a decrease in DFW rate (Figs. 2 and 3 and Table 3) for all courses except Bio I; that is, if the proportion of 3DL points increases (as measured by the 3D-LAP), then we generally observe concomitant positive changes in average grade and DFW rate. While improvements to student grades and DFW rates are not necessarily an expected outcome of instructors increasing the use of 3DL-focused assessments, we provide these data first and foremost as evidence of having done no harm to the students. Further, it is reasonable to hypothesize that increasing the use of 3DL-focused assessments in a course might correlate with the class meetings being more active (as seen in chemistry and biology in Fig. 4); active learning in STEM courses has been shown to improve student performance on concept inventories and exams and to decrease failure rates (*6*).

Chemistry

Figure 1A shows a marked increase in 3DL items in the two general chemistry courses over the 4-year period. The structure and organization of the curriculum for these large-enrollment general chemistry courses differ from those in biology or physics. The chemistry courses are highly coordinated and organized by the director of General Chemistry; the learning outcomes and course exams are common, and all instructors teach approximately the same material at the same pace over the same time period. A transformed general chemistry course sequence [a National Science Foundation–supported project called CLUE (Chemistry, Life, the Universe and Everything) (*38*)] was piloted in years 1 and 2, and by year 3, the entire general chemistry course structure was transformed. In year 1, only a small fraction of the total students (11 and 24% in Chem I and Chem II, respectively) were tested using 3DL items, whereas by year 3, all students (3450 in Chem I and 1250 in Chem II) were tested using nearly 50% 3DL items in both courses.

Simple exposure to 3DL assessment items is no guarantee of more effective teaching and learning; however, the transformation efforts were also studied as part of the CLUE project. The evidence from CLUE indicates that students who were enrolled in the transformed courses had significantly improved understanding of core ideas such as structure-property relationships (*39*, *40*) and that this improvement

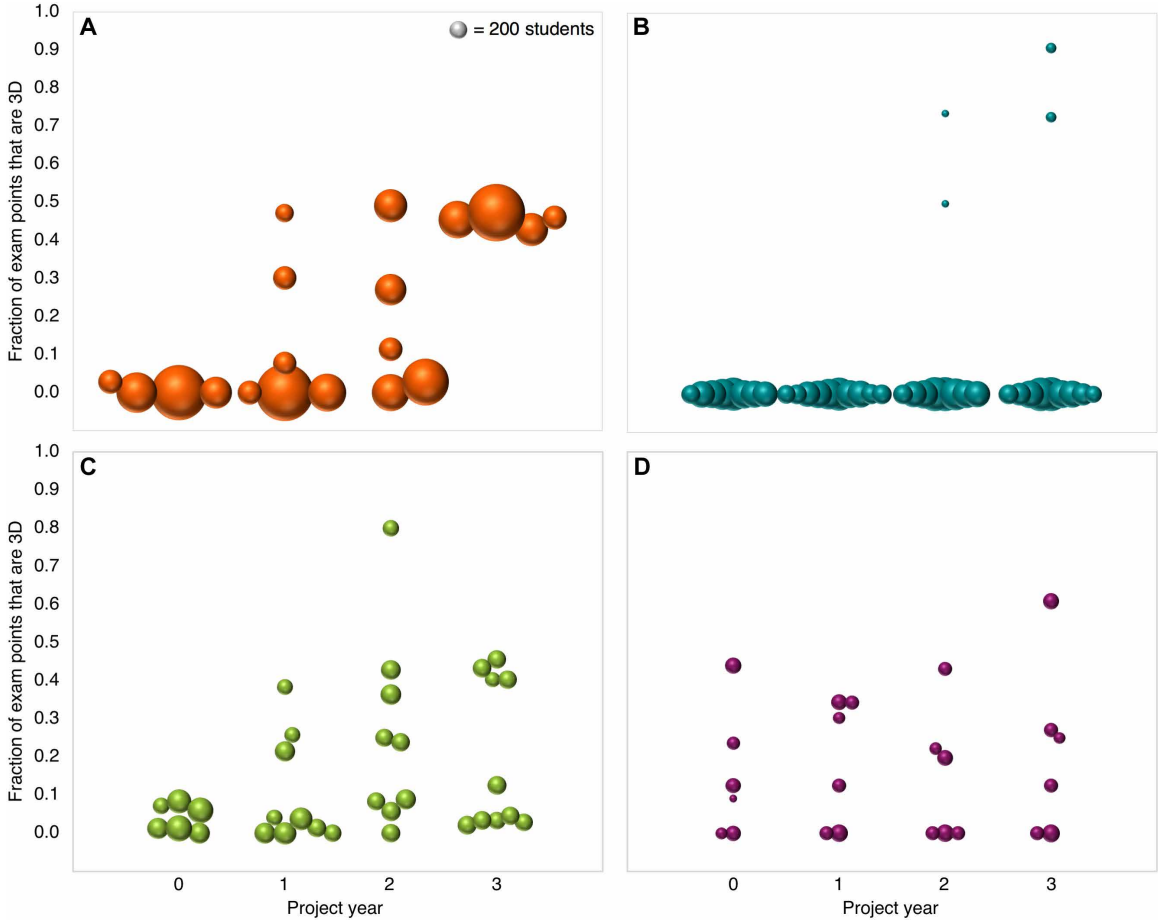

Fig. 1. 3D assessment items over time. Fraction of exam points that reflect the three dimensions over time in (**A**) Chem I and II, (**B**) Phy-A and Phy-C I and II, (**C**) Bio I, and (**D**) Bio II. Each data point (bubble) represents either a final exam or, when the final exam was unavailable, two or more midterm exams. Each data point is scaled by the number of students who took each exam; the largest points represent common exams. The scale is consistent across all four panels.

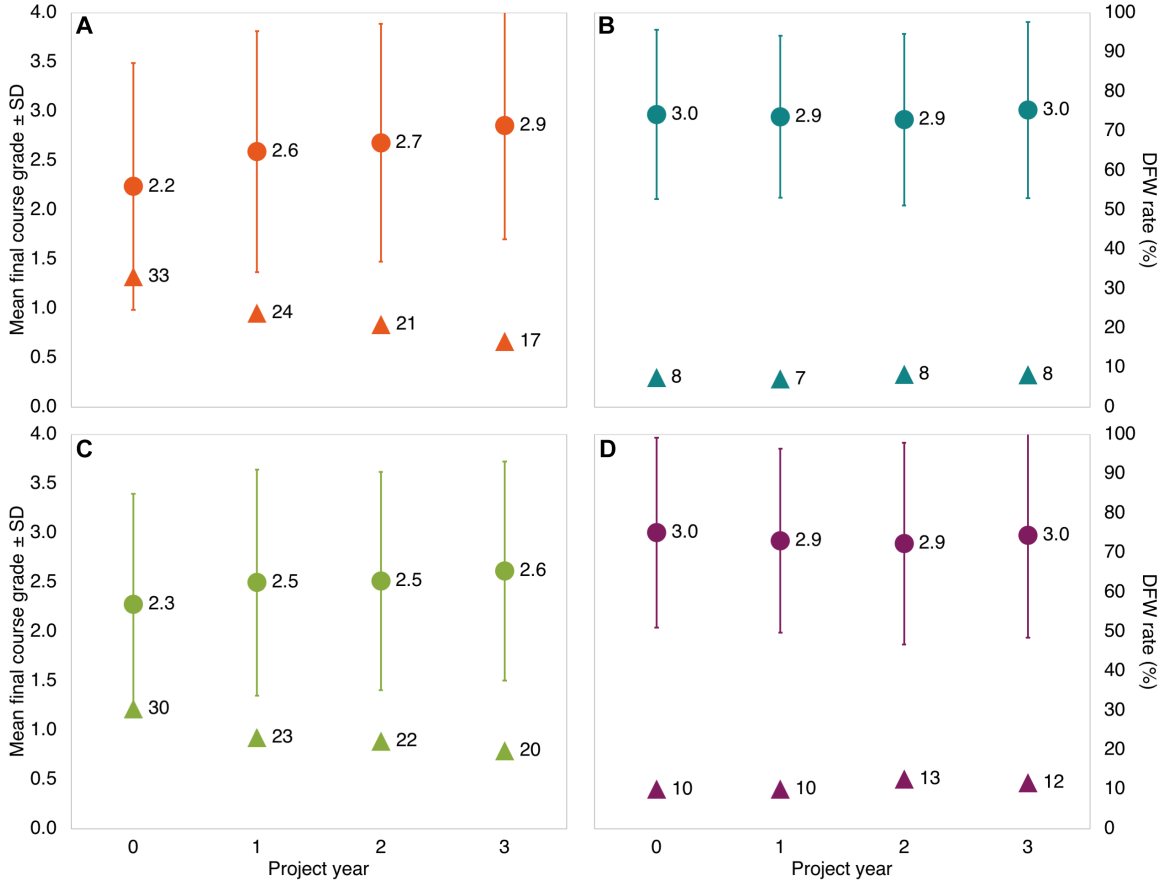

Fig. 2. Final course grade and DFW rate over time. Mean final course grade ± SD on a 4.0 scale (circles) and DFW rate (triangles) over time in (**A**) Chem I and II (average annual enrollment = 4650 students), (**B**) Phy-A and Phy-C I and II (4390), (**C**) Bio I (2160), and (**D**) Bio II (930).

persisted at least through the next year of study in organic chemistry (*41*, *42*). After two semesters, students in the transformed sections performed above the national average on a nationally normed American Chemical Society exam, although this exam was not developed with the intention to measure 3DL. In addition, the average grade across all students increased, and the overall DFW rate decreased by approximately 16% (Fig. 2A). This decrease in DFW rate practically translates to approximately 740 more students earning a grade of 2.0 or above in year 3 compared with year 0.

Physics

The situation in physics remains largely unchanged in terms of both 3DL assessment items (Fig. 1B) and student outcomes (Fig. 2B). Similar to chemistry, these courses are often coordinated across sections, using the same syllabus and textbook. However, unlike the chemistry and biology courses, the physics courses typically draw assessment items from a large, established pool of randomized formative and summative assessment items generated by an online course management system with a long history in the department (*43*). Although many physics faculty contributed regularly to discussions of physics core ideas and practices, there was no observed change in the assessment items used by the large introductory physics courses over the course of the project.

One interpretation of these findings is that the reliance on test bank–generated items undermined any transformation efforts that faculty may have intended with regard to their assessments; another is that the assessment system has too much institutional inertia behind it to support change. In years 2 and 3 of the project, however, four small course sections of Phy-C I appear that incorporate assessments with high percentages of 3DL assessment items. These data represent completely transformed sections that use a different curriculum and approach than the rest of the physics sections, and evidence (*44*) shows that students in these transformed sections improve their understanding of physics concepts compared with traditional sections as measured by the Force and Motion Conceptual Evaluation (*45*).

Biology

Figure 1 (C and D) shows the change in the proportion of exam points allocated to 3DL assessment items over 4 years for the two courses in the gateway biology sequence. Trends in the data for the two courses are notably different, which can be attributed, in part, to the two courses being in different stages of transformation.

Bio I: At the start of the project, there was little coordinated transformation effort in Bio I, although some individual faculty members designed and used learner-centered instructional activities. The different sections all used the same textbook but did not have a common set of learning goals or use common exams. Rather, instructors developed their own teaching materials based on an agreed-upon set of topics. Over the course of this project, 20 different instructors taught or cotaught in the 34 offered course sections of Bio I, 10 of whom were engaged in coordination efforts centered on the 3DL approach. For example, a course committee focused on Bio I met during year 1

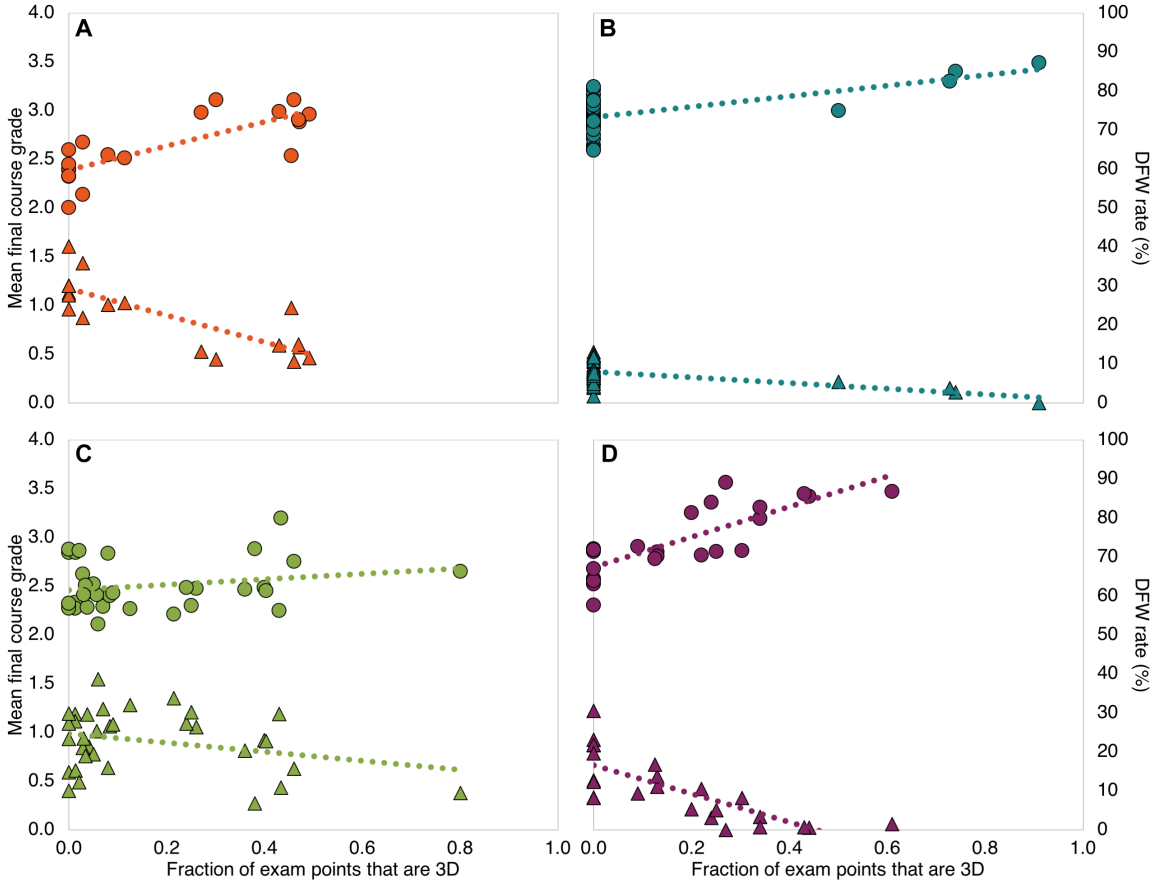

Fig. 3. Final course grade and DFW rate versus fraction of 3D exam points. Mean final course grade on a 4.0 scale (circles) and DFW rate (triangles) versus fraction of exam points that reflect the three dimensions in (**A**) Chem I and II, (**B**) Phy-A and Phy-C I and II, (**C**) Bio I, and (**D**) Bio II.

Table 3. Correlations (r_s) between the fraction of exam points that are 3D, and mean final course grade and DFW rate (%).

Correlate with 3D exam points	Chem I and II	Phy-A and Phy-C I and II	Bio I	Bio II
Mean final course grade	0.74*	0.44*	0.10	0.71*
DFW rate	−0.78*	−0.47*	−0.07	−0.83*

*$P < 0.01$ (two tailed).

as part of the Biology Initiative, an internally funded effort to improve undergraduate biology education through the investment of new resources in the biology departments and the Biological Sciences Program, which administers the introductory biology courses. A major outcome for this committee was identifying the core ideas and scientific practices most relevant to Bio I. In addition, three members of the Bio I teaching team were involved in the STEM Gateway Fellowship program, and several members attended professional development workshops offered on campus about 3DL assessment and instruction. During this time frame, we observe an increase in the fraction of points allocated to 3DL items, although there is still a spread across sections (Fig. 1C), a concomitant increase in average course grade, and a decrease in DFW rate (Fig. 2C).

Bio II: In contrast, Bio II had already undergone a series of transformations associated with an earlier project (*46*–*48*) that was not initially associated with 3DL per se but was based on implementing teaching and learning strategies that promote student use of scientific practices including data analysis, argumentation, and modeling, all while emphasizing collaboration. Figure 1D demonstrates that many exam points in Bio II were already allocated to 3DL assessment items at the beginning of this project, and a similar level was observed at the end. However, it is also meaningful that there was little change in the proportion of 3DL questions in the nontransformed sections over the same time frame, suggesting that obstacles to adoption remain for some faculty. The average grade and DFW rate did not change (Fig. 2D), consistent with the observation

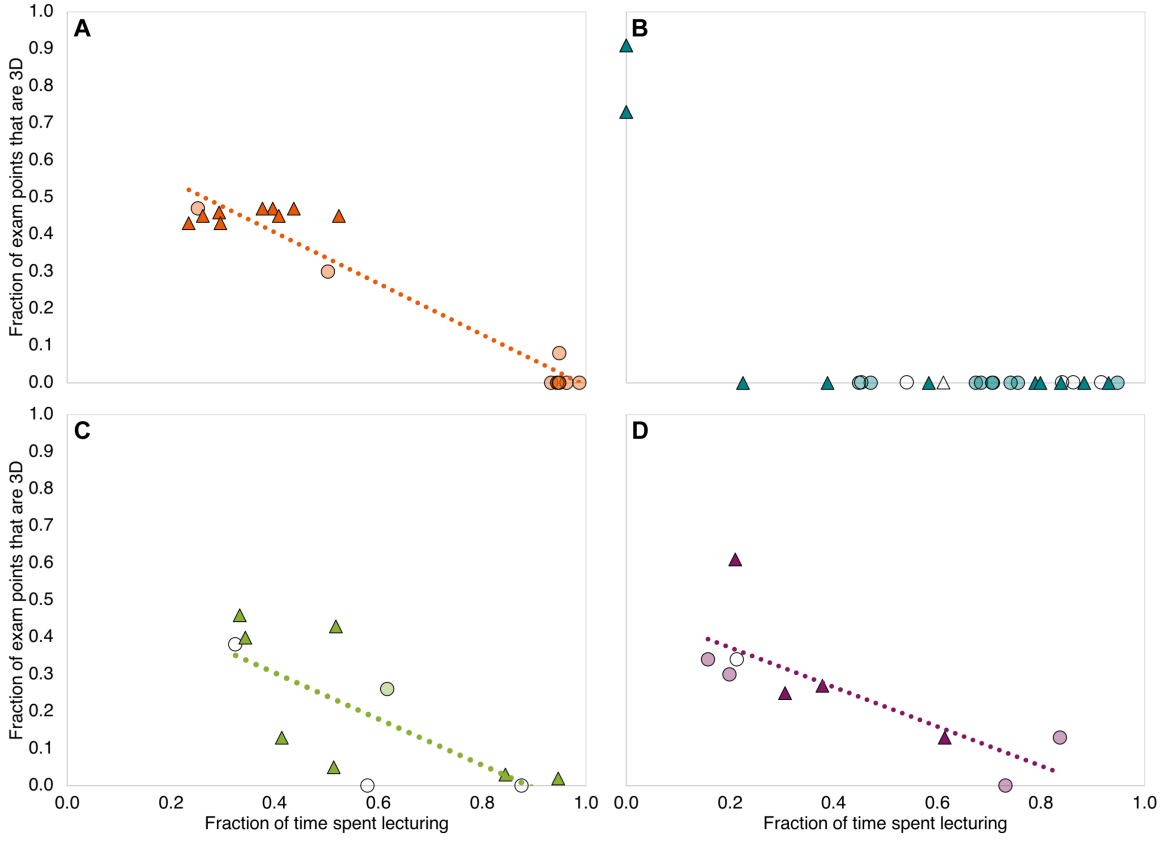

Fig. 4. 3D assessment items by time spent lecturing. Fraction of exam points that reflect the three dimensions as a function of time spent lecturing averaged over three videos in year 1 (light circles) and year 3 (dark triangles) course sections of (**A**) Chem I and II, (**B**) Phy-A and Phy-C I and II, (**C**) Bio I, and (**D**) Bio II. Open symbols denote course sections where fewer than three video recordings were collected in year 1 (open circles) and year 3 (open triangles). No trendline is provided for (B) because of the bimodal nature of the data.

that there was little change in the overall nature of the assessments in Bio II.

RQ2: Relationship between time lecturing and use of 3DL assessment items

Video recordings of class meetings from the 65 unique course sections (out of 81 possible sections) were collected during the transformation effort across the gateway chemistry, physics, and biology courses. In Fig. 4, the fraction of 3D exam points is shown against the time spent lecturing. For the biology and chemistry courses, there appears to be an inverse correlation between the time spent lecturing and the use of 3DL items used in the exams for that section. This relationship is similar although the course structures for chemistry and biology are quite different, supporting the idea that incorporating 3DL assessment items also promotes active engagement in the classroom. However, the situation in physics looks different in that many of the physics faculty used active learning techniques such as clicker questions in their class meetings, although no 3DL items were used on most exams—active learning instructional strategies do not inherently engage students in 3DL. The exceptions correspond to two low-enrollment sections of the transformed Phy-C I course, which has no lecturing during class meetings and a high fraction of 3DL assessment items. Further, in contrast with the idea that class size affects instructor teaching practices (*49*), here, we find no relationship between class size and time spent lecturing (Fig. 5).

Three disciplines, four outcomes

The results presented in Fig. 1 show that although the transformation efforts in each discipline were guided by the 3DL approach, and the discussions of core ideas and scientific practices were attended by many faculty from each discipline, the measurable changes appear to be highly dependent on other factors including the course organizational structure, perceived ownership of the course, departmental culture, available resources, faculty expertise, and the power dynamics between faculty and those calling for change.

For example, the courses of interest here are quite different in the ways that they are organized and "owned" by faculty. The introductory courses in physics and chemistry are each housed within a disciplinary department, yet the results from these courses lie at the limits of measurable impact of the project. Both the chemistry and physics courses are coordinated, with common syllabi and common exams, but the course ownership is different. In chemistry (as in many large chemistry departments), the course is organized and coordinated by a full-time director of General Chemistry, relatively few tenure-track faculty rotate through the course, and most of the course sections are taught by non-tenure-track (although long-term) instructors. The decisions about changes to the course structure, textbook, and homework system are relatively centralized with the director, department chair, and the faculty who regularly teach the course. When the decision was made to move from the original curriculum to the new 3DL curriculum, the text, teaching materials, and online assessment system

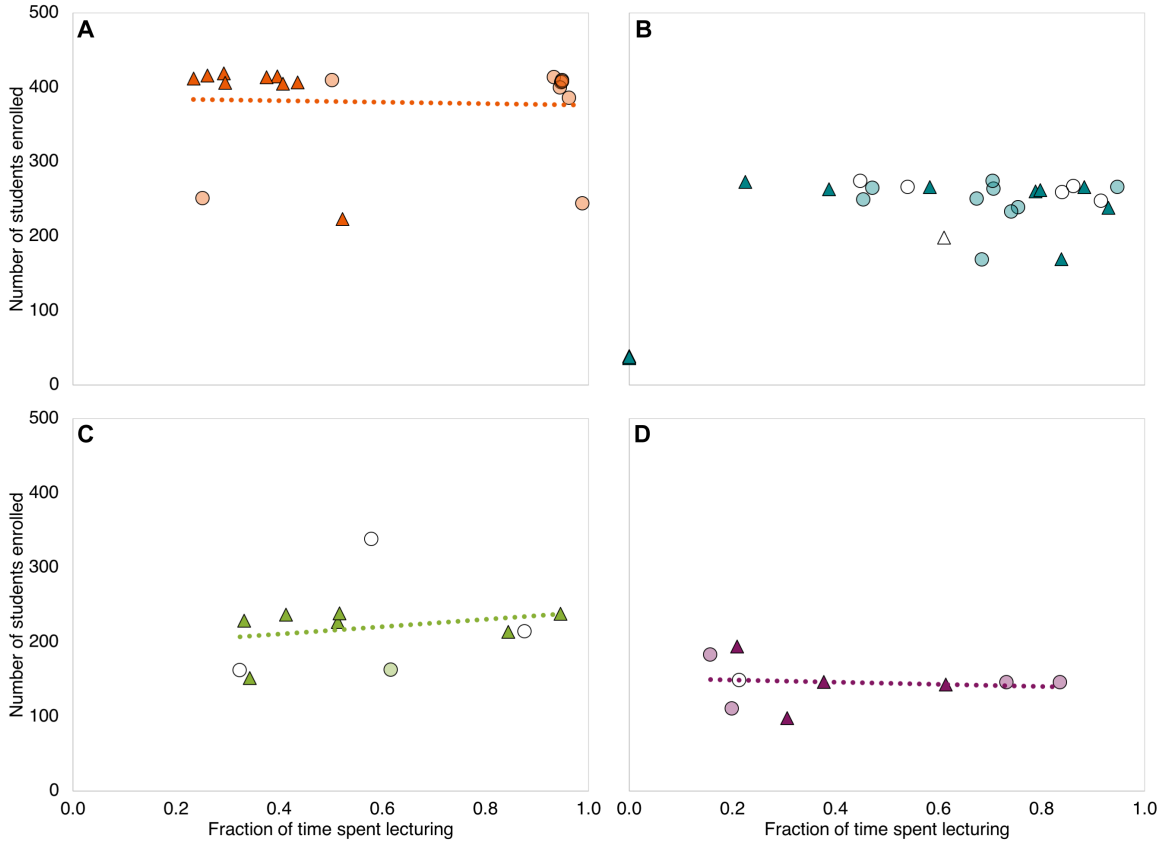

Fig. 5. Comparison of student enrollments and time spent lecturing. Number of students enrolled as a function of time spent lecturing averaged over three videos in year 1 (light circles) and year 3 (dark triangles) course sections of (**A**) Chem I and II, (**B**) Phy-A and Phy-C I and II, (**C**) Bio I, and (**D**) Bio II. Open symbols denote course sections where fewer than three video recordings were collected in year 1 (open circles) and year 3 (open triangles). No trendline is provided for (B) because of the bimodal nature of the data.

were all changed to the integrated materials developed as part of the CLUE system. Adoption of the CLUE curriculum was also accompanied by change in the format of assessment items from entirely multiple choice to a mix of multiple choice and constructed response. Increased administrative support [in the form of extra graduate teaching assistant (TA) and undergraduate learning assistant (LA) time] was used in the transformed sections to support grading constructed-response exam questions and in-class activities.

In contrast, most faculty in the physics department, both tenure-track and non-tenure-track, teach in the gateway courses and contribute to changes in the curriculum, consistent with the general expectation in physics that faculty be able to teach any undergraduate course. There is no director who facilitates this process, and during the period of this study, there was no wholesale change in the commercially available text and peripherals or in the online assessment system. The assessment system was locally developed, has been used for 25 years, and incorporates personalized multiple-choice questions and calculations generated by many former and current faculty (*50*). However, this system does not immediately lend itself to generation of 3DL tasks and may have been an impediment to change overall, despite the support and interest of the department and departmental administration. Extra support for LAs was directed toward the laboratories; thus, limited support for grading constructed-response exams may have been another impediment to change.

The biology courses differ from both chemistry and physics in organization and administration. The Biological Sciences Program administers the introductory biology courses, which are taught by a rotating committee of faculty from multiple different biology-related departments and colleges. The project described herein overlapped with the Biology Initiative, an internal effort to improve undergraduate biology education (*51*). Early Biology Initiative investments were focused on Bio I and led to an increased number of faculty engaged in teaching Bio I, in turn decreasing the number of students per section from approximately 425 to less than 250. TAs and LAs were added to the teaching teams, enabling the use of both formative and summative constructed-response assessments. A course curriculum coordinator is now responsible for working with faculty to maintain a shared vision for the course based on a common set of learning outcomes that blend core ideas, scientific practices, and crosscutting concepts. In addition, multiple Bio I faculty participated in the STEM Gateway Fellowship, including the first course curriculum coordinator. These new resources and coordination efforts appear to have facilitated increased use of 3DL assessment items in Bio I.

On the other hand, we observed little change over time in the Bio II course sections. The sections that were already transformed maintained their status, but we saw no increase in the use of 3DL assessment items in other sections. The previous transformation efforts focused on infusing scientific practices into the curriculum were

embraced by a group of faculty located primarily in one department; faculty from other units generally did not participate. While the Bio II faculty participated in discussions centered on disciplinary core ideas during this project, the Biology Initiative funds were allocated over a span of 5 years, and investment in Bio II did not take place until this study was completed. The absence of a course curriculum coordinator and the lack of additional TAs during the project time frame may have hindered the type of transformation efforts that led to increased use of 3DL assessments in Bio I.

In sum, while this work shows that change is happening in most courses, it is occurring at different rates. In chemistry, the transformation appears to be relatively rapid, but that does not consider the 10 years of prior development and pilot work with the transformed CLUE curriculum. Once this curriculum was available, the structure and organization of general chemistry at MSU allowed for a rapid transition.

Although there seems to be little change in physics, it may be because they are at the very beginning of development and pilot work. The small physics sections with high proportions of 3DL assessments that appear in years 2 and 3 correspond to pilot sections of a transformed Phy-C I course that does not rely on existing texts and homework systems. The analyses carried out in this project have informed plans to expand these completely transformed sections, rather than to "shoehorn" transformation into an existing curriculum, and the 3DL approach is being expanded to the Phy-C II course as well. In the current academic year, 140 more seats in the transformed courses are available to students compared with those available in year 3, and ongoing research is focused on student learning of and participation in scientific practices. These pilot courses and variants, which integrate laboratory work, have been accepted by the faculty and are being expanded across all course offerings within the next 5 years as the department aligns its entire curriculum with a 3DL approach. It may well be that 10 years is a reasonable time frame to bring about complete transformation (development and scaling up) (*52*), although it should be noted that typical funding periods for these efforts, both external and internal, are much shorter.

It may also be that Bio I is in transition at a slower pace because the transformation is occurring within an existing curriculum that is supported by commercial textbooks and enacted independently in each section, whereas materials were developed and piloted to support the new curricula in the transformed chemistry and physics courses. The continuing efforts in Bio I to develop and use common learning goals, some common assessments, and 3DL materials such as whole-class modeling activities (*53*) may facilitate further increases in use of 3DL assessment items. Another possibility, however, is that changes in Bio I will stall in the same way that Bio II appears to have stalled. That is, faculty who are open to these changes have changed, and others who prefer more traditional curricula and assessments will not, although a recent shift to a new governance structure for the introductory biology courses emphasizes collective oversight and is intended to assign faculty generally supportive of 3DL to these courses (*51*).

CONCLUSION

There are many ongoing efforts to transform STEM education, but few tools to measure change, particularly in curricular materials such as assessments. Here, we demonstrate that the 3D-LAP can be used to measure change in assessments across multiple disciplines and across multiple instructors within a given discipline. The protocol is sensitive enough to distinguish a complete curricular transformation (as in Chem I and II) from an ongoing, more slowly progressing effort (as in Bio I). We have also used this protocol to demonstrate that active learning approaches do not necessarily result in changes to assessments that align with 3DL. Further, assessments, not just instructional activities, must elicit evidence of 3DL to communicate to students that 3DL is highly valued (*35*). Therefore, we recommend that a combination of tools, including the 3D-LAP, be used to evaluate and understand STEM transformation efforts, particularly those in science disciplines. Change in assessments indicates a change in instructor focus and, consequently, student focus; these results can be viewed as a measure of both instructional change and student learning.

In addition to showing that the 3D-LAP is a valuable tool for measuring transformation efforts, this study reveals key curricular and programmatic elements that should be considered in transformation efforts in higher education:

1) Highly structured and coordinated efforts can facilitate transformation and sustainability; alternatively, grass roots efforts by groups of faculty in a supportive department can accelerate course transformation.

2) Developing and adapting transformed instructional materials and online assets can hasten transformation, while using solely traditional instructional materials and online assets, particularly those with a local history and pattern of resource investment, can hinder transformation.

3) Institutional investments, such as support for graduate TA and undergraduate LA time, can be used to leverage change, but how those investments are used affects transformation.

Although there is no single recipe for sustainable change within the same institution or even discipline, our work suggests that these factors can serve as levers and obstacles. Successful and sustainable transformation efforts must leverage institutional investment, centralized structural support, and educational leadership and expertise to engage faculty in meaningful change.

Last, this work raises additional questions that are the subjects of ongoing and future work: (i) What characterizes 3DL instruction? (ii) What supports and hinders faculty in applying the 3DL approach to their courses, and how can barriers to adoption be addressed? (iii) In what ways does participating in 3DL-focused courses change outcomes for students? The fundamental purpose of any transformation project is, of course, to improve outcomes for students. Our ultimate goal is to establish how 3DL courses affect student understanding and use of knowledge, and how these courses can equip students with the knowledge and skills that support them in becoming scientifically literate citizens and successful scientists and engineers.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/4/10/eaau0554/DC1

Protocol for coding video data for teaching activities

Table S1. Summary of unique attendees at 3DL activities.

Table S2. Complete tabulation of exam data.

Table S3. Complete tabulation of video data.

Raw exam, grade, and video data

REFERENCES AND NOTES

1. National Research Council, *Transforming Undergraduate Education in Science, Mathematics, Engineering, and Technology* (The National Academies Press, 1999).
2. V. K. Otero, D. E. Meltzer, 100 years of attempts to transform physics education. *Phys. Teach.* **54**, 523–527 (2016).

3. National Academy of Sciences, National Academy of Engineering, Institute of Medicine, *Rising above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future* (The National Academies Press, 2007).
4. National Research Council, *Reaching Students: What Research Says about Effective Instruction in Undergraduate Science and Engineering* (The National Academies Press, 2015).
5. President's Council of Advisors on Science and Technology, *Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics* (2012); https://files.eric.ed.gov/fulltext/ED541511.pdf.
6. S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8410–8415 (2014).
7. M. Stains, J. Harshman, M. K. Barker, S. V. Chasteen, R. Cole, S. E. DeChenne-Peters, M. K. Eagan Jr., J. M. Esson, J. K. Knight, F. A. Laski, M. Levis-Fitzgerald, C. J. Lee, S. M. Lo, L. M. McDonnell, T. A. McKay, N. Michelotti, A. Musgrove, M. S. Palmer, K. M. Plank, T. M. Rodela, E. R. Sanders, N. G. Schimpf, P. M. Schulte, M. K. Smith, M. Stetzer, B. Van Valkenburgh, E. Vinson, L. K. Weir, P. J. Wendel, L. B. Wheeler, A. M. Young, Anatomy of STEM teaching in North American universities. *Science* **359**, 1468–1470 (2018).
8. Association of American Universities, *Progress Toward Achieving Systemic Change: A Five-Year Status Report on the AAU Undergraduate STEM Education Initiative* (Association of American Universities, 2017); www.aau.edu/sites/default/files/AAU-Files/STEM-Education-Initiative/STEM-Status-Report.pdf.
9. National Research Council, *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (The National Academies Press, 2012).
10. M. M. Cooper, M. D. Caballero, D. Ebert-May, C. L. Fata-Hartley, S. E. Jardeleza, J. S. Krajcik, J. T. Laverty, R. L. Matz, L. A. Posey, S. M. Underwood, Challenge faculty to transform STEM learning. *Science* **350**, 281–282 (2015).
11. M. M. Cooper, It is time to say what we mean. *J. Chem. Educ.* **93**, 799–800 (2016).
12. C. Henderson, A. Beach, N. Finkelstein, Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *J. Res. Sci. Teach.* **48**, 952–984 (2011).
13. M. Borrego, C. Henderson, Increasing the use of evidence-based teaching in STEM higher education: A comparison of eight change strategies. *J. Eng. Educ.* **103**, 220–252 (2014).
14. A. Kezar, *How Colleges Change: Understanding, Leading, and Enacting Change* (Routledge, ed. 1, 2014).
15. J. T. Laverty, S. Tessmer, M. M. Cooper, M. D. Caballero, Engaging physics faculty in course transformation, paper presented at the Physics Education Research Conference, American Association of Physics Teachers, 30 July 2014; www.compadre.org/per/items/detail.cfm?ID=13470.
16. J. T. Laverty, S. M. Underwood, R. L. Matz, L. A. Posey, J. H. Carmel, M. D. Caballero, C. L. Fata-Hartley, D. Ebert-May, S. E. Jardeleza, M. M. Cooper, Characterizing college science assessments: The three-dimensional learning assessment protocol. *PLOS ONE* **11**, e0162333 (2016).
17. M. M. Cooper, L. A. Posey, S. M. Underwood, Core ideas and topics: Building up or drilling down? *J. Chem. Educ.* **94**, 541–548 (2017).
18. C. E. Wieman, Large-scale comparison of science teaching methods sends clear message. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8319–8320 (2014).
19. B. A. Couch, T. L. Brown, T. J. Schelpat, M. J. Graham, J. K. Knight, Scientific teaching: Defining a taxonomy of observable practices. *CBE Life Sci. Educ.* **14**, ar9 (2015).
20. M. T. Hora, A. Oleson, J. J. Ferrare, *Teaching Dimensions Observation Protocol (TDOP) User's Manual* (Wisconsin Center for Education Research, University of Wisconsin-Madison, 2013).
21. D. Sawada, M. D. Piburn, E. Judson, J. Turley, K. Falconer, R. Benford, I. Bloom, Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *Sch. Sci. Math.* **102**, 245–253 (2002).
22. M. K. Smith, F. H. M. Jones, S. L. Gilbert, C. E. Wieman, The classroom observation protocol for undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *CBE Life Sci. Educ.* **12**, 618–627 (2013).
23. M. Dancy, C. Henderson, C. Turpen, How faculty learn about and implement research-based instructional strategies: The case of Peer Instruction. *Phys. Rev. Phys. Educ. Res.* **12**, 010110 (2016).
24. D. Ebert-May, T. L. Derting, J. Hodder, J. L. Momsen, T. M. Long, S. E. Jardeleza, What we say is not what we do: Effective evaluation of faculty professional development programs. *BioScience* **61**, 550–558 (2011).
25. D. Ebert-May, T. L. Derting, T. P. Henkel, J. M. Maher, J. L. Momsen, B. Arnold, H. A. Passmore, Breaking the cycle: Future faculty begin teaching with learner-centered strategies after professional development. *CBE Life Sci. Educ.* **14**, ar22 (2015).
26. R. Kane, S. Sandretto, C. Heath, Telling half the story: A critical review of research on the teaching beliefs and practices of university academics. *Rev. Educ. Res.* **72**, 177–228 (2002).
27. C. Henderson, M. H. Dancy, Impact of physics education research on the teaching of introductory quantitative physics in the United States. *Phys. Rev. ST Phys. Educ. Res.* **5**, 020107 (2009).
28. J. K. Hubbard, M. A. Potts, B. A. Couch, How question types reveal student thinking: An experimental comparison of multiple-true-false and free-response formats. *CBE Life Sci. Educ.* **16**, ar26 (2017).
29. H.-S. Lee, O. L. Liu, M. C. Linn, Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Appl. Meas. Educ.* **24**, 115–136 (2011).
30. T. J. Crooks, The impact of classroom evaluation practices on students. *Rev. Educ. Res.* **58**, 438–481 (1988).
31. N. J. Entwistle, Approaches to learning and perceptions of the learning environment: Introduction to the special issue. *High. Educ.* **22**, 201–204 (1991).
32. J. Momsen, E. Offerdahl, M. Kryjevskaia, L. Montplaisir, E. Anderson, N. Grosz, Using assessments to investigate and compare the nature of learning in undergraduate science courses. *CBE Life Sci. Educ.* **12**, 239–249 (2013).
33. K. Scouller, The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *High. Educ.* **35**, 453–472 (1998).
34. B. R. Snyder, *The Hidden Curriculum* (The MIT Press, 1973).
35. S. M. Underwood, L. A. Posey, D. G. Herrington, J. H. Carmel, M. M. Cooper, Adapting assessment tasks to support three-dimensional learning. *J. Chem. Educ.* **95**, 207–217 (2018).
36. J. T. Laverty, M. D. Caballero, Analysis of the most common concept inventories in physics: What are we assessing? *Phys. Rev. Phys. Educ. Res.* **14**, 010123 (2018).
37. R. L. Matz, J. T. Laverty, S. E. Jardeleza, C. M. Morrison, Z. D. Nusbaum, S. A. Ly, S. M. Underwood, C. L. Fata-Hartley, L. A. Posey, D. Ebert-May, J. S. Krajcik, M. D. Caballero, M. M. Cooper, Investigating change in classroom instruction of scientific practices, crosscutting concepts, and core ideas, paper presented at the National Association for Research in Science Teaching Annual International Conference, Chicago, IL, 13 April 2015.
38. M. Cooper, M. Klymkowsky, Chemistry, life, the universe, and everything: A new approach to general chemistry, and a model for curriculum reform. *J. Chem. Educ.* **90**, 1116–1122 (2013).
39. M. M. Cooper, S. M. Underwood, C. Z. Hilley, M. W. Klymkowsky, Development and assessment of a molecular structure and properties learning progression. *J. Chem. Educ.* **89**, 1351–1357 (2012).
40. M. M. Cooper, H. Kouyoumdjian, S. M. Underwood, Investigating students' reasoning about acid–base reactions. *J. Chem. Educ.* **93**, 1703–1712 (2016).
41. L. C. Williams, S. M. Underwood, M. W. Klymkowsky, M. M. Cooper, Are noncovalent interactions an Achilles heel in chemistry education? A comparison of instructional approaches. *J. Chem. Educ.* **92**, 1979–1987 (2015).
42. S. M. Underwood, D. Reyes-Gastelum, M. M. Cooper, When do students recognize relationships between molecular structure and properties? A longitudinal comparison of the impact of traditional and transformed curricula. *Chem. Educ. Res. Pract.* **17**, 365–380 (2016).
43. G. Kortemeyer, W. Bauer, D. Kashy, E. Kashy, C. Speier, The LearningOnline Network with CAPA initiative, paper presented at the 31st Annual Frontiers in Education Conference, Institute of Electrical and Electronics Engineers Computer Society, Washington, DC, 10 October 2001; dl.acm.org/citation.cfm?id=1253531.1254661.
44. P. W. Irving, M. J. Obsniuk, M. D. Caballero, P^3: A practice focused learning environment. *Eur. J. Phys.* **38**, 055701 (2017).
45. R. K. Thornton, D. R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula. *Am. J. Phys.* **66**, 338–352 (1998).
46. E. B. Speth, J. L. Momsen, G. A. Moyerbrailean, D. Ebert-May, T. M. Long, S. A. Wyse, D. Linton, 1, 2, 3, 4: Infusing quantitative literacy into introductory biology. *CBE Life Sci. Educ.* **9**, 323–332 (2010).
47. J. T. Dauer, J. L. Momsen, E. B. Speth, S. C. Makohon-Moore, T. M. Long, Analyzing change in students' gene-to-evolution models in college-level introductory biology. *J. Res. Sci. Teach.* **50**, 639–659 (2013).
48. T. M. Long, J. T. Dauer, K. M. Kostelnik, J. L. Momsen, S. A. Wyse, E. B. Speth, D. Ebert-May, Fostering ecoliteracy through model-based instruction. *Front. Ecol. Environ.* **12**, 138–139 (2014).
49. M. K. Smith, E. L. Vinson, J. A. Smith, J. D. Lewin, M. R. Stetzer, A campus-wide study of STEM courses: New perspectives on teaching practices and perceptions. *CBE Life Sci. Educ.* **13**, 624–635 (2014).
50. G. Kortemeyer, E. Kashy, W. Benenson, W. Bauer, Experiences using the open-source learning content management and assessment system LON-CAPA in introductory physics courses. *Am. J. Phys.* **76**, 438–444 (2008).
51. R. L. Matz, S. E. Jardeleza, Examining the role of leadership in an undergraduate biology institutional reform initiative. *CBE Life Sci. Educ.* **15**, ar57 (2016).

52. A. Kezar, Bottom-up/top-down leadership: Contradiction or hidden phenomenon. *J. High. Educ.* **83**, 725–760 (2012).
53. N. B. Hubbs, K. N. Parent, J. R. Stoltzfus, Models in the biology classroom: An in-class modeling activity on meiosis. *Am. Biol. Teach.* **79**, 482–491 (2017).

Acknowledgments: We thank the many faculty who provided us their exams, allowed us to video record their class meetings, and participated in the disciplinary discussions, STEM Gateway Fellowship, and our workshops and seminars. Figure1 was initially conceptualized by K. Fredlund in the Data Inquiry Lab at Grand Valley State University. S. E. Jardeleza contributed to the development of the teaching activities video coding protocol, and our team of undergraduate research assistants (S. R. Luba, S. A. Ly, C. M. Morrison, K. L. Noyes, and Z. D. Nusbaum) helped anonymize the exam data and code the videos for teaching activities. Helpful feedback on paper drafts was provided by K. R. Bain, E. M. Duffy, J. R. Stoltzfus, R. D. Sweeder, S. H. Tessmer, M. Urban-Lurain, and the anonymous reviewers. **Funding:** Research funding was provided by the Association of American Universities, the MSU Office of the Provost, an MSU CREATE for STEM Institute LPF-CMP 2 Innovation Grant, and the National Science Foundation [DUE 0816692 (1359818) and 1725520]. **Author contributions:** M.M.C. conceived the study and supervised the project. R.L.M. and M.M.C. co-wrote the manuscript. R.L.M. managed data collection and co-supervised the research assistants with J.T.L. All authors coded and analyzed data, discussed the results, and commented on the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 2 May 2018
Accepted 19 September 2018
Published 24 October 2018
10.1126/sciadv.aau0554

Citation: R. L. Matz, C. L. Fata-Hartley, L. A. Posey, J. T. Laverty, S. M. Underwood, J. H. Carmel, D. G. Herrington, R. L. Stowe, M. D. Caballero, D. Ebert-May, M. M. Cooper, Evaluating the extent of a large-scale transformation in gateway science courses. *Sci. Adv.* **4**, eaau0554 (2018).

Evaluating the extent of a large-scale transformation in gateway science courses

Rebecca L. Matz, Cori L. Fata-Hartley, Lynmarie A. Posey, James T. Laverty, Sonia M. Underwood, Justin H. Carmel, Deborah G. Herrington, Ryan L. Stowe, Marcos D. Caballero, Diane Ebert-May and Melanie M. Cooper

Sci Adv **4** (10), eaau0554.
DOI: 10.1126/sciadv.aau0554