Classroom sound can be used to classify teaching practices in college science courses

Melinda T. Owens,1 Shannon B. Seidel,1 Mike Wong,1 Travis E. Beijnes,1 Susanne Lietz,2 Joseph R. Perez,2 Shangheng Sit,2 Zahur-Saleh Subedar,2 Gigi N. Acker,2 Susan F. Akana,3 Brad Balukjian,9 Hilary P. Benton,9 J. R. Blair,9 Segal M. Boaz,9 Katheryn E. Boyer,2 Jason B. Bram,2 Laura W. Burrous,3 Dana T. Byrd,2 Natalia Caporable,2 Edward J. Carpenter,2 Yee-Hung Mark Chan,2 Lily Chen,2 Amy Chovnicki,2 Diana S. Chu,2 Bryon K. Clarkson,1 Sara E. Cooper,2 Catherine Creech,2,9 Karen D. Crow,1 José R. de la Torre,1 Wilfred F. Denedclaw,5 Kathleen E. Duncan,5 Amy S. Edwards,5 Karen L. Erickson,9 Megumi Fuse,6 Joseph J. Gorgan,8 Brinda Govindan,8 L. Jeanette Green,9 Paul Z. Hankamp,10 Holly E. Harris,9 Zheng-Hui He,12 Stephen Ingalls,2 Peter D. Ingmire,5 J. Rebecca Jacobsh,3 Mark Kamakea,2 Rhea R. Kimpo,11 Jonathan D. Knight,3 Sara K. Krause,1 Lori E. Krueger,5 Terry L. Light,5 Lance Lund,12 Leticia M. Márquez-Magaña,3 Briana K. McCarthy,9 Linda J. McPherson,8 Vanessa C. Miller-Sims,3 Christopher A. Moffatt,6 Pamela C. Muick,2,11 Paul H. Nagami,9,12 Gloria L. Nusse,3 Kristine M. Okimura,8 Sally G. Pasion,9 Robert Patterson,10 Pleuni S. Pennings,13 Blake Riggs,4 Joseph Romeo,14 Scott W. Roy,9 Tatiana Russo-Tait,1b Lisa M. Schultheis,13 Lakshmikanta Sengupta,1 Rachel Small,62 Greg S. Spiker,6 Jonathon H. Stillman,13 Andrea Swaei,2 Jennifer M. Wade,14,15 Steven B. Waters,14 Steven L. Weinstein,14 Julia K. Willise,15 Diana W. Wright,2e Colin D. Harrison,13 Loretta A. Kelley,2g,9 Gloriana Trujillo,5h,11 Carmen R. Domingo,9 Jeffrey N. Schinskey,1h,15 and Kimberly D. Tanner12

1Department of Biology, San Francisco State University, San Francisco, CA 94132; 2Department of Biology, Pacific Lutheran University, Tacoma, WA 98447; 3Center for Computing for Life Sciences, San Francisco State University, San Francisco, CA 94132; 4Department of Biology, De Anza College, Cupertino, CA 95014; 5Nutrition, Food Science, and Packaging Science, San Francisco State University, San Francisco, CA 94132; 6Biological Department, City College of San Francisco, San Francisco, CA 94112; 7Biological Department, Laney College, Oakland, CA 94607; 8Department of Biology, Foothill College, Los Altos Hills, CA 94022; 9Biological Department, Las Positas College, Livermore, CA 94551; 10Romberg Tiburon Center for Environmental Studies, San Francisco State University, Tiburon, CA 94920; 11Department of Neurobiology, Physiology, and Behavior, University of California, Davis, CA 95616; 12Department of Biological Science, Diablo Valley College, Pleasant Hill, CA 94523; 13Department of Biology, Portland Community College, Portland, OR 97219; 14Math and Sciences Department, Diablo Valley College, San Ramon, CA 94582; 15Science and Technology Division, Cañada College, Redwood City, CA 94061; 16Biological Department, College of San Mateo, San Mateo, CA 94042; 17Division of Undergraduate Education and Academic Planning, San Francisco State University, San Francisco, CA 94132; 18Life Science Department, Chabot College, Hayward, CA 94545; 19Science/Mathematics/Technology Division, Skyline College, San Bruno, CA 94066; 20Life Sciences Department, Palomar College, San Marcos, CA 92069; 21Biological Department, Solano Community College, Fairfield, CA 94534; 22Department of Biological Sciences, California State University, Sacramento, CA 95819; 23Biological Department, Los Medanos College, Pittsburg, CA 94565; 24Science Department, Berkeley City College, Berkeley, CA 94704; 25Biological Sciences Department, Contra Costa College, San Pablo, CA 94806; 26Department of Biological Science, Holy Names University, Oakland, CA 94619; 27Department of Earth and Climate Sciences, San Francisco State University, San Francisco, CA 94132; 28Department of Curriculum and Instruction, STEM Education, University of Texas at Austin, Austin, TX 78712; 29Department of Chemistry and Biochemistry, San Francisco State University, San Francisco, CA 94132; 30Department of Biology, University of San Francisco, San Francisco, CA 94117; 31Biological, Health & Environmental Sciences Division, DeAnza College, Cupertino, CA 95014; 32School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332; 33Kelley, Petterson, and Associates, Inc., San Francisco, CA 94127; and 34Office of the Vice Provost for Teaching and Learning, Stanford University, Stanford, CA 94305

Edited by Bruce Alberts, University of California, San Francisco, CA, and approved January 31, 2017 (received for review November 20, 2016)

Active-learning pedagogies have been repeatedly demonstrated to produce superior learning gains with large effect sizes compared with lecture-based pedagogies. Shifting large numbers of college science, technology, engineering, and mathematics (STEM) faculty to include any active teaching in their mind retains and more effectively educate far more students than having a few faculty completely transform their teaching, but the extent to which STEM faculty are changing their teaching methods is unclear. Here, we describe the development and application of the machine-learning-derived algorithm Decibel Analysis for Research in Teaching (DART), which can analyze thousands of hours of STEM course audio recordings quickly, with minimal costs, and without need for human observers. DART analyzes the volume and variance of classroom recordings to predict the quantity of time spent on single voice (e.g., lecture), multiple voice (e.g., pair discussion), and no voice (e.g., clicker question thinking) activities. Applying DART to 1,486 recordings of class sessions from 67 courses, a total of 1,720 h of audio, revealed varied patterns of lecture (single voice) and nonlecture activity (multiple and no voice) use. We also found that there was significantly more use of multiple and no voice strategies in courses for STEM majors compared with courses for non-STEM majors, indicating that DART can be used to compare teaching strategies in different types of courses. Therefore, DART has the potential to systematically inventory the presence of active learning with ~90% accuracy across thousands of courses in diverse settings with minimal effort.

Active-learning | evidence-based teaching | science education | lecture | assessment

Current college STEM (science, technology, engineering, and mathematics) teaching in the United States continues to be lecture-based and is relatively ineffective in promoting learning (1, 2). Undergraduate instructors continue to struggle to engage, effectively teach, and retain postsecondary students, both generally and particularly among women and students of color (3, 4). Federal analyses suggest that a 10% increase in retention of undergraduate STEM students could address anticipated STEM workforce shortfalls (5). Replacing the standard lecture format with more active teaching strategies has been shown to increase

PNAS | March 21, 2017 | vol. 114 | no. 12 | 3085–3090

www.pnas.org/cgi/doi/10.1073/pnas.1618693114


This article is a PNAS Direct Submission. Freely available online through the PNAS open access option.

M.T.O., S.B.S., and M.W. contributed equally to this work.

1To whom correspondence should be addressed. Email: kdtanner@sfsu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1618693114/DCSupplemental.

Published by AAAS on September 28, 2021.

Copyright © 2021 by The National Academy of Sciences, USA.
Significance

Although the United States needs to expand its STEM (science, technology, engineering, mathematics) workforce, United States postsecondary institutions struggle to retain and effectively teach students in STEM disciplines. Using teaching techniques beyond lecture, such as pair discussions and reflective writing, has been shown to boost student learning, but it is unknown what proportion of STEM faculty use these active-learning pedagogies. Here we describe DART: Decibel Analysis for Research in Teaching, a machine-learning-derived algorithm that analyzes classroom sound to predict with high accuracy the learning activities used in classrooms, and its application to thousands of class session recordings. DART can be used for large-scale examinations of STEM teaching practices, evaluating the extent to which educators maximize opportunities for effective STEM learning.

Results

Our key insight from observations of classroom environments was that nonlecture activities are typically associated with either unusually high noise levels (e.g., pair discussions, small group discussions) or unusually low noise levels (e.g., individual clicker question response, minute paper writing). This suggests that variation in the sound level of a classroom may indicate variation in teaching strategies. To test this hypothesis, an initial 45 audio recordings from 8 instructors teaching different courses (Table 1, pilot group) were analyzed by extracting audio decibel levels at a 2-Hz sampling rate (every 0.5 s) and graphing sound waveforms. To analyze DART’s performance in diverse teaching settings, these instructors were purposefully drawn from an atypically consistent from people from many different institutions who had undergone over 40 h of professional development in scientific teaching. To determine if patterns of variation in waveforms correlated with activity types, a three-person team listened to all recorded class sessions and individually annotated them using six emergent annotation codes (lecture with Q/A, discussion, silent, video, transition, and other) (Table S1). Sound-level patterns in class sessions primarily using lecture with Q/A were visibly different from the patterns in class sessions with varied learning activities (Fig. 1 A and C).

Developing an Algorithm to Automate the Classification of Classroom Noise. To develop DART, human annotations were used to design and optimize a machine-learning–based algorithm that reports what types of activities are going on in a classroom based on sound waveforms. To do this task, we applied methods from the field of audio segmentation, which applies machine learning to classify sound into different categories based on statistical characterizations (20). Because some of the human annotation categories yielded waveforms that were statistically similar to each other, we collapsed the six human annotation categories into four activity prediction modes with distinct waveform profiles: single voice, multiple voice, no voice, and other. Lecture with Q/A and video were aggregated into the mode “single voice”; discussion and transition were aggregated into the mode “multiple voice”; silent was assigned to the mode “no voice”; and other was assigned to the mode “other” (Table S1).

To prepare the classroom audio-recording waveforms for the optimization procedure, we tagged each 0.5-s sample of sound from each recording from the pilot group (640,152 samples in total) with three pieces of data: its label from human annotation (S for single voice, M for multiple voice, or N for no voice), the normalized mean volume of the 15-s window of audio around it, and the normalized SD in that window’s volume (Fig. S1A). Both the mean volume and the SD of the volume of each sample were normalized with respect to their class session.

Then, to sort the samples into the four prediction modes (single voice, multiple voice, no voice, and other), we used an ensemble of binary decision trees comprised of four nodes connected serially. A binary decision tree is a series of decisions to either sort or not sort a given input into a certain category based on the values of the input. Here, the inputs were the 0.5-s samples of classroom audio, and the sorting decisions were based on each sample’s normalized mean volume and SD of the volume. In our tree, each node represented one activity prediction mode, and the nodes for each mode were connected in order of decreasing frequency from the pilot data, so

Table 1. Overview of DART study participants

<table>
<thead>
<tr>
<th>Group</th>
<th>Instructors</th>
<th>Courses</th>
<th>Class sessions</th>
<th>Recorded hours (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilot group</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Community college</td>
<td>8</td>
<td>8</td>
<td>45</td>
<td>65</td>
</tr>
<tr>
<td>Large-scale analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Community college</td>
<td>27</td>
<td>35</td>
<td>712</td>
<td>970</td>
</tr>
<tr>
<td>Four-year university</td>
<td>22</td>
<td>32</td>
<td>774</td>
<td>750</td>
</tr>
<tr>
<td>All</td>
<td>49</td>
<td>67</td>
<td>1,486</td>
<td>1,720</td>
</tr>
</tbody>
</table>

| Number of instructors, courses, class sessions, and hours recorded in each group.
that the dominant class activity (single voice) was detected first, and less-frequent class activities follow (multiple voice, no voice, and other, in that order) (Fig. S1B). This ordering emphasized the importance of predicting the common activities correctly while allowing some prediction flexibility for the less-frequent activities.

Next, we optimized the selection parameters that would determine which audio samples were sorted into which activity modes. To accomplish this, we used machine learning, specifically grid search (Fig. S1C and D). Grid search is a brute-force method to select the optimal selection parameters for each mode by first evaluating each possible combination of the two selection parameters, the normalized average volume and the normalized average SD, and then choosing the pair of parameter values that yielded the model with the best match to human annotation, defined as the fewest number of errors. This grid search process was conducted three times—once each for single voice, multiple voice, and no voice—to find the optimal parameters for each activity prediction mode. For more details of the development of the DART algorithm, refer to SI Methods, Development of DART Algorithm with Machine Learning.

We found that the resulting algorithm, DART, is able to classify each 0.5-s sample of a recording into one of three DART prediction modes: single voice, multiple voice, or no voice. (The final algorithm never categorizes samples as other, probably because the human annotation “other” was assigned only 0.9% of the time to a variety of instances that were difficult to categorize in the pilot data.) Single-voice samples, characterized by one person speaking at a time (e.g., lecture, question/answer, and so forth), were of average volume but high variance. Single voice typically indicated nonactive teaching strategies given that only a single active voice was heard, with all other individuals passively listening. In contrast,
multiple-voice samples, characterized by many people speaking simultaneously (e.g., pair discussions), were of high mean volume and low variance. No-voice samples, characterized by quiet throughout the classroom (e.g., silent writing), were of low mean volume and low variance. As verified by human annotations, multiple and no voice generally indicated active learning because many or all students actively were engaged in a task.

DART CLASSIFIES CLASSROOM NOISE WITH HIGH ACCURACY. To assess the accuracy of DART, we compared DART’s classifications of classroom noise to the human annotations in various ways, both in the original dataset of 45 class sessions collected from 8 instructors and a new, larger dataset comprised of 1,486 class sessions collected from 49 instructors, representing 67 courses taught across 15 community colleges and a 4-y university, a total of 1,720 h of recordings (Table 1). Qualitatively, we saw that DART was able to differentiate between lecture and nonlecture classroom activities. For example, DART predicted a class session that was annotated as 98% lecture with Q/A to be solely single voice (Fig. 1 A and B) and a class session with varied activities, like silent writing and discussion, to have a variety of modes (Fig. 1 C and D). DART identification of varied learning activities was robust in both small and large classes (Fig. 1 E and F). Its predictions reveal that waveform “footprints” are indicative of specific teaching techniques (Fig. 1 G). For example, the common active learning technique “think-pair-share” actually consists of three distinct activities in response to an instructor’s question to the class: first students silently think or write about the answer, then they discuss it in pairs or small groups, and finally some students share their responses individually with the class. A human would annotate these three phases, in order, as silent, discussion, and lecture with Q/A. Similarly, DART assigns no voice (think), multiple voice (pair), and single voice (share) (Fig. 1 G).

We also assessed DART’s accuracy quantitatively by measuring how often DART predictions matched the human annotations. In the original dataset used for optimizing the algorithm, DART classification matched the human annotations 90% of the time across all modes. In comparison, human annotators agreed with each other only 95% of the time, showing that DART was almost as accurate at identifying classroom activities as human annotators were. To see if this high rate of accuracy was retained in a new context, we randomly chose one class session from each of the 67 courses recorded as part of the new, larger dataset, performed human annotation, and compared DART’s classifications to the human annotation. We again obtained a very high accuracy of 87%, suggesting that DART can accurately applied to many different classroom contexts.

To further assess DART’s ability to discern the presence of activities that may indicate active learning or traditional lecture, we used signal-detection theory to analyze DART’s accuracy by mode. In the original dataset, we used signal-detection theory to discriminate for each mode (single voice, multiple voice, and no voice) between correct inclusions (hits) and incorrect exclusions (misses) (21). We also used this method to determine the rates of correct exclusions (correct rejections) and incorrect inclusions (false alarms) for each of the three modes (21). The results are given in Fig. 2. DART correctly identifies nearly all instances of lecture and Q/A as single voice (hit rate = 98.0%) (Fig. 2 A). In addition, the false-alarm rates for multiple voice and no voice are low (2.3% and <0.1%, respectively) (Fig. 2 B and C). Combined, these rates mean that most errors over- rather than underestimate lecture, minimizing the potential for falsely indicating the presence of active learning in class sessions.

DART can be used to perform large-scale analysis of classrooms. We sought to explore how DART could be used to analyze classroom audio recordings on a larger scale, so we performed DART analysis on the larger dataset consisting of 1,720 h of recordings of 67 courses. DART analysis revealed that in these courses, a range of instructional strategies were represented. Although all courses (n = 67) used single voice a majority of the time, ranging from 69 to 100%, among individual class sessions (n = 1,486), time spent in single voice ranged from 15 to 100% (Fig. 3 A and B). Within a course, we observed that the time spent in single voice could vary from 15% in one class session to 90% in another class session (Fig. 3 C). In addition, some instructors that had no multiple or no voice in some class sessions nevertheless spent up to 37% of the time in these categories in another class session within the same course (Fig. 3 D). This within-course variability highlights the need for a tool that can efficiently analyze every class session of a course.

To determine the likelihood a student experienced active learning in any one of these courses, we calculated the percentage of class sessions within each course that included any multiple or no voice (<100% single voice). Whereas only 31% of the courses had multiple or no-voice activities in all class sessions, 88% of courses had multiple or no-voice activities in at least half of their class sessions (Fig. 3 D), indicating that many of these instructors are using active-learning strategies, which is likely unusual among undergraduate STEM instructors.

DART also has the potential to reveal differences in how courses are taught across instructors and courses in particular departments or institutions. In this course sample, we found that the percentage of time spent in multiple or no voice did not vary by instructor gender (n = 36 female, n = 26 male; P = 0.10) but was significantly higher in courses for biology majors (n = 32) than nonbiology majors (n = 35; P = 0.01) (Fig. 3 B and E).

Discussion
In summary, we have described the development and validation of DART, an analytical tool that uses sound levels to predict classroom activities, as well as results from applying DART to 67 STEM courses. We show that DART is robust to varying class sizes and can determine the presence and quantity of single-voice (e.g., lecture), multiple-voice (e.g., pair or group discussion), or no-voice (e.g., clicker question, thinking, or quiet writing) learning activities with ~90% accuracy. At this level of accuracy, easy, and time efficiency (~5 min per 2-h class session), one could analyze and draw broad conclusions about millions of hours of class sessions at periodic intervals over time. Because DART only analyzes sound
levels, it protects the anonymity of instructors and students. Furthermore, because DART detected differences in the extent of nonlecture in courses for nonbiology majors versus biology majors, DART additionally promises to reveal differences among other types of courses, instructors, disciplines, and institutions that were previously not feasible for study.

DART is relevant to many educational stakeholders, from individual instructors to institutions and researchers. For individual instructors, DART holds additional promise as a tool for individual instructor professional development. Although previous studies have shown that many STEM faculty aspire to change their teaching (22), detailed observations of classroom videos suggest that instructors overestimate the extent to which they have integrated reformed teaching practices in their classrooms (23). DART could provide instructors with quick and quantitative evidence for instructor self-study. DART can quickly identify those class sessions with minimal to no learning activities for students and enable faculty to specifically target how they spend their limited time for pedagogical innovation. For disciplinary programs or departments and the faculty developers that support their teaching efforts, DART could supplement ongoing program assessment, providing insight into the nature of the learning activities happening in different courses with varying student outcomes. It could quickly reveal differences in the teaching strategies used across a department, allowing faculty to have discussions of teaching goals across the curriculum. For institutions, DART may provide a means for describing to prospective students and skeptical parents the added value of a STEM education at their particular campus. Increasingly, parents and students seek information about the added value of an education at particular institution, going beyond academic reputation and research profile, and DART could help institutions make transparent the extent to which their students experience active engagement and their faculty use pedagogically effective teaching methods in their courses. Finally, for federal and private agencies attempting to foster change in STEM faculty teaching practices, DART has the potential to dramatically increase systematic measurement of classroom practices and expand insights being gained from current evaluation approaches through self-report, occasional classroom observations, and time-consuming videotape analyses. In addition, although DART emerged from studies of STEM classrooms, DART also has the potential to address similar inquiries about university classrooms in other subjects or about precollege settings. DART’s efficiency could allow for studying correlations between DART’s quantitative metrics and a variety of variables associated with STEM courses, including positive outcomes, such as overall learning gains, pass rates, and success in follow-on courses, as well as negative outcomes, such as persistent achievement gaps correlated with student gender or cultural background. It is important to note that DART is not suitable for ranking or evaluating individual instructors, both because of the possibility of errors and because DART is not intended to measure the quality of teaching. Although much research has established that any form of active learning appears to produce higher learning gains than lecture alone (9), it is not known how much or what patterns of active learning may be adequate or optimal for learning.

So, what proportion of STEM instructors in the United States and internationally regularly use teaching strategies beyond lecture? What is the probability that an undergraduate STEM student would have the opportunity to speak, write, or discuss their ideas with peers in every class session? Analyzing classroom noise can quickly and anonymously reveal what is happening in classrooms, making DART a measurement tool with the potential to systematically inventory the presence of active learning across all types of higher education institutions. Given pressing needs to expand and diversify STEM workforces in the United States and beyond, DART can also be used to characterize the extent to which educators are maximizing opportunities for effective STEM learning. Because DART will be available online at dart.sfsu.edu, thousands of instructors, students, or other stakeholders could soon perform DART analyses, opening a variety of new lines of research and inquiry.

Methods

Audio Recording. Each audio recording analyzed as part of this paper was obtained from Sony audio recorder model ICD-PX333. Decibel analysis has also been completed using recordings made on the iPhone Voice Memo App.
as well as live-recording the sound levels in the classroom using the iPhone Decibel 10th App. Instructors were given audio recorders and asked to record every class session of at least one of the courses they were teaching. They were instructed to place the audio recorders at the front of the classroom (e.g., on a lectern) with the microphone pointing in the general direction of students. Before analysis, recordings were trimmed by hand at the beginning and end to exclude noise associated with student arrival and departure.

Instructor Population. Courses analyzed in this study were taught by collaborators on the Talk Matters Project, an advanced collaborative scientific teaching research project. Participating instructors in this project were drawn from two faculty development programs focusing on scientific teaching: Community College Biology Faculty Enhancement through Scientific Teaching (CCB FEST), for community college biology faculty; and Biology Faculty Explorations in Scientific Teaching (Biology FEST), for biology faculty in a single 4-y university. They included part-time, full-time, and tenured/tenure-track faculty teaching a variety of biology courses, including lower- and upper-division courses and courses for biology majors and nonbiology majors. Course enrollments ranged from 4 to 287 students with a median course size of 48 students. Faculty were recruited in two phases, a pilot phase in Spring 2014 and a large-scale analysis phase in Spring 2015. The research done was a collaboration between dozens of faculty instructors, and as a result there were no human subject protections and no informed consent. Each instructor who contributed recordings has a letter of collaboration on file with San Francisco State University’s Institutional Review Board, which approved the research described in this report in exempt protocols #E14-141a-d. For more information about faculty recruitment and participation rates, see SI Methods, Participant Recruitment and Table S2.

Human Annotation of Pilot Data. The development of annotation codes was an iterative process. A team of three people annotated a total of 45 class session recordings split between the 8 instructors in the pilot group. Initially, human annotation was unstructured, and coders were charged to individually listen to audio files, observe audio waveforms, and develop codes that correlated with the types of activities occurring in class sessions. For each new activity lasting more than 15 s, annotators indicated a start time (minutes and seconds) and a code (e.g., Q/A, lecture). Codes from all three annotators were compared and collapsed into six categories (lecture with Q/A, discussion, individually listen to audio files, observe audio waveforms, and develop codes)

Common DART errors are described in Table S3.

**Measurement of DART’s Accuracy.**

**Pilot Data.** In the final model used for DART, model prediction accuracy was found to be 89.5% accurate overall on the pilot data. The accuracy was found by calculating the percentage of time the prediction mode matched the human annotation for all 66 annotations (of 45 class sessions; some classes were annotated by multiple people). As noted above, by the same metric, the human annotators achieved an accuracy of 87.7%, because human annotators did not always agree. We also analyzed the accuracy of DART with signal-detection theory (Fig. 2). Signal-detection theory calculations of hit, miss, false positive, and correct rejection rates used equations outlined in Stanislaw and Todorov (21) and were calculated in Excel.

For further analyses of DART’s accuracy on the pilot group data, see SI Materials and Methods, Further DART Accuracy Measures and Fig. S2. Common DART errors are described in Table S3.

**Large-scale analysis data.** To calculate DART’s accuracy on the large-scale analysis data set, one class session from each of this dataset’s 67 courses was randomly chosen and annotated by a new two-person team trained in annotation using the previous annotation team’s codes and files. We compared how often the human annotations matched DART’s predictions, obtaining an accuracy of 87%.

**DART Analysis of a Large Set of Courses.** Fifty-seven instructors recorded at least one class session in 78 distinct courses. Of these 78 courses, we only included nonlaboratory biology courses where at least 30% of class sessions were recorded. Therefore, we excluded three courses for being laboratories and eight courses for having low numbers of recordings, giving an inclusion rate of 67 of 78 = 85.9%.

DART was used to calculate the time spent in single voice, multiple voice, and no voice for each class session. To compare DART data between different groups of courses, we used t tests in Excel on logit-transformed DART data, to correct for using percentage data.

**ACKNOWLEDGMENTS.** We thank John Coley, Heidi McLaughlin, Sarah Bissonnette, Kristin de Nee, the National Science Foundation-funded Community College Biology Faculty Enhancement through Scientific Teaching community, and the Howard Hughes Medical Institute-funded Biology Faculty Enhancement through Scientific Teaching community for insightful discussion and support. We also thank the Center for Computing for Life Sciences at San Francisco State University for extensive support. This work was funded by Howard Hughes Medical Institute Undergraduate Science Education Award 52007556 and National Science Foundation Transformation Undergraduate Education in Science, Technology, Engineering, and Mathematics Award DUE-1226361.

6. Eddy SL, Hogan KA (2014) Getting under the hood: How and for whom does in-