Journal of Learning Spaces
Volume 9, Number 2. 2020

ISSN 21586195

The Practitioner's Panacea for Measuring Learner-Centeredness?

Ashley B. Heim
University of Northern Colorado

Emily A. Holt
University of Northern Colorado

The Decibel Analysis for Research in Teaching (DART; Owens et al., 2017), a sound-based metric of learner-centeredness, is highly accessible, requires no training, and can be conducted with minimal classroom observations; yet, DART has not been evaluated in comparison with other validated metrics or in consideration of potentially confounding classroom characteristics (e.g. enrollment, classroom size, number of doors). We analyzed recordings from 42 class sessions of an undergraduate biology course with DART, the Reformed Teaching Observation Protocol (RTOP), and nine classroom characteristics. We found that enrollment was the best single predictor of the DART output of learner-centeredness, percent Multiple Voice.

Introduction

What is Learner-Centeredness and the Challenges in Measuring It?

Learner-centeredness is characterized by how actively students are engaged in the learning process as they interact with their peers and instructor (Fahraeus, 2013). Often, but not always, active learning is necessary to foster a learner-centered classroom. Learner-centeredness has many suggested benefits for students, including lower failure rates (Freeman et al., 2014), increased student performance (Armbruster, Patel, Johnson, & Weiss, 2009; Freeman et al., 2014; Kahl Jr. & Venette, 2010; Walker, Cotner, Baepler, & Decker, 2008), and increased critical thinking skills, metacognitive abilities, and content knowledge (Bransford, Brown, & Cocking, 1999; Crouch & Mazur, 2001; Hake, 1998; Shepard, 2000). Given these benefits, instructors and researchers have sought reliable measures of learner-centeredness for reflection and to guide reform.

Observation rubrics objectively measure the quality or quantity of teaching strategies or tasks and student contributions in a classroom, thus they tend to be more accurate for education research studies (Cohen & Goldhaber, 2016; Shavelson, Webb, & Burstein, 1986). One of the most heavily used observation protocols, the Reformed Teaching Observation Protocol (RTOP; Amrein-Beardsley & Popp, 2012; Sawada et al., 2002), requires time-intensive training, which precludes its accessibility by practitioners. Even observation protocols that require less intensive training (e.g., Classroom Observation Protocol for Undergraduate STEM, Smith et al., 2013; Practical Observation Rubric To Assess Active Learning, Eddy, Converse, & Wenderoth, 2015; Teaching Perspectives Inventory, Pratt & Collins, 2000) are still time-intensive to conduct, or cannot be conducted with just a few observations (Measurement Instrument for Scientific Teaching-Observable, Durham et al., 2018). Thus, a more automated method of objectively classifying learner-centeredness in undergraduate courses is necessary for accessible and accurate feedback.

Reformed Teaching Observation Protocol

While the RTOP, with its extensive training requirements, is not accessible to all users, it is also considered the standard in observation protocols for discipline-based education research. RTOP has been used across science fields, including biology (e.g., Ebert-May et al., 2011, 2015; Gormally, Brickman, Hallar, & Armstrong, 2011; Heim & Holt, 2018), physics (e.g., MacIsaac & Falconer, 2002; Falconer, Wyckoff, Joshua, & Sawada, 2001), and chemistry (e.g., Rushton, Lotter, & Singer, 2011). Additionally, RTOP is versatile across education levels—including K-12 (Kilday & Kinzie, 2008; Sawada et al., 2002; Tarr et al., 2008) and college (Amrein-Beardsley & Popp, 2012; MacIsaac & Falconer, 2002; Ebert-May et al., 2011, 2015; Gormally et al., 2011; Heim & Holt, 2018). Researchers have used this instrument to study both longitudinal changes (Ebert-May et al., 2011, 2015) in classroom teaching practices as well as single time points or multiple RTOP scores averaged for individual class

Ashley B. Heim was a Ph.D. candidate in the School of Biological Sciences at the University of Northern Colorado when this manuscript was submitted. She is currently a postdoctoral researcher in Ecology & Evolutionary Biology at Cornell University.

Emily A. Holt is an Associate Professor in the School of Biological Sciences at the University of Northern Colorado.

sections (Amrein-Beardsley & Popp, 2012; Heim & Holt, 2018; Rushton et al., 2011). RTOP has been used to inform classroom reform (Gormally et al., 2011; Kilday & Kinzie, 2008; MacIsaac & Falconer, 2002) and for professional development (Ebert-May et al., 2011, 2015; Singer, Lotter, Feller, & Gates, 2011). The breadth and adaptability of RTOP make it an ideal instrument for objectively measuring learner-centered teaching practices, and represent the standard against which other instruments have been compared (Heim & Holt, 2018).

Classroom Sound as a Measure of Learner-Centeredness

Studies suggest that types of classroom learning activities can be categorized based on vocal classroom discourse and sound. Kranzfelder et al. (2019) developed the Classroom Discourse Observation Protocol to characterize teacher discourse moves in an undergraduate biology course. Wang, Pan, Miller, and Cortina (2014) reported that the Language Environment Analysis system, originally designed for infants and pre-schoolers, can distinguish among lecturing, whole class discussion, and group work in an elementary school math class. Li and Dorai (2006) describe two types of vocal discourse: question-and-answer between instructors and students, and group discussions engaging multiple students.

Owens et al. (2017) developed the Decibel Analysis for Research in Teaching (DART), which analyzes audio recordings from a classroom session to estimate the percentage of the session dedicated to active versus passive learning strategies, based on an algorithm that outputs the number of voices (i.e., Single, Multiple, or None) extracted from the recording. For a given audio file, DART outputs waveform visualizations and percent ratios of Single Voice, Multiple Voices, and No Voice for each class session, each with a possible range of 0-100%, with the assumption that Multiple Voice and No Voice correlate most with active learning components of learner-centered classroom practices (Owens et al., 2017).

The DART instrument represents an exciting tool to potentially address the need for a universally available, low-cost method for practitioners and researchers alike to categorize the learner-centeredness of undergraduate science classrooms. To date, no study has compared DART estimates to other standard measures of learner-centeredness to describe its validity in reference to other reliable metrics. While DART is accessible and easy to use, it is unclear if the data it provides overlap with elements of learner-centered practices that prior instruments also measure. Hence, we sought to explore whether DART could provide accurate measurements of learner-centeredness comparable to other available metrics, thus clarifying the potential of DART to be used by everyday practitioners in the classroom.

External Factors that Contribute to Learner-Centeredness and Classroom Sound

While our first goal was to investigate the alignment of DART with RTOP, we also sought to explore other potential factors that could affect the noise levels of a classroom that may subsequently bias a sound-based metric such as DART. Specifically, we speculated that physical aspects of the classroom itself and the types and background of the people in the classroom may alter both the sound during a class and its learner-centeredness, biasing estimates from DART.

Classroom characteristics

We predicted that numerous physical characteristics of a classroom could affect its learner-centeredness, but these same characteristics also may contribute to noise, unrelated to the quality and frequency of learner-centered activities. For example, some higher education institutions have redesigned their classroom spaces to support active learning (Harvey & Kenyon, 2003) by moving away from a fixed-seat lecture hall (Oblinger, 2006). Despite these redesigns, large classroom sizes, in terms of both enrollment and square footage, still exist and may limit students' motivation to participate in discussions or activities (Abdullah, Bakar, & Mahbob, 2012), minimize support from instructors (Loh Epri, 2016), increase challenges in classroom management (Ayeni & Olowe, 2016) and hinder large-scale active learning activities. Obviously, although greater enrollment of students in large lecture halls may increase background noise, high enrollment classrooms may lead to decreased engagement (Bradley, 2005; Seep Glosemeyer, Hulce, Linn, & Aytar, 2000).

Additionally, movable seating and flexible writing surfaces have been found to support more active learning classroom practices (Lombardi & Wall, 2006; Sanders, 2013). For example, flat seating with movable furniture may be more conducive to learner-centered practices when desks are arranged into small groups for discussion (Park & Choi, 2014). The number of doors and windows in a classroom may also influence student engagement. While some suggest that open doors and windows may act as distractors for students and instructors alike by allowing entry of sound from outside the lecture space (Lei, 2010; Veltri et al., 2006), others emphasize the importance of windows in maintaining a positive and comforting learning environment (Chism, 2006; Montgomery, 2008).

Table 1. Categorical predictors of % Multiple Voice in the classroom.

	Predictor	Counts for Category	How does it contribute to learner-centeredness?	How does it contribute to classroom sound?
Demographic	**Instructor gender**	**Males** *n=5* **Females** *n=4*	Female instructors may encourage increased participation among female students. Student gender may also influence teaching and learning practices in a class.	Female students may be more likely to vocally participate when they have a female instructor. In the absence of a female instructor, only a proportion of the class (i.e., males, who generally have deeper, louder voices) may be speaking rather than all students, contributing to an overall noisier classroom.
Classroom Characteristics	**Chair type**	**Fixed** *n=6* **Non-fixed** *n=5* **Mixed** *n=2*	Physically larger classroom spaces tend to be louder, making it difficult for students to engage in learner-centered practices. However, some classroom attributes such as movable furniture may be more conducive to active learning practices (e.g. group discussions).	Chairs and tables that are non-fixed may be noisier than their fixed counterparts, as students reposition during class. Physically larger classroom spaces with wooden or plastic furniture and stadium seating tend to amplify noise.
	Table type	**Moveable** *n=6* **Fixed** *n=7*		
	Chair material	**Fabric** *n=8* **Plastic** *n=5*		
	Table connectivity	**Individual** *n=7* **Shared** *n=6*		
	Seat arrangement	**Stadium** *n=9* **Flat** *n=4*		

Student and instructor demographics

Beyond the physical characteristics of a classroom, student and instructor demographics may also influence learner-centeredness and classroom noise. Female students are more likely to vocally participate when they have a female instructor (Cornelius-White, 2007; Fassinger, 1996; Pearson & West, 1991); therefore, instructor demographics can influence class engagement. Reciprocally, student gender may influence how students interact with one another and perform (Eddy, Brownell, & Wenderoth, 2014; Eddy, Brownell, Thummaphan, Lan, & Wenderoth, 2015). Male students tend to participate more than their female counterparts and dominate classroom discussions (Howard & Henney, 1998; Pearson & West, 1991), so a class with more male students may be louder than the same-sized class with a lower male tofemale ratio. Further, because first-generation, low socioeconomic status students, and older non-traditional students tend to experience more social and academic challenges than traditional students (Bowl, 2001; Crosnoe & Muller, 2014; Schuetze & Slowey, 2002; Wilbur & Roscigno, 2016), students in these populations may be less inclined to engage in discussions or collaborative in-class activities (Pike & Kuh, 2005).

Research Goals and Questions

To our knowledge, no research has yet explored the relationships between recorded sound in a classroom using DART, other valid metrics of learner-centeredness (i.e., RTOP), physical characteristics of the classroom, or instructor and student demographics. Many studies have characterized learning activities from audio recordings in a classroom setting, yet these have almost exclusively been conducted at the K-12 level and have generally been implemented only in small classes (Donnelly et al., 2016; Donnelly et al., 2017; Wang et al., 2014), excepting the study conducted by Owens et al. (2017). Specifically, there is a need for an accurate, accessible instrument that can be implemented by everyday practitioners in the classroom. Thus, our research questions were: (1) Does a validated

metric of learner-centeredness—the RTOP—predict percent Multiple Voice from DART? (2) Do external variables such as classroom characteristics and demographics of instructors and students predict percent Multiple Voice from DART?

Methods

Ethics Statement

The procedures for this study were approved by the Institutional Review Boards of the Utah Valley University (IRB# 01103) and University of Northern Colorado (IRB #932641-1) . Written informed consent was obtained by all participating faculty and students at the beginning of the study.

Participants, Classrooms, and Variables

We conducted this observational study within a non-majors introductory biology course at a public 4-year university in the western United States. Nine instructors collectively taught thirteen sections of this introductory biology course during Fall 2013 and Spring 2014. Our instructor sample included four females and five males.

The thirteen class sections in our study varied by several factors. We coded instructor gender into two categories (Table 1). De-identified student demographic information was retroactively obtained from the institution's office of institutional research, including gender, first-generation status, age, and Pell Grant eligibility (used as a proxy for students' socioeconomic status), in accordance with our IRB approval. Unfortunately due to considerable missing data, first-generation status and Pell Grant eligibility were not used in our final models. In our analyses, student gender was represented as the proportion of males in a course section; student age was represented by the mean age of students in a course section (Table 2).

Our 13 participating sections were scheduled in 9 locations across the same campus. Classroom characteristics were described by an outside observer or from institutional facilities statistics. The classroom characteristics we captured included square footage, number of doors, number of windows, chair type (fixed, non-fixed, or a combination of these two types), chair material (plastic or fabric), table type (fixed or moveable), table connectivity shared with peers or individual), seat arrangement (stadium or flat seating), and section enrollment (Tables 1 and 2).

Video Recordings

During Fall 2013 and Spring 2014, 42 class sessions were randomly recorded throughout the semester across the 13 course sections. A video recording device was situated on a tripod at the back of the lecture space; each instructor was instructed to secure a wireless lapel microphone and battery pack to their person. The number of class sessions filmed within each course section ranged from three to four. Generally, the instructor was not given advance notice that their lecture would be video-recorded on filming days. These video recordings were used to analyze: (1) audio recordings with the Decibel Analysis for Research in Teaching (DART) instrument and (2) video recordings for the Reformed Teaching Observation Protocol (RTOP).

Decibel Analysis for Research in Teaching (DART)

We converted all video files to .wav audio files compatible with DART using Audacity (Version 2.2.1; Audacity Team, 2017). We also used Audacity to trim each audio file to limit background noise from before/after class and during breaks to ensure that the visualizations and predictions generated by DART were solely based on instructional time. Trimmed audio files were individually uploaded onto the publicly available DART software page (Version 1; sepaldart.herokuapp.com; Science Education Partnership & Assessment Laboratory, San Francisco State University).

In this study, our response variable was percent Multiple Voice predicted by DART for each audio file. For a given audio file, DART outputs waveform visualizations and percent ratios of Single Voice, Multiple Voices, and No Voice for each class session, with the assumption that Multiple Voice and No Voice correlate most with active learning components of learner-centered classroom practices (Owens et al., 2017). The No Voice DART category was not detected in any of our audio recordings, thus our use of Multiple Voice percent alone as a response for learner-centeredness is appropriate.

To ensure the validity of DART, we used human annotation on 17% of the data to measure the accuracy of DART, according to Owens et al. (2017). We annotated the two class session recordings with the highest percent Multiple Voice, the two recordings with the lowest percent Multiple Voice, and three random recordings with varying 'moderate' percent Multiple Voice output from within our sample. These annotations consisted of two trained annotators independently coding the length of time spent lecturing with question-and-answer, silent working, discussing in pairs or small groups, or other activities not represented as a prior code, using codings for human annotation described by Owens et al. (2017). Our inter-rater reliability, the Pearson correlation between the two raters across the seven video recordings, of 0.96 was high; Cohen's alpha was inappropriate because our data were continuous rather than categorical.

Table 2. Continuous predictors of % Multiple Voice in the classroom.

	Predictor	Minimum	Maximum	Mean	How does it contribute to learner-centeredness?	How does it contribute to classroom sound?
Demographic	**Student age**	16	63	22.7	Students in these demographic populations (including older, non-traditional students) may disengage from in-class learning activities more so than other students.	Students in these demographic populations may be less inclined to engage in discussions or collaborative in-class activities and quieter in the classroom overall. Hence, we predicted a greater percentage of non-traditional and female students may contribute to a less noisy classroom due to fewer voices being expressed.
	% Female students in a section	32.7	64.2	48.0		
Classroom Characteristics	**Enrollment**	30	391	94.6	Large classroom sizes may make learning more difficult and active learning practices less effective in larger classroom spaces due to physical constraints of the classroom and a high quantity of students.	High enrollment of students in large lecture halls may increase background noise, contributing to a louder classroom.
	Room size (sq ft)	691.0	5173.0	1966.1		
	Number of doors	1	16	4.5	Increased lighting may positively affect students and increase their willingness to engage in active learning exercises, though many doors and windows in a classroom could also lead to higher potential for distractions.	More doors or windows in a classroom may increase classroom noise if used frequently.
	Number of windows	0	16	4.6		
RTOP	**Mean RTOP scores per section**	30.2	54.4	38.8	Higher RTOP scores indicate greater learner centered practices by students and the instructor.	Higher RTOP scores could indicate both a noisier classroom (e.g., lots of interactive active learning occurring) or quieter classroom (e.g., silent reflective/thinking exercises).
	Mean Classroom Culture scores per section (RTOP subcategory)	9.0	26.5	15.8		

Reformed Teaching Observation Protocol (RTOP)

The RTOP, considered both valid and reliable (Amrein-Beardsley & Popp, 2012; Marshall, Smart, Lotter, & Sirbu, 2011; Piburn & Sawada, 2000; Sawada et al., 2002), allows experts to objectively quantify learner-centeredness in classrooms based on observations. In our study, we had eight trained raters who differed from the DART annotators, and differing combinations of two of these raters individually scored each of the 46 video-recorded class sessions (Generalizability Coefficient = 0.787; see Holt, Young, Keetch, Larsen, & Mollner, 2015) and their scores were averaged for each class session. The RTOP is composed of three scales—lesson design and implementation, content, and classroom culture—from 25 items. Items are scored on a scale from zero (absent) to four (present; Sawada et al., 2002), and scores across all items are then summed to calculate a final RTOP score ranging from 0-100. Thus, a higher RTOP score indicates a more learner-centered classroom. In addition to total RTOP score, we also chose to include the score (ranging from 0-20) from the "Classroom Culture: Student/Teacher Relationships" scale in our models, which

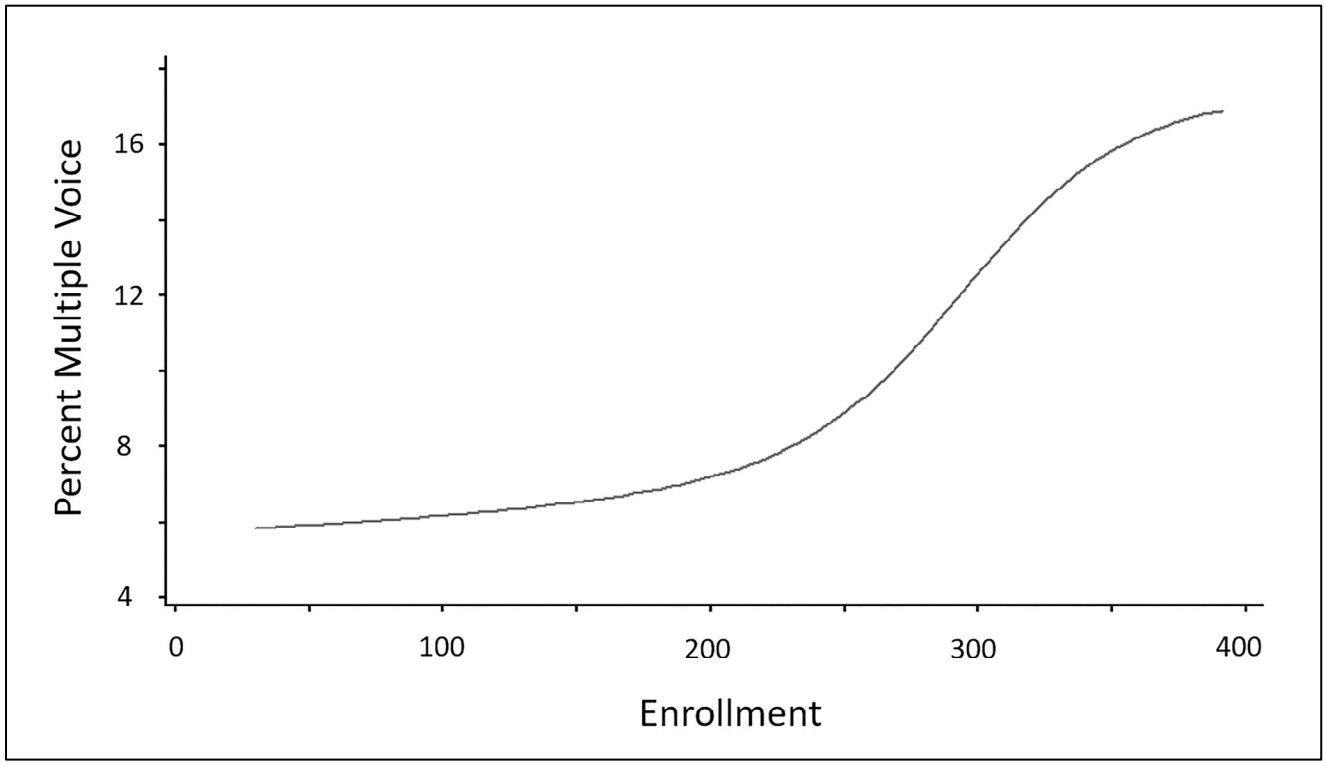

Figure 1. **Two-dimensional fit response curve from NPMR, modelling section enrollment as a single predictor of percent Multiple Voice**. Enrollment was the single variable in the best one-predictor NPMR model.

is a 5-item scale within RTOP focused on student and instructor interactions that we felt might be more relevant for predicting DART due to its potential alignment with learner-centeredness in the classroom.

As multiple video recordings for one course section (i.e., different meetings from the same class) included redundant data for the instructor, students enrolled in the course section, and classroom characteristics, we were cognizant about the inherent pseudoreplication problem within our dataset and sought to minimize its impact. Thus, we ran each of the models described below with a random subset of 13 individual sessions from the 13 class sections; the variance explained by these models changed drastically when an additional predictor variable was included in the model, suggesting that a single-class subset was a poor approach due to the small sample size. All analyses and results below, therefore, represent the full 42 class sections.

Statistical Analyses

Initial analyses included descriptive statistics to describe participants and classroom characteristics, interpret distributions of the data, and assess suitability of potential variables to be included in our models. Bivariate correlations (Pearson correlations for relationships between continuous data) were conducted in SPSS (IBM Corp., 2017) to measure relationships only between significant predictors in our models. We visually inspected scatterplots for the Pearson correlations to ensure that these data were generally linear in nature. The sample units for our data analyses were individual recordings (n = 42 class sessions) rather than course sections. In recognition of pseudoreplication mentioned previously within our data, we included both instructor and section number in our models to better understand how this redundancy affected our findings.

We used nonparametric multiplicative regression (NPMR) modeling to identify potential predictors of percent Multiple Voice in the classroom. NPMR is a flexible method of regression that allows for complex interactions that are not possible to analyze with general linear regression models (Berryman & McCune, 2006). NPMR models predict quantitative response variables using a smoothing function and Gaussian local mean estimators and are assessed with a leave-one-out cross-validated R^2 (xR^2). Further, predictors in NPMR models are considered multiplicatively; thus, multicollinearity is not a concern when running these analyses. Scree plots incorporating xR^2 and predictor variables of interest were used to select a final model. We ran our NPMR models in HyperNiche (2009, Version 2.0, MjM Software) with medium overfitting controls, deleting all but the best predictors in the final models.

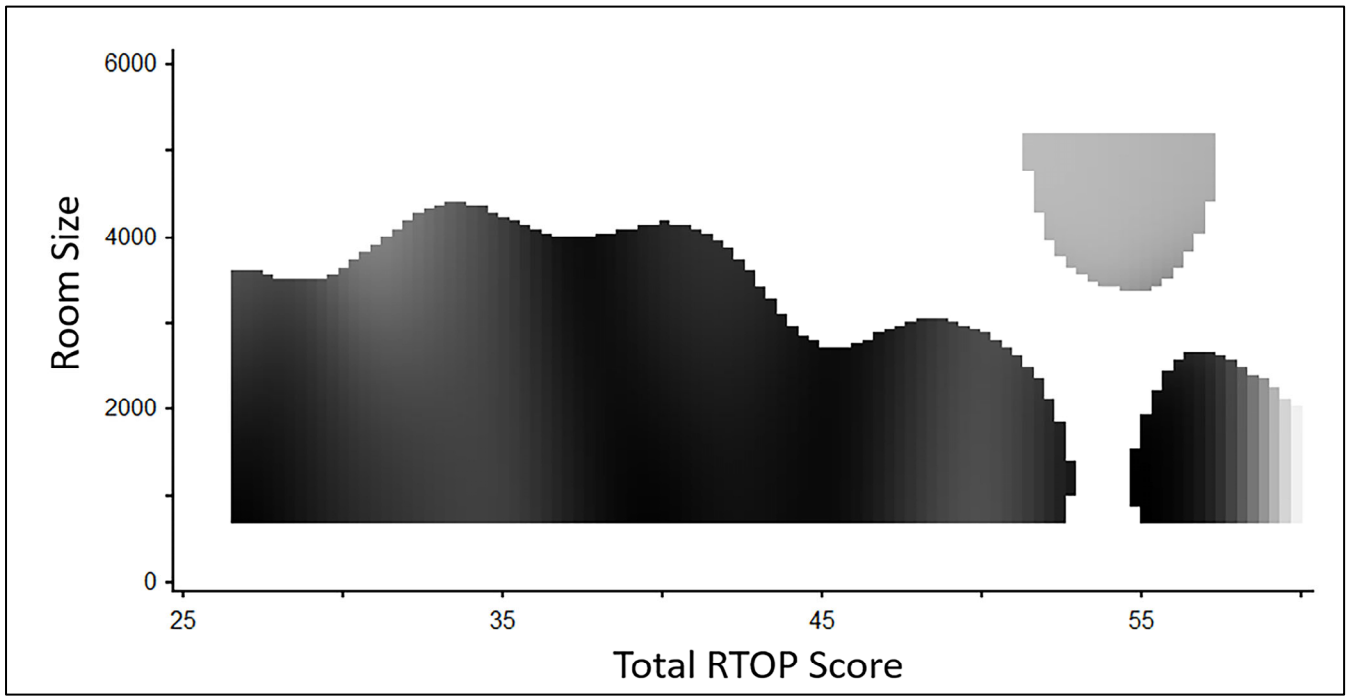

Figure 2**. Three-dimensional contour response curve modelling total RTOP score and room size as the strongest predictors of percent Multiple Voice in a two-predictor model.** Room size (square footage) and RTOP were the two variables in the best two-predictor NMPR model. The lightest colors represent the highest percent Multiple Voice detected by DART, grading into the darkest black that corresponds to the lowest percent Multiple Voice.

We developed NPMR models to predict the average percent Multiple Voice based on 16 possible predictors. Our full predictor set included 3 demographic predictors (e.g., instructor gender, student gender, student age), 9 classroom characteristic predictors (e.g., chair type and material, square footage of classroom, number of doors and windows in class, section enrollment, table type, table connectivity, and seat arrangement), as well as both total RTOP score and Classroom Culture scale score from RTOP. We also included instructor identity number and course section as predictors to detect the effect of pseudoreplication.

Results

Human Annotations of Classroom Activities

Comparing DART output and human annotations of classroom activities (Table 3), the majority of time in each classroom session was spent lecturing (with the exception of Classroom Session 7), yet this value does not specifically align with the percent Single Voice output by DART (other than for Classroom Session 1). Further, even in Classroom Session 7, where nearly 70% of class time was spent in pair or small group discussions as noted by human annotation, the 30.8% Multiple Voice DART output—though the highest value across all recorded sessions—was rather low. However, this inconsistency at the higher end may have been partially due to microphone issues. Various instances of pair/small group discussions observed through annotation of this class were categorized as a single voice by DART, likely because a single voice of the instructor or a student immediately adjacent to the instructor was louder than the overall student oral discourse in the background.

Additionally, other sessions were inversely mismatched, when DART detected moderate levels of percent Multiple Voice annotations consisted primarily of lecture and lecture with question-and-answer (Table 3, Classes 3-6). Misalignment of DART output with our human annotations suggests that many instances of lecture with question-and-answer included background student discussions beyond the individual student or instructor asking or answering questions. Hence, this may have been a weakness in our activity categories for annotation (i.e., some question-and-answer time may be more active than we expected), or may suggest that DART was able to better parse out background noise and side discussions among students during lecture with or without question-and-answer.

Descriptive Analyses

Percent Multiple Voice across our 42 sampled class meetings, as predicted by DART, ranged from 0% to 30.8%, with a mean of 7.14% across all recordings. Percent Single Voice across recordings ranged from 81.86% to 100%, with a

Table 3. Comparison of DART output (in the form of % Multiple Voice) and human annotations of classroom activities. Annotations were conducted on the two class session recordings with the highest percent Multiple Voice, the two recordings with the lowest percent Multiple Voice, and three random recordings with varying 'moderate' percent Multiple Voice output from within our sample. The Pearson correlation for each class session represents the agreeability between the two raters' annotations across each of the five annotation categories listed in the table. The difference in DART and Human Annotation highlights where the two measures agree or not and degree of agreement.

			DART Scoring	Human Annotation Scoring (% of time in class session spent performing each activity)					
				Coded as SV[2]		Coded as NV[2]	Coded as MV[2]	Coding Unkn[2]	Pearson r (inter-rater reliability)
Class Session	Total RTOP Score	Difference in DART and Human Annotation[1]	% Multiple Voice	Lecture without Q&A	Lecture with Q&A	Silent working	Pair/small group discussion	Other	
1	29.5	0	0	93.86	5.88	0.00	0.00	0.26	0.9986
2	43.5	0	0	52.38	31.92	0.00	0.00	15.69	0.7877
3	34	10.5	10.5	68.72	31.23	0.00	0.00	0.05	0.9911
4	33.5	12.2	12.2	85.75	14.25	0.00	0.00	0.00	0.9819
5	32	20.3	20.3	88.03	9.17	0.00	0.00	2.80	0.9994
6	34	22.7	22.7	66.83	30.09	0.00	0.00	3.08	0.9786
7	60	-37.4	30.8	12.94	5.15	0.00	68.20	13.71	0.9967

[1]DART Multiple Voice (MV) % minus the human annotation % of activities assumed to correspond to times with multiple voices. Positive values represent instances where DART overestimates MV, and negative values are when DART underestimates MV.

[2]SV: single voice, NV: no voice, MV: multiple voice, Unkn: human annotations did not fit into categories that obviously align with DART categories, so it is labelled as 'unknown'.

mean of 93.75%. DART did not detect any instances of 'No Voice' in our sample. Across the 42 class recordings, the mean total RTOP score was 38.8 (i.e., teacher-centered lecture with limited demonstrations and student participation). Ranges and means of continuous classroom characteristics and student and instructor demographics are reported in Table 2.

Nonparametric Multiplicative Regression

In our NPMR models, the best predictors of percent Multiple Voice based on DART output were enrollment (i.e., the best one-predictor model; $xR^2 = 0.140$; Figure 1) and total RTOP score and room size (i.e., the best two-predictor model; $xR^2 = 0.2043$; Figure 2). Models with more than two predictors are not further discussed, as additional variables contributed minimally to the cross-validated R^2.

Total RTOP, enrollment, and room size were all significantly correlated, and room size and enrollment were the two most highly correlated predictors in our study ($r = 0.974$), thus effectively representing an equal measure of class size and capacity (Table 4). We found that the highest percentages of Multiple Voice were recorded in both: a) small classrooms taught by instructors with our highest values of total RTOP scores; and b) large classrooms taught by instructors with moderate to high total RTOP scores (Figure 2). The best one-predictor model, where we forced the single predictor to be total RTOP score, explained little variance in Multiple Voice ($xR^2 = 0.0234$).

Discussion

DART's Misalignment

There could be multiple reasons why DART did not align well with an established measure of learner-centeredness (i.e., RTOP), and often underestimated the level of learner-centeredness for instructors scoring higher on the RTOP in our sample. Perhaps the singular focus of DART on sound within a classroom versus the more integrated focus of RTOP on both audio and visual observations within a classroom, caused misalignment in the output between these two instruments. Potentially DART captures different aspects of learner-centeredness than measured by RTOP, a phenomenon reported elsewhere for other instruments (Heim & Holt, 2018). Owens et al. (2018) even suggest that while DART may be a good indicator of general learner-centeredness, future work could investigate alignment of

DART with other observations rubrics (e.g., Smith et al., 2013; Durham et al., 2018).

As there is a need for instruments that accurately gauge learner-centeredness of classrooms—which can easily be implemented by the "common educator"—and a need for undergraduate biology classrooms to be more active (Woodin, Carter, & Fletcher, 2010), observation protocols may provide benefits over other learner-centered instruments in that they utilize a more objective vantage point to both quantify learner-centered instruction and provide meaningful feedback to practitioners (Amrein-Beardsley & Popp, 2012; Durham et al., 2018; Eddy, Converse, & Wenderoth., 2015; Heim & Holt, 2018; Pratt & Collins, 2000; Sawada et al., 2002; Smith et al., 2013). While we initially expected DART would provide an effective and novel solution to the problem of practitioners' need for an accurate, off-the-shelf measure of learner-centeredness, this was not the case in our study. Calibration activities could have potentially improved the accuracy of DART (K. Tanner, pers. comm.); however, best practices and research on necessary calibration tasks are not widely available, further complicating the accessibility of DART for practitioners.

Additionally, the use of lapel microphones by the instructors in our study may have interfered with how effectively student discussion in the classroom was detected by the audio recording devices, and represent a limitation of our study. If the microphones were mainly recording the instructor's voice because of their proximity to the instructor, this may explain why variance in percent Multiple Voice was fairly low (min = 0%, max = 30.8%). While this low variance was a limitation in our study, it also suggests a possible limitation in using DART among practitioners. Others have also found that to accurately capture students' voices in a classroom, multiple audio recording devices need to be set up throughout the room as to avoid singly capturing the instructor's voice simply due to proximity (Su, Dzodzo, Wu, Liu, & Meng, 2019). The positioning of audio recording devices in the classroom appears to be important for DART to collect sound accurately, yet further work is needed to clarify the optimal type of recording device and/or the placement of that device for everyday use by practitioners.

Big, Large Enrollment Classes Confounds DART's Signal

We found that as enrollment increased, as the single best predictor, so did percent Multiple Voice categorized by DART (Figure 1). Ultimately, more students in a classroom lead to more noise, whether from discourse related to course content or more individuals moving about the classroom; however, this finding also suggests that DART may be biased in detecting learner-centeredness across classes of variable enrollments. Our best two-predictor model including room size (Figure 2) further suggests that these large classes may bias DART's estimation of learner-centeredness, particularly since physically larger classroom spaces often amplify noise (Bradley, 2005; Seep Glosemeyer, Hulce, Linn, & Aytar, 2000). While large enrollment classes can offer learner-centered environments (Knight, Wise, & Southard, 2013; Zagallo, Meddleton, & Bolger, 2016), it is unclear if DART can untangle these two sources of sound.

Table 4. Pearson correlations between continuous variables in our models.

	% Multiple Voice (DART)	Total RTOP score	Room size (sq ft)	Enrollment
Total RTOP score	0.315*			
Room size (sq ft)	0.440**	0.381*		
Enrollment	0.460**	0.390*	0.974**	

Encouragingly, the contribution of RTOP in our best two-predictor model was a near 50% increase over the variance explained in the one-predictor model by enrollment alone. While the overall variance explained by these two predictors was low, the addition of RTOP as a secondary predictor and its interaction with enrollment indicates that DART's prediction of learner-centeredness, at least minorly, aligns with another objective measure of learner-centeredness. Unfortunately, total RTOP score alone was not a good predictor of percent Multiple Voice ($xR^2 = 0.0234$). Although total RTOP scores had moderately low variance in our dataset, we argue that there was sufficient variance for our study (coefficient of variance = 24.67) to detect differences. While Bernstein (2018) suggests that DART could be a helpful tool in quantifying active learning in a classroom if further validated, many have found that observation protocols continue to provide the most accurate measurements of learner-centeredness in classroom (Amrein-Beardsley & Popp, 2012; Durham et al., 2018; Eddy, Converse, & Wenderoth., 2015; Heim & Holt, 2018; Pratt & Collins, 2000; Sawada et al., 2002; Smith et al., 2013). Overall DART's minor and interactive role in predicting learner-centeredness, and its misalignment with hand annotations in our study weakens hope that it could be the panacea tool for practitioners.

Many Classroom Characteristics May Not Interfere with DART's Signal

We included classroom characteristics in our models because we felt that some of these factors may unnecessarily distract from a signal of learner-centeredness. While enrollment and room size are clearly confounding factors with using DART, no other physical attributes of a classroom nor demographic factors were selected in the best models, which suggests that they were not contributing as much to classroom noise as we originally predicted.

Limitations of our Sample

We were mindful of pseudo replication in our study, but neither instructor nor section identifiers were top predictors, thus this inherent redundancy was clearly not driving the overarching patterns we noticed in our models. Nine instructors teaching thirteen course sections were included in our sample to ensure consistency in course content being covered; however, greater variance in the classroom characteristic and demographic predictors, which could potentially be attained by increasing the number of course sections, instructors, and students sampled, could improve the fit of the models and allow us to measure which variables were most predictive of percent Multiple Voice with greater accuracy.

Conclusions

We found that enrollment was the best single predictor of percent Multiple Voice, and that total RTOP score and room size also predicted percent Multiple Voice when combined multiplicatively with one another, albeit weakly. Specifically in regard to our research questions, we found that (1) DART did not align well with an established measure of learner-centeredness (i.e., RTOP) and often underestimated the level of learner-centeredness for instructors scoring higher on the RTOP in our sample, and that (2) only certain external variables (i.e., enrollment and room size) predicted DART output. We suggest that additional research is needed to clarify the types and positioning of audio recording devices necessary for effective DART analysis. Finally, RTOP and DART may be measuring distinct aspects of learner-centeredness, so the inclusion of other measures of learner-centeredness will be important to employ in future iterations of this research to determine whether DART is generally aligned with other instruments of learner-centeredness.

Acknowledgements

We would like to thank Kimberly Tanner for a friendly review of our manuscript in its earlier stages. Further, we acknowledge the undergraduate researchers who assisted in data collection, as well as all participating instructors and students. This study was not funded by any grants.

References

Abdullah, M. Y., Bakar, N. R. A., & Mahbob, M. H. (2012). Student's Participation in Classroom: What Motivates them to Speak up? *Procedia-Social and Behavioral Sciences*, *51*, 516-522.

Amrein-Beardsley, A., & Popp, S. E. O. (2012). Peer observations among faculty in a college of education: investigating the summative and formative uses of the Reformed Teaching Observation Protocol (RTOP). *Journal of Personnel Evaluation in Education*, *24*(1), 5-24.

Armbruster, P., Patel, M., Johnson, E., & Weiss, M. (2009). Active learning and student-centered pedagogy improve student attitudes and performance in introductory biology. *CBE-Life Sciences Education*, *8*(3), 203-213.

Audacity Team (2017). Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 2.2.1. Retrieved December 20, 2017, from: https://audacityteam.org/.

Ayeni, O. G., & Olowe, M. O. (2016). The Implication of Large Class Size in the Teaching and Learning of Business Education in Tertiary Institution in Ekiti State. *Journal of Education and Practice*, *7*(34), 65-69.

Bernstein, D. A. (2018). Does active learning work? A good question, but not the right one. *Scholarship of Teaching and Learning in Psychology*, *4*(4), 290.

Berryman, S., & McCune, B. (2006). Estimating epiphytic microlichen biomass from topography, stand structure and lichen community data. *Journal of Vegetation Science*, *17*(2), 157-170.

Biggs, J., Kember, D., & Leung, D. Y. (2001). The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology*, *71*(1), 133-149.

Bradley, J. S. (2005). Does the classroom assist or impede the learning process? *Canadian Association of Principals Journal*, *13*, 32-34.

Bowl, M. (2001). Experiencing the barriers: Non-traditional students entering higher education. *Research Papers in Education*, *16*(2), 141-160.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (eds.) and Committee on Developments in the Science of Learning, National Research Council (1999). *How People Learn: Brain, Mind, Experience, and School.* Washington, DC: National Academies Press.

Chism, N. V. N. (2006). Challenging traditional assumptions and rethinking learning spaces. *Learning Spaces*, 2-1.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*(6), 378-387.

Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of Educational Research*, *77*(1), 113-143.

Crosnoe, R., & Muller, C. (2014). Family socioeconomic status, peers, and the path to college. *Social Problems*, *61*(4), 602-624.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics*, *69*(9), 970-977.

Davis, Z. T. (1987). The Effect of Time-of-Day of Instruction on Eighth-Grade Students' English and Mathematics Achievement. *The High School Journal*, *71*(2), 78-80.

Donnelly, P. J., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., ... & D'Mello, S. K. (2016, July). Automatic teacher modeling from live classroom audio. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization* (pp. 45-53). ACM.

Donnelly, P. J., Blanchard, N., Olney, A. M., Kelly, S., Nystrand, M., & D'Mello, S. K. (2017, March). Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 218-227). ACM.

Durham, M. F., Knight, J. K., Bremers, E. K., DeFreece, J. D., Paine, A. R., & Couch, B. A. (2018). Student, instructor, and observer agreement regarding frequencies of scientific teaching practices using the Measurement Instrument for Scientific Teaching-Observable (MISTO). *International Journal of STEM Education*, *5*(1), 31.

Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of Research in Education*, *32*(1), 268-291.

Ebert-May, D., Derting, T. L., Henkel, T. P., Middlemis Maher, J., Momsen, J. L., Arnold, B., & Passmore, H. A. (2015). Breaking the cycle: Future faculty begin teaching with learner-centered strategies after professional development. *CBE—Life Sciences Education*, *14*(2), ar22.

Ebert-May, D., Derting, T.L., Hodder, J., Momsen, J.L., Long, T.M., & Jardeleza, S.E. (2011). What we say is not what we do: effective evaluation of faculty professional development programs. *BioScience, 61*(7), 550-558.

Eddy, S. L., Brownell, S. E., & Wenderoth, M. P. (2014). Gender gaps in achievement and participation in multiple introductory biology classrooms. *CBE—Life Sciences Education*, *13*(3), 478-492.

Eddy, S. L., Brownell, S. E., Thummaphan, P., Lan, M. C., & Wenderoth, M. P. (2015). Caution, student experience may vary: social identities impact a student's experience in peer discussions. *CBE—Life Sciences Education*, *14*(4), ar45.

Eddy, S. L., Converse, M., & Wenderoth, M. P. (2015). PORTAAL: a classroom observation tool assessing evidence-based teaching practices for active learning in large science, technology, engineering, and mathematics classes. *CBE—Life Sciences Education*, *14*(2), ar23.

Entwistle, N., McCune, V., & Hounsell, J. (2002). Approaches to studying and perceptions of university teaching-learning environments: Concepts, measures and preliminary findings. *Occasional Report*, *1*.

Fahraeus, A. (2013). Research Supports Learner-Centered Teaching. *Journal of the Scholarship of Teaching and Learning, 13*(4), 126-131.

Falconer, K., Joshua, M., Wyckoff, S., & Sawada, D. (2001, March). Effect of reformed courses in physics and physical science on student conceptual understanding. Paper presented at the annual meeting of the National Association of Research in Science Teaching, St. Louis, MO.

Fassinger, P. A. (1996). Professors' and students' perceptions of why students participate in class. *Teaching Sociology*, 25-33.

Freeman, S., Eddy, S.L., McDonough, M., Smith, M.K., Okoroafor, N., Jordt, H., & Wenderoth, M.P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410-8415.

Goldstein, D., Hahn, C. S., Hasher, L., Wiprzycka, U. J., & Zelazo, P. D. (2007). Time of day, intellectual performance, and behavioral problems in morning versus evening type adolescents: Is there a synchrony effect? *Personality and Individual Differences*, *42*(3), 431-440.

Gormally, C., Brickman, P., Hallar, B., & Armstrong, N. (2011). Lessons Learned About Implementing an Inquiry-Based Curriculum in a College Biology Laboratory Classroom. *Journal of College Science Teaching*, *40*(3).

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*(1), 64-74.

Harvey, E. J., & Kenyon, M. C. (2013). Classroom seating considerations for 21st century students and faculty. *Journal of Learning Spaces*, *2*(1).

Heim, A. B., & Holt, E. A. (2018). Comparing student, instructor, and expert perceptions of learner-centeredness in post-secondary biology classrooms. PLoS ONE, 13(7): e0200524. https://doi.org/10.1371/journal.pone.0200524

Holloway, J. (1999). Giving our students the time of day. *Educational Leadership*, *57*(1), 87-88.

Holt, E. A., Young, C., Keetch, J., Larsen, S., & Mollner, B. (2015). The Greatest Learning Return on Your Pedagogical Investment: Alignment, Assessment or In-Class Instruction? PLoS ONE, 10(9), e0137446.

Howard, J. R., & Henney, A. L. (1998). Student participation and instructor gender in the mixed-age college classroom. *The Journal of Higher Education*, *69*(4), 384-405.

HyperNiche. (2009). Nonparametric Multiplicative Habitat Modeling. Version 2.0. MjM Software, Gleneden Beach, Oregon, USA.

IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.

Kahl Jr, D. H., & Venette, S. (2010). To lecture or let go: A comparative analysis of student speech outlines from teacher-centered and learner-centered classrooms. *Communication Teacher*, *24*(3), 178-186.

Kilday, C. R., & Kinzie, M. B. (2009). An analysis of instruments that measure the quality of mathematics teaching in early childhood. *Early Childhood Education Journal*, *36*(4), 365-372.

Klein, J. (2001). Attention, scholastic achievement and timing of lessons. *Scandinavian Journal of Educational Research*, *45*(3), 301-309.

Knight, J. K., Wise, S. B., & Southard, K. M. (2013). Understanding clicker discussions: student reasoning and the impact of instructional cues. *CBE—Life Sciences Education*, *12*(4), 645-654.

Knight, J. K., & Wood, W. B. (2005). Teaching more by lecturing less. *Cell Biology Education*, *4*(4), 298-310.

Kranzfelder, P., Bankers-Fulbright, J. L., García-Ojeda, M. E., Melloy, M., Mohammed, S., & Warfa, A. R. M. (2019). The Classroom Discourse Observation Protocol (CDOP): A quantitative method for characterizing teacher discourse moves in undergraduate STEM learning environments. *PLoS ONE*, *14*(7), e0219019.

Lei, S. A. (2010). Classroom physical design influencing student learning and evaluations of college instructors: A review of literature. *Education*, *131*(1), 128-135.

Lemke, J. L. (1990). *Talking science: Language, learning, and values*. Norwood, NJ: Ablex Publishing Corporation.

Li, Y., & Dorai, C. (2006). Instructional video content analysis using audio information. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(6), 2264-2274.

Loh Epri, M. (2016). A case study on the impact of large classes on student learning. *Contemporary PNG Studies*, *24*, 95.

Lombardi, M. M., & Wall, T. B. (2006). *Learning Spaces*, 17-1. Duke University: Perkins Library.

MacIsaac, D., & Falconer, K. (2002). Reforming physics instruction via RTOP. *The Physics Teacher*, *40*(8), 479-485.

Marshall, J. C., Smart, J., Lotter, C., & Sirbu, C. (2011). Comparative Analysis of Two Inquiry Observational Protocols: Striving to Better Understand the Quality of Teacher-Facilitated Inquiry-Based Instruction. *School Science and Mathematics*, *111*(6), 306-315.

McCombs, B. (2003, April). *Defining tools for teacher reflection: The assessment of learner-centered practices*. Paper presented at the Annual Meeting of the American Educational Research Association.

Millar, K., Styles, B. C., & Wastell, D. G. (1980). Time of day and retrieval from long-term memory. *British Journal of Psychology*, *71*(3), 407-414.

Montgomery, T. (2008). Space matters: Experiences of managing static formal learning spaces. *Active Learning in Higher Education*, *9*(2), 122-138.

Oblinger, D. G. (2006). Space as a change agent. *Learning Spaces*, *1*.

O'Connor, C., Michaels, S., Chapin, S., & Harbaugh, A. G. (2017). The silent and the vocal: Participation and learning in whole-class discussion. *Learning and Instruction*, *48*, 5-13.

Onyper, S. V., Thacher, P. V., Gilbert, J. W., & Gradess, S. G. (2012). Class start times, sleep, and academic performance in college: A path analysis. *Chronobiology International*, *29*(3), 318-335.

Owens, M. T., Seidel, S. B., Wong, M., Bejines, T. E., Lietz, S., Perez, J. R., ... & Balukjian, B. (2017). Classroom sound can be used to classify teaching practices in college science courses. *Proceedings of the National Academy of Sciences*, https://doi.org/10.1073/pnas.1618693114

Owens, M. T., Trujillo, G., Seidel, S. B., Harrison, C. D., Farrar, K. M., Benton, H. P., ... & Byrd, D. T. (2018). Collectively improving our teaching: attempting biology department–wide professional development in scientific teaching. *CBE—Life Sciences Education*, *17*(1), ar2.

Park, E. L., & Choi, B. K. (2014). Transformation of classroom spaces: Traditional versus active learning classroom in colleges. *Higher Education*, *68*(5), 749-771.

Pearson, J. C., & West, R. (1991). An initial investigation of the effects of gender on student questions in the classroom: Developing a descriptive base. *Communication Education*, *40*(1), 22-32.

Piburn, M., & Sawada, D. (2000). Reformed Teaching Observation Protocol (RTOP) Reference Manual. Technical Report.

Pike, G. R., & Kuh, G. D. (2005). First-and second-generation college students: A comparison of their engagement and intellectual development. *The Journal of Higher Education*, *76*(3), 276-300.

Pratt, D. D., & Collins, J. B. (2000). The Teaching Perspectives Inventory (TPI). Paper presented at the *Adult Education Research Conference*, Vancouver, BC.

Rands, M. L., & Gansemer-Topf, A. (2017). 'The room itself is active': How classroom design impacts student engagement. *Journal of Learning Spaces*, *6*(1).

Richardson Jr, R. C., & Skinner, E. F. (1992). Helping first-generation minority students achieve degrees. *New Directions for Community Colleges*, *1992*(80), 29-43.

Rushton, G. T., Lotter, C., & Singer, J. (2011). Chemistry teachers' emerging expertise in inquiry teaching: the effect of a professional development model on beliefs and practice. *Journal of Science Teacher Education*, *22*(1), 23-52.

Sanders, M. (2013). Classroom design and student engagement. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2013 (57)*, 496.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics, 102*(6), 245-253.

Schuetze, H. G., & Slowey, M. (2002). Participation and exclusion: A comparative analysis of non-traditional students and lifelong learners in higher education. *Higher Education*, *44*(3-4), 309-327.

Seep, B., Glosemeyer, R., Hulce, E., Linn, M., & Aytar, P. (2000). Classroom Acoustics: A Resource for Creating Environments with Desirable Listening Conditions. https://files.eric.ed.gov/fulltext/ED451697.pdf

Shavelson, R. J., Webb, N. M. & Burstein, L. (1986). Measurement of Teaching. In M. C. Wittrock (Ed.), *Handbook of Research on Teaching*, (pp. 50–91). New York, NY: Macmillan.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4-14.

Singer, J., Lotter, C., Feller, R., & Gates, H. (2011). Exploring a model of situated professional development: Impact on classroom practice. *Journal of Science Teacher Education*, *22*(3), 203-227.

Smith, M. K., Jones, F. H., Gilbert, S. L., & Wieman, C. E. (2013). The Classroom Observation Protocol for Undergraduate STEM (COPUS): a new instrument to characterize university STEM classroom practices. *CBE-Life Sciences Education*, *12*(4), 618-627.

Su, H., Dzodzo, B., Wu, X., Liu, X., & Meng, H. (2019). Unsupervised Methods for Audio Classification from Lecture Discussion Recordings. *Proceedings of Interspeech 2019*, 3347-3351.

Tarr, J. E., Reys, R. E., Reys, B. J., Chavez, O., Shih, J., & Osterlind, S. J. (2008). The impact of middle-grades mathematics curricula and the classroom learning environment on student achievement. *Journal for Research in Mathematics Education*, 247-280.

Trigwell, K., & Prosser, M. (2004). Development and use of the approaches to teaching inventory. *Educational Psychology Review, 16*(4), 409-424.

Veltri, S., Banning, J. H., & Davies, T. G. (2006). The community college classroom environment: Student perceptions. *College Student Journal*, *40*(3), 517-528.

Walker, J. D., Cotner, S. H., Baepler, P. M., & Decker, M. D. (2008). A delicate balance: integrating active learning into a large lecture course. *CBE-Life Sciences Education*, *7*(4), 361-367.

Wang, Z., Pan, X., Miller, K. F., & Cortina, K. S. (2014). Automatic classification of activities in classroom discourse. *Computers & Education*, *78*, 115-123.

Wilbur, T. G., & Roscigno, V. J. (2016). First-generation disadvantage and college enrollment/completion. *Socius*, *2*, https://doi.org/10.1177/2378023116664351

Woodin, T., Carter, V. C., & Fletcher, L. (2010). Vision and change in biology undergraduate education, a call for action—initial responses. *CBE—Life Sciences Education*, *9*(2), 71-73.

Zagallo, P., Meddleton, S., & Bolger, M. S. (2016). Teaching real data interpretation with models (TRIM): Analysis of student dialogue in a large-enrollment cell and developmental biology course. *CBE—Life Sciences Education*, *15*(2), ar17.