Large-enrollment STEM Learning Environments: Cultural and experimental perspectives

Benjamin L Wiggins

A dissertation
submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington
2015

Reading Committee:
Philip Bell, Chair

Megan Bang

Kenneth Zeichner

Scott Freeman

Program Authorized to Offer Degree:
College of Education
Education: Learning Sciences

University of Washington

Abstract

Large-enrollment STEM Learning Environments: Cultural and experimental perspectives

Benjamin L Wiggins

Chair of the Supervisory Committee: Dr. Philip Bell (UW Learning Sciences & Human Development)

Post-secondary STEM education forces students through an experiential bottleneck in the form of large-enrollment lecture courses. At colleges and universities around the world, talented students who have already passed through several academic filtering processes are challenged with coursework in environments that are impersonal, intimidatingly large, and typically taught using traditional passive lecture techniques. This is no way to educate the next generation of scientists, researchers, doctors, and innovators. Calls for improvement are frequent and loud, but the research on improvement in this unique arena is seldom conclusive useful for practitioners. This dissertation focuses on the large-enrollment STEM classroom as an opportunity for inspection and change.

This dissertation seeks a wide-spectrum view across several key issues. Comprised of peer-reviewed articles, this work will examine the large-enrollment STEM classroom from four angles. First, an analysis of interactions between different types of teaching assistants will inform the student-instructor relationship and speak to important attributes in instructor teams. Secondly, the methods for modeling large education environments through social network analysis are developed. Third, a qualitative deep dive into student experiences is used to develop a survey instrument focused on student engagement with active teaching practices in a large-enrollment classroom; these mixed methods triangulate similar conclusions about the important factors for students. Lastly, predictions at the apex of the ICAP framework for active learning

practices developed by Chi and Wylie (2014) are validated through an experimental use of split large-lecture courses as a model system.

This dissertation is intentionally broad. As this learning environment is yet poorly understood, the author intends to address change in future work through a wider scope on multiple avenues for positive change. While this does not address any particular issues in the depth of multiple peer reviewed papers, it is agreed by the committee that a group-authored and collaborative look at multiple issues is a useful education for potential needs of post-secondary STEM education overall.

Dear Thesis Committee,

This memo is intended to frame and give context for my dissertation. Especially given the unusual nature of this dissertation when compared to the standard single-paper publication format. We decided at my general exam that it would be a good idea to write out this short overall description.

My intention through this doctoral work is to position myself as a well-rounded consultant for issues and practices related to large-enrollment STEM lecture courses. I foresee opportunities in research and professional development around this unique learning environment: possible jobs include assessor on research grants, consultant for instructional staff, advisor to STEM university administrators or acting as an administrator for large STEM departments. The last of these is a good description of my current job in which I will remain for the near future. These jobs require a broad understanding of the research literature that might influence teaching and learning in large STEM classes. This broad education is likely to be more useful than a deep and singular experience in a particular research motif; this is why my dissertation is a collection of four papers rather than a single narrative.

As of my general exam, our plan was for me to pursue publication of four papers as a stretch goal, with a more realistic goal of 2-3 papers in progress and a first-author credit on at least one. Paper #3 has been submitted and Paper #4 was pre-approved. Overall, I will end up with four papers published (of which I have first author status on three). Below I will describe those four papers in brief.

The first paper is an exploration of teaching roles and identities using a survey data set from several large biology courses. This is relatively early work for me, and was accepted for publication in 2013. Here I'm exploring the relationships of STEM undergraduates and their near-peer instructors with whom they have the most direct contact (either graduate or undergraduate teaching assistants). Undergraduate success is not statistically influenced by the identity of the TA, indicating that different types of teaching assistants may have the ability to do similarly effective teaching work. Undergraduate students do note several aspects of near-peers that may assist in their learning even in spite of the perceived gap in their own scientific education compared to graduate students.

The second paper, published in 2014, is a methods paper covering the use of social network analysis in education environments. This came from a collaboration and research project which was unique at the time: using longitudinal social network surveys to visualize the cultural progression of student lives both in formal classrooms and the informal spaces surrounding those STEM courses. This paper lays out how to use ERGMs to model and test hypotheses. We develop a new (though simple) method for collecting data and building on that student data with existing institutional statistics to build a network model. I'm happy to note that it has already been cited 10 times. At least two more papers are in production from our groups using this method (neither of which are included in this dissertation).

The third paper is the most complex. This work, which is in submission now, is a mixed methods analysis of introductory biology courses. I spearheaded a qualitative deep dive into student experiences in active learning. This large data set (27 interview subjects on transcripts of 45 minutes or more each) was progressively coded into emergent themes. These themes were used as the basis to develop a survey that was iteratively validated for use with these students. Why is this paper valuable? Coding developed a three-part structure to define student experiences. That three-part structure was closely matched independently by principle

component statistical analysis of the final survey. This is a strong triangulation of mixed methods. This kind of matched conclusion should be able to convince both STEM researchers (who traditionally downplay the relevant of qualitatively-derived conclusions) and Education researchers (who are skeptical of the real-world significance of p-values and other publishable statistics). Both sides of the interdisciplinary line I straddle can find support for student experiences being based in the Value of Group Work, the Instructor Effect on the Experience, and in Personal Effort and Motivation. This was the work in which I was most able to develop and practice a broad range of education research skills. This paper has been reformatted since I sent a previous draft to you in July, but the conceptual takeaways and conclusions remain the same.

The last of four papers is in draft form, though it has already been pre-approved and requested by journal editors after submission of the abstract (note that this does not guarantee acceptance). This is experimental work. The experimental design was long and involved, but the resulting paper is actually quite short. This paper validates an important prediction in Chi and Wylie's ICAP framework for active learning (2014). We used a carefully controlled experimental system to demonstrate that interactive learning results in improved outcomes when compared to constructive learning. This isn't an earth-shattering result, but the confirmation of this prediction had not been previously made in a real classroom (let alone a classroom of 350+ students). The implications and discussion are interesting, because this result naturally must be balanced with the other limited resources that instructors have.

As discussed, this work is highly collaborative. This allows the work to take advantage of skills that I do not have and do not wish to pursue such as regression modeling. I expect my future research work to be similarly collaborative, so this is good practice for me. However, it does mean that I did not do this work alone. In an attempt to be as transparent as possible, I have written (below) an outline of all of the roles of my collaborators for all four papers included in this dissertation. I hope this outline gives you what you need to make the determination that I have done enough work to be considered for a PhD, and I am happy to expand or clarify any parts herein.

My future plans are fairly straightforward. I will continue in my role as the Manager of Instruction for the UW Biology Department. I leverage my research and education backgrounds for the support of the 60+ faculty and many staff and teaching assistants who teach here. I will continue as Lecturer for large molecular biology introductory courses twice per year. I will also start teaching a new upper-division lab class in Autumn 2016. Assuming I pass my dissertation, I am likely to go through a promotion process hopefully resulting in moving to the title of Senior Lecturer (for which I am currently missing only the actual PhD as a qualification). I am involved with a group submitting an NSF IUSE grant through WWU for which I'll work roughly three weeks per year. Most importantly, from a research standpoint, I will continue to have access to a uniquely controllable dual-section large-enrollment Biology course. This is where my own research will likely center in future years.

Thank you for taking the time to read and for your feedback. I am most interested in your feedback on Paper #4, as it is the one paper yet unsubmitted. Feedback in any format is appreciated, so please use whatever mode is easiest for you.

Sincerely,
Ben Wiggins

Description of Roles Throughout Ben Wiggins' Dissertation

Paper 1:

- Wiggins' Role: co-first author with Dr. Chapin, developed project, data collection, data analysis, wrote education sections
- Other Roles:
    - Dr. Hannah Chapin (UW Biochemistry): co-first author with Wiggins, wrote literature review and primary driver on manuscript completion
    - Dr. Linda Martin-Morris (UW Biology): last and corresponding author, main editor

Paper 2:

- Wiggins' Role: second author, data development and collection, education-focused parts primarily in introduction and methods
- Other Roles:
    - Daniel Grunspan (UW Anthropology): primary driver on project, first author, primary writer
    - Dr. Steven Goodreau (UW Statistics): last and corresponding author, wrote data analysis pieces and oversaw editing

Paper 3:

- Wiggins' Role: first and corresponding author, primary driver on paper, co-PI on original grant, developed project, large parts of data collection, qualitative data analysis, majority of survey development
- Other Roles:
    - Dr. Alison Crowe (UW Biology): last author, co-PI on original grant, survey development and validation, co-writing of introduction, oversaw editing
    - Dr. Sarah Eddy (UT-Austin Biology Education): statistical analysis including core work on principal component analysis and wrote statistical portions of paper, development of group clustering items in previous qualitative work, observation data protocol development and data collection
    - Dr. Sara Brownell (Arizona State Biology Education): development of group clustering items in previous qualitative work, helped with survey development and validation
    - Leah Wener-Fligner (UW Islandwood): qualitative data analysis and interviewing (as undergraduate worker on original grant)
    - Dr. Jerry Timbrook (Dayton Psychometrics): item validation, wrote face validation portion of paper
    - Dan Grunspan (UW Anthropology): survey development, editing
    - Dr. Karen Freisem (UW Center for Teaching and Learning): outside reviewer for grant, assistance with observation protocols and observation data collection

Paper 4:

- Wiggins' Role: first author, co-PI on original grant, developed project, primary activity developer, majority of data collection
- Other Roles:

- Dr. Alison Crowe (UW Biology): last author, co-PI on original grant, survey development and validation, oversaw editing
- Dr. Amanda Schivell (UW Biology): activity development
- Dr. Sarah Eddy (UT-Austin Biology Education): statistical analysis including core work on principal component analysis, observation data protocol development and data collection
- Dan Grunspan (UW Anthropology): data collection and activity editing
- Dr. Karen Freisem (UW Center for Teaching and Learning): outside reviewer for grant, assistance with observation protocols and observation data collection

Undergraduate Science Learners Show Comparable Outcomes Whether Taught by Undergraduate or Graduate Teaching Assistants

By Hannah C. Chapin, Benjamin L. Wiggins, and Linda E. Martin-Morris

Peer educators can be a powerful addition to classroom learning environments. Traditionally, the university science teaching model relies on graduate teaching assistants (GTAs) to provide instruction in laboratory class sessions, but there is increasing evidence that undergraduate TAs (UTAs) can fill an equivalent role. A comparison of student performance in a series of two introductory biology classes and one third-year class shows that students with GTA and UTA leaders earn comparable final course grades. Additionally, both UTAs and GTAs are considered effective at encouraging a positive attitude toward science and fostering a positive laboratory environment, though the UTAs receive slightly higher scores on two of the assessments of attitude toward science. These results demonstrate that equally well-trained UTAs and GTAs have equivalently positive impacts on laboratory learners. Without diminishing the value of teaching by and for the GTA, this data demonstrates the comparable value of UTA laboratory educators.

There is no categorical definition of who can be a teacher within the university educational system; the position can be filled by a professor, instructor, graduate student, or undergraduate peer. The position of teaching assistant (TA) traditionally carries some authority and respect in the undergraduate science classroom, and the person holding that position is often in charge of weekly laboratory sessions that coordinate with lecture classroom education. TAs are frequently pulled from the ranks of graduate students in the given department. As discussed in this article, there is substantial evidence that undergraduates can just as effectively fill the role of TA. Despite this evidence, there has been no direct comparison of the effect that graduate and undergraduate TAs have on their students' course grades or the laboratory section environment. This study seeks to fill this gap and finds that the two TA groups are functionally equivalent, as quantified by our measures. This equivalence suggests that undergraduates can reap the benefits of filling TA positions while not negatively affecting the experience of students in the class.

Many undergraduate science classes are structured to contain both lecture and laboratory components. The entire class meets together several times a week for lecture-based learning, and then students are broken into smaller groups that meet for one extended session of laboratory-based learning per week. In large classes, it is impractical for a single instructor to teach both the lecture and laboratory sections, so a common solution is to have graduate students lead the laboratory sections. The primary role of these graduate teaching assistants (GTAs) is to help students learn the scientific material. A secondary benefit is that the experience gives the GTAs practice in teaching, and teaching is one of a number of professional development activities required of graduate students (Park, 2004).

GTAs, by definition, do not have the same educational background or ultimate authority as the primary instructor; however, students rate TAs and peer instructors highly on competence and instructional ability (Weyrich et al., 2008; Weyrich et al., 2009; Zijdenbos, De Haan, Valk, & Ten Cate, 2010). When asked to assess GTAs, students rated GTAs as more "relatable" and "engaging" than the primary instructors, who were seen as more "formal" and "boring," though competent and knowledgeable (Kendall & Schussler, 2012). The

student perception of GTA ability may vary between TAs and between classes or even universities, because GTAs receive widely varying levels of departmental or institutional support for teaching (Luft, Kurdziel, Roehrig, & Turner, 2004; Park, 2004). Training GTAs in teaching methods and supporting them during their teaching can dramatically improve their confidence and success in the classroom, as shown by the fact that trained GTAs receive higher scores than untrained GTAs on student-given ratings of effective teaching, respecting students, and being prepared than those without training (Marbach-Ad et al., 2012). GTAs are, in sum, a common instructional resource that is used to varying levels of success.

The other source of supplemental instruction available in classrooms is the students themselves, often students who have already taken the class. There is a substantial literature on the use of peer educators in medical education, a field that turned to peers given the lack of graduate students and the relative maturity and intellectual ability of their student population. In most studies conducted in medical settings, peer teaching is received favorably by both students and instructors and believed to have a positive effect on student learning, although a challenge commonly cited by authors of meta-analyses is that these published studies are often poorly designed and are qualitative rather than quantitative (Santee & Garavalia, 2006; Secomb, 2007; Ten Cate & Durning, 2007). What is clear is that medical peer TAs, like GTAs, benefit from some guidance about how to teach. When students in anatomy classes are simply asked to teach content to their peers without guidance or preparation, neither the teachers nor learners feel they benefit from the experience (Johnson, 2002). Conversely, providing even minimal preparation to the teachers increases satisfaction for both sides and can increase learning (Evans & Cuffe, 2009; Krych et al., 2005; Nnodim, 1997).

Peer education is also successful in the undergraduate classroom. Although some instructors remain doubtful of undergraduates' abilities to guide learning (Luft et al., 2004, and the authors' personal observations), undergraduates have played a role in peer instruction for decades, and the practice has benefits for both undergraduate teachers and learners (Fremouw, Millard, & Donahoe, 1979; Goldschmid & Goldschmid, 1976; Topping, 1996). One way that undergraduates are often asked to guide the learning process for their peers is through heavily structured learning activities in which groups of students work through in-class or near-class instructional activities together (Crouch & Mazur, 2001; Mazur & Somers, 1999; Otero, Pollock & Finkelstein, 2010). The combination of active learning and peer guidance is powerful, and this type of peer education has been shown to be effective in multiple disciplines (Eberlein et al., 2008; Lewis & Lewis, 2005; Tien, Roth, & Kampmeier, 2004; Wamser, 2006). A slightly different set of challenges and opportunities becomes relevant when undergraduates are asked to fill in the role of a course TA. Undergraduate TAs (UTAs) are sometimes used in the same classroom with GTAs (Chandler & Sweller, 1991; Otero et al., 2010), though there is increasing willingness to let UTAs lead their own sections (Hogan, Norcross, Cannon, & Karpiak, 2007; Romm, Gordon-Messer, & Kosinski-Collins, 2010; Sana, Pachai, & Kim, 2011). Having UTAs in a course offers more personalized instruction than the main instructor could provide alone, while avoiding increased demands on existing GTA resources. This is especially appealing to institutions that are dealing with increasingly limited budgets to support an expanding undergraduate class enrollment. Given their potential lack of teaching skills or experience, UTAs are most successful as teachers when they are provided with organized instruction in pedagogy and sometimes content concurrent with or preceding their time in the classroom (Hogan et al., 2007; Otero et al., 2010; Roderick, 2009; Romm et al., 2010; Sana et al., 2011). The teaching experience also benefits the UTAs themselves. A well-structured peer teaching experience allows the UTAs to gain self-confidence, communication skills, scientific understanding, and exposure to education as a career (Gafney & Varma-Nelson, 2007; Otero et al., 2010; Schalk, McGinnis, Harring, Hendrickson, & Smith, 2009; Tien et al., 2004). This program is analogous to a university-level adaptation of the UTeach program for early guided induction of new teachers into real learning environments (Mervis, 2007).

We set out to compare the effectiveness of GTAs and UTAs in identical roles leading majors' biology laboratory sections. Building on the evidence about how best to prepare TAs, we designed instruction to prepare and support the TAs and allowed them to be the primary instructional authority in the sections. We found UTAs to be as good as or better than GTAs in a number of measures, contributing data on the benefit of UTAs in university undergraduate science classrooms. Given the benefits to the UTAs themselves and the equivalent grade outcomes for their learners, this data supports an expanded role

for UTAs in undergraduate science laboratories.

Methods

The effects of TAs on students were evaluated for three courses in the majors' curriculum at a large state university in the Pacific Northwest. For two introductory courses (Course 1 and Course 2), results were collected over a 3-year span, and for one junior-level course (Course 3), results were collected for 2 years (Table 1). The structure of both Courses 1 and 2 involves a professor-led lecture given to the entire class that meets for four 1-hour sessions per week, and students are divided into small groups to attend a weekly 3-hour laboratory section taught by a single teacher, either a UTA or GTA with occasional support staff help when requested. Course 3 is more focused on laboratory work, meeting once a week for lecture and then having students attend two 3-hour labs per week run by a GTA or UTA.

The GTAs and UTAs were hired from different sources. As is common in universities, GTAs were assigned to classes as a primary method of funding support during the graduate studies of the GTA. Some GTAs had input into their desired courses to teach, whereas others were assigned with little or no choice. GTAs were generally first- and second-year life science graduate students from strong academic undergraduate science backgrounds. Classroom experience varied between GTAs, with some having served as TAs previously but only in rare exceptions having had experience teaching independent classes. The UTAs were advanced undergraduates who had previously performed well in the class, volunteered to be TAs, and received credit for their time. UTAs were typically a year or more removed from their own time in the course and were predominantly in their third year or later of undergraduate studies. UTAs were selected from an applicant pool on the basis of high GPA and perceived potential for ambitious and equitable teaching by course staff. UTA teaching experience varied, but most UTAs used the experience as their introduction to classroom teaching. Although their position of authority could be seen to elevate them above the role of a true peer educator, anecdotal evidence suggests that their status as undergraduates distinguishes them from GTAs in the eyes of both students and professors, and so the peer educator literature informs our analysis in the absence of a suitable body of UTA literature.

All TAs had the same training for the classes. The regimen included initial discussions and guidelines for role definitions and ethical considerations as well as weekly meetings to learn the lab exercises and discuss class progress. All TAs were given opportunities for feedback on their own classroom instructional technique throughout the term. All TAs were responsible for the same workload: leading labs, coming to all lectures, facilitating in-lecture activities, grading exam questions, grading weekly lab worksheets, and holding office hours. All TAs typically taught two lab sections per week, although a few TAs taught a single lab or as many as four labs per week in this study.

For Courses 1 and 2, class sizes ranged from 200 to 700 students in a single course. Students in these courses were primarily sophomores and had completed prerequisite courses in chemistry and mathematics. Students registered into labs of 20–24 students meeting once per week for 3 hours of lab time under the direction of a single TA. Student registration did not give the choice of a UTA or GTA, although demand for particular lab times made TA assignment predominantly arbitrary but not random. Students in Course 3 were juniors and seniors in a class focused on laboratory technique, and students enrolled into laboratory sections prior to knowing whether the section would be led by a UTA or GTA.

Grades for biology courses are calculated using numerical scores and no direct subjective assessment of student progress. For Courses 1 and 2, the final grades combined scores from both lecture and laboratory. The grades given for laboratory work were normalized within a course

TABLE 1

Summary of students included in this study.

Course	Number of students with a TA	Number of students with a GTA	Total number of students	Lab section size
Course 1	1,020	4,222	5,242	20
Course 2	1,100	1,680	2,780	24
Course 3	35	64	99	15

Note: Student data used in this summary were collected from three biology classes: Introductory Evolution, Diversity and Ecology (Course 1), Introductory Molecular, Cellular and Developmental Biology (Course 2), and Lab Techniques in Cell and Molecular Biology (Course 3). TA = teaching assistant; GTA = graduate teaching assistant.

to adjust for overly harsh or lenient TA grading, but TA-graded lab work accounted for a maximum of 15% of any overall grade and had little to no effect on final grade variability. The vast majority of grade variability in Courses 1 and 2 resulted from exam scores, which could be directly compared without adjustment to normalize between TA sections. For Course 3 the grade was focused on performance in the laboratory. The use of identical evaluative measures for students in this class made it possible to directly compare grades between sections run by UTAs and GTAs.

The effect of TA standing on student outcomes was measured in three ways. A large cadre of student end-of-course grades within a 3-year period was classified as "students of GTAs" or "students of UTAs" on a course-by-course basis. These course grades were then compared in aggregate using a *T*-test (Figure 1). A systemic difference between course means for students of UTAs versus students of GTAs would indicate nonequivalence of these different teaching assistants.

The second effect of TA standing looked at student evaluations of their TA, as assessed using a survey in a single term of Course 1 containing 630 students. Students were asked to rate their own TA on eight different statements of TA demeanor, effectiveness, attitude, and content knowledge (Figure 2). Comparison of these survey responses was done by a chi-square test, and if GTAs or UTAs were perceived by their students to be systematically different in one of these measures, we would expect to see a significant difference between TA types.

The effect of TAs on student perceptions of science was further measured in Course 2. In a standard postcourse evaluation, students were asked to choose between six answers indicating their outlook on science for their own future plans, and their responses were categorized as follows:

- Positive attitude: "I am more excited to continue with Biology," or "I am more likely to continue with Biology."
- Neutral attitude: "I am equally likely to continue with Biology," or "I never planned on continuing in Biology."
- Negative attitude: "I am more discouraged with Biology now," or "I am less likely to continue with Biology."

Student evaluations suggest that professors can have dramatically varying impacts on students' interest in continuing with biology, as shown by the wildly varying differences between two equivalent evaluations of professors from winter 2010 (N = 364) and summer 2012 (N = 184). To assess any differences in attitude correlated with TA type, we analyzed the responses of this survey from 1,120 students enrolled from winter 2010 to summer 2011 (Figure 3). Because the large majority of instructor–student contact time in this course is between students and TAs in lab, it is reasonable to assume that there may be an effect of TA on student outlook on science even though students were not specifically primed to think about their TA in this survey question. Statistical similarity was evaluated using a chi-square test. For all statistical comparisons, a p-value equal to or less than .05 was considered significant, and values above that threshold indicated no difference between the compared groups.

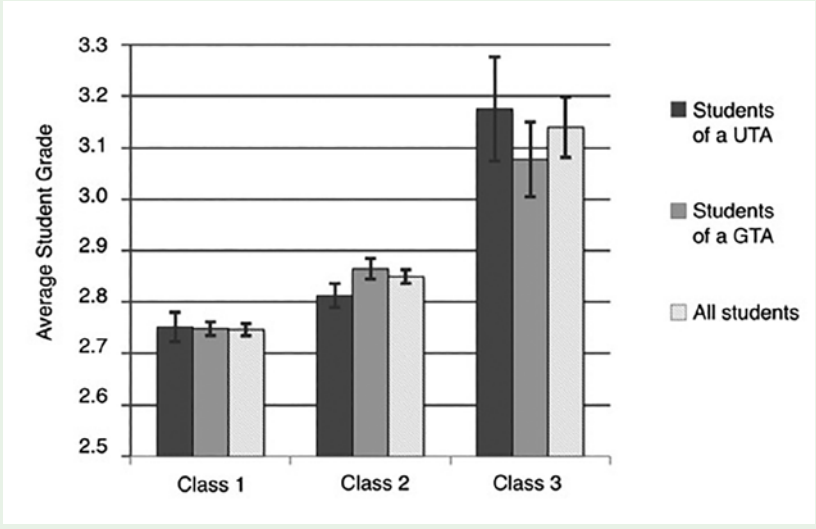
Results

To assess student outcomes on the basis of TA type, we evaluated grade-based outcomes and attitudinal outcomes. The first assessment we performed was to compare the final course grade of students with GTAs versus those with UTAs. For Courses 1 and 2, the course grading is structured so that overall grade primarily reflects student performance on exams, and points given for lab work are designed to encourage effort rather than assess knowledge gain. Students who complete the 10-week course typically receive 90% or more of all available points for laboratory work, and labs account for a very small portion of overall class grade variability. For this reason, lab scores are independent of overall grade, and we therefore focused on final course grade for these courses. In Course 3, laboratory performance is the centerpiece of grades, consisting of quizzes, lab reports, and lab practical. The average grade for each course was normalized to allow comparisons between terms and years. There was no difference in the course grade between students who had UTAs and those who had GTAs for any of our courses (Figure 1). The average for each group of students was also statistically similar to the overall class average (data not shown), suggesting that outliers were not throwing off the spread. Both of these comparisons were indistinguishable for students in Course 1, Course 2, and Course 3.

Given that there was no difference in class grade, we wanted to find out whether there was a qualitative difference in the laboratory environment established by each type of TA. We examined subjective student experience in two ways: student perception of instructor ability and student outlook on biology. An existing survey data set provided an opportunity for insight in a course from within the

FIGURE 1

Grades of students of UTAs and GTAs. The average grades of students taking Course 1, Course 2, and Course 3 were categorized on the basis of the type of TA (undergraduate or graduate) that led the section. There was no statistically significant difference between TA categories in each of the three classes. Error bars represent standard error of the mean.



scope of our grade data. To address student perception of TA ability we looked at a survey taken during a single quarter of Course 1 in which the 700 students were asked to complete an electronic survey after the completion of the course using an eight-question survey to measure each student's perception of the TA's instructional ability and attitude. Although logistical constraints limited this data collection to one academic quarter, the academic performance of these students and the composition of the TA pool were indistinguishable from that of other quarters, suggesting that data collected from this quarter is representative of students as a whole. In the survey, six of the questions directly addressed the TA's characteristics, and two questions asked about the relevance of the laboratory to the classroom exams. The questions were asked using a Likert scale, allowing the students to rate their TAs on a 5-point scale in which the higher score indicated greater satisfaction. Students reported that UTAs were more effective than GTAs at encouraging students to ask questions and making all students feel respected (Figure 2). In all other scores the UTAs and GTAs were statistically indistinguishable (Figure 2).

We assessed the impact of the class on the students' attitude toward science in an attempt to determine how similar the GTAs and UTAs were on this measure. Students in Course 2 were asked, as part of a routine course evaluation after the conclusion of the course, about their attitude toward science and the likelihood of their continuing to take biology classes in the future. The questions were scored as indicating positive, neutral, or negative attitude toward science. It is clear that there can be significant differences between classes of students, which correlates with the professor/course year, with some professors leading classes that have students with much more positive attitudes toward science (Figure 3A). When looking at the difference between student attitudes as grouped by TA type, however, the UTAs were as likely as the GTAs to have students with positive attitudes toward science courses (Figure 3B). This data contained the student responses from three separate quarters, but the lack of significant difference held true when comparing within and between quarters, suggesting a lack of seasonal effects. The majority of students stated that the course either increased their desire to pursue education in biology or had a neutral effect on that decision, and this was true for both students with UTAs and those with GTAs.

Discussion

The economic downturn in 2008 forced emergency-level measures throughout our university. The use of UTAs was massively increased to preserve the lab components of biology courses in the face of slashed budgets. Although we originally operated under the assumption that decreased use of GTAs would be detrimental for students, this "necessary evil" afforded an opportunity to observe the results of TA type in a well-controlled environment. The effects of TA roles, lecturer skills, student populations, and training are complex variables that are impossible to completely account for in a large study of student academic achievement; however, this time period in which course mechanics remained unchanged while adding UTAs to the instructor pool offered a uniquely controlled situation for assessing TA effect on student academic achievement. The inclusion of UTAs was initially a response to expanding enrollment in conditions of budgetary

constraints but was recognized as an opportunity to compare the effects of UTAs and GTA.

The student learners in all three courses were ably instructed by both UTAs and GTAs as evidenced by measures of academic achievement and subjective perceptions. This is consistent with the qualitative outcomes of previous researchers (Hogan et al., 2007; Secomb, 2007; Smith, 2008; Stanger-Hall, Lang, & Maas, 2010; Ten Cate & Durning, 2007) and adds to these studies by providing quantitative measures. The comparability of UTA and GTA, with no significant differences favoring GTAs, suggests that students without advanced degrees can be valuable educational partners in our classes. Incorporating peer teaching as part of the instructional corps used in large lecture courses may be a way to leverage motivated undergraduates in productive teaching roles. GTAs have benefited from the service learning opportunities of their roles: Because of our findings, we advocate for providing these opportunities to UTAs as well.

By comparing the grades of students with UTAs and those with GTAs, we hoped to address a fundamental, but often unspoken, assumption about undergraduates: that they cannot fulfill the educational duties of the TA as well as graduate students can. If course grade is one of the commonly recognized measures of educational progress, presumed UTA inferiority would result in lower course grades for students taught by UTAs, but we found no difference between the two TA groups for any of the courses (Figure 1). Given the significant differences in TA types in terms of academic background, this result may be surprising. On the other hand, UTAs may be more closely situated to their students in development or experience, as further discussed next.

Teaching is a complex profession in which the effect of a successful teacher can be massive. Balancing perceived strengths and weaknesses of different teaching types, however, makes for a more complicated system in which grades may not tell the entire story. If we wanted to assess TA success, we knew that further investigation of the student experience on the basis of TA type was necessary.

Given that grades can vary independently from the experience of the educational environment, we asked about the students' qualitative perception of the TAs and found that undergraduates were slightly better than graduate students at encouraging students to ask questions and making them feel respected, but that in other

FIGURE 2

Quality of student laboratory experience compared by TA types in Course 1. Students in the autumn 2010 of Course 1 labs were asked for their agreement with eight different statements about the attributes of their TA on a scale from 5 (*strongly agree*) to 1 (*strongly disagree*). Averages +/- standard error of the mean is shown. Asterisk indicates UTA scores significantly higher than GTA scores (*T*-test $p < .02$), and there were a total of 630 student responses included in the analysis. All questions were asked on a scale of 1 to 5, except for the Respect question, which was asked on a scale of 1 to 10 and then normalized to 5 points for consistency. The statements used for Likert agreement were worded as follows: (1) The TA effectively explains the material covered in lab. (2) The TA adjusts the pace of the lab to match student abilities and progress. (3) The TA encourages students to ask questions. (4) The TA answers student questions clearly and thoroughly. (5) The TA shows enthusiasm for the material she or he is presenting. (6) The TA treats EVERY student with respect. (7) The labs helped prepare me to successfully use the concepts on exams. (8) The TA's explanation of concepts during lab helped during exams.

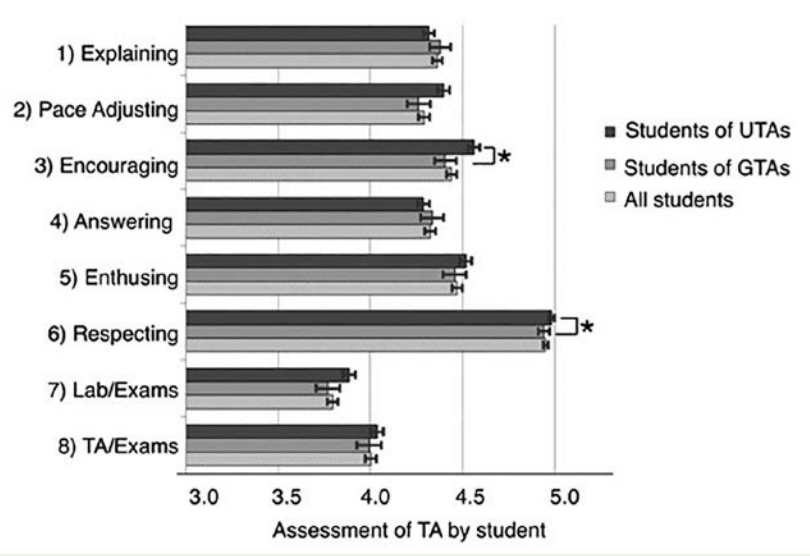
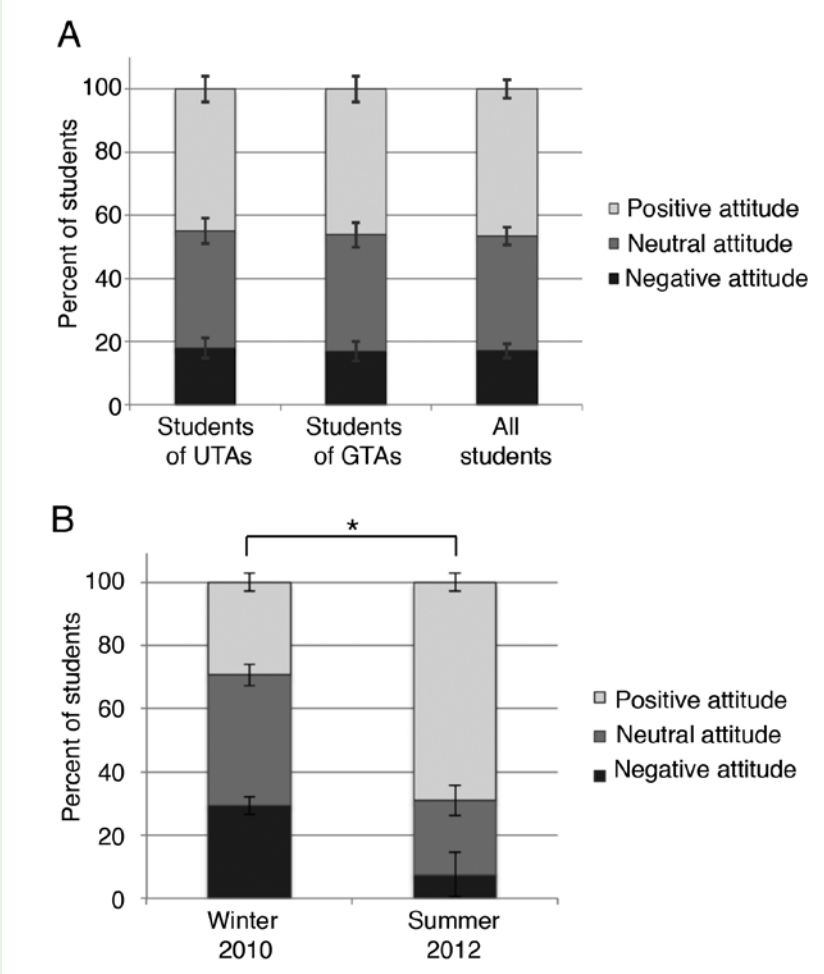


FIGURE 3

Student attitude toward science based on professor or TA. Students in Course 2 were asked, via an electronic survey after the conclusion of the course, about their attitude toward science and the likelihood of their continuing to take biology classes in the future. Responses from the multiple-choice question were categorized as detailed in the Methods section. (A) To analyze the correlation with TA type, we looked at responses from winter 2010 to summer 2011, a total of 588 students with a UTA and 560 students with a GTA. Bars show percentage of the total for each category, with error bars representing the 95% confidence interval. There were no differences observed between the attitudes of students taught by UTAs and graduate TAs, with a chi-square $p = .87$. (B) Grouping the responses of two different courses by course rather than TA type (winter 2010, $N = 364$; summer 2012, $N = 184$) shows the possible range of responses. In this case, significant differences were seen between quarters (asterisk indicates chi-square $p = .01$).



measures the TAs received equivalent ratings (Figure 2). The statistically significant difference suggests that the undergraduates were connecting with their UTAs differently from how they connected with their GTAs. One possible explanation of this is the social congruence between the groups, given that the undergraduates are all similar in age and experience (Lockspeiser, O'Sullivan, Teherani, & Muller, 2008). Accordingly, undergraduate TAs might see themselves reflected in the students, and vice versa, allowing an easier connection than was formed between the GTAs and students. Though significant, the UTA advantage in these measures was small and suggests that neither group has a large, overall advantage in our assessment of quality.

In addition to affecting how students feel about the daily experience of being in the classroom, teachers can also affect how students feel about the discipline as a whole. An interview-based study showed that teachers' comments or attitudes can sway a student to stick with, or abandon, engineering (Hong & Shull, 2010). To evaluate our students' attitudes toward science, we used an online questionnaire to gauge positive, neutral or negative attitude toward continuing with biology classes. Our comparison of students' attitudes toward biology showed no difference between students with UTAs and those with GTAs (Figure 3A). Students' attitudes did not differ between groups, either because both TA types were similarly effective in supporting students' disciplinary attitude or because instructors have little impact. We therefore examined these same attitude responses categorized by quarter instead of by TA type. A select comparison of responses from two different quarters (different faculty lecturers)

showed that student attitude is not uniform and can vary significantly between quarters (Figure 3B). This could suggest that an instructor has a sizeable impact on student attitudes; however, there are other parameters that changed between the quarters in question, and therefore the cause of this shift is not easily determined. Nevertheless, such a large difference between quarters strongly suggests that the lack of difference between UTA and GTA is noteworthy.

The lack of significant difference between UTAs and GTAs brings up questions about whether differences between TAs can be minimized through training. All reasonable efforts were made to have the groups be as functionally equivalent as possible, and the instructional support we provided to all TAs might have leveled the playing field and could explain their equivalent performance. Our training was based on the evidence that peer teachers benefit from explicit instruction about how to teach and how to help students learn the material in question (Roderick, 2009; Romm et al., 2010; Sana et al., 2011; Tien et al., 2004) and included role-playing exercises, public speaking practice, thought modeling games, peer feedback from other TAs, etc. We look forward to reports from other research groups analyzing the efficacy of different elements of TA preparation. Another possibility, that TAs with previous teaching experience in our program would be more effective, was not suggested in our findings (data not shown). Finally, it is worth noting that we could identify no categorical differences in instructional style or classroom authority between GTAs and UTAs, as anecdotally assessed by department lecturers who visited each section regularly while they were in session (A. Crowe and A. Schivell, personal communications).

Despite the shared TA training, GTAs and UTAs as individuals were different in possibly significant ways. As far as recruitment or selection, UTAs self-selected and applied to participate in the TA program, suggesting a high level of commitment to leading these classes. Graduate TAs were also self-selected in the sense that students enrolling in graduate programs are aware of the teaching requirement and have some input into the class for which they TA, though they may or may not value that teaching experience. This creates an obvious difference in potential motivations, which was a confounding factor but also likely to be common across universities. The two groups of TAs also had different content knowledge. Each UTA recently took the class in which he or she was teaching, which is a near-term selection for strength in understanding of the specific material. Graduate students could be expected to have higher overall content knowledge, but had not attended the class lecture until their teaching assistantship. Our results suggest that that none of these differences provides a uniquely significant advantage, although more targeted studies will be needed to address the relative value of the many components of an effective TA.

We do not suggest that institutions should replace GTAs with UTAs under their existing training regimens. Although this might be seen as an easy cost-saving measure, we hope our data help make the case that effective training can create an environment where the best possible teachers can be selected from both levels and supported well. Furthermore, tools for assessing in-class teaching ability are noticeably lacking. Without rigorous examinations of our TAs, we cannot fully evaluate how much extra benefits might be attainable but unrealized using our current training schema. Departments should assess and collect data wherever possible to address this issue and continue to improve the training of all TAs as a major factor in educating our future workforce.

In summary, our data provide quantitative evidence that UTAs can be effective as independent teachers. The extent to which TAs succeed in helping students learn content and appreciate science does not directly correlate with their years spent in school, as GTAs and UTAs are equally effective. The benefits of having well-trained UTAs goes far beyond simply increasing the number of nonfaculty instructors in a department. Students in the laboratory sections benefit by receiving high-quality instruction, and both GTAs and UTAs gain confidence and instructional skills. Although this training requires significant work on the part of course instructors, enhancing the quality of laboratory TAs could provide far-reaching benefits for students and peer teachers in biology and other laboratory-based sciences. ■

Acknowledgments

We thank Miles McDonough and Scott Freeman for supplying the data about attitude toward science and Alison Crowe and Amanda Schivell for their laboratory observations. We also thank them and other members of the Biology Education Research Group at the University of Washington for their feedback and helpful discussions.

References

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8,* 293–332.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of*

Physics, 69, 970–977.

Eberlein, T., Kampmeier, J., Minderhout, V., Moog, R. S., Platt, T., Varma-Nelson, P., & White, H. B. (2008). Pedagogies of engagement in science. *Biochemistry and Molecular Biology Education, 36,* 262–273.

Evans, D. J. R., & Cuffe, T. (2009). Near-peer teaching in anatomy: An approach for deeper learning. *Anatomical Sciences Education, 2,* 227–233.

Fremouw, W. J., Millard, W. J., & Donahoe, J. W. (1979). Learning-through-teaching: Knowledge changes in undergraduate teaching assistants. *Teaching of Psychology, 6,* 30–32.

Gafney, L., & Varma-Nelson, P. (2007). Evaluating peer-led team learning: A study of long-term effects on former workshop peer leaders. *Journal of Chemical Education, 84,* 535–539.

Goldschmid, B. G., & Goldschmid, M. L. (1976). Peer teaching in higher education: A review. *Higher Education, 5,* 9–33.

Hogan, T. P., Norcross, J. C., Cannon, J. T., & Karpiak, C. P. (2007). Working with and training undergraduates as teaching assistants. *Teaching of Psychology, 34,* 187–190.

Hong, B. S. S., & Shull, P. J. (2010). A retrospective study of the impact faculty dispositions have on undergraduate engineering students. *College Student Journal, 44,* 266–278.

Johnson, J. H. (2002). Importance of dissection in learning anatomy: Personal dissection versus peer teaching. *Clinical Anatomy, 15,* 38–44.

Kendall, K. D., & Schussler, E. E. (2012). Does instructor type matter? Undergraduate student perception of graduate teaching assistants and professors. *CBE—Life Sciences Education, 11,* 187–199.

Krych, A. J., March, C. N., Bryan, R. E., Peake, B. J., Pawlina, W., & Carmichael, S. W. (2005). Reciprocal peer teaching: Students teaching students in the gross anatomy laboratory. *Clinical Anatomy, 18,* 296–301.

Lewis, S. E., & Lewis, J. E. (2005). Departing from lectures: An evaluation of a peer-led guided inquiry alternative. *Journal of Chemical Education, 82,* 135–139.

Lockspeiser, T. M., O'Sullivan, P., Teherani, A., & Muller, J. (2008). Understanding the experience of being taught by peers: The value of social and cognitive congruence. *Advances in Health Sciences Education, 13,* 361–372.

Luft, J. A., Kurdziel, J. P., Roehrig, G. H., & Turner, J. (2004). Growing a garden without water: Graduate teaching assistants in introductory science laboratories at a doctoral/research university. *Journal of Research in Science Teaching, 41,* 211–233.

Marbach-Ad, G., Schaefer, K. L., Kumi, B. C., Friedman, L. A., Thompson, K. V., & Doyle, M. P. (2012). Development and evaluation of a prep course for chemistry graduate teaching assistants at a research university. *Journal of Chemical Education, 89,* 865–872.

Mazur, E., & Somers, M. D. (1999). Peer instruction: A user's manual. *American Journal of Physics, 67,* 359–361.

Mervis, J. (2007). UTexas tells science majors: We want U (to) teach. *Science, 316*(5829), 1275.

Nnodim, J. O. (1997). A controlled trial of peer-teaching in practical gross anatomy. *Clinical Anatomy, 10,* 112–117.

Otero, V., Pollock, S., & Finkelstein, N. (2010). A physics department's role in preparing physics teachers: The Colorado learning assistant model. *American Journal of Physics, 78,* 1218–1224.

Park, C. (2004). The graduate teaching assistant (GTA): Lessons from North American experience. *Teaching in Higher Education, 9,* 349–361.

Roderick, C. (2009). Undergraduate teaching assistantships: Good practices. *Mountain Rise: The International Journal for the Scholarship of Teaching and Learning, 5*(2). Available at http://mountainrise.wcu.edu/index.php/MtnRise/article/view/109

Romm, I., Gordon-Messer, S., & Kosinski-Collins, M. (2010). Educating young educators: A pedagogical internship for undergraduate teaching assistants. *CBE—Life Sciences Education, 9,* 80–86.

Sana, F., Pachai, M., & Kim, J. A. (2011). Training undergraduate teaching assistants in a peer mentor course. *Transformative Dialogues, 4*(3).

Santee, J., & Garavalia, L. (2006). Peer tutoring programs in health professions schools. *American Journal of Pharmaceutical Education, 70*(3).

Schalk, K. A., McGinnis, J. R., Harring, J. R., Hendrickson, A., & Smith, A. C. (2009). The undergraduate teaching assistant experience offers opportunities similar to the undergraduate research experience. *Journal of Microbiology and Biology Education, 10,* 32–42.

Secomb, J. (2007). A systematic review of peer teaching and learning in clinical education. *Journal of Clinical Nursing, 17,* 703–716.

Smith, T. (2008). Integrating undergraduate peer mentors into liberal arts courses: A pilot study. Innovative *Higher Education, 33,* 49–63.

Stanger-Hall, K. F., Lang, S., & Maas, M. (2010). Facilitating learning in large lecture classes: Testing the "teaching team" approach to peer learning. *CBE—Life Sciences Education, 9*, 489–503.

Ten Cate, O., & Durning, S. (2007). Dimensions and psychology of peer teaching in medical education. *Medical Teacher, 29,* 546–552.

Tien, L. T., Roth, V., & Kampmeier, J. A. (2004). A course to prepare peer leaders to implement a student-assisted learning method. *Journal of Chemical Education, 81,* 1313–1321.

Topping, K. J. (1996). The effectiveness of peer tutoring in further and higher education: A typology and review of the literature. *Higher Education, 32,* 321–345.

Wamser, C. C. (2006). Peer-led team learning in organic chemistry: Effects on student performance, success, and persistence in the course. *Journal of Chemical Education, 83,* 1562–1566.

Weyrich, P., Celebi, N., Schrauth, M., Möltner, A., Lammerding-Köppel, M., & Nikendei, C. (2009). Peer-assisted versus faculty staff-led skills laboratory training: A randomised controlled trial. *Medical Education, 43,* 113–120.

Weyrich, P., Schrauth, M., Kraus, B., Habermehl, D., Netzhammer, N., Zipfel, S., . . . Nikendei, C. (2008). Undergraduate technical skills training guided by student tutors—Analysis of tutors' attitudes, tutees' acceptance and learning progress in an innovative teaching model. *BMC Medical Education, 8*(1), 18.

Zijdenbos, I. L., De Haan, M. C., Valk, G. D., & Ten Cate, O. T. (2010). A student-led course in clinical reasoning in the core curriculum. *International Journal of Medical Education, 1,* 42–46.

Hannah C. Chapin *is a senior fellow in the Department of Biochemistry,* ***Benjamin L. Wiggins*** *is an instructional supervisor in the Department of Biology and a graduate student in the College of Education, and* ***Linda E. Martin-Morris*** (lmorris@u.washington.edu) *is a principal lecturer in the Department of Biology, all at the University of Washington in Seattle. Authors Hannah C. Chapin and Benjamin L. Wiggins contributed equally to this work.*

CBE—Life Sciences Education
Vol. 13, 167–178, Summer 2014

Research Methods

Understanding Classrooms through Social Network Analysis: A Primer for Social Network Analysis in Education Research

Daniel Z. Grunspan,* Benjamin L. Wiggins,† and Steven M. Goodreau*

*Department of Anthropology and †Department of Biology, University of Washington, Seattle, WA 98185

Submitted August 20, 2013; Revised January 22, 2014; Accepted January 23, 2014
Monitoring Editor: Erin Dolan

Social interactions between students are a major and underexplored part of undergraduate education. Understanding how learning relationships form in undergraduate classrooms, as well as the impacts these relationships have on learning outcomes, can inform educators in unique ways and improve educational reform. Social network analysis (SNA) provides the necessary tool kit for investigating questions involving relational data. We introduce basic concepts in SNA, along with methods for data collection, data processing, and data analysis, using a previously collected example study on an undergraduate biology classroom as a tutorial. We conduct descriptive analyses of the structure of the network of costudying relationships. We explore generative processes that create observed study networks between students and also test for an association between network position and success on exams. We also cover practical issues, such as the unique aspects of human subjects review for network studies. Our aims are to convince readers that using SNA in classroom environments allows rich and informative analyses to take place and to provide some initial tools for doing so, in the process inspiring future educational studies incorporating relational data.

INTRODUCTION

Social relationships are a major aspect of the undergraduate experience. While groups on campus exist to facilitate social interactions, the classroom is a principle domain wherein working relationships form between students. These relationships, and the larger networks they create, have significant effects on student behavior. Network analysis can inform our understanding of student network formation in classrooms and the types of impacts these networks have on students. This set of theoretical and methodological approaches can help to answer questions about pedagogy, equity, learning, and educational policy and organization.

DOI: 10.1187/cbe.13-08-0162
Address correspondence to: Daniel Z. Grunspan (grunny@uw.edu).

Social networks have been successfully used to test and create paradigms in diverse fields. These include, broadly, the social sciences (Borgatti *et al.*, 2009), human disease (Morris, 2004; Barabási *et al.*, 2011), scientific collaboration (Newman, 2001; West *et al.*, 2010), social contagion (Christakis and Fowler, 2013), and many others. Network analysis entails two broad classes of hypotheses: those that seek to understand what influences the formation of relational ties in a given population (e.g., having the same major, having relational partners in common), and those that consider the influence that the structure of ties has on shaping outcomes, at either the individual level (e.g., grade point average [GPA] or socioeconomic status) or the population level (e.g., graduation rates or retention in science, technology, engineering, and mathematics [STEM] disciplines). A growing volume of research on social influences at the postsecondary level exists, examining outcomes such as overall GPA and academic performance (Sacerdote, 2001; Zimmerman, 2003; Hoel *et al.*, 2005; Foster, 2006; Stinebrickner and Stinebrickner, 2006; Lyle, 2007; Carrell *et al.*, 2008; Fletcher and Tienda, 2008; Brunello *et al.*, 2010), cheating (Carrell *et al.*, 2008), drug and alcohol use (Duncan *et al.*, 2005; DeSimone, 2007; Wilson, 2007), and job choice (Marmaros and Sacerdote, 2002; De Giorgi *et al.*, 2009). The impacts are often significant, perhaps not surprisingly; this

research has many implications, including the importance that randomly determined relationships such as roommate or lab partner can have on undergraduates' behavioral choices and, consequently, their college experiences.

One key direction for education researchers is to study network formation within classrooms, in order to elucidate how the realized networks affect learning outcomes. Network analysis can give a baseline understanding of classroom network norms and illuminate major aspects of undergraduate learning. Educators interested in changing curriculum, introducing new teaching methods, promoting social equity in student interactions, or fostering connections between classrooms and communities can obtain a more nuanced understanding of the social impacts different pedagogical strategies may have. For example, we know active learning is effective in college classrooms (Hake, 1998; O'Sullivan and Copper, 2003; Freeman *et al.*, 2007; Haak *et al.*, 2011), but the full set of causal pathways is unclear. Perhaps one important change introduced by active learning is the facilitation of student networks to be stronger, less centralized, or structured in some other new way to maximize student learning. Social network analysis (SNA) can help us assess these types of hypotheses.

Recent research in physics education has found that a student's position within communication and interaction networks is correlated with his or her performance (Bruun and Brewe, 2013). An informal learning environment was found to be facilitative in mixing physics students of diverse backgrounds (Fenichel and Schweingruber, 2010; Brewe *et al.*, 2012). However, these exciting initial steps into network analysis in STEM education still leave many hypotheses to explore, and SNA provides a diverse array of tools to explore them.

The goal of this paper is to enable and encourage researchers interested in biology education, and education research more generally, to perform analyses that use relational data and consider the importance of learning relationships to undergraduate education. In doing so, we first introduce some of the many basic concepts and terms in SNA. We outline methods and concerns for data collection, including the importance of gaining approval from your local institutional review board (IRB). We briefly discuss a straightforward way to organize data for analysis, before performing a brief analysis of a classroom network along three avenues: descriptive analysis of the network, exploration of network evolution, and analysis of network position as a predictor of individual outcomes. This paper is aimed at serving as an initial primer for education researchers rather than as a research paper or a comprehensive guide. For the latter, see *Further Resources*, where we provide a list of additional resources.

INTRODUCTION TO THE CASE STUDY

In introducing network analysis, we draw our example from a subset of a 10-wk introductory biology course with 187 students who saw the course to completion as an example. Each student in this course attended either a morning or afternoon 1-h lecture of ~90 students four times a week and attended one of eight student labs of ~24 students each, which met once a week for 3 h and 20 min. This course used a heavy regimen of active learning, including a significant amount of guided student–student interaction in both lecture and lab. The total percentage of active-learning activities used in this lecture course was greater than 65% of classroom time, including audience response–device questions. The data we collected included who students studied with for the first three exams, all of their class grades, the lecture and lab sections to which they belonged, and general demographic information from the registrar.

Network Concepts

In this section, we lay out some of the foundations of SNA and introduce concepts and measurements commonly seen in network studies.

Social Network Basics. SNA aims to understand the determinants, structure, and consequences of relationships between actors. In other words, SNA helps us to understand how relationships form, what kinds of relational structures emerge from the building blocks of individual relationships between pairs of actors, and what, if any, the impacts are of these relationships on actors. *Actors*, also called *nodes*, can be individuals, organizations, websites, or any entity that can be connected to other entities. A group of actors and the connections between them make up a network.

The importance of relationships and emergent structures formed by relationships makes SNA different from other research paradigms, which often focus solely on the attributes of actors. For example, traditional analyses may separate students into groups based on their attributes and search for disproportional outcomes based on those attributes. A social network perspective would focus instead on how individuals may have similar network positions due to shared attributes. These similar network positions may present the same social influences on both individuals, and these social influences may be an important part of the causal chain to the shared outcome. In situations in which a presence or absence of social support is suspected to be important to outcomes of interest, such as formal learning within a classroom, the SNA paradigm is appealing.

Network Types. One way to categorize networks is by the number of types of actors they contain. Networks that consist of only one type of actor (e.g., students) are referred to as *unipartite* (or sometimes *monopartite* or *one-mode*). While not discussed in detail here, *bipartite* (or sometimes *two-mode*) networks are also possible, linking actors with the groups to which they belong. For example, a bipartite network could link scholars to papers they authored or students to classes they took, differing from a unipartite network, which would link author to author or student to student.

Networks can also be categorized by the nature of the ties they contain. For example, if ties between actors are inherently bidirectional, the network would be referred to as *undirected*. A network of students studying with one another is an example of an undirected network; if student A studies with student B, then we can be certain that student B also studied with student A, creating an undirected tie. If the relational interest of a network has an associated direction, such as student perceptions of one another, then it is referred to as a *directed* network; if student A perceives student B as smart, it does not imply that student B perceives student A as smart; without the latter, we would have one directed tie from A to B.

Ties can also be *binary* or *valued*. Binary ties represent whether or not a relation exists, while valued ties include additional quantitative information about the relation. For example, a binary network of student study relations would indicate whether or not student A studied with student B, while a valued network would include the number of hours they studied together. Binary networks are simpler to collect and analyze. Valued networks include a trade-off of more information in the data versus increased analytical and methodological complexity. Using the example of a study network, the added complexity of valued networks would allow an investigation regarding a threshold number of study hours necessary for a peer impact on learning gains, while a binary network would treat any amount of study time with a peer equally.

Network Data Collection. Collecting network data requires deciding on a time frame for the relationships of interest. Real-world networks are rarely static; ties form, break, strengthen and weaken over time. At any given time, however, a network takes on a given cross-sectional realization. Network data collection (and subsequent analyses) can be categorized, then, by whether it considers a static network, a cross-sectional realization of an implicitly dynamic network, or an explicitly dynamic network. The last of these may take the form of multiple cross-sectional snapshots or of some form of continuous data collection. Measuring and analyzing dynamic networks introduces a host of new challenges. Because the set of actors in a classroom population is mostly static for a definite period of time (i.e., a semester or quarter), while the relational ties among them may change over that period, all three options are feasible in this setting. The type of collection should, of course, be driven by the research question at hand. For example, our interest in the evolution of study networks inspired a longitudinal network collection design. Examining the impact of network ties on subsequent classroom performance, on the other hand, could be done with a single network collection.

Beyond considering the time frame of collection, it is also important to consider how to sample from a population. *Egocentric* studies focus on a sample of individuals (called "egos") and the local social environment surrounding them without explicitly attempting to "connect the dots" in the network further. Typically, respondents are asked about the number and nature of their relationships and the attributes of their relational partners (called "alters"). In some fields, the term "egocentric data collection" implies that individual identifiers for relational partners are not collected, while in other fields this is not part of the definition. By either definition, egocentric studies tend to be easier to implement than other methods, both in terms of data collection and ethics and human subjects review. Egocentric data are excellent first descriptors of a sample and, in many situations, may be the only form of data available. A wide range of important hypotheses can be tested using egocentric data, although questions about larger network structure cannot. Asking a sample of college freshmen to list friends and provide demographic information about each friend listed would represent egocentric network collection.

At the other end of the spectrum, *census* networks, sometimes referred to as *whole* networks, collect data from an entire bounded population of actors, including identifiable information about the respondents' relational partners. These alters are then identified among the set of respondents, yielding a complete picture of the network. This results in more potential hypotheses to be tested, due to the added ability to look at network structures. In our classroom study, we asked students to list other students in that same classroom with whom they studied; this is an example of a census network whose population is bounded within a single classroom.

High-quality census networks are rare, due to the exhaustive nature of the data collection, as well as the need for bounding a population in a reasonable way. It is worth noting that census networks may lack information on potentially influential relations with actors who are not a part of the population of interest; for example, important interactions between students and teaching assistants will be absent in a census network interested in student–student interactions, as would any students outside the class with whom students in the class studied. In the case of longitudinal studies, an added challenge arises—handling students who withdraw from the class or who join after the first round of data collection has been conducted. Census data collection also presents a nonresponse risk, which may result in a partial network. Nonresponse is more acute in complete network studies than other kinds of data collection because many of the commonly used analytical methods for complete networks consider the entire network structure as an interactive system and assume that it has been completely observed. Educational environments such as classrooms are fairly well bounded and have unique and important cultures between relatively few actors; they are thus prime candidates for census data collection, although the above issues must still be attended to.

Network Level Concepts and Measures. Network analysis entails numerous concepts and measurements absent in more standard types of data analyses. Perhaps the most basic measurement in network analysis is network *density*. The density of a network is a measurement of how many links are observed in a whole network divided by the total number of links that could exist if every actor were connected to every other actor. These measurements are frequently small but vary by the type and size of the network. Density measurements are often hard to interpret without comparable data from other similar networks.

Density is a global metric that simply indicates *how many* ties are present. A long list of network concepts are further concerned with the patterns of *who is connected with whom*. One pervasive concept in the latter realm is *homophily* (McPherson *et al.*, 2001), a propensity for similar actors to be disproportionately connected in a relation of interest. If we are interested in who studies with whom, and males disproportionately studied with other males and females with other females, this would exemplify some level of homophily by gender. Likewise, we could see homophily by ethnicity, GPA, office-hours attendance, or any other characteristic that can be the same or similar between two students. Understanding and researching homophily in classroom and educational networks may be central for several reasons. For example, two reasonable hypotheses are that relationships of social support in classrooms are more likely to be seen between students with similar backgrounds and that having sufficient social support is important for STEM retention. Testing these hypotheses by looking for homophily in networks with relation

to STEM retention would provide valuable information regarding the lower STEM retention rates of underrepresented groups. Confirming these hypotheses, then, would inform improved classroom behavioral strategies for educators to emphasize.

Finding a pattern of homophily for certain research questions is interesting on its own. Note, however, that a pattern of homophily can emerge from multiple processes. Two examples of these are *social selection* and *social influence*. Social selection occurs when a relationship is more likely to occur due to two actors having the same attributes, while social influence occurs when individuals change their attributes to match those of their relational partners, due to influence from those partners. As an example, we can imagine a hypothetical college class in which a network of study partners reveals that students who received "A's" disproportionately studied with other students receiving "A's." If "A"-level students seek out other "A"-level students to study with, this would be social selection; if studying with an A-level student helps raise other students' grades, this would be social influence. Depending on the goals of a study, disentangling between these two possibilities may or may not be of interest. Doing so is most straightforward when one has longitudinal data, so that event sequences can be determined (e.g., whether student X became an "A" student before or after studying with student Y).

Analyzing ties between two individuals independently, such as in studies of homophily, falls into the category of dyad-level analysis. When one has a census network, however, analysis at higher levels such as triads is possible. Triads have received considerable interest in network theory (Granovetter, 1973; Krackhardt, 1999) due to their operational significance. *Triads* are any set of three nodes and offer interesting structural dynamics, such as one node brokering the formation of a tie between two other nodes, or one node acting as a conduit of information from one node to the other. One version of classifying triads in an undirected network (commonly called the undirected Davis-Leinhardt triad census) is shown in Figure 1.

In a study network, a class exhibiting many complete triads may indicate a strong culture of group study compared with a class that exhibits comparatively few complete triads. One way to examine this would be a *triad census*—a simple count of how many different triad types exist in a network. Another way to measure this would be to look at *transitivity*, a value representing the likelihood of student A being tied to C, given that A is tied to B and B is tied to C. Transitivity is a simple, local measure of a more general set of concepts related to clustering or cohesion, which may extend to much larger groups beyond size three.

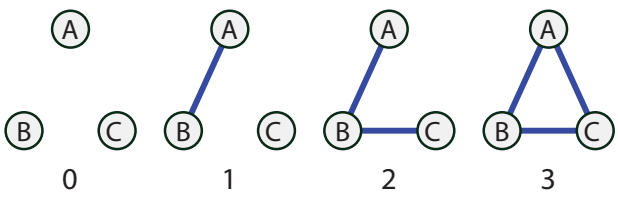


Figure 1. Davis and Leinhardt triad classifications for undirected networks.

In directed networks, transitivity can take on a different meaning, pointing to a distinct pair of theoretical concepts. When three actors are linked by a directed chain of the form A→B→C, then there are two types of relationships that can close the triad: either A→C or C→A (or, of course, both). The first option creates a structure called a *transitive triad*, and the latter a *cyclical triad*. For many types of relationships (i.e., those involving giving of goods or esteem), a preponderance of transitive triads is considered an indicator of hierarchy (with A always giving and C always receiving), while a preponderance of cyclical triads is an indicator of egalitarianism (with everyone giving and everyone receiving). If asking students about their ideal study partners, the presence of transitive triads would reflect a system wherein students agree on an implicit ranking of best partners, presumably based on levels of knowledge and/or helpfulness. Cyclic triads (as well as other longer cycles) would be more likely to appear if students believed that other factors mattered instead or as well; for instance, that it is most useful to study with someone from a different lab group or with a different learning style so as to maximize the breadth of knowledge.

Actor-Level Variables. Nodes within a network also have their own set of measurements. These include the exogenously defined attributes with which we are generally familiar (e.g., age, race, major), but they also include measures of position of nodes in the network. Within the latter, a widely considered cluster of interrelated metrics revolves around the concept of *centrality*. Several ways of measuring centrality have been proposed, including *degree* (Nieminen, 1974), *closeness* (Sabidussi, 1966), *betweenness* (Freeman, 1977), and *eigenvector* centrality (Bonacich, 1987). Degree centrality represents the total number of connections a node has. In networks in which relations are directional, this includes measures of *indegree* and *outdegree*, or the number of edges pointing to or away from an actor, respectively. Degree centrality is often useful for examining the equity or inequity in the number of ties between individuals and can be done by looking at the degree distribution, which shows the distribution of degrees over an entire network. Betweenness centrality focuses on whether actors serve as bridges in the shortest paths between two actors. Actors with high betweenness centrality have a high probability of existing as a link on the shortest path (*geodesic*) between any two actors in a network. If one were to look at an airport network (airports connected by flights), airports serving as main hubs, such as Chicago O'Hare and London Heathrow, would have high betweenness, as they connect many cities with no direct flights between them. Closeness centrality focuses on how close one actor is to other actors on average, measured along geodesics. It is important to keep in mind that closeness centrality is poorly suited for disconnected networks (networks in which many actors have zero ties or groups of actors have no connection to other groups). Eigenvector centrality places importance on being connected to other well-connected individuals; having well-connected neighbors gives a higher eigenvector centrality than having the same number of neighbors who are less well connected. Easily the most famous metric based upon eigenvector centrality is the PageRank algorithm used by Google (Page *et al.*, 1999). Because the interpretation of what centrality is actually measuring depends on the metric selected and the type of

network at hand, careful consideration is advised before selecting one or more types of centrality for one's study.

Network Methods: Data Collection

In this section, we provide guidance for collecting network data from classrooms. Our discussion is based on existing literature as well as personal experience from our previously described network study.

Both relational and nodal attribute data can be collected using surveys. Designing an effective survey is a more challenging task than often anticipated. There are excellent resources available for writing and facilitating survey questions (Fink, 2003; Denzin and Lincoln, 2005). This section highlights some of the issues unique to surveys for educational network data.

Survey fatigue, and its resulting problems with data quality (Porter *et al.*, 2004), can be an issue for any form of survey research; however, for network studies, it can be especially challenging, given that students are reporting not only on themselves but also on each of their relational partners. For our project, we avoided overuse of surveys in several ways. Routine administrative information such as lab section, lecture section, student major, course grades, and exam grades was easily collected from instructor databases. Data about student demographics, educational background, and standardized testing were obtained through a request to our university's registrar's office (with accompanying human subjects approval).

We strongly suggest pilot studies with your survey, as scheduling a single high-value data collection as the first use of a survey instrument can be risky. The delay in waiting for the next term or the next class for a more vetted collection is worthwhile. Data processing time and effort can be greatly reduced by streamlined data collection, and analysis will be strengthened by iterative improvement of survey questions. With adequate design preparation, brief surveys can easily collect relational data. It is important to keep questions clear and compact. Guidance into the form of the data can make data collected from both closed- and open-ended questions much simpler to clear and process (Wasserman *et al.*, 1990; Scott and Carrington, 2011).

Relational data collected in a closed-ended format such as lists, drop-down menus, or autocomplete forms can limit errors that come with open-ended data collection and are often easier to process. While these streamline student choices, they also come with a downside: they can introduce name confusions (e.g., in our class, nine students share the same first name) and are most problematic when students use nicknames. List data should always allow for both a "Nobody" answer choice and a default "I prefer not to answer" answer choice. An example of data collection with a closed list is shown below:

> ***Question 11:*** *We are interested in learning how in-class study networks form in large undergraduate classes. Over the next few pages is a class roster with two checkboxes next to each student—one which says "Pre-class friend" and one which says "Strong student".* ***For each student, evaluate whether they fit the description for each box*** *(immediately below this paragraph), and check the box if they do.*
>
> ***Pre-class friend****: A student that you would consider a friend from BEFORE the term of this class. If you have met someone in this class that you would consider a friend now but not before this class, do not list them as a pre-class friend.*
>
> ***Strong student****: A student you believe is good at understanding class material.*
>
> ***If you are not exactly sure of a name, mark your best guess.*** *The next question in this survey will allow you to write in a name if you don't see one or aren't sure.*
>
> ****Please know that your response is completely confidential. All names will be immediately re-coded so we will have no idea who studied with whom. This information will never be used for any class purpose, grading purpose, or anything else before the end of the class. Also, please note that students that you list will not know that you listed them in this survey, and you will not know if anyone listed you.****

	Pre-class friend	***Strong student***
Curie, Marie	☐ *Pre-class friend*	☐ *Strong student*
Darwin, Charles	☐ *Pre-class friend*	☐ *Strong student*
Einstein, Albert	☐ *Pre-class friend*	☐ *Strong student*
Franklin, Rosalind	☐ *Pre-class friend*	☐ *Strong student*
If no checkmarks:	☐ *Nobody fits descriptions above*	☐ *Nobody fits descriptions above*
	☐ *I prefer not to answer*	☐ *I prefer not to answer*

The number of possible choices given to subjects is an area of intense interest to survey writers in other fields (Couper *et al.*, 2004). Limiting respondents to a given number of answers has a variety of purposes; e.g., in egocentric studies in which a respondent will be asked many questions about each partner, it can help to limit respondent fatigue. For census network data, this is not an issue because we will not need to ask students a long list of questions about the attributes of their alters; we will have that information from the alters themselves, who are also students in the class. It can also help avoid a subject with a broad definition of friendship or collaboration from dominating the data set. We chose to avoid limits on numbers of student nominations, which have the potential to induce subjects to enter data to fill up their perceived quota. In our experience, individual student responses are typically few; no student listed so many friends or study partners that it drowned out other signals significantly.

Open-ended data collection should also include a means for students to indicate that no choices fit the question, to differentiate between nonrespondents and null answers. The largest source of respondent error in open-ended data is again name confusion between students. However, errors can be minimized by providing concise instructions for student-answer formatting. For one of our projects, one example of an open-ended relational survey question was:

> *We are interested in how networks form in classes. Please list first and last names if possible. If this is not possible, last initials or any description of that person would be appreciated (ie: "they are in the same lab as me", "really tall" or "sits in the second row").*
>
> *If no one fits one of these descriptions, simply write "none."*

> ***Your response is completely confidential. All names will be re-coded so we will have no idea who listed whom. This information will never be used for any class purpose, grading purpose, or anything else before the end of the class. Also, please note that students that you list will not know you listed them in this survey, and you will not know if anyone listed you.***
>
> *There are no right or wrong answers for this. We will ask you similar questions a few times this term. These data are incredibly valuable, so we truly appreciate your answers!*
>
> ***Please list any people in the class that you know are strong with class material. If you do not list anybody, please type either "No one fits description" OR "I prefer not to answer".*** *(separate multiple students with a comma, like "Jane Doe, John Doe").*

Finally, it may be appropriate in smaller classes, communities with less online capability, or in particularly well-funded studies to collect relational data by interviews. This brings along greater privacy concerns but may be necessary for some hypotheses. Open-ended questions allow for greater breadth of data collection but come with intrinsic complexity in processing. For example, a valued network describing the amount of respect that students have for various faculty might be best collected in a private interview. In this format, the interviewer could more thoroughly describe "respect" by using repeated and individualized questioning to ascertain the amount of respect a student has for each faculty member.

Timing of Survey Administration

Timing of survey questions throughout a class is important. For classroom descriptions consisting of a single network, data should be collected at the earliest possible time that all students have had the experiences desired in the research study. This limits the loss of data due to students forgetting particular ties, dropping or switching classes, or failing to complete the assignment as submission rates inevitably drop toward the end of the term. For longitudinal studies involving several collections, relational data can be collected either at regular intervals or around important classroom events. In either case, we strongly suggest implanting relational survey questions in already existing assignments, if permitted, to maximize data collection rates.

For our project, we collected data throughout the 10-wk term of an introductory biology course. We surveyed for student study partnerships after each exam, spread at semiregular intervals throughout the term (weeks 3, 5, 8, and 10). It will come as no surprise to instructors that attempts to administer an additional, nongraded survey gave lower response rates from already overworked and overscheduled undergraduates. Instead, we appended ungraded survey questions to existing graded online assignments. Depending on your research question, it may be appropriate to repeat some collections to allow for redundancy or for longitudinal analyses. Friendships, for example, are subjectively defined and temporal (Galaskiewicz and Wasserman, 1993). In some of our projects, we ask students for friendship relational data at both the beginning and end of the term as an internal measure of this natural volatility.

Given high response rates, anecdotal accounts of student study groupings that corroborated with the relational data, and limited extra work placed on students to provide data, we have a high level of confidence in the efficacy of our data collection methods, and others interested in network research with similar populations may also find these methods effective.

IRB and Consent

Data used solely for curricular improvement and not for generalizable research often do not require consent, but any use of the data for generalizable research does (Martin and Inwood, 2012). Social network data include the unique issue of one individual reporting on others in some form or other, even if it is only on the presence of a shared relationship. They also often describe vulnerable populations; this can be especially true for educational network research, when researchers are often also acting as instructors or supervisors to the student subjects and are thus in a position of authority. This may create the impression in students' minds that research participation is linked to student assessment. Because of this, early and frequent conversations with your local human subjects division are useful, illuminating, and should take priority (Oakes, 2002).

The nature of network data not only allows subjects to report information on other subjects but may allow recognizability of even anonymized data (called *deductive disclosure*), especially in small networks. This makes larger data sets typically safer for subjects. It also means that some network data fields must be stripped of information (Martin and Inwood, 2012). A relatively common example is in networks of mixed ethnicity in which one ethnic group is extremely small. In these cases, ethnicities may need to be identified by random identifiers rather than specific names. In many scenarios, researchers must plan on anonymizing or removing identifiers on data (Johnson, 2008). Your IRB will determine the best fit of plan for any given population of subjects.

Obtaining consent makes networks exciting and problematic at the same time. Complete inclusion of all subjects gives fascinating power to network statistics. Incomplete networks are far less compelling. More so than simpler unstructured data, networks may hinge on a small group of centralized actors in a community. The twin goals of subject protection and data set completion may compete (Johnson, 2008).

In our experience, conversations with IRB advisors led to an understanding of opt-in and opt-out procedures. For example, a standard opt-in procedure would use an individual not involved with the course to talk students through a consent script, answer questions, and retrieve signed consent forms from consenting subjects. An opt-out procedure would provide the same opportunities for student information and questions but ask subjects to opt out by signing a centrally located and easily accessible form kept confidential from researchers until after the research is completed. While the opt-in procedures are more common and foreground subject protection, they tend to omit data with a bias toward underserved and less successful populations. For this reason, we used an opt-out procedure, which commonly leads to higher rates of data return. Balancing research goals and appropriate protection of subject rights and privacy is critical (Johnson, 2008). By minimizing the risk to our subjects via confidential network collection, the use of an opt-out procedure was justified.

Table 1. Example of nodal attributes held in a matrix

	Gender	Major	Lab section	Grade
Marie	1	Chemistry	2	3.5
Charles	0	Theology	1	2.6
Rosalind	1	Biophysics	4	3.8
Linus	0	Biochemistry	5	4.0
Albert	0	Physics	5	3.3
Barbara	1	Botany	1	3.1
Greg	0	Pre-major	3	3.0

Data Management

Matrices are a powerful way to store and represent social network data. Common practice is to use a combination of matrices, one (or more) containing nodal attributes (see Table 1) and one (or more) containing relational data. A common form for the latter is called a *sociomatrix* or *adjacency matrix* (see Table 2); another is as an *edgelist*, a two-column matrix with each row identifying a pair of nodes in a relationship. For our study, we compiled several sociomatrices taken longitudinally at key points in the class, as well as one matrix with data of interest about our students.

A unipartite sociomatrix will always be square, with as many rows and columns as there are respondents. For undirected networks, the sociomatrix will be symmetric along the main diagonal; for undirected, the upper and lower triangles will instead store different information. Matrices for binary networks will be filled with 1s and 0s, indicating the existence of a tie or not, respectively. In cases of nonbinary ties (e.g., how many hours each student studied together) the numbers within the matrix may exceed one. The matrix storing nodal attribute information need not be square; it will have a row for each respondent and a column for each attribute measured.

It is important to understand the value of keeping rows of attribute data linkable to, and in the same order as, sociomatrices—this will ensure the relational data of a student are paired properly to his or her other data. The linkage can be done through unique names; more typically it will be done using unique study IDs.

The amount of effort and time spent cleaning the data will depend on how the data were collected and the classroom population. For this reason, it is advisable to plan the amount and means of collecting data around your ability to process them. Recently, we collected a large relational data set via open-ended survey online. To process these data into sociomatrices we created a program capable of doing more than 50% of the processing (Butler, 2013), leaving the rest to simple data entry. For data collected using a prepopulated computerized list, it may even be possible for all data processing to be automated.

Table 2. Example of a small sociomatrix

	Marie	Charles	Rosalind	Linus	Albert	Barbara	Greg
Marie	–	0	1	0	1	0	1
Charles	0	–	0	1	0	0	0
Rosalind	0	0	–	0	0	0	0
Linus	0	0	0	–	0	0	0
Albert	1	0	0	0	–	0	0
Barbara	0	0	0	1	0	–	0
Greg	0	0	0	0	0	0	–

Data Analysis

Many different questions can be addressed with SNA, and there are nearly as many different SNA tools as there are questions. As an example, we will look at the change in student study networks over the span of two exams from our previously described study. Our main interest in these analyses will be how study networks form in a classroom and the impacts these networks have on students. To generate testable hypotheses, we will first perform exploratory data analysis, taking advantage of sociographs. These informative network visualizations offer an abundance of qualitative information and are a distinguishing feature of SNA. It is important to note that, while SNA lends itself well to exploratory analyses, it is often judicious to have a priori hypotheses before beginning data collection. The exploratory data analysis embedded below is used to provide a more complete tutorial rather than to model how research incorporating relational data must be performed.

Starting Analyses

Most familiar statistical methods require observations to be independent. In SNA, not only are the data dependent among observations, but we are fundamentally interested in that dependence as our core question. For these reasons, the methods must deal with dependence. As a result, analyses may occasionally seem different from familiar methods, while at other times they can seem familiar but have subtle differences with important implications. This point should be kept in mind while reading about or performing any analysis with dependent data.

There are a number of proprietary software packages available for performing SNA, and interested investigators should weigh the pros and cons of each for their own purposes before choosing which to use. We use the statnet suite of packages (Handcock *et al.*, 2008; Hunter *et al.*, 2008) in R, primarily its constituent packages network and sna (Butts, 2008). R is an open-source statistical and graphical programming language in which many tools for SNA have been, and continue to be, developed. The learning curve is steeper than for most other software packages, but it comes with arguably the most complex statistical capabilities for SNA. Other network analysis packages available in R are RSiena (Ripley *et al.*, 2011), and igraph (Csardi and Nepusz, 2006). Other software packages commonly used for analysis for academic purposes include UCINet (Borgatti *et al.*, 1999), Pajek (Batagelj and Mrvar, 1998), NodeXL (Smith *et al.*, 2009; Hansen *et al.*, 2010), and Gephi (Bastian *et al.*, 2009).

We include R code for step-by-step instructions for our analysis in the Supplemental Material for those interested in using statnet for analyses. The Supplemental Material also includes instructions for accessing a mock data set to use with the included code, as confidentiality needs and corresponding IRB agreements do not allow us to share the original data.

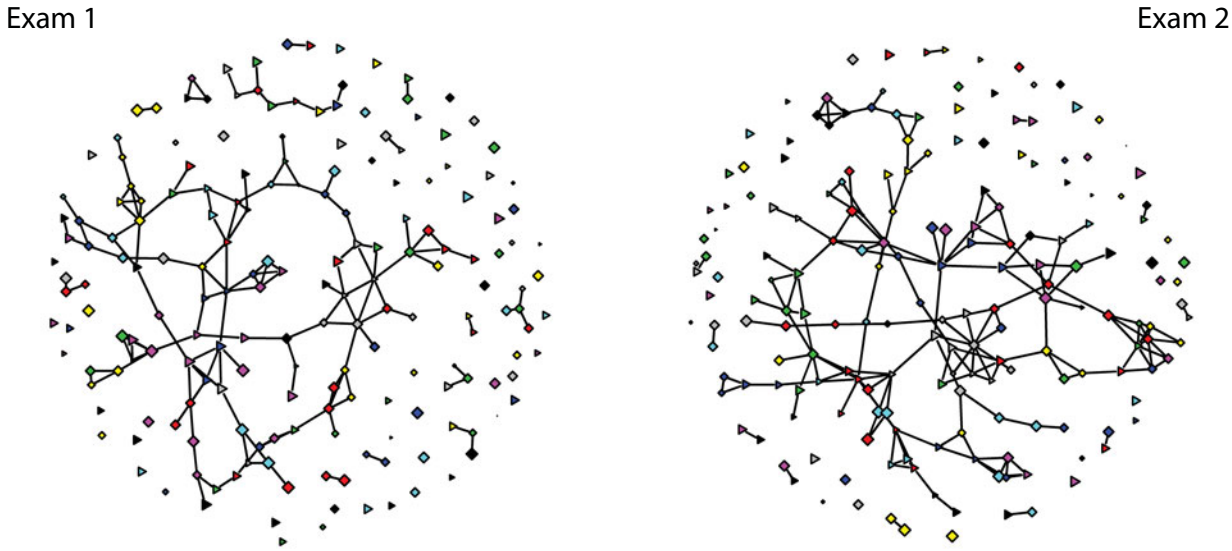


Figure 2. Sociographs representing study networks for the first and second exam. Male students are represented as triangles and females as diamonds. The color of each node corresponds to the lab section each student was in. Edges (lines) between nodes in the networks represent a study partnership for the first and second exam, respectively.

Exploratory Data Analysis

In performing SNA, visualizing the network is often the first step taken. Using sociographs, with nodal attributes represented by different colors, shapes, and sizes, we will be able to begin qualitatively assessing a priori hypotheses and deriving new hypotheses. We hypothesize that students who are in the same lab are more likely to study together, due to their increased interaction. We also think students with fewer study partners, and thus less group support in the class, are less likely to perform well in the class.

Figure 2 contains two sociographs visualizing the study networks for the first and second exam. Each shape represents a student, and a line between two shapes represents a study relationship. In these graphs, each color represents a different lab section, shape represents gender, and the size of each shape corresponds to how well the student performed in the class.

While no statistical significance can be drawn from sociographs, we can qualitatively assess our hypotheses. Judging by the clustering of colors, it seems as though same-lab study partnerships were rarer in the first exam than the second exam, for which several same-color clusters exist. This provides valuable visual evidence, but more rigorous statistical methods are important, particularly if policy depends on results.

There does not seem to be any strong visual evidence for an association between classroom performance and number of study partners. If this were true, we would see isolated nodes (those with zero ties) and nodes with few connections to be smaller on average than well-connected nodes. Visually, it is hard discern whether this is the case, and more rigorous tests can help us test this hypothesis. We first explore structural changes in study networks between the first two exams before statistically testing for an association between test scores and social studying.

Network Changes over Time

We can compare the study networks from the first and second exams using network measures such as density, triad censuses, and transitivity. These measurements allow us to assess whether the number of study partnerships are increasing or decreasing and whether any changes affect larger network structures such as triads.

Examining Table 3, a few things become clear. First, 34 more study partnerships exist in the second exam compared with the first, a 22.5% increase in network density. This increase in study partnerships does not distinguish between students moving from studying alone to studying with other students and students who have study partners adopting more study partners. One way to gain a better understanding of the increase in overall study partnerships is to look at the degree distribution for the first two exams, seen in Table 4.

There are fewer students without study partners on the second exam, several students exhibiting extreme sociality in their study habits, and an overall trend toward more students with upwards of five study partners. Unfortunately, the degree distribution does not completely illuminate the social

Table 3. General measurements taken from study networks of the first two exams

Measure	First exam study network	Second exam study network
Edges	151	185
Density	0.00868	0.01064
Triad (0)	1,044,790	1,038,672
Triad (1)	27,407	33,384
Triad (2)	216	326
Triad (3)	32	63
Transitivity	0.3077	0.3670

Table 4. Degree distribution from the study networks of the first two exams

	Degree									
	0	1	2	3	4	5	6	7	8	9
First	57	45	32	34	8	7	4	0	0	0
Second	51	43	24	33	17	12	1	3	2	1

mobility of students between the first and second exam. One way to view general trends is to use a parallel coordinate plot using the degree data from the first and second study networks.

The plot in Figure 3 seems to indicate that the overall increase in study partnerships is not dominated by a few individuals and is instead an outcome of an overall class increase in social study habits. While we see many isolated students studying alone on the first and second exam, we also find many branching off and studying socially in the second exam. At the same rate, many students studied with partners in the first exam and become isolated on the second.

Not only are there more overall connections, but we see higher transitivity and a trend toward complete triads. This increase in both measures indicates how students find their new study partners; they become more likely to study with their study partner's study partner, resulting in more group studying.

Ties as Predictors of Performance

Understanding study group formation and evolution is both interesting and important, but we are not limited to questions focused on network formation. As educators, we are inherently interested in what drives student learning and the kinds of environments that maximize the process. We can start addressing this broad question by integrating student performance data with network data.

As an example, we will test for an association between exam scores and both degree centrality and betweenness centrality. Studying with more students (indicated by degree centrality) and being embedded centrally in the larger classroom study network (indicated by betweenness centrality) may be a better strategy than studying alone or only with socially disconnected students. If we think of each edge in the study network as representing class material being discussed in a bidirectional manner, then more social students may have a leg up on those who are not grappling with class material with peers.

Owing to the dependent nature of centrality measures, testing for an association between network position and exam performance is not completely straightforward. One way around the dependence assumption is to use a permutation correlation test. The general idea is to create a distribution of correlations from our data by randomly sampling values from one variable and matching them to another. In effect, we will assign each student in the study network a randomly selected exam score from the scores in the class 100,000 times. This creates a null distribution of correlation coefficients (ρ) for the correlation between exam score and centrality measure for the set of exam scores found in our data, as seen in Table 5. We can then test the null hypothesis that $\rho = 0$ using this created distribution.

With a one-tailed test, we see no significant correlation for either centrality measure for the first exam but find a significant correlation between both betweenness centrality and degree centrality and exam performance on the second exam. With our understanding of how students changed their studying patterns between the first and second exam, this finding is rather interesting. Given the opportunity to revise their network positions after some experience in the course, we find a social influence on exam performance.

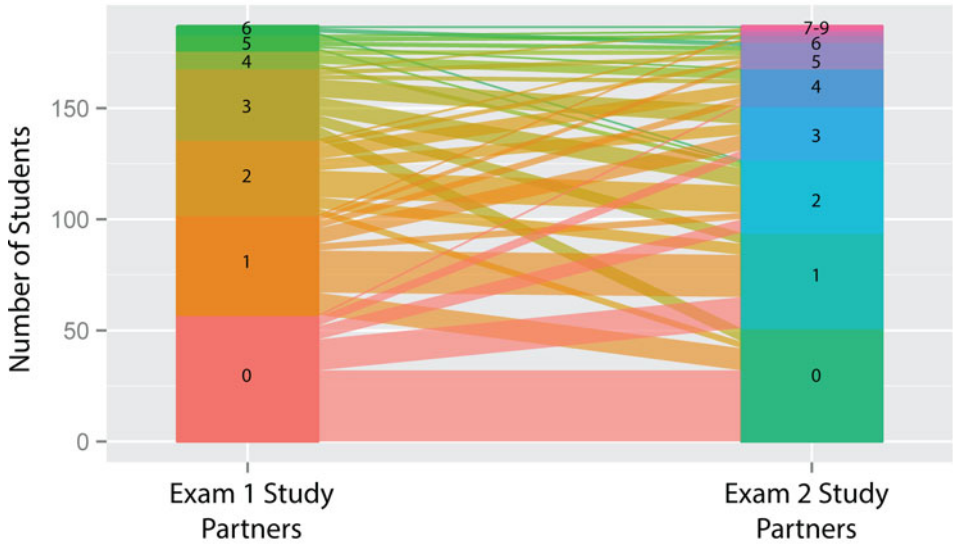


Figure 3. A parallel coordinate plot tracking changes in number of study partners from the first and second exam. The number of students whose number of study partners changes from exam 1 to exam 2 is denoted by the line widths.

Table 5. Results from a permutation correlation test between degree and betweenness centrality and student exam performance

Centrality measure	Exam[a]	Correlation	Pr ($\rho \geq$ obs)
Degree centrality	Exam 1	0.072	0.164
	Exam 2	0.212	0.001
Betweenness centrality	Exam 1	0.031	0.337
	Exam 2	0.117	0.048

[a]Significance is seen between both types of centrality for the second exam, but not the first.

Because we are unable to control for student effort (a measure notoriously hard to capture), we are unable to discern whether study effort confounds our finding and makes causality vague. Regardless, the association is interesting and exemplifies the sort of direction researchers can take with SNA.

More Complex Models of Network Formation

The methods we present here only scratch the surface of those available and largely focus on fairly descriptive techniques. A variety of approaches exists to explore the structure of networks, to infer the processes generating those structures, and to quantify the relationships among those structures and the flow of entities on them, with a recent trend away from description and toward more inferential models. For instance, past decades saw great interest in specific models for network structure (e.g., the "small-world" model) and their implications (Watts and Strogatz, 1998). A host of methods exist for identifying endogenous clusters in networks (e.g., study groups) that are not reducible to exogenous attributes like major or lab group; these have evolved over the decades from more descriptive approaches to those involving an underlying statistical model (Hoff *et al.*, 2002). Recently, more general approaches for specifying competing models of network structure within the framework and performing model selection based on maximum likelihood have become feasible. These include actor-oriented models, implemented in the RSiena package (Snijders, 1996), and exponential-family random graph models, implemented in statnet (Wasserman and Pattison, 1996; Hunter *et al.*, 2008). One recent text that covers all of these and more, using examples from both biology and social science and with a statistical orientation, is *Statistical Analysis of Network Data* by Kolaczyk (2009).

FUTURE DIRECTIONS

Within education research, we are just beginning to explore the kinds of questions that can benefit from these methods. Correlating student performance (on any number of measures) to network position is one clear area of research possibility. Specific experiments in pedagogical strategies or tactics, beyond having effects on student learning, may be assessable by differential effects on student network formation. For example, three groups of students could be required to perform a classroom task either by working alone, by working in pairs, or by working in larger groups. Differential outcomes might include grade results, future self-efficacy, or understanding of scientific complexity. The outcomes could be correlated with significant differences in the emergent network structures, strength of ties, and number of ties that emerge in a network of studying partnerships. Controlled experimentation with social constraints and network data would provide insight on advantages or disadvantages of intentional social structuring of class work.

Educational networks are not exclusive to students; relational data between teachers, teacher educators, and school administrators may reveal how best teaching practices spread and explain institutional discrepancies in advancing science education.

Beyond correlational studies, major questions of equity and student peer perceptions will be a good fit for directed network analysis. Conceivably, network analysis can be used to describe the structure of seemingly ethereal concepts such as reputation, charisma, and teaching ability through the social assessment of peers and stakeholders. With a better understanding of the formation and importance of classroom networks, instructors may wish to understand how their teaching fosters or hinders these networks, potentially as part of formative assessment. Reducing the achievement gaps along many demographic lines is likely to involve social engineering at some granular level, and the success or failure of interventions represents rich opportunities for network assessment.

SUMMARY

In this primer, we have analyzed two study networks from a single classroom. We have discussed collection of both nodal and relational data, and we specifically focused on keeping surveys brief and simple to process. We transitioned these data to a sociomatrix form for use with SNA software in a statistical package. We analyzed and interpreted these data by visualizing network data with sociographs, looking at some basic network measurements, and testing for associations between network position and a nodal attribute. Data were interpreted both as a description of a single network and as a longitudinal time lapse of community change. For this project, data collection required a single field of data from the institution registrar and a single survey question asked longitudinally on just two occasions. With a relatively small investment in data collection we can rigorously assess hypotheses about interactions within our educational environments.

It bears repeating: this primer is intended as a first introduction to the power and complexity of educational research aims that might benefit from SNA. Your specific research question will determine which parts of these methods are most useful, and deeper resources in SNA are widely available.

In short, networks are a relatively simple but powerful way of looking at the small and vital communities in every school and college. Empirical research of undergraduate learning communities is sparse, and instructors are thus limited to anecdotal evidence to inform decisions that may impact student relations. We hope this primer helps to guide educational researchers into a growing field that can help investigate classroom-scale hypotheses, and ultimately inform for better instruction.

FURTHER RESOURCES

For readers whose interest in SNA has been piqued, there are numerous resources to use in learning more. We provide some of our favorites here:

Carolan, Brian V. *Social Network Analysis and Education: Theory, Methods & Applications*. Los Angeles: Sage, 2013.

Kolaczyk, Eric D. *Statistical Analysis of Network Data: Methods and Models*. Springer Series in Statistics. New York: Springer, 2009.

Lusher, Dean, Johan Koskinen, and Garry Robbins. *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Structural Analysis in the Social Sciences 35. Cambridge, UK: Cambridge University Press, 2012.

Prell, Christina. *Social Network Analysis: History, Theory & Methodology*. Los Angeles: Sage, 2012.

Scott, John, and Peter J. Carrington. *The Sage Handbook of Social Network Analysis*. London: Sage, 2011.

Wasserman, Stanley, and Katherine Faust. *Social Network Analysis: Methods And Applications*. Structural Analysis in the Social Sciences 8. Cambridge, UK: Cambridge University Press, 1994.

Other resources include the journals *Social Network Analysis* and *Connections*, both published by the International Network for Social Network Analysis; the SOCNET listserv; and the annual Sunbelt social networks conference.

ACKNOWLEDGMENTS

We thank Katherine Cook, Sarah Davis, Arielle DeSure, and Carrie Sjogren for fast and fastidious data-cleaning work. We thank Carter Butts for allowing us to use code originally written by him in our analyses. We also thank our funders at the National Science Foundation (NSF), IGERT Grant BCS-0314284 and NSF-DUE #1244847, for supporting this line of research. Finally, we greatly appreciate the discussions and moral support of the University of Washington Biology Education Research Group.

REFERENCES

Barabási A-L, Gulbahce N, Loscalzo J (2011). Network medicine: a network-based approach to human disease. Nat Rev Genet *12*, 56–68.

Bastian M, Heymann S, Jacomy M (2009). Gephi: an open source software for exploring and manipulating networks. ICWSM.

Batagelj V, Mrvar A (1998). Pajek-program for large network analysis. Connections *21*, 47–57.

Bonacich P (1987). Power and centrality: a family of measures. Am J Sociol *92*, 1170–1182.

Borgatti SP, Everett MG, Freeman LC (1999). UCINET 6.0, Version 1.00, Lexington, KY: Analytic Technologies.

Borgatti SP, Mehra A, Brass DJ, Labianca G (2009). Network analysis in the social sciences. Science *323*, 892–895.

Brewe E, Kramer L, Sawtelle V (2012). Investigating student communities with network analysis of interactions in a physics learning center. Phys Rev Spec Top Phys Educ Res *8*, 010101.

Brunello G, De Paola M, Scoppa V (2010). Peer effects in higher education: does the field of study matter. Econ Inq *48*, 621–634.

Bruun J, Brewe E (2013). Talking and learning physics: predicting future grades from network measures and Force Concept Inventory pretest scores. Phys Rev Spec Top Phys Educ Res *9*, 020109.

Butler D 2013. Sociomatrix Reader. https://github.com/djbutler/sociomatrix-reader (accessed 10 August 2013).

Butts CT (2008). Social network analysis with sna. J Stat Softw *24*, 1–51.

Carrell SE, Fullerton RL, West JE (2008). Does your cohort matter? Measuring peer effects in college achievement. National Bureau of Economic Research Working Paper 14032, Cambridge, MA.

Christakis NA, Fowler JH (2013). Social contagion theory: examining dynamic social networks and human behavior. Stat Med *32*, 556–577.

Couper MP, Tourangeau R, Conrad FG, Crawford SD (2004). What they see is what we get —response options for web surveys. Soc Sci Comput Rev *22*, 111–127.

Csardi G, Nepusz T (2006). The igraph software package for complex network research. InterJ, Complex Systems *1695*.

De Giorgi G, Pellizzari M, Redaelli S (2009). Be as careful of the company you keep as of the books you read: peer effects in education and on the labor market. National Bureau of Economic Research Working Paper 14948, Cambridge, MA.

Denzin NK, Lincoln YS (eds.) (2005). The Sage Handbook of Qualitative Research, Sage.

DeSimone J (2007). Fraternity membership and binge drinking. J Health Econ *26*, 950–967.

Duncan GJ, Boisjoly J, Kremer M, Levy DM, Eccles J (2005). Peer effects in drug use and sex among college students. J Abnormal Child Psychol *33*, 375–385.

Fenichel M, Schweingruber HA (2010). Surrounded by Science: Learning Science in Informal Environments, Washington, DC: National Academies Press.

Fink A (2003). The Survey Kit, Thousand Oaks, CA: Sage.

Fletcher JM, Tienda M (2008). High school peer networks and college success: lessons from Texas. Discussion Paper Series DP 2008-07, University of Kentucky Center for Poverty Research, Lexington.

Foster G (2006). It's not your peers, and it's not your friends: some progress toward understanding the educational peer effect mechanism. J Public Econ *90*, 1455–1475.

Freeman LC (1977). A set of measures of centrality based on betweenness. Sociometry, 35–41.

Freeman S, O'Connor E, Parks JW, Cunningham M, Hurley D, Haak D, Dirks C, Wenderoth MP (2007). Prescribed active learning increases performance in introductory biology. CBE Life Sci Educ *6*, 132–139.

Galaskiewicz J, Wasserman S (1993). Social network analysis—concepts, methodology, and directions for the 1990s. Sociol Method Res *22*, 3–22.

Granovetter MS (1973). The strength of weak ties. Am J Sociol *78*, 1360–1380.

Haak DC, HilleRisLambers J, Pitre E, Freeman S (2011). Increased structure and active learning reduce the achievement gap in introductory biology. Science *332*, 1213–1216.

Hake RR (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. Am J Phys *66*, 64.

Handcock MS, Hunter DR, Butts CT, Goodreau SM, Morris M (2008). statnet: software tools for the representation, visualization, analysis and simulation of network data. J Stat Softw *24*, 1548.

Hansen D, Shneiderman B, Smith MA (2010). Analyzing Social Media Networks with NodeXL: Insights from a Connected World, San Francisco, CA: Morgan Kaufmann.

Hoel J, Parker J, Rivenburg J (2005). Peer effects: do first-year classmates, roommates, and dormmates affect students' academic success. Paper presented at the Higher Education Data Sharing Consortium Winter Conference, Santa Fe, NM (accessed 14 January 2005).

Hoff PD, Raftery AE, Handcock MS (2002). Latent space approaches to social network analysis. J Am Stat Assoc *97*, 1090–1098.

Hunter DR, Handcock MS, Butts CT, Goodreau SM, Morris M (2008). ergm: a package to fit, simulate and diagnose exponential-family models for networks. J Stat Softw *24*, nihpa54860.

Johnson TS (2008). Qualitative research in question—a narrative of disciplinary power with/in the IRB. Qual Inq *14*, 212–232.

Kolaczyk ED (2009). Statistical Analysis of Network Data: Methods and Models, New York: Springer.

Krackhardt D (1999). The ties that torture: Simmelian tie analysis in organizations. Res Sociol Organizat *16*, 183–210.

Lyle DS (2007). Estimating and interpreting peer and role model effects from randomly assigned social groups at West Point. Rev Econ Statist *89*, 289–299.

Marmaros D, Sacerdote B (2002). Peer and social networks in job search. Eur Econ Rev *46*, 870–879.

Martin DG, Inwood J (2012). Subjectivity, power, and the IRB. Prof Geogr *64*, 7–15.

McPherson M, Smith-Lovin L, Cook JM (2001). Birds of a feather: homophily in social networks. Annu Rev Sociol *27*, 415–444.

Morris M (2004). Network Epidemiology: A Handbook for Survey Design and Data Collection, Oxford, UK: Oxford University Press.

Newman ME (2001). The structure of scientific collaboration networks. Proc Natl Acad Sci USA *98*, 404–409.

Nieminen J (1974). On the centrality in a graph. Scand J Psychol *15*, 332–336.

Oakes JM (2002). Risks and wrongs in social science research—an evaluator's guide to the IRB. Evaluation Rev *26*, 443–479.

O'Sullivan DW, Copper CL (2003). Evaluating active learning: a new initiative for a general chemistry curriculum. J Coll Sci Teach *32*, 448–452.

Page L, Brin S, Motwani R, Winograd T (1999). The PageRank citation ranking: bringing order to the Web. Stanford InfoLab technical report, Stanford University, Stanford, CA.

Porter SR, Whitcomb ME, Weitzer WH (2004). Multiple surveys of students and survey fatigue. New Dir Inst Res *2004*, 63–73.

Ripley R, Snijders T, Preciado P (2011). Manual for SIENA Version 4.0 (version December 11 2011), Oxford, UK: University of Oxford, Department of Statistics, Nuffield College. www.stats.ox.ac.uk/siena.

Sabidussi G (1966). The centrality index of a graph. Psychometrika *31*, 581–603 (accessed 10 August 2013).

Sacerdote B (2001). Peer effects with random assignment: results for Dartmouth roommates. Q J Econ *116*, 681–704.

Scott J, Carrington PJ (2011). The SAGE Handbook of Social Network Analysis, London: Sage.

Smith MA, Shneiderman B, Milic-Frayling N, Mendes Rodrigues E, Barash V, Dunne C, Capone T, Perer A, Gleave E (2009). Analyzing (social media) networks with NodeXL. Proceedings of the Fourth International Conference on Communities and Technology, 255–264.

Snijders TA (1996). Stochastic actor-oriented models for network change. J Math Sociol *21*, 149–172.

Stinebrickner R, Stinebrickner TR (2006). What can be learned about peer effects using college roommates? Evidence from new survey data and students from disadvantaged backgrounds. J Public Econ *90*, 1435–1454.

Wasserman S, Faust K, Galaskiewicz J (1990). Correspondence and canonical-analysis of relational data. J Math Sociol *15*, 11–64.

Wasserman S, Pattison P (1996). Logit models and logistic regressions for social networks. I. An introduction to Markov graphs and *p*. Psychometrika *61*, 401–425.

Watts DJ, Strogatz SH (1998). Collective dynamics of "small-world" networks. Nature *393*, 440–442.

West JD, Bergstrom TC, Bergstrom CT (2010). The Eigenfactor Metrics^{TM}: a network approach to assessing scholarly journals. Coll Res Libr *71*, 236–244.

Wilson J (2007). Peer effects and cigarette use among college students. Atl Econ J *35*, 233–247.

Zimmerman DJ (2003). Peer effects in academic outcomes: evidence from a natural experiment. Rev Econ Statistics *85*, 9–23.

HIGHLIGHT:

The authors introduce basic concepts in SNA, along with methods for data collection, data processing, data analysis, and conduct analyses of a study relationship network. Also covered are generative processes that create observed study networks and practical issues, such as the unique aspects of human subjects review for network studies.

Research in Science Education

Development of a grounded survey to measure student engagement in large active-learning classrooms

--Manuscript Draft--

Manuscript Number:	RISE-D-15-00190
Full Title:	Development of a grounded survey to measure student engagement in large active-learning classrooms
Article Type:	Manuscript
Keywords:	STEM education, qualitative and quantitative research, sociocultural learning, active learning, survey, engagement
Corresponding Author:	Benjamin Wiggins, MS University of Washington Seattle, WA UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Washington
Corresponding Author's Secondary Institution:	
First Author:	Benjamin Wiggins, MS
First Author Secondary Information:	
Order of Authors:	Benjamin Wiggins, MS
	Sarah L Eddy, PhD
	Leah S Wener-Fligner, BS
	Daniel Z Grunspan, MS
	Jerry P Timbrook, PhD
	Karen Freisem
	Alison J Crowe, PhD
Order of Authors Secondary Information:	
Funding Information:	National Science Foundation (US) (DUE1244847) — Mr Benjamin Wiggins
Abstract:	We have developed a survey to analyze student experiences in a large undergraduate biology course using a mixed-method approach. Survey items address themes that emerged from direct student talk and that were identified by students as essential to their learning in this setting. Face validation was conducted through a series of qualitative methods with direct student input. Statistical analysis of survey results provide support for two qualitatively-derived narratives of student experiences within large active learning classrooms. In one narrative, student engagement in active-learning STEM courses was explained by three major factors: value of group activities, personal effort, and instructor influences. These three scales emerged independently through open-ended inquiry, and from exploratory factor analysis. In the second narrative, quantitative cluster analysis of survey data found that students' self-described roles in groupwork fell into four distinct groups: Disengaged, Listener-Only, Leader/Explainer, and Broadly Engaged. Both narratives identify important characteristics of sociocultural learning in large-enrolment active learning STEM courses.

Development of a grounded survey to measure student engagement in large active-learning classrooms

Abstract

We have developed a survey to analyze student experiences in a large undergraduate biology course using a mixed-method approach. Survey items address themes that emerged from direct student talk and that were identified by students as essential to their learning in this setting. Face validation was conducted through a series of qualitative methods with direct student input.

Statistical analysis of survey results provide support for two qualitatively-derived narratives of student experiences within large active learning classrooms. In one narrative, student engagement in active-learning STEM courses was explained by three major factors: value of group activities, personal effort, and instructor influences. These three scales emerged independently through open-ended inquiry, and from exploratory factor analysis. In the second narrative, quantitative cluster analysis of survey data found that students' self-described roles in groupwork fell into four distinct groups: Disengaged, Listener-Only, Leader/Explainer, and Broadly Engaged. Both narratives identify important characteristics of sociocultural learning in large-enrolment active learning STEM courses.

Keywords

STEM education, qualitative and quantitative research, sociocultural learning, active learning, survey, engagement

Introduction

STEM undergraduates navigate complex social learning environments as they develop expertise in tasks crucial to careers in science and technology. The design and efficacy of these

environments are of increasing interest to educators as social, practice-based, and active learning techniques are called upon to improve STEM learning and diversity(Woodin et al. 2010; Holdren 2013; Universities 2011). A number of observation tools exist to measure instructor and student behaviour in the classroom (Hora 2014; Smith 2013; Piburn 2000; Sawada 2002; Eddy 2015; Lane and Harris 2015). However, while methods are available for the study of student outcomes and instructor practices, a tool is not available for analysis of the student experience of collaborative active learning within STEM classrooms environments. We report a series of investigations into the complicated social learning environment of a large-enrolment active-learning STEM undergraduate course. Specifically, we focus on student experiences with activities based in small group work. Beginning with open-ended and deep qualitative approaches, our research questions evolved through a combination of interacting mixed methods. Our end result is a Student Engagement Survey for Active Learning Environments (SESALE) grounded in real student experiences that has been validated for use in a large active learning STEM classroom. In this paper, we provide a description of the process undertaken and the triangulated understanding achieved by analyzing these learning environments from a mixed methods standpoint.

Review of Relevant Literature

Post-secondary STEM education

Current post-secondary STEM education methods and institutions have great room for improvement in guiding, motivating and mentoring all students in the complex tasks needed to negotiate science careers (Sciences 2013). STEM education in colleges and universities is often competitively oriented, stocked with high-achieving students, and expected to produce science majors capable of taking on highly technical jobs. While training of these top students is a major factor in a national economy(Holdren 2013), the ‘leaky pipeline’ of STEM education indicates that

far too many talented and interested students are lost along the way (Dasgupta and Stout 2014; Committee on Science 2007; Drew 2011). These students come disproportionately from backgrounds historically underrepresented in STEM and report social threats and concerns as major factors behind their decisions to leave STEM education (Steele 1997; London et al. 2012; Graham et al. 2013; Seymour and Hewitt 1999). This system has overseen the professional training of skilled nurses, doctors, researchers and scientists; however, it is also reasonable to view the traditional STEM classroom as a selector of talented individuals instead of an incubator for emerging STEM talent (Blickenstaff 2005). The traditional mode of STEM education in Western colleges and universities is passive 'sage on a stage' lecturing. Recently, in response to national calls to adopt curricula focusing on skills and identity development, there has been a shift toward using more student-centered instructional approaches collectively referred to as "active learning". (Universities 2011; Wieman 2014; Suchman 2014; Waldrop 2015).

Active learning and STEM education

Active learning is a diverse set of practices intended to engage students, scaffold professional skills, and create opportunities for feedback to students. Active learning is not new (Dewey 1938) but has only recently been well defined as a continuum of practices involving student creation and co-organization of knowledge (Chi and Wylie 2014). Recent studies have demonstrated the effectiveness of active learning both in terms of learning outcomes (Freeman et al. 2014) and important issues like retention and identity creation (Kvam 2000; McConnell et al. 2003; Haak et al. 2011). The call to activate post-secondary teaching practices has been loud and uniform (Linn 2003; Mayer 2011; Committee on the Status 2012; Rosenberg et al. 2006; Wieman 2014). Active learning is a well-established best practice in K-12 STEM education (Engle 2002; Minner 2010; Council

2000, 2013). It is the study of these pedagogies applied to post-secondary learning contexts and student communities that intrigues us.

The social context of active learning and teaching

Introductory STEM classrooms are often larger than the hometowns of some of the students they serve. Within classrooms, students navigate diverse social opportunities and threats as they grow towards complex and sometimes competing goals (Grunspan et al. 2014). These social links can include student-instructor relationships (D'avanzo 2013), outside friendships that bridge into the classroom, working interactions with peer strangers, and often navigation of multiple levels of authority (from teaching assistants to co-instructors to sole professors). Social links may impact learning more in active learning classrooms that put the onus on students and student groups to do the work of education (Esmaeili 2014; Lorenzo et al. 2006). These working relationships may be complicated by diverse student assumptions and assessments of group work; this issue is not well studied in active learning contexts. Increasing the diversity of student tasks within the classroom may benefit students, but there is little literature to inform how diversity influences the learning environment(Nasir and Hand 2006; Kurth et al. 2002). To understand the influence of student experiences on classroom outcomes, research tools must assess the impact of multiple levels of social contact on individual students. These complicated questions are likely to inform curricular design as well as best practices for instructor language, goals, and training.

Motivation and Student Engagement

There is a growing recognition that student learning in the post-secondary classroom is influenced by non-cognitive factors such as motivation, attitude and a sense of belonging (Lovelace and Brickman 2013). Although several different models of motivation have been developed, one

shared feature is that perception of a task's value influences level of engagement (CS. Hulleman et al. 2008; Svinicki 2004). The expectancy-value model provides a theoretical framework to explain the positive relationship between perceived value of a task and level of engagement (J. Eccles 1983; J. S. Eccles and Wigfield 2002). In an educational context, this model predicts that tasks perceived as having value by the student will promote higher levels of engagement (J. S. Eccles and Wigfield 2002; Hug et al. 2005) as well as academic performance (C.; Hulleman et al. 2010). For example, students will be more motivated to engage in tasks that they consider either enjoyable or useful and relevant to their lives (J. Eccles 2005). Importantly, individual student characteristics such as prior interest in a topic (Durik 2007) or low performance expectations (C.; Hulleman et al. 2010) can impact how much value students place on an activity and therefore their motivation.

A number of excellent inventories have been developed for measuring student engagement, motivation and attitude in the post-secondary biology classroom (Handelsman et al. 2005; Pintrich 1993; Semsar et al. 2011). However, these surveys are either focused on only one aspect of a student's experience such as personal motivation or interest (Pintrich 1993) or geared toward assessing student engagement in a traditional, lecture-based college classroom (Handelsman et al. 2005).

There are also several classroom observation tools that measure instructional practices or student behaviour (Eddy 2015; Sawada 2002; Smith 2013; Hora 2014; Lane and Harris 2015). These tools are designed to assess what the instructor and/or students are doing in the classroom. For example, the RTOP includes items such as "students were reflective about their learning" and "the lesson promoted strongly coherent conceptual understanding" (Sawada 2002). PORTAAL measures whether or not instructors are implementing evidence-based best practices such as giving students opportunities to practice logic development (Eddy 2015). By necessity, these tools are limited to measuring only those classroom features that can be viewed by an external observer. An

observer may perceive that students are actively engaged in a thought-provoking activity, but those overt signs of engagement may not accurately reflect a student's self-perceived level of engagement (Pritchard 2008).

We have intentionally taken a different approach. Our goal was to develop a survey that would quantify student perceptions regarding in-class group work. In the long term, this tool could then be used to gain perspective on how student's unique characteristics, such as gender, ethnicity, and incoming grade point average influence their experience during an active learning exercise. The extent to which students do or do not engage, and why they engage or fail to do so, is important for classroom design, instructor best practices, and for comprehensive assessment of the values of different active learning strategies.

Goals and Theoretical Framework

Active learning improves student outcomes, but the modes and implementation of active learning vary widely, even within very similar classrooms (Freeman et al. 2014). For whom, with what instructional practices, and in what ways does active learning work best remain open areas for needed research? Our goal was to develop a tool that could be used to explore how the implementation of active learning influences the student experience with a particular activity. Our scope is at the ecological level of the classroom community, and the intent of this tool is not to delve into a deeper psychometric analysis of engagement as a mental construct. Specifically, we employ a mixed-methods approach to: 1) Identify what aspects of an in-class exercise influence student experiences with the activity and their willingness to engage in the task, and 2) Develop and validate a survey based on these ideas to measure factors that influence student engagement. Grounded qualitative work gave rise to survey creation to better assess themes that emerged from

student interviews and focus groups. The final survey created is a tool that can be used to explore aspects of active learning and their influence on student experience.

We take a sociocognitive framework into this work. Learning is implicitly social: STEM students learn, live and balance disparate goals within complex communities that include STEM classrooms in addition to families and other social architectures. We attempt to conceptualize our learners as being partially apprenticed into a particular cultural expertise; in this case, modern biological science (Nasir and Hand 2006). While distantly abstracted from the master-and-student model, large lecture classrooms are still a gathering point for apprenticeship of learners and a way for them to practice a set of cultural tools (Lave and Wenger 1991). Our sociocultural lens does not ignore the individual learning that often occurs with motivated students, but rather frames it within the greater community of people, problems, roles, and practices that gives meaning to that learning.

While quantitative analyses of individualized student learning are relatively rapid and methodically objective, they inherently instill bias towards existing stereotypes and cultural norms. Large-scale statistics like gender, ethnicity, race, and socioeconomic status frequently mask the diversity and interplay of human practices (Gutiérrez and Rogoff 2003; Hawkins and Pea 1987; Geertz 1973). Furthermore, cultural aspects of the classroom like instructor language, solo status in group work, identity/intersectionality, or group work niches may have profound impacts on students' willingness and inspiration to attempt difficult STEM learning tasks. These aspects may be ignored without framing the learner as a community member first. A better understanding of these social linkages that impact learning is an important goal for us. In this study, we seek to balance the power and scope of quantitative methods that individually position the learner with qualitative methods that more deeply frame learners within their own cultures. Listening to learner narratives and perceptions as a starting point is an incomplete but useful step in this direction.

Methodology

Here we describe the series of operations through which we developed and gathered validity evidence for our survey. Our process was guided by the 3-stage validation framework established by Benson (1998), and included substantive, structural and external validation stages as described below. While our goal was to develop a tool for both researchers and instructors interested in measuring student perception of group work in similar environments, the process of qualitative survey development and validation was itself a rich source of data about what our students are thinking and doing during small group work in the classroom.

Figure 1. Overall development scheme for the SESALE survey. Final item wording was approached through a series of validation, testing and development steps. As the project progressed (downwards, in this diagram) the SESALE reached final form.

Phase I: Development of constructs to be measured in survey

In this phase, consistent with the substantive stage described by Benson (1998), we define the theoretical basis for the constructs to be measured by the survey.

Individual interviews and focus groups

The students who participated in this study were enrolled in the second course of a three-course introductory biology sequence at a research university in the Pacific Northwest over a three-year period. Class size ranged from 150 (summer courses?) to over 770 (W2015?). Students enrolled in this course were primarily sophomores and juniors, with females making up about 60% of the classroom population. Of the students enrolled in the course over the period of this study, 40-

45% were identified through the University registrar as Asian Americans, 40-45% White Americans, 6-7% International students, 5-6% Latin@ Americans, and the remainder of the population was comprised of Black Americans, Hawaiians, Pacific Islanders and Native Americans. Different iterations of this course were taught by different instructors, but always included high levels of active learning in the form of small group and whole class discussions. In addition, about once a week students worked in self-selected groups of 2-3 to complete longer 30-40 minute activities focused on a particular topic in molecular or cellular biology. These longer activities included a variety of active learning strategies but all involved students working collaboratively.

We initially used direct student interviews to answer our early research questions centered around how students experience different active learning approaches in the classroom. Students who had recently completed one of the longer collaborative activities were recruited to provide feedback on their experience. We hoped that students would describe which specific aspects of the different active learning strategies were most engaging or most helped with their learning. Initially, interviewers posed questions focused on worksheet wording, worksheet enjoyment, instructor identity, and student confidence. However, students in these interviews and focus groups kept returning to different themes. These student-generated themes focused on group dynamics, instructor language, and process-oriented goals. We assumed that the lack of student interest in our original questions indicated the small relative influence of our original topics on student experience. Therefore, we changed focus and implemented an open-ended interview strategy (Rubin HJ & Rubin 2005) to follow these new prevalent themes.

Open-ended interviews were used to elicit student perceptions of their experience in large, active STEM classrooms. A typical 50-minute interview included a maximum of three short, intentionally-broad questions (for example: What was important about today's class? What helped your learning? Did anything make learning harder?). Interview prompts were iteratively improved

through the series of interviews. Students were encouraged to speak on any topic that they felt relevant, and our primary goals were a) to increase the percentage of interview time containing free student talk and b) to encourage broadly framed discussion of students' experiences and perceptions in order to broadly capture themes that were important to students. Follow-up questions were unscripted but consistently required that students explain their reasoning as deeply as possible. Individual interviews were replaced by small focus group interviews of 2-5 students, helping to create micro-communities of review and reflection from which common experiences could be analyzed.

In total, 25 students were involved in a series of nine interviews and focus groups over the course of Autumn 2012 and Winter 2013. The numbers of student participants in each of nine different interview or focus group were (in temporal order from earliest to last): 1, 1, 4, 2, 3, 5, 3, 3, and 3. Students were recruited randomly by email from the class of 750+ students. Research participants encompassed similar characteristics to the class as a whole in terms of ethnicity, race, gender and final course grades (from 0.8 to 3.8 on a 4.0 grade scale). Interviews and focus groups were transcribed using pseudonyms for further coding and exploration, using a grounded theory approach(Glaser and Strauss 1967; Corbin and Strauss 2014).

Coding and identification of emergent themes

While interviews and focus groups were a rich source of students' experiences, it was only through iteratively developed coding that cohesive themes emerged from the data. An initial set of codes was used to analyze early transcripts. Coders met to discuss consensus definitions of codes as well as any student experiences that did not fit into any particular code. Codes were adjusted, older transcripts were recoded, and newer transcripts subjected to the latest coding regime. Eventually, a steady state set of codes developed. For example, 'grade motivation' was amalgamated into

'motivation' after a single iteration, and then finally positioned within 'motivators' when 'motivation' was subdivided into a motivator/motivating factor dichotomy. This better fit the majority of student talk about grades as an influence on classroom experience. The final codes used are shown in Table 1. Lines of text were used as an imperfect approximation of frequency. If a single statement on Code 1 spanned 3 lines of text, then this statement was counted as '3' in the category of Code 1. Three major strands emerged from this coding: statements that focused on the Instructor, the Value of the Group Activity, and Personal Involvement.

Table 1. Descriptions and Examples of Emergent Codes.

Category	Code Title	Description	Prevalence (out of 3,267 lines of text)	Representative Quote
Instructor	Instructor Effort	Describes student perceptions of the effort spent by instructors both in- and outside of the classroom	270 (8.3%)	"I appreciate how he tries to make it [a] less-than-500 person class… I introduced myself, and he remembered my name every single time after that, didn't forget. And I think just those little things… show that he's really invested in teaching and invested in helping us succeed too."
Instructor	Modes of Exam Practice	Involves the multiple pathways of preparation for difficult high-stakes summative assessments	366 (11.2%)	"Gets me used to seeing that type of question…where it's just like 'answer these' and being scared because it's like a 3 page thing…it's terrifying. But it gets that first terrifying 3 page thing out of the way."
Instructor	Motivators	Student goals or potential negative consequences that influence motivation to engage in the course.	334 (10.2%)	"…my other classes, there aren't reading quizzes so I'm less motivated to keep up … when [the instructor] has the reading quizzes it kind of forces you to know the material."
Value of the group activity	Sociocognition	Awareness of and/or actions based on the perceived thoughts of peers	1089 (33.3%)	"I personally struggle with the clickers, because I always sit by people who don't want to talk to me… and I don't follow through [by] asking"
Value of the group activity	Language Barriers	Difficulties in classrooms related to language background and usage	144 (4.4%)	"For example, one of my classmates … he talks in a more understandable language for us. But when he answers the questions in class, and he answers them a lot, he'll pull out terms that weren't even in the reading...I think he's just trying to seem impressive."
Personal	Metacognition	Awareness and cultivation of one's own thoughts and thought processes.	1179 (36.1%)	"I'm also more of a slow thinker… I need to really read through the question, I don't like to be rushed…So a lot of times it is a time crunch for me, where I rush and I start making more and more mistakes."
Personal	Motivational effectors	Factors that influence the force and/or applicability of motivators.	1134 (34.7%)	"I've been putting so much time in… I honestly have been putting all my time into bio and forgetting my other classes…That's my weak point, because I can't see it being applied for me personally."
Personal	Ownership	Factors that regulate whether aspects of the course fall within the students' domain of influence and obligation.	803 (24.6%)	"My teacher said I should read this, but I don't think I'm going to…but with this you're really forced to focus more during lecture for the clicker questions"

Phase II: Initial development and face validation of survey items

In this phase we develop the initial items for the survey and gather evidence supporting the face validity of the constructs we identified in Phase I. See Figure 1 for an overall development scheme, and see Figure 2 for an example of the development of a single item.

Initial survey design

Intrigued by the depth of information captured by the interviews and focus groups, we pursued a scalable survey to examine the prevalence of these themes in large courses. Our motivation was not to simplify the data, but rather to search for broader application of these themes in the larger student population. We developed a survey in order to quantify these qualitatively-derived insights.

A subset of our research team wrote an initial set of twenty-six survey items through a short series of group-writing tasks and editing sessions (Dillman 2014). We focused our items on those themes that were most prevalent in student talk (see Table 1). Similar to the CLASS instrument, the questions included in the Student Engagement Survey for Active Learning Environments (SESALE) are based on themes that arose from student focus groups, not questions pre-determined by the researchers (Semsar et al. 2011).

Think-alouds for coherence with student language

Face validation of the initial questions was provided through seven individual think-alouds. Think-alouds are structured interviews intended to elucidate participant talk as they complete a task or access a text (Gubrium and Holstein 2002). We recruited student participants to engage in guided readings of the survey. An interviewer asked the students to read these nascent survey items silently,

then read them aloud. Students were asked to answer the survey item out loud as well as to justify their reasoning for their answer. For each item, the interviewer prompted further student explanation to elucidate their thought processes in reading the survey item. Care was taken to avoid invalidating student processes (for example, students were neither chided nor praised if their explanations went beyond the written text or indicated that they did not read the entire item). Finally, students were asked to identify any problematic parts of the item and to suggest alternative language if applicable. These think-aloud responses were collected for all items and returned to an editing group comprised of a subset of our researchers. The editing group altered, removed, or split items with mutual goals of student authenticity and fidelity to the original qualitative emergent themes. Each iteration of the think-aloud-and-editing process accessed new student participants. Three versions of the survey were created and analyzed before a subjective equilibrium on item wording was reached. Significant editing based on direct student talk was an important step in developing our survey to reflect student language and experiences.

Initial Survey piloting

The initial survey was piloted in a large enrolment Biology course. The survey was given twice to all students in the course; 227 and 218 of the 264 enrolled students completed the first and second iterations, respectively. Students received a nominal number of course points for completing the survey. The same survey was completed by students within 18 hours after, and with reference to, extremely different class sessions in terms of instructor, topic, activity, and time in the quarter. Our hope was to show some dynamic range in the answers that students gave, and that seemed to be supported in that 12 of the 39 items showed significant differences at the $p < 0.01$ level by two-tailed Chi-squared testing of binary answer choices. However, these initial survey items were problematic when analyzed by factor analysis. Instead of items factoring by reasonable topics, they

factored by the way in which the individual question was asked. For example, questions asked in the negative factored together even on seemingly unrelated topics. Anecdotally, students reported confusion with items and an inability to accurately express their views. In addition, several items appeared to be answered with nearly identical responses even though they address relatively unrelated topics. To address these issues we revised and refined the SESALE using best practices in survey design (Dillman et al. 2014) as described below.

Face Validity

Through consultation with a collaborator experienced in survey design, we undertook several steps to improve our survey items (Figures 1 and 2). Together, we identified and addressed a few face validity issues. For example, we corrected double-barrelled questions to ensure that we were asking respondents about only one construct per survey item. In addition, questions with ambiguous wording were made more explicit. To improve the richness of the data collected by the survey, we increased the number of response alternatives for each question. For example, simple binary responses (Yes/No or Agree/Disagree) were changed to a 6-point Likert scale (Strongly Agree to Strongly Disagree). We were also able to shorten the survey by removing items that measured the same general construct. For example, the initial draft of the survey asked respondents to rate the question: *Explaining the material to my group improved my understanding of it.* Later, respondents were also asked: *During the PCR activity in class today I benefited from explaining ideas to other students*. We removed the latter.

The revised survey contained five items asking about the instructor, eight items asking about the student experience with the group activity, and seven items asking about student engagement with the material during the activity. Additionally, to gain information on the various roles students played during the group work, we included five Roles Questions. Three demographics questions

were also included at the beginning of the survey to control for group size, prior experience with active learning, and friendships within the working group. With this newly reformatted survey in hand, we initiated cognitive testing and face validation of open-ended student responses.

Cognitive testing

The goal of cognitive testing was to identify any conflicting or confusing interpretations of survey items that might lead to students giving the same answer for multiple reasons (Willis 2005). Six participants were involved, and all had sophomore/junior standing. Participants were randomly recruited by email from a large list of similar biology students. Students took the entire survey in paper form, and then afterwards worked as a group to 1) read each item aloud and come to a consensus meaning, then 2) discuss any possible alternative interpretations with the help of interviewer's repeated follow-up prompts. For a few items with multiple significant interpretations, participants voted on which was more salient. Facilitator notes were kept on participant interpretations, and participants also submitted their paper surveys on which they highlighted words that were problematic for them individually. The research team discussed all interpretations item-by-item.

All SESALE items were understood by participants during cognitive testing. Seventeen of the 23 items were understood clearly enough that participants did not indicate any significant possible alternative interpretations indicating a need for any editing. Of the six items indicated for some review, only one ('*One group member dominated discussion during today's group activity'*) was indicated for multiple and contradicting interpretations. Several students assumed that the question implied that domination of a group was a negative experience for other students in the group; this view was challenged by two participants. The question was not changed, as researchers had previously identified negative consequences of group domination as the intended focus of the

item. Five other survey items were identified by students as having possible alternative interpretations, but in all cases the participants unanimously agreed on the most salient interpretation. These salient interpretations matched the goals and intentions of the item in each case. One item (*The instructor put a good deal of effort into my learning for today's class*) was indicated by participants to be a conglomeration of several different constructs. This item was intentionally of large scope, so the inclusion of multiple constructs into 'effort' was appropriate.

Figure 2. Development of an example item. This question was iteratively improved through the qualitative steps discussed above. Several examples of specific improvements are noted.

Large-scale face validation of survey items

As an additional measure to ensure that students were interpreting the final questions as intended, we asked students to provide written explanations for choosing their selected answers after completing an online version of the survey. All 397 students in a large-enrolment Biology course completed the online survey. Questions were randomized so that each student received the questions in a different order. At the end of the survey, students were asked to explain their reasoning for answering two randomly selected questions from the survey that they had just completed. 383 students completed these open-ended questions, providing us with approximately 40 open ended responses per item.

Three researchers independently coded the open-ended responses. For items in which cognitive testing had previously shown multiple interpretations, these established categories were used as initial codes. After independent coding by three researchers, consensus was reached on whether or not each item was asking about the themes that the researchers intended and that previous qualitative data had indicated most important. The large majority of student responses to

each of the items assessed had been previously described in cognitive testing. Only one survey item was identified as problematic. This item ("I engaged in critical thinking during today's group activity") was removed due to ambiguity in student open-ended responses, indicating that students had variable interpretations of the term "critical thinking". Student explanations of why they answered this question the way they did ranged from not understanding the activity (e.g. "A lot of it was confusing") to making connections with previous course content (e.g., "we taught ourselves with what we learned throughout the quarter" and "it caused me to try and delve more into my understanding of the material being taught this week"). As there is continued debate even among experts as to the definition of critical thinking, we decided to remove this item from further analysis.

Phase III: Piloting and Exploratory Factor Analysis

In this phase, consistent with Benson's (1998) structural stage, we examine correlations between survey items and perform exploratory factor analysis to assess the internal consistency of the constructs measured by the survey. Data for these analyses came from the online administration of the survey in a large introductory biology classroom described in Phase II.

Refinement of scales

The survey contained 20 Likert-like items intended to measure three constructs of interest: Influence of (a) the instructor, (b) the activity, and (c) groupmates. As described in the section on large-scale validation above, one item ("I engaged in critical thinking during today's group activity") was removed due to ambiguity in students' open-ended responses.

Next, a pair-wise comparison of the correlations between the questions was conducted to identify non-useful items. Consistently low inter-item correlations indicate low potential to contribute to

measurement of underlying constructs as part of a cohesive multi-item scale. Thus, items with consistently low inter-item correlations (Pearson's correlation coefficient $r < 0.3$ for at least 80% of correlations) were removed (Tabachnick and Fidell 2001). This resulted in the deletion of one item ("One group member dominated discussion during today's group activity"). We conducted several iterations of exploratory factor analysis with the remaining 19 items. All exploratory factor analyses were conducted using the psych package in R (Revelle 2015). An oblique (promax) rotation was used, as we hypothesized our three constructors were correlated with each other. Students with missing responses were excluded from the analysis. This resulted in 17 students (4.5% of sample) being dropped from the analysis.

There was evidence in support of both a three and four factor solution. We ultimately chose to focus on the 3-factor solution because the 4-factor solution had (1) multiple instances of strong crossloading of items and (2) no clear theoretical basis for distinguishing two of the factors that are collapsed into one factor in the 3-factor solution. In both solutions one item did not seem to load on any factor ("I knew what I was expected to accomplish during the … activity") and one item loaded only weakly in the 4-factor solution and not at all in the 3 ("I felt comfortable with my group"). These two items were removed from the final factor analysis. If they are of particular interest to an instructor, we recommend analyzing them individually along with the third item that was not correlated with the rest of the survey: "One group member dominated discussion during today's group activity".

We conducted a final iteration of the 3- factor solution for the exploratory factor analysis with the remaining 16 items. Factor loadings for the final survey were consistently above the suggested minimum cutoff of 0.32 (Tabachnick and Fiddell, 2001) with 14 of the items exhibiting factor loadings above 0.6. Cronbach's alpha for each of the three scales was calculated with the result of $\alpha > 0.78$. Together these findings provide evidence that the SESALE is measuring three

distinct constructs and that these constructs are aligned with the themes that emerged from student focus groups in Phase I.

The three-factor solution consisted of:

- **Value of the Group Activity**: The first factor consisted of nine items exploring student's perception of the activity's value for learning (e.g. "Explaining the material to my group improved my understanding of it") or other reasons (e.g., "I had fun during today's group activity"). Cronbach's α for this scale was 0.91.
- **Personal Effort**: The second factor consisted of three items that measured how much individual effort a student put into the activity (e.g., "I worked hard during today's group activity" and "I made a valuable contribution to my group today"). Cronbach's α for this scale was 0.84.
- **Instructor Effort**: The final factor measured how much effort the students perceived that the instructor put into the activity (e.g., "The instructor put a good deal of effort into my learning for today's class" and "The instructor's enthusiasm made me more interested in the group activity"). Cronbach's α for this scale was 0.78.

Figure 3. Thematic representation of factor analysis of the Engagement Survey items 1-16. Items were loaded onto three factors: Value of Group Activity, Personal Effort and Instructor Effort Thematic titles are based on the interpretation of the items that load onto each factor. Numbers in brackets indicate coefficient of loading onto the factor in which they are placed. Vertical spacing of items correlates loosely with the size of the loading coefficient for that factor. All coefficients noted are above the 0.32 cutoff for likely loading onto that factor.

Loading coefficients:	VoGA	PE	IE
Item Text:			
Explaining the material to my group improved my understanding of it.	**0.80**	0.11	-0.13
Having the material explained to me by my group members improved my understanding of the material.	**0.78**	-0.11	0.00
Group discussion during the [insert Topic] activity contributed to my understanding of the course material.	**0.79**	0.00	0.04
I had fun during today's [Topic] group activity.	**0.65**	0.04	0.14
Overall, the other members of my group made valuable contributions during the [Topic] activity.	**0.41**	0.05	0.03
I would prefer to take a class that includes this [Topic] activity over one that does not include today's group activity.	**0.63**	-0.01	0.11
I am confident in my understanding of the material presented during today's [Topic] activity.	**0.70**	0.04	-0.04
The [Topic] activity increased my understanding of the course material.	**0.83**	-0.02	0.04
The [Topic] activity stimulated my interest in the course material.	**0.71**	-0.07	0.14
I made a valuable contribution to my group today.	0.07	**0.73**	-0.04
I was focused during today's [Topic] activity.	0.12	**0.71**	-0.05
I worked hard during today's [Topic] activity.	-0.12	**0.91**	0.07
The instructor's enthusiasm made me more interested in the [Topic] activity.	0.18	-0.7	**0.71**
The instructor put a good deal of effort into my learning for today's class.	0.02	0.00	**0.75**
The instructor seemed prepared for the [Topic] activity.	-0.11	0.14	**0.72**
The instructor and TAs were available to answer questions during the group activity.	0.06	0.03	**0.45**

Table 2. Loading coefficients onto three factors for factor analysis of survey items. Items are considered to be a good fit for loading onto a factor if the loading coefficient is >0.4 and also <0.3 on all other factors. Coefficients are bolded/underlined in the column pertaining to the factor on which they loaded best. Questions are reorganized for ease of reading of each factor (numbers in parentheses indicate order on the final survey).

Scale Reliability

The Cronbach's alpha coefficients (Cronbach, 1951) range observed for each of the three factors described above (0.78-0.91) indicates that students have a similar response pattern for the items within a given factor. To further assess the internal consistency of the scales identified in the exploratory factor analysis, we administered the SESALE to a similar population of introductory biology students in a consecutive quarter of the same course in which we had performed the exploratory factor analysis. Histograms of student responses are available (Online Resource 2). Cronbach's alpha coefficients for each scale ranged from 0.81 to 0.91 providing evidence that the survey reliably measures the same constructs in a similar population (Online Resource 3).

Phase IV: External validity

In this phase, consistent with Benson's (1998) external validation stage, we gather further evidence for construct validity by assessing whether students' survey answers vary predictably in response to different class activities.

To be a useful research tool, the SESALE survey must be sensitive to changing levels of student engagement with different activities. To test its ability to discriminate between activities, we compared the SESALE responses of a similar population of students during two different activity

types: 1) a regular class day with a series of series of 8-10 short activities centered around instructor-posed clicker questions, and 2) an activity day in which students completed one long (~30 min activity) interspersed with clicker questions. We hypothesized that students would perceive the instructor putting in more effort on a typical class day, because the instructor more frequently provides feedback to the entire class than is typical on a class day with a long activity. Based on student focus groups and our analysis of student open-ended responses to items on the SESALE, we also hypothesized that students would place more value in the short activities compared to the one long activity because students often voiced frustration regarding infrequent instructor feedback during the long activities. We did not have an a priori hypothesis abut which context would be perceived to require more work.

We first tested whether the questions on the SESALE still captured the same three constructs in this new population that had completed the short activities by calculating the Cronbach's alpha for each scale (Cronbach, 1951). Since the SESALE was designed to capture student opinion about in-class activities, we reasoned that the same scales should be observed when students reflect on these short instructor-directed activities as were found when surveying students after days with long in-class activities. This was supported by our finding that Cronbach's α values for each scale on a regular class day again fell between 0.78 and 0.91 (Online Resource 3). We then used a linear mixed-effect model to calculate the effect of the two different activity types on each of the three factors that make up the SESALE: Value of group activity, Personal effort and Instructor effort. The mixed effects model was necessary because we had a repeated measures design (the same students took the survey in two different contexts) and these models can handle this non-independence of outcomes by including a random effect term for student (Zurr et al. 2009). Thus the results of this analysis can be interpreted as the change in the SESALE response of the

same student when they working on the single long activity (reference level, 0) vs the multiple short activities (treatment,1). Our final models were: SESALE factor ~ Activity type + (1 | Student ID).

We find that, all else being equal, a student perceives more value (β=1.57, $p<0.001$) and increased instructor effort (β=1.10, $p<0.001$) in a regular class day relative to a long activity day. No statistically significant difference in perceived personal effort was observed. This provides evidence that the value and instructor element constructs are able to distinguish between student self-reported engagement for two different types of activities. This discrimination ability implies that this survey may be a useful tool for studies comparing different activities.

Characterization of Roles Students Play During In-Class Group Activities

In addition to characterizing student experience during active learning, we were also interested in how students' self-selected roles might vary during different small group in-class activities. This research question stemmed from earlier work (Eddy et al., submitted) exploring student preference for different roles during collaborative learning. In that study, which was performed on a similar population of introductory biology students as described here, students were asked an open-ended question regarding role preference. The wording of that question was: "My preferred role in the group is...". Through iterative coding of student responses Eddy et al. classified student preferences into four major role categories: 1) Leader 2) Collaborator 3) Listener and 4) Recorder (Eddy et al., submitted). The majority of students fell into the first two categories, both of which included a strong preference for playing the role of explainer in their group. To expand on these findings and determine what roles students self-reported actually playing during the in-class activities, we developed a set of five "Roles Questions" asking students how much of the time they engaged in one of five different actions most commonly reported in the earlier study: 1) Leading, 2) Explaining 3) Asking Questions 4) Listening and 5) Writing (Online Resource 1). Each of these

five items provided the same answer choices for participants based on frequency from 'None of the time' to 'All of the time'.

Cluster Analysis of Roles Data:

To assess the interplay of group roles, these items were analyzed by hierarchical cluster analysis. This allowed us to assess whether combinations of answers were overrepresented and might indicate common correlated suites of behaviours during groupwork. Using the FactoMineR package in R (Husson 2008) four clusters of student roles during in-class activities were identified: 1) Leader/Explainer 2) Broadly Engaged 3) Listener and 4) Disengaged. The first three clusters aligned closely with three of the categories identified in the previous study, whereas the fourth cluster of "Disengaged" students did not map onto any of the previously identified groups. As shown in Table 2, the Leader/explainers had high self-reported rates of Leading and Explaining, but below-average rates of listening and asking questions, aligning closely with the category of Leader. The Broadly Engaged students reported above average participation in all five roles, correlating well with students in the previously characterized Collaborator category. The Listeners in our study reported higher than average rates of listening, but below average rates for leading, explaining and writing notes, very similar to students who fell into the Listener category. However, unlike the previous study that identified a small group of students (5%) who preferred to act as a recorder, our analysis did not find a distinct cluster of students who engaged predominantly in writing. This finding suggests either that there is a small group of students who prefer being the recorder, but in actuality take on additional roles during group work, or that our hierarchical analysis approach could not distinguish these students from those in other categories. We did identify a fourth cluster that we termed "Disengaged" who self-reported below average participation in all five group work roles. The strong overall alignment between the two studies indicates that the Roles Questions can

be used to gather data on the distinct ways students see themselves participating during a collaborative in-class activity.

Action:	Across Entire Data Set: *(Mean ± SD)*	Cluster 1: Disengaged	Cluster 2: Listening -only	Cluster 3: Leader/Explainer	Cluster 4: Broadly Engaged
Leading	3.0 ± 0.93	2.5 ± 0.86	2.4 ± 0.60	4.2 ± 0.67	3.2 ± 0.68
Explaining	3.0 ± 0.92	2.3 ± 0.72	2.4 ± 0.58	3.9 ± 0.70	3.4 ± 0.73
Recording	2.7 ± 1.3	1.8 ± 0.77	2.0 ± 0.82	Not different from overall mean	3.8 ± 0.89
Questioning	3.0 ± 1.0	2.2 ± 0.68	Not different from overall mean	2.4 ± 0.87	3.6 ± 0.82
Listening	3.7 ± 0.95	2.7 ± 0.68	4.4 ± 0.56	3.2 ± 0.86	3.9 ± 0.81

Table 2. Means for frequency students engaging in each role in each cluster. Green numbers are significantly above the means for the entire sample and red numbers are significantly below the mean.

Figure 4. Radar plot of the four clusters by actions in group work. Each color represents one of the four clusters, and more of a particular action is denoted by farther stretch towards that arm. Each frequency is compared to the reference level of zero, which is determined by the average response for each item.

Summary of Results

Qualitative work was used to identify important themes and to develop a survey capturing student engagement with in-class group-based activities. The survey was built and iteratively validated to investigate these elements of student experience, and item data from 382 participants was analyzed using factor analysis and cluster analysis. Analysis of survey results gave two big-picture findings. First, variability in the student experiences with in-class activities was mediated by three major factors: Personal factors like individual effort, experiences that influenced the Value of the Group Activity for students, and Instructor factors like instructor-displayed enthusiasm. These factors closely matched the major themes that emerged in the original qualitative research.

Secondly, student group work clustered into four common roles. Students primarily took up a role in groups as one of: Listener students who did little else, Leader/Explainers, Broadly Engaged students who engaged in all noted roles heavily or Disengaged students who reported doing very little during group work. These roles align closely with the roles that originally surfaced in analysis of open-ended questions about students' preferred roles in group work (Eddy et al., submitted). In total, these two narratives indicate strong triangulation among very different kinds of data to quantify the important aspects of the student experience in large active-learning STEM classrooms.

The final version of the SESALE and Roles Questions are available (Online Resource 1). It is important to note that these items were developed using a population of students in a single STEM major at one university, and that the applicability of these items may not be universal. Using this as a basis for a new instrument more fitting to a different environment and using input from real student experience is likely to be fruitful, so we encourage researchers to make use of this survey as an initial step and then adjust items to meet the needs of local contexts and learning cultures after appropriate validation in the new population and learning environment

Discussion & Conclusion

The SESALE, combined with the Roles Questions, provides a means of analyzing student experiences in higher education STEM active learning classrooms. The SESALE differs from other instruments that measure student experiences during active learning (Visschers-Pleijers et al. 2005; Pazos et al. 2010)) in that our survey is designed to be completed by students and is not specific to a particular type of active learning such as problem-based learning. The 20-item survey and 5 roles questions can be administered electronically. Cognitive testing groups demonstrated that students complete these 25 questions in 6-7 minutes on average. Together these tools provide a rapid way to

systematically gather student opinion on specific aspects of their experience and level of engagement during an in-class group activity.

We report two aspects of a mixed methods study into the student experience of active learning classrooms. In both cases, deep qualitative work with a small number of students informed the development of quantitative tools for large-scale study. In both cases, subsequent analysis of the quantitative data gave support for the broad applicability of quantitatively derived emergent themes in the wider student population. The internal consistency of our findings using these two different approaches provides increased confidence that our conclusions are meaningful for students in a large undergraduate STEM classroom.

The SESALE was designed to elicit student perception of three key aspects of the student experience of active learning: 1) utility and intrinsic value of a group activity 2) personal effort invested during the activity, and 3) instructor contribution to the activity and to student learning. Factor analysis supports the assumption that the SESALE is measuring three discrete scales that align closely with these three constructs, providing support for validity of the observed scales. We also provide evidence regarding the reliability of the three scales as measured by Cronbach's alpha coefficient. We did not find evidence for a separate scale related to group dynamics. Instead, in the 3-factor solution, three of the items relating to group function clustered with Value of Group Activity and the fourth item, "I felt comfortable in my group" was weakly correlated (0.22-0.26) with multiple factors. If a group is not functioning well, or if a student does not perceive value from the group interaction, this may negatively influence their overall perception of the activity's value.

Underlying this mixed methods research is an overall desire to understand where, how and for whom there is room for improvement of active learning methods. The ICAP framework (Chi and Wylie 2014) provides a strong theoretical basis for categorizing active learning processes. Whether changes along the ICAP spectrum result in equal gains for all types of students is an open question.

As more classrooms turn to cutting edge active learning techniques, the improvement of the student experience will require better and finer tools to gauge improvement. In particular, our grounded analysis has identified primary themes that can serve as pathways towards improvement. The resulting combination of different kinds of data on student experiences suggests that active learning classrooms can prioritize design effort on three key issues:

- Improvement and facilitation of the group work experience
- Encouraging and supporting the perceptions of self-efficacy and personal effort
- Making more obvious the efforts of teaching faculty for students

It is striking that these three salient themes all suggest that development of faculty soft skills, rather than specific curriculum development, may be the most profitable avenues for classroom improvement.

This work leaves interesting areas unexplored. The use of the SESALE as a tool for assessing classroom culture is probably most appropriate for experimental comparisons between classes with different teaching techniques. It might reasonably be used to analyze differences between techniques with the same instructor or even to differentiate the student experience between instructors who teach the same course. Cluster analysis of group roles may be best fit for assessing interventions into group work and equity within classrooms. Student motivations are likely to be dynamic, and are likely to depend on sustainable best practices for instructors. Each of these questions has obvious importance in STEM classrooms, especially those in which diverse scientists are being trained to take on complicated new challenges in increasingly diverse professional environments. We hope to follow up on these avenues for future research, using the tools that have emerged from mixed research.

The mixed methods used here allow an internal triangulation between statistical analysis and direct narrative data from student experiences. Quantitative studies of education reify inherent

biases and are an abstraction of true student engagement, motivation, etc. Qualitative studies provide depth but are practically difficult to scale to the size of modern STEM classrooms. We hope that the combined signal from these types of data helps us? to understand deeply student learning while concurrently venturing towards ecological populations of our students.

Implications

Student motivations are key to learning and eventual successful outcomes (Svinicki 2004). Personal motivations may be influenced by pedagogical choices. Interventions into group work design or usage may help to alleviate stereotype threat and solo concerns (Steele 1997; Sekaquaptewa et al. 2007). Instructor influences on student motivations are important factors for faculty development and professional development of young scientists who will take on teaching roles; more interpersonal and emotive talent may prove more beneficial than increased access to classroom technology or even active learning methods.

We compared student responses on the SESALE after a normal class day consisting of clicker questions and after a day in which students engaged in a long in-class activity. Using a linear regression model we found that students as a whole placed less value on long active-learning exercises than on instructor-guided discussions during short activities on a normal class day. These findings have implications for how active-learning is implemented in the classroom, suggesting potential benefit of more frequent interruptions of longer active learning exercises to provide instructor feedback.

Limitations:

This study relies heavily on leveraging subjective expertise in qualitative coding and interpretation. This is acknowledged in the literature on qualitative methods, and rigorous categorization of best practices exist to mitigate this limitation. For our example, interdisciplinary links, including established methods from the Learning Sciences, helped inform our discipline-based work. In all cases, the familiarity with local context and issues is crucial. In this case, researchers have long and deep experience working with introductory biology students in large, small and lab classes from standpoints as instructors, researchers, teacher evaluators and students. This study relies heavily on student self-reporting on surveys, although sociometric and face validity measures have been taken to diminish possible avenues for biases in data collection.

The tool developed here is not intended as a psychometric analysis of the mental construct of student engagement, for which additional validation beyond the scope of our observations would be necessary. Instead, our interest was in developing a way to systematically collect students' opinions about in-class activities. Student perception of self-engagement is just one measure that can be used in conjunction with formative and summative assessment data to inform instructional choices in the classroom. Further, because social learning tasks are so deeply situated in student culture, significant validation was necessary to ensure that survey items remained true to student language in order to capture the effects of social and active learning on engagement (Lave and Wenger 1991). Best use of this survey will require validation and qualitative triangulation in new classroom environments to ensure that the student language and experiences in the observed population are reflected in the instrument used.

The analyses in this work do not include demographic factors as controls. Future studies may find that underlying cultural factors like gender spectra, race/ethnicity, and/or socioeconomic backgrounds may contribute to or even predict roles and responses. These demographic descriptions are oversimplifications of the heterogeneity of culture, development and code switching of

identities. Further work may reveal correlations with simplified cultural indicators, and this work may open these spaces for deeper grounded research. Similarly, we have not attempted to experimentally determine baselines for these survey items from participants in classrooms that are primarily passive in nature and do not use active learning techniques.

Future studies will be required to provide evidence that the three factors identified here (Value of Group Activity, Personal Effort and Instructor Effort) are positively correlated with student learning. There is evidence in the literature that the value students place on an activity as well as how much personal effort they are investing are linked to both motivation and performance (Eccles & Wigfield 2002; Hulleman et al. 2010; Hidi & Renninger 2006; Lee & Anderson 1993). Our work on this survey incites many questions for us about sociocognitive learning in STEM classrooms (for which this survey might be a useful tool). For instance; do student identities within the majority cultural population of a classroom impact uptake of active learning practices? Do students enculturate active learning more deeply when their own cultural histories are similar to those used in the classroom? Do students of dissimilar cultural traditions provide useful habits of mind for group work, or do they require extra scaffolding to maximize learning? To what extent are socially learned concepts engaging students into skills that translate towards success on exams or long-term understanding? Further research needs to be done to explore possible relationships between student perceptions of active learning and student learning.

References

Blickenstaff, J. C. (2005). Women and science careers: leaky pipeline or gender filter? *Gender and Education, 17*(4), 369-386, doi:10.1080/09540250500145072.

Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219-243

Committee on Science, E., and Public Policy (2007). Rising above the gathering storm: Energizing and employing America for a brighter economic future. In N. A. Press (Ed.).

Committee on the Status, C. a. F. D. o. D.-B. E. R. (2012). Discipline-Based Education Rsearch: Understanding and improving learning in undergraduate science and engineering. In B. o. S. E. o. t. N. R. Council (Ed.). National Academies Press: National Research Council.

Corbin, J., & Strauss, A. (2014). *Basics of qualitative research: Techniques and procedures for developing grounded theory*: Sage publications.

Council, N. R. (Ed.). (2000). *How People Learn: Brain, Mind, Experience, and School*. Washington, D.C.: National Academies Press.

Council, N. R. (2013). Framework for K-12 Science Education. In N. R. C. Press (Ed.), (Vol. 1). http://www.nextgenscience.org.

D'avanzo, C. (2013). Post-Vision and Change: Do we know how to change? *Cbe-Life Sciences Education, 12*(3), 373-382, doi:Doi 10.1187/Cbe.13-01-0010.

Dasgupta, N., & Stout, J. G. (2014). Girls and Women in Science, Technology, Engineering, and Mathematics: STEMing the tide and broadening participation in STEM careers. *Policy Insights from the Behavioral and Brain Sciences, 1*(1), 21-29

Dewey, J. (1938). *Experience and Education*.

Dillman, D. A. S., Jolene D.; Christian, Leah M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (4th ed.). Hoboken, NJ.

Drew, C. (Nov 6, 2011) Why science majors change their minds (it's just so darn hard). *New York Times*.

Durik, A. M., & Harackiewicz, J. M. (2007). Different strokes for different folks: How individual interest moderates the effects of situational factors on task interest. *Journal of Educational Psychology, 99*, 597-610.

Eccles, J. (1983). *Expectancies, values, and academic behaviors* (Achievement and achievement motives: Psychological and soiological approaches). San Francisco, CA: WH Freeman.

Eccles, J. (2005). Subjective task value and the Eccles et al. model of achievement-related choices. In C. Dweck (Ed.), *Handbook of Competence and Motivation* (pp. 105-121). New York, NY: Guilford.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109-132.

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).

Engle, R. A., Conant, Faith R. (2002). Guiding Principles for Fostering Productive Disciplinary Engagement: Explaining an emergent argument in a community of learners classroom. *Cognition and Instruction, 20*(4).

Esmaeili, M. E., Ali (2014). *The Relationship of Active Learning Based Courses and Student Motivation for Pursuing STEM Classes.* Paper presented at the ASEE 2014 Zone I Conference, Bridgeport, CT, April 3-5

Freeman, S. R., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., et al. (2014). Active learning increases student performance across STEM disciplines. *Proc Natl Acad Sci U S A*.

Geertz, C. (1973). *The interpretation of cultures*.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory; strategies for qualitative research* (Observations). Chicago,: Aldine Pub. Co.

Graham, M. J., Frederick, J., Byars-Winston, A., Hunter, A. B., & Handelsman, J. (2013). Science education. Increasing persistence of college students in STEM. *Science, 341*(6153), 1455-1456, doi:10.1126/science.1240487.

Grunspan, D. Z., Wiggins, B. L., & Goodreau, S. M. (2014). Understanding classrooms through Social Network Analysis: A primer for social network analysis in education research. *Cbe-Life Sciences Education, 13*, ?

Gubrium, J. F., & Holstein, J. A. (2002). *Handbook of interview research: Context and method*: Sage.

Gutiérrez, K. D., & Rogoff, B. (2003). Cultural Ways of Learning: Individual traits or repertoires of practice. *Educational Researcher, 32*(5), 19-25, doi:10.3102/0013189x032005019.

Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased structure and active learning reduce the achievement gap in introductory biology. *Science, 332*(6034), 1213-1216, doi:10.1126/science.1204820.

Handelsman, M. M., Briggs, W. L., Sullivan, N., & Towler, A. (2005). A measure of college student course engagement. *The Journal of Educational Research, 98*(3), 184-192, doi:10.3200/JOER.98.3.184-192.

Hawkins, J., & Pea, R. D. (1987). Tools for bridging the cultures of everyday and scientific thinking. *Journal of Research in Science Teaching, 24*(4), 291-307, doi:Doi 10.1002/Tea.3660240404.

Holdren, J. a. L., E (2013). Engage to Excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics. Executive Office of the President: President's Council of Advisors on Science and Technology.

Hora, M. F., J. (2014). The teaching dimensions observation protocol (TDOP) 2.0.

Hug, B., Krajcik, J., & Marx, R. (2005). Using innovative learning technologies to promote learning and engagement in an urban science classroom. *Urban Education, 40*, 446-472.

Hulleman, C., Durik, A., Schweigert, S., & Harackiewicz, J. (2008). Task values, achievement goals, and interest: An integrative analysis. *Journal of Educational Psychology,, 100*(2), 398-416.

Hulleman, C., Godes, O., & Hendricks, B. H., J.M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology, 102*(4), 880-896.

Husson, F. J., Julie; Le, Sebastien; Mazet, Jeremy (2008). FactoMineR: An R Package for Multivariate Analysis. . *Journal of Statistical Software* (Vol. 25, pp. 1-18).

Kurth, L. A., Anderson, C. W., & Palincsar, A. S. (2002). The case of Carla: Dilemmas of helping all students to understand science. *Science Education, 86*(3), 287-313, doi:Doi 10.1002/Sce.10009.

Kvam, P. H. (2000). The effect of active learning methods on student retention in engineering statistics. *American Statistician, 54*(2), 136-140.

Lane, E. S., & Harris, S. E. (2015). A new tool for measuring student behavioral engagement in large university classes. *Journal of College Science Teaching, 44*(6), 83-91.

Lave, J., & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. New York: Cambridge University Press.

Linn, M. C. (2003). Technology and science education: starting points, research programs, and trends. *International Journal of Science Education, 25*(6), 727-758, doi:Doi 10.1080/0950069032000076670.

London, B., Downey, G., Romero-Canyas, R., Rattan, A., & Tyson, D. (2012). Gender-based rejection sensitivity and academic self-silencing in women. *J Pers Soc Psychol, 102*(5), 961-979, doi:10.1037/a0026615.

Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics, 74*(2).

Lovelace, M., & Brickman, P. (2013). Best practices for measuring students' attitudes toward learning science. *CBE Life Sci Educ, 12*(4), 606-617, doi:10.1187/cbe.12-11-0197.

Mayer, R. E. (2011). Applying the science of learning to undergraduate science education. In N. A. B. o. S. Education (Ed.), *Committee on the Status, Contributions and Future Directions of Discipline Based Education Research* (pp. 21).

McConnell, D. A., Steer, D. N., & Owens, K. D. (2003). Assessment and active learning strategies for introductory Geology courses. *Journal of Geoscience Education, 51*(2), 205-216.

Minner, D. D., Levy, Abigail Jurist, Century, Jeanne (2010). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching, 47*(4), 474-496, doi:10.1002/tea.20347.

Nasir, N. S., & Hand, V. M. (2006). Exploring sociocultural perspectives on race, culture, and learning. *Review of Educational Research, 76*(4), 449-475, doi:Doi 10.3102/00346543076004449.

Piburn, M. S., D.; Falconer, K.; Turley, J.; Benford, R.; Bloom, I. (2000). Reformed Teaching Observation Protocol (RTOP). In A. IN-003 (Ed.).

Pintrich, P. R. S., David A. F.; Garcia, Teresa; Mckeachie, Wilbert J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (Mslq). *Educational and Psychological Measurement, 53*, 801-813.

Revelle, W. (2015). psych: Procedures for Personality and Psychological Research. (1.5.1 ed.). Northwestern University, Evanston, Illinois.

Rosenberg, J. L., Lorenzo, M., & Mazur, E. (2006). Peer Instruction: Making science engaging. *Handbook of College Science Teaching*, 25.

Rubin HJ & Rubin, I. (2005). Qualitative Interviewing: The art of hearing data. In. Thousand Oaks: Sage Publications.

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.

Sciences, N. A. o. (2013). Next Generation Science Standards: For States, by States.

Sekaquaptewa, D., Waldman, A., & Thompson, M. (2007). Solo status and self-construal: being distinctive influences racial self-construal and performance apprehension in African American women. *Cultur Divers Ethnic Minor Psychol, 13*(4), 321-327, doi:10.1037/1099-9809.13.4.321.

Semsar, K., Knight, J. K., Birol, G., & Smith, M. K. (2011). The Colorado Learning Attitudes about Science Survey (CLASS) for use in Biology. *CBE Life Sci Educ, 10*(3), 268-278, doi:10.1187/cbe.10-10-0133.

Seymour, E., & Hewitt, N. M. (1999). *Talking about leaving : why undergraduates leave the sciences*. Boulder, Colo.: Westview Press.

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.

Steele, C. M. (1997). A Threat in the Air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*(6), 613-629.

Suchman, E. L. (2014). Changing academic culture to improve undergraduate STEM education. *Trends in Microbiology, 22*(12), 657-659, doi:10.1016/j.tim.2014.09.006.

Svinicki, M. (2004). *Learning and Motivation in the Postsecondary Classroom*. Bolton, MA: Anker Publishing Company.

Tabachnick, B. G., & Fidell, L. S. (2001). Using multivariate statistics.

Universities, A. o. A. (2011). STEM Education Initiative.

Waldrop, M. M. (2015). Why we are teaching science wrong, and how to make it right. *Nature, 523*, 272-274, doi:10.1038/523272a.

Wieman, C. E. (2014). Large-scale comparison of science teaching methods sends clear message. *Proceedings of the National Academy of Sciences, 111*(23), 8319-8320, doi:10.1073/pnas.1407304111.

Willis, G. B. (2005). *Cognitive Interviewing: A Tool for Improving Questionnaire Design.* .

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, doi:Doi 10.1187/Cbe.10-03-0044.

Zurr, A., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effect models and extensions in ecology with R* (Springer Science and Business Media).

Development of a grounded survey to measure student engagement in large active-learning classrooms

Benjamin L Wiggins
Biology and Learning Sciences, University of Washington, USA

Sarah L Eddy
College of Natural Sciences, University of Texas, USA

Leah S Wener-Fligner
Biology and Drama, University of Washington, USA

Daniel Z Grunspan
Anthropology, University of Washington, USA

Jerry P Timbrook
Academic Affairs and Learning Initiatives, University of Dayton, USA

Karen Friesem
Center for Teaching and Learning, University of Washington, USA

Alison J Crowe
Biology, University of Washington, USA

Abstract

We have developed a survey to analyze student experiences in a large undergraduate biology course using a mixed-method approach. Survey items address themes that emerged from direct student talk and that were identified by students as essential to their learning in this setting. Face validation was conducted through a series of qualitative methods with direct student input.

Statistical analysis of survey results provide support for two qualitatively-derived narratives of student experiences within large active learning classrooms. In one narrative, student engagement in active-learning STEM courses was explained by three major factors: value of group activities, personal effort, and instructor influences. These three scales emerged independently through open-ended inquiry, and from exploratory factor analysis. In the second narrative, quantitative cluster analysis of survey data found that students' self-described roles in groupwork fell into four distinct groups: Disengaged, Listener-Only, Leader/Explainer, and Broadly Engaged. Both narratives identify important characteristics of sociocultural learning in large-enrolment active learning STEM courses.

Compliance with Ethical Standards

All ethical standards, participant consent procedures and data security measures for this work have been carefully reviewed, coordinated and approved under UW IRB Plan #44438.

Acknowledgements

We thank Katherine Cook, Sarah Davis, Aaron Rosen and Carrie Sjogren for specific and detail-oriented data cleaning tasks. We express our gratitude and admiration for our colleagues in the UW Biology Education Research Group. This work is supported by open-minded faculty and administrators at UW Biology and the UW Center for Teaching and Learning. We express our continuing thanks for the hard work and helpful comments from the members of the UW Institutional Review Board who have helped us to develop and oversee protection of student data and interests. This work was funded by the National Science Foundation's Transforming Undergraduate Education in Science program (#DUE1244847).

Conflicts of Interest

We declare that we have no conflicts of interest related to this work.

Keywords: *STEM education, qualitative and quantitative research, sociocultural learning, active learning, survey, engagement*

Fig1 Overall Development Scheme

Click here to download Colour figure: Fig1 Overall Development Scheme.pptx

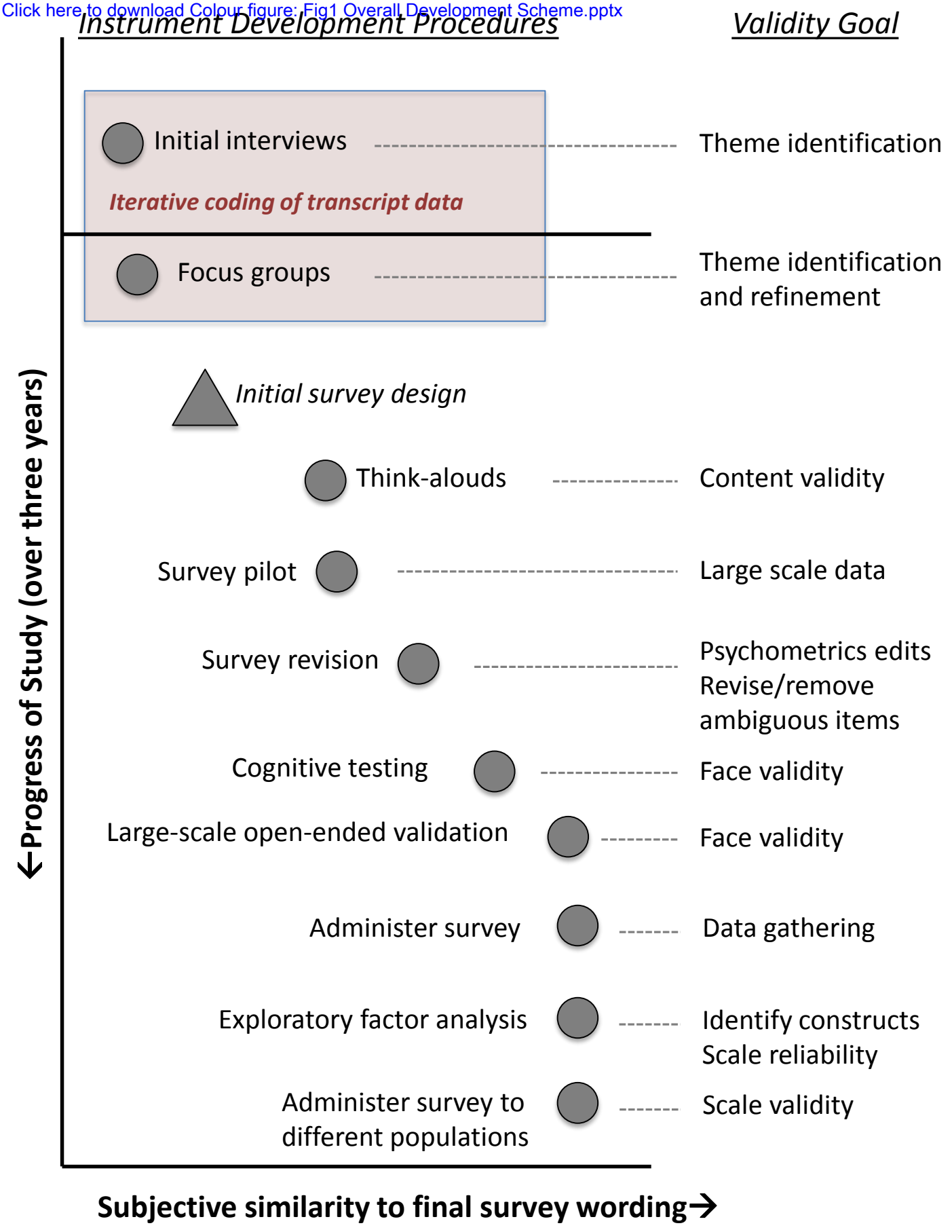


Fig2 Item Development Example

Click here to download Colour figure: Fig2 Item Development.pptx

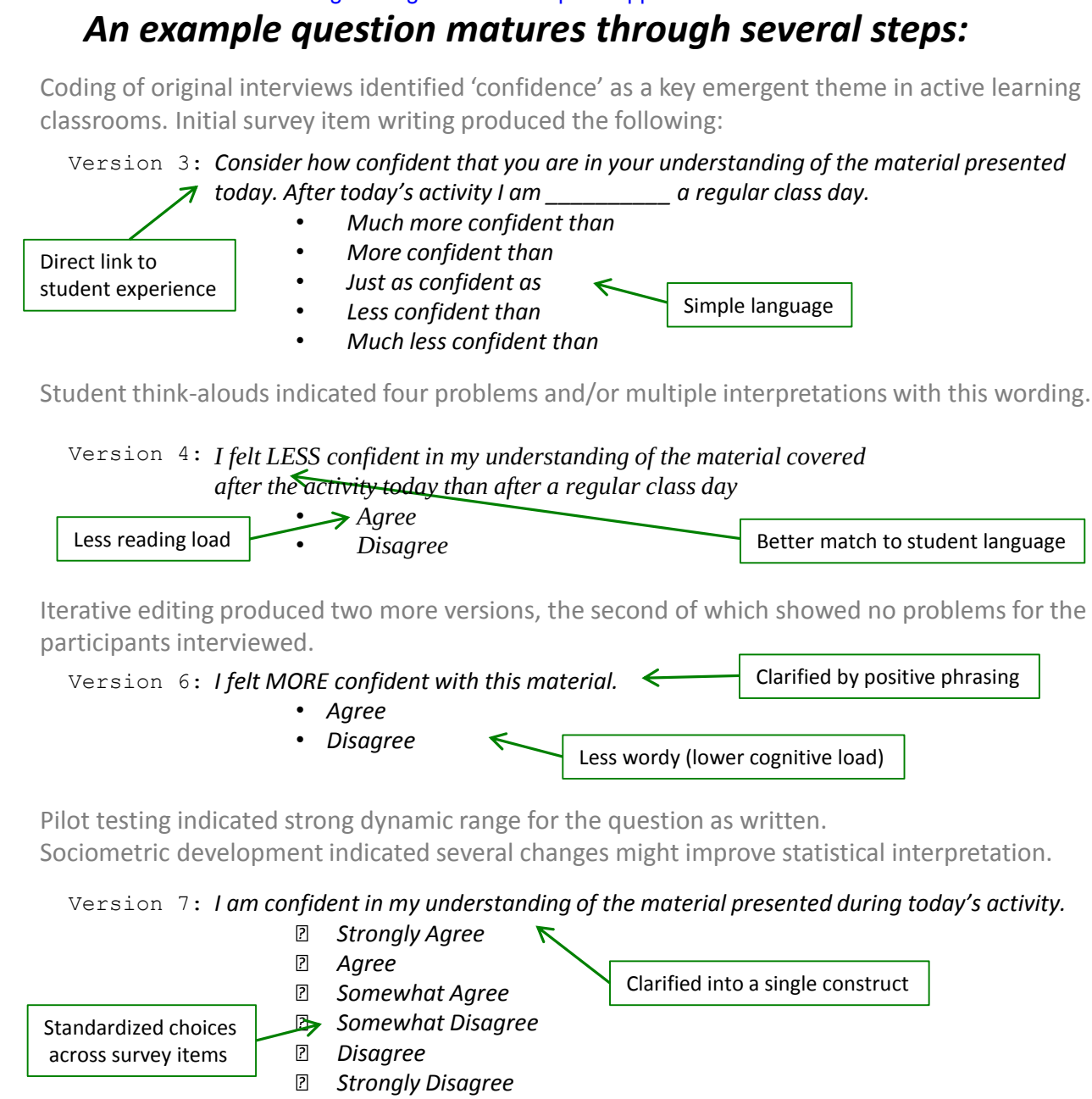


Fig3 Factor Analysis Representation

Click here to download Colour figure: Fig3 Factor Analysis Representation.pptx

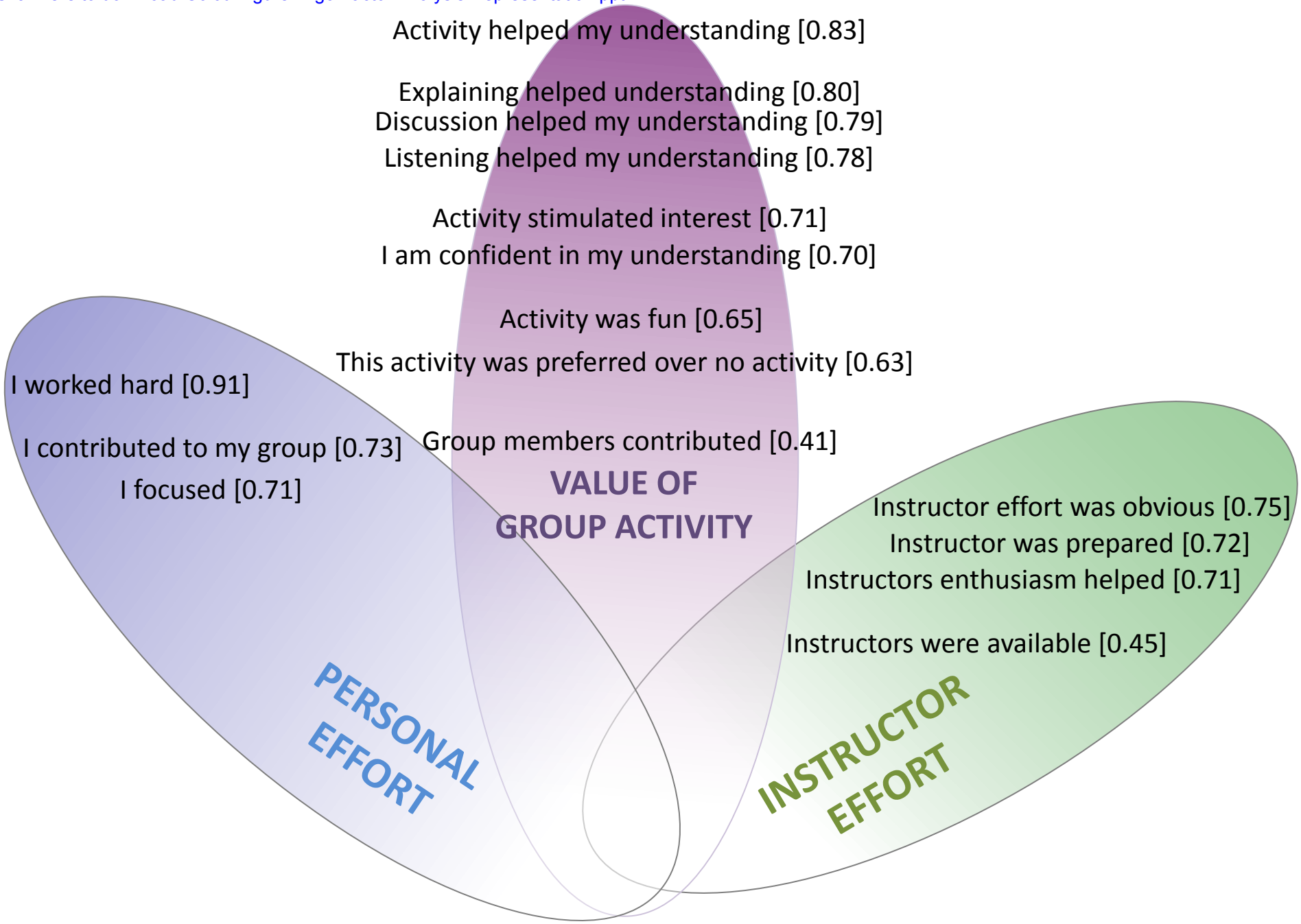
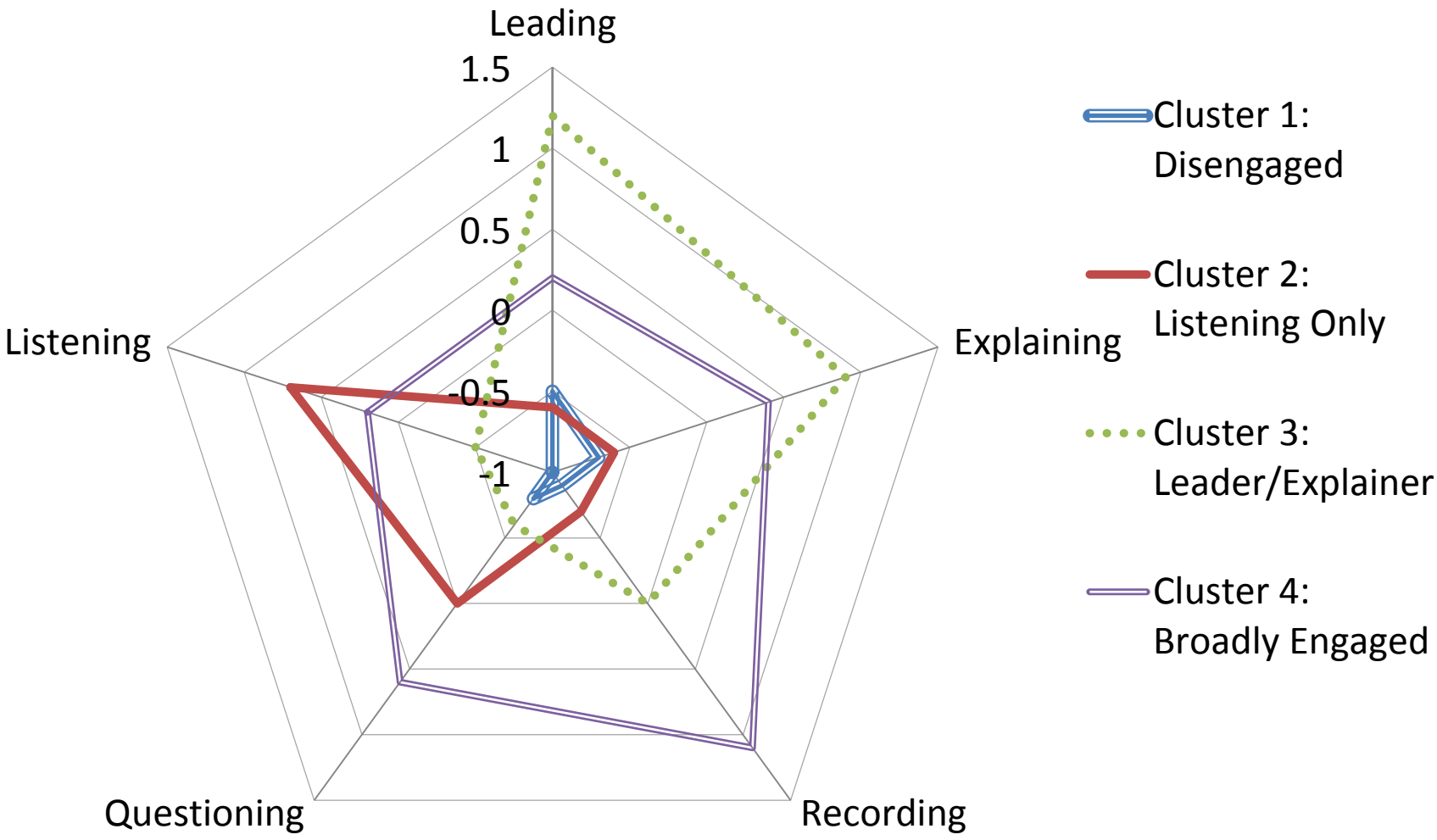


Fig4 Radar Plot of Roles Clusters

Click here to download Colour figure: Fig4 Radar Plot of Roles Clusters.pptx



Electronic Supplementary Materials 1)
Final SESALE wording for days with a single-group, individually active, or jigsaw* workflow

Instructions for students:
All questions in this survey refer to today's class in which you completed an activity on [topic name]. Your responses on this survey will be used to evaluate how we teach this topic in future Bio 200 classes. Your instructor will not know whether you completed this survey or how you answered the questions, but your effort will impact the experience of future students in this series.

Demographic questions:
A) During class today, you and your classmates completed a [topic name] activity in a group. How many students (including you) worked in your group?

Possible Answers: *1 (just me)*
2
3
4
More than 4

B) Are you friends with at least one person that was in your group?

Possible Answers: *Yes*
No

C) As a college student, have you been asked to work with other students during class time in large lecture courses (over 100 students)? Do not include this class in your answer.

Possible Answers: Yes
No, other large lecture courses I have taken have not asked me to do this
No, this is the first large lecture course I have taken

**For jigsaw workflow: In questions A and B, replace "your group" with "your SECOND group".*

For questions 1-20, students answered on a 6-point Likert Scale from Strongly Agree to Strongly Disagree (questions 1 and 3 had an additional "This did not happen today" answer included).
Question abbreviations are in parentheses
Instructions for students
The following questions ask you about your experience with the [topic name] activity that you completed today. Please rate how strongly you agree or disagree with each of the following statements.
1) Explaining the material to my group improved my understanding of it. (Explain.Understand)
2) The instructor's enthusiasm made me more interested in the group activity. (Enthusiasm)
3) Having the material explained to me by my group members improved my understanding of the material. (Listening)
4) Group discussion during the activity contributed to my understanding of the course material. (Disc.Understand)
5) The instructor put a good deal of effort into my learning for today's class. (Inst.Effort)
6) I had fun during today's group activity. (Fun)
7) Overall, the other members of my group made valuable contributions during the group activity. (Group.Contrib)
8) I knew what I was expected to accomplish during the group activity. (Expected)
9) The instructor seemed prepared for the group activity. (Prepared)
10) I would prefer to take a class that includes this [topic name] group activity over one that does not include this [topic name] group activity. (Prefer.Act)
11) One group member dominated discussion during today's group activity. (Dominate)

12) I am confident in my understanding of the material presented during today's group activity. (Confident)
13) I made a valuable contribution to my group today. (Per.Contrib)
14) The instructor and TAs were available to answer questions during the group activity. (Avail)
15) The group activity increased my understanding of the course material. (Act.Understand)
16) I engaged in critical thinking during today's group activity. (Crit.Think)*
17) I felt comfortable with my group. (Comfort)
18) I was focused during today's group activity. (Focus)
19) The group activity stimulated my interest in the course material. (Stimulate)
20) I worked hard during today's group activity. (Work.Hard)

*This item was removed from the survey during the Phase II face validation step due to ambiguity in student interpretation of the term "critical thinking"

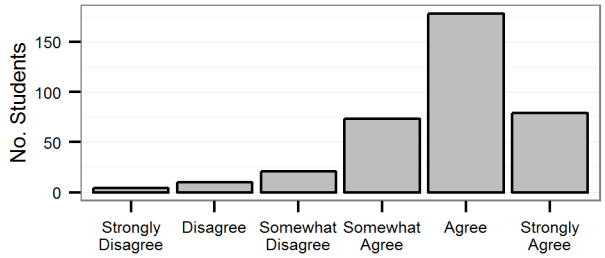
Roles Questions:
For questions 21-25, students were instructed:
Select the option that best describes the amount of time YOU spent performing each of the following actions during the [topic name] activity.

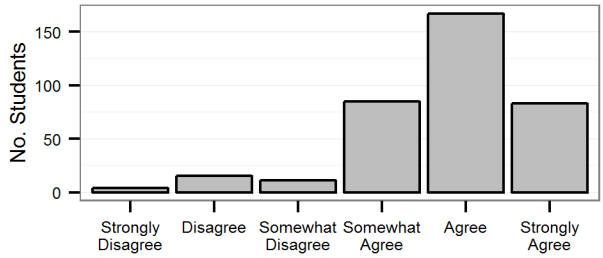
21) Writing down what was said during group discussion
22) Listening to what other group members had to say
23) Leading the group discussion
24) Explaining concepts to other group members
25) Asking questions about the [topic name] topic
For each, the possible answers were:
Nearly all of the time
A lot of the time
About half of the time
A little of the time
None of the time

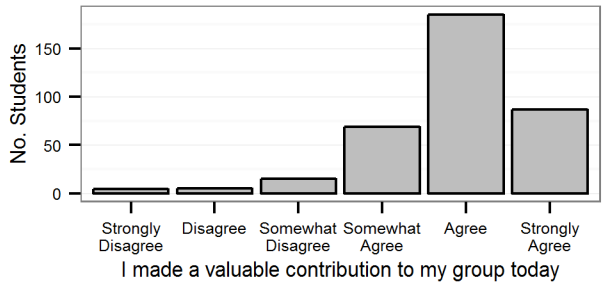
Online Resource 2 (edited to remove identifiers)
Click here to download Colour figure: ESM_2Histograms of Student Responses to SESALE Items.pdf

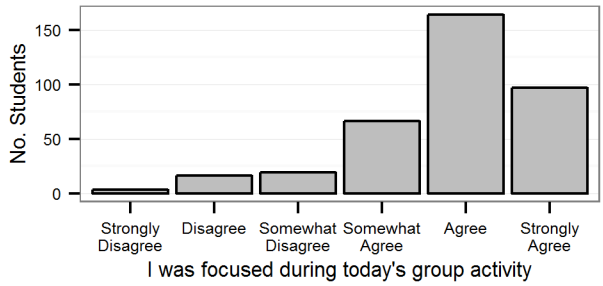
Electronic Supplementary Materials 2)
Histograms of student responses on all SESALE items.
These histograms were used to assess if responses to any items were problematically narrow and lacking in distribution. Each set of diagrams corresponds to a single factor or to those three questions that did not load onto a factor well.

Items that loaded onto the Personal Effort factor:

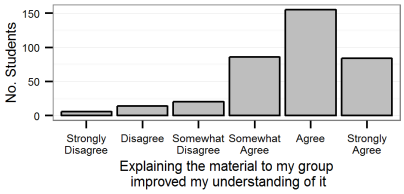


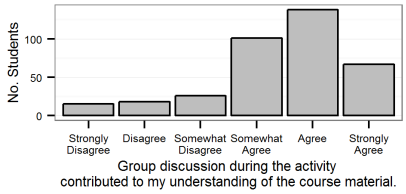


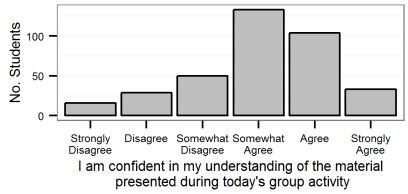


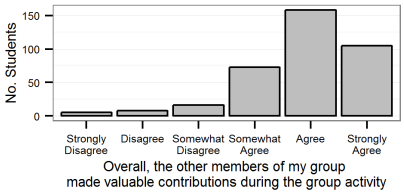


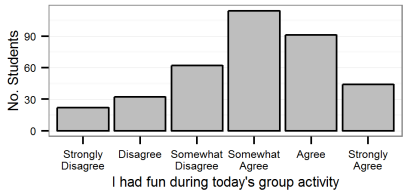
Items that loaded onto the Value of Group Activity factor:

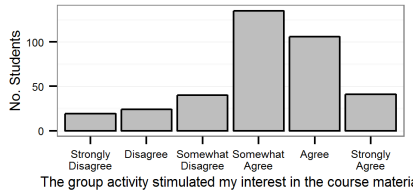


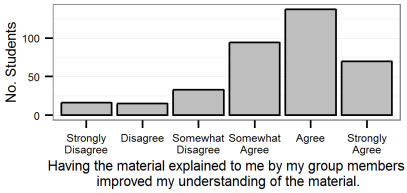


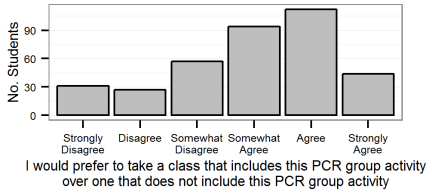


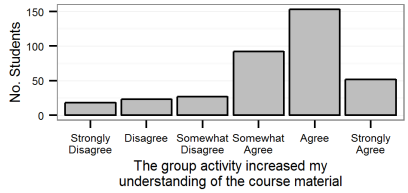




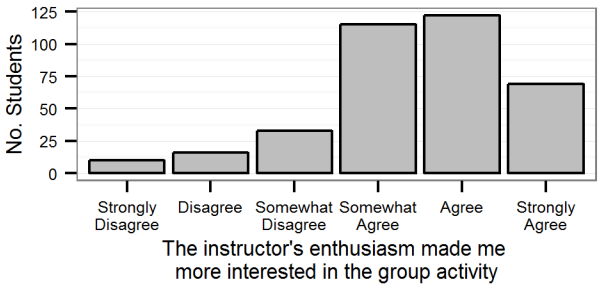


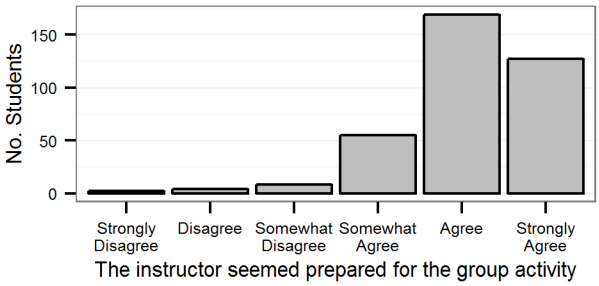


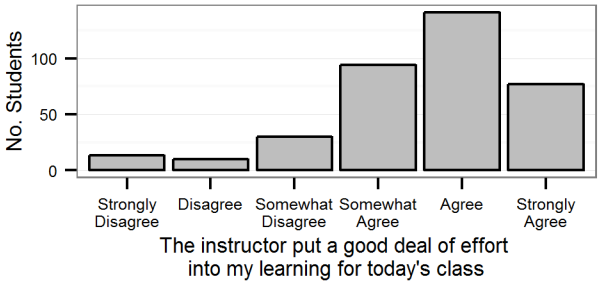


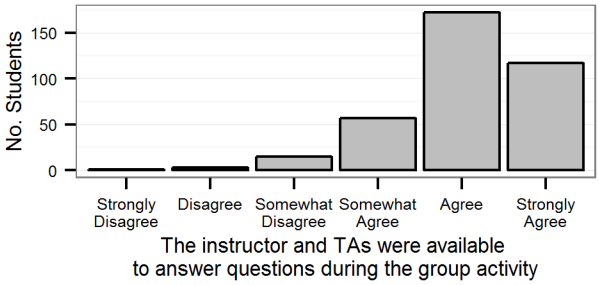


Items that loaded onto the Instructor factor:

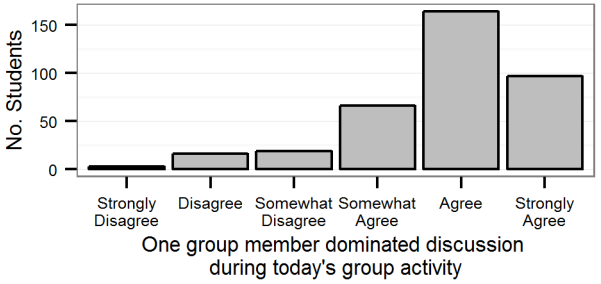


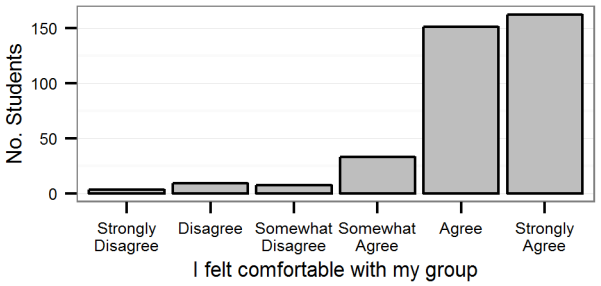


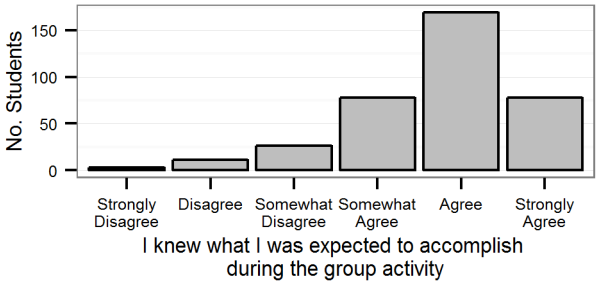




Other three items that did not load well onto the three factors:







Electronic Supplementary Materials 3)

Cronbach's α values resulting from exploratory factor analysis

The following table provides a comparison of Cronbach's α values resulting from exploratory factor analysis of student responses to the SESALE given to the indicated populations.

Factor	**Fall 2014 In-class Activity**	**Winter 2015 In-class Activity**	**Winter 2015 Regular Class**
Value of Group Activity	0.91	0.91	0.91
Personal Effort	0.84	0.81	0.78
Instructor Effort	0.78	0.84	0.82

The ICAP active learning framework predicts experimentally assessed learning gains for intensely active classroom experiences

Abstract (accepted in principal by AERA ONE):

While traditionally taught using ineffective passive lectures, STEM classrooms in higher education are rapidly improved by the proper use of active learning techniques (Waldrop 2015; Holdren 2013). Active classrooms retain more students and improve performance on assessments (S. R. Freeman et al. 2014). Active learning techniques often incorporate peer interaction as well as data- and task-based instruction to give opportunities for students to practice dynamic skills instead of simple memorization. These techniques occupy a descriptive spectrum that transcends passive teaching toward active, constructive, and finally interactive methods (Chi and Wylie 2014). Based on this ICAP framework, the highly active STEM classrooms of tomorrow, would benefit from well-designed instructional techniques that go beyond knowledge construction into knowledge co-creation (Committee on the Status 2012). While aspects of this framework have been examined experimentally, no large-scale or actual classroom based data exists to inform higher education STEM instructors about possible learning gains (Chi and Wylie 2014).

We describe the results of an experimental study to test the apex of the ICAP framework in this ecological classroom environment. A series of contrasting in-class activities were designed that encouraged either constructive or interactive behaviors. We experimentally implemented these activities in two sections of a split, large introductory Biology course. Student learning outcomes were assessed by validated pre-post assessments within a 24-hour period around instruction. Assessments test higher-order cognitive skills, requiring conceptual understanding rather than simple memorization. Comparisons were performed on three different class days allowing a repeated measures analysis of learning outcomes.

Controlling for prior demonstrated ability, instructor ability, course material and demographic factors, students in interactive classrooms demonstrate significantly improved learning outcomes relative to students in constructive classrooms. This improvement in learning is relatively subtle; similar experimental designs without repeated measures would be unlikely to have the power to observe this significance. This improvement equates to a 25% chance that a particular student will correctly answer at least one additional question correctly on an 8-question assessment. We discuss the importance of seemingly small learning gains that might propagate throughout a course or departmental curriculum. We also discuss the balance of these improvements with the necessity for faculty to develop and implement similar intensively active classroom materials.

Introduction:

Improvement goals in post-secondary education at the classroom level have largely focused on the opportunities provided by instructors for student engagement with course content (National calls citations). While post-secondary classrooms traditionally employ passive delivery techniques like lecturing, a range of cognitive engagement activities is available to instructors. This range of activity is well-described at a theoretical level in the ICAP framework (Chi and Wylie 2014).This framework predicts that the least engaging classroom strategies will give

students opportunities only to Passively approach the material. Subsuming and surpassing passive strategies, Active methods provide more opportunities for individual engagement. Continuing along the spectrum, Constructive activities create opportunities for students to generate beyond those outputs provided by instruction. Ultimately, Interactive methods use collaborative generative learning with significant dialogue to provide the most engaging learning opportunities. The ICAP framework predicts that categories of higher engagement will result in increased student learning. Specifically, ICAP predicts that interactive activities will support increased learning when compared even to the constructive activities in which students are extremely engaged. Based on ecologically realistic experiments in a post-secondary course, we present evidence to further validate the increased learning predicted by the ICAP framework in the most intensely active student activities.

A large amount of empirical evidence supports this framework. Within discipline-based education research, much of that work has been focused on demonstrating the post-secondary need for instructional methods beyond passive lecturing. Here we focus on an example of a 2nd-generation study in which post-secondary active instructional methods are compared and analyzed (Linton et al. 2014a; Stockwell et al. 2015; Linton et al. 2014b). Previous experiments analyze self-explaining, concept mapping, and note taking as examples of experimental conditions embodying one of four parts of the ICAP spectrum. Furthermore, studies conducted using classroom environments use pairwise comparisons of classroom techniques validate the predictions of the ICAP framework (Chi and Wylie 2014). To date, these experiments do not include an 'ecological' setting in which multiple experimental manipulations are performed in the context of a real classroom. This is an important step to inform best practices: instructors need to know that a) theoretical gains will be reflected in the complex cultural environments of classrooms, and b) that the magnitude of the benefit for students make economic sense when compared with the limited resources of time, training and cost inherent to real classrooms. These are necessary outputs from research if practitioners are to be expected to uptake the practices and findings of education research. As a natural successor to previous experimentation, we undertook an ecologically relevant experimental model to more closely examine the least-well understood difference of the ICAP framework between intensely active classroom experiences.

Methods: Experimental setting

The classroom used in this ecological experimentation was an introductory science course at a large public university. A diverse enrollment of more than 750 undergraduate students were taught in equally large back-to-back sections for 50 minutes 4x per week. Sections were taught by the same instructor over the same topic over the same concepts. Students self-selected one of the two class sections during open enrollment. Associated 2.5-hr labs were held each week of the 10-week course. Students were evaluated through a number of assignments, with the large majority of the grading variation coming from four non-cumulative exams. Grades were a strong motivating factor for students (Wiggins et al 2015, in review) as the median grade was 2.9 while a 2.0 was required to continue on in the course series. For each section, demographic statistics were procured from the university registrar to describe categorical features and past numerical academic outcomes.

As described by registrar statistics, this classroom was comprised of 61% female students, 6% community college transfer students, 56% non-Caucasian students, 46% First Generation students, and 18% underrepresented minorities. Students in the class were predominantly of sophomore and junior standing and had declared a wide range of majors,

typically in the natural sciences. The average SAT scores of this population were 549 for Math and 515 for Verbal. All classroom data and student consent was protected and managed under IRB protocols. No students opted out of the final consented student population for use of their outcome and demographic data.

The split-section environment allows for rigorously controlled experimentation with teaching techniques or interventions. By manipulating aspects of instruction, we can examine learning outcomes that are largely controlled for topic, instructor, classroom environment, time on task, instruction wording, and motivating factors. Previous academic success for the sections could not be controlled due to the lack of random assignment, but we describe measures below to control for differences in historical student abilities within the topic in specific and academics in general. It is within this controlled experimental system that instruction was manipulated between interactive and constructive student activities.

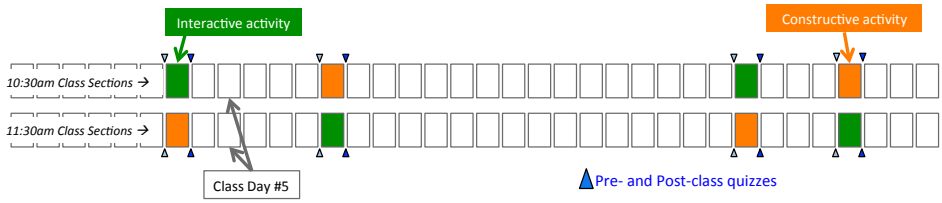


Fig 1. **Timing of experimental class sections and pre/post quiz administration**. For each of the four experimental topics, one section of the class used an interactive classroom activity while the other used a constructive activity. The choice of section was rotated to allow for a repeated-measures analysis. Pre-quizzes were administered as part of a daily reading quiz for each experimental day to students in both section. Post-quizzes were administered on the following reading quiz. Quizzes were open as of the previous class afternoon and closed early on the morning of the relevant class so that all learning-based outcome data was collected in the near term within 24 hours of the experiment. Note that in the final statistical analysis the first of four comparisons was removed (explanation below).

Four topics were chosen for use in constructive-vs-interactive experiments. These four commonly taught topics were chosen for their ubiquity among introductory science coursework and applicability of designed activities to similar courses at other institutions. Usefully, these four topics are generally not included in high school science courses, which simplifies interpretation by removing a confounding variable. For two of these topics, an existing activity was designed into activities along constructive or interactive lines of student action. For the other topics, interactive and constructive activities were designed *de novo*.

Design of activities was conducted using principles described in the ICAP framework (Chi and Wylie 2014). No real-world activity is likely to be a perfect fit into a single ICAP category, and the activities we designed here are no exception; our goal was to create experimental treatments that were predominantly within a single category and for which the differences otherwise were minimal. In designing constructive versions of activities, we focused on providing opportunities for students to generate outputs of their own understanding that went beyond the answers provided. Students were asked to integrate concepts across texts, to compare and contrast mechanisms, and to predict outcomes for new situations using conceptual understandings from relevant but distinct examples. Throughout our constructive activities, students worked in small groups but interaction with group members was neither required by the

tasks given nor implied by written or instructor guidance. In designing interactive versions of activities, those interactions between students to co-generate new understanding were prioritized. Typically, this was done through adaptation of a 'jigsaw' model in which each student in a small group was given different source information. While peer teaching does not inherently define interactive in the ICAP framework, these adapted jigsaws required groups to solve high-level cognitive tasks requiring information from each of the students and thinking significantly beyond what any one student was given. Completion of the activity required some combination of scientific debate and co-developing a solution or model.

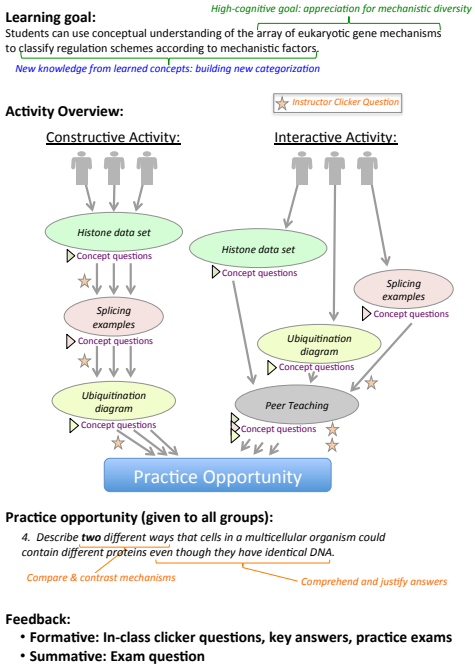


Figure 2. **Example of the difference between interactive and constructive strategies for a single learning goal**. In the constructive strategy, students collaboratively work through three mechanisms. They then use this conceptual understanding to build new knowledge in the practice opportunity that goes beyond the initial three mechanisms. In the interactive strategy, each student becomes a 'micro-expert' in one of the mechanisms. The practice opportunity can only be completed successfully by students that interact through debate and justification to parse a correct answer.

Re-design of classroom activities to better fit the ICAP framework was conducted iteratively across several prior quarters of the same. Student feedback through online written evaluations, personal conversations and intentional focus groups was used to develop better activities. Re-design tasks within our research team and with input from colleagues helped to incrementally center each activity into predominantly-constructive and predominantly-interactive domains.

	Protein Translation	Eukaryotic Gene Regulation	Cancer	PCR
Constructive	6	6	3	5
Interactive	2	7	3	3

Table 1. **Iterations in classroom use of activity versions**. For each version of each activity, student feedback was obtained after each classroom use. For example, the constructive version of the Cancer activity was used in three different quarters of the same course. A combination of student feedback, design team edits and outside suggestions by education research colleagues helped to iteratively develop activities that fit into *interactive* or *constructive* models within the ICAP framework. This table should make clear that design efforts were extensive and were initiated long before data collection for this study.

Methods: Measures

Pre/post tests

To assess student understanding of the key concepts introduced in the in-class activities, we designed multiple choice tests aligned with each activity's learning goals. Each test contained 8 items focused primarily on assessing higher-order cognitive skills (Blooms 1956; Crowe et al. 2008). As a measure of content validity, each question was reviewed by at least three experts in cell and molecular biology. To test the assumption that each question was measuring the same construct as the rest of the items on the test, we assessed item fit (Bond and Fox 2001) using the eRM package in R (Mair and Hatzinger 2007). Significant p values for item fit indicate an item is measuring a different construct. Based on these analyses, three questions were revised to more closely align with the learning goals of the activities. Rasch analysis of post-test results confirmed that there was a range of item difficulty on each of the tests allowing discrimination between students. For each activity, students completed the 8-item test on-line on the night prior to the activity as part of a daily reading quiz and then repeated the same 8-item test the night of the activity as part of the daily reading quiz for the subsequent day's lecture.

In-class observations

To monitor the level of student engagement during the two treatments, we performed two independent types of observations: 1) overall student interaction and 2) small groups. To monitor overall student interaction, two experienced observers from the Center for Teaching and Learning attended each of the six class sessions in which the constructive and interactive activities were implemented. The observers counted the number of students talking [Note: Insert Karen's description of protocol here including timing of the two observations, 1st while students are learning the concepts and 2nd during completion of the integration questions]. Interrater reliability calculated using the Intraclass Correlation ranged from moderate to strong depending on the observer pair (0.62-0.85) (Portney and Watkins 2000; Landis and Koch 1977). One of the observers was present at all six class sessions and performed the count of total number of students per row at the beginning of each class session. We chose to use that single observer's observation data for maximum consistency across treatment and topic. We also made the assumption that the total number of students per row remained constant within a single class session. Observation data was pooled across all three topics for each treatment set. A similar number of students were observed for each of the treatments (Constructive, n=660; Interactive, n=647). Consistent with the increased structure and role assignments built into the interactive activities, we found significantly more students talking to each other during the interactive

activities than the constructive activities (Fisher's Exact Test, $p < 0.001$, $p = 0.01$, respectively for the two observation time points).

Statistical Analysis of Student performance on Post-Tests

This study utilized repeated measures of the same student's performance in two different contexts: Constructive vs Interactive Activities. Thus, statistical analyses had to account for the non-independence of the post-test scores (post-test scores of the same student are more likely to be similar to each other than post-test scores of different students). In addition, post-test scores on any individual activity were not normally distributed instead they were left-skewed. Both of these properties made the typical linear regression analysis inappropriate. Instead, we employed a generalize mixed effects model with ordinal regression using the ordinal Package in R (Christensen 2010). Mixed effect models include a random effect term that can account for hierarchical structure in the data (in this case multiple post-test scores per student). Ordinal regressions treats the post-test score as if it was an ordered-categorical measure which is a reasonable approach in this case because the possible scores on the post-test are tightly bounded (ranging from 0-8) and partial credit was not possible. Ordinal regressions models the odd of getting at least one additional question correct on the post-test with an increase in an explanatory variable (i.e., as GPA increases the odds that a student will get at least one additional question correct on the post-test). In addition, our study design was quasi-random (students in two different classes), we included a variable to control for potential differences in student ability between the two classes in our model. This control was cumulative college GPA at the point of entry into the class. The majority of students in the study were sophomores and, thus, this control accounted for their performance in their first year of college courses. This measure has been shown in prior studies at this institution to be a strongly predict student performance in the introductory biology series (S. Freeman et al. 2011; Eddy 2015).

In a preliminary analysis we explored where the treatment effect was consistent across the 3 activities or whether it varied by activity topic (a treatment x activity interaction term). We did not find support for this interaction term ($p > 0.28$ for each comparison) and, thus, did not include it in the final model. Thus, our final model was:

$$\text{Post-test score} \sim \text{Cumulative GPA} + \text{Pre-test Score} + \text{Treatment} + (1|\text{Stu.ID})$$

In addition to documenting the average effect of the two activities on student performance, we also explored whether different student groups showed different patterns of responses on the post-test. In these analyses, we tested whether adding a proxy for socio-economic status (a binary variable indicating whether a student was eligible for the Education Opportunities Program), Gender (represented as a binary as we did not have the sample size to test the impact of activities on students who did not identify as male or female), or ethnicitiy/Race/nationality improved the explanatory power of our base model by step-wise adding in the student group variable and comparing the AIC values to the base model using an F-test [Note: Citation needed.]

Results:

We utilized an experimental system to test learning gains between interactive and

constructive teaching strategies in a classroom learning environment. This system controls for instructor, topic, learning environment, collaborative interaction and models a correction for student ability. Through a repeated-measures statistical analysis, we conclude that student learning was significantly improved by an interactive teaching strategy as compared to a constructive strategy. This is strong evidence in support of the prediction made within the ICAP framework (Chi and Wylie 2014).

On an eight-item content quiz in a pre/post format, a student taught with an interactive strategy was 25% more likely to answer at least one additional question correctly on the post-test than that same student taught with a constructive strategy. This change is similar in magnitude to the difference we would expect on the post-test for a student who has a cumulative GPA that is a quarter point higher than another student.

We did not find support for the hypothesis that the impact of completing a constructive vs. interactive activity varied between student groups. Using likelihood ratio tests we did not see that adding a main effect group or interaction term between treatment and group increased the fit of the model to the data for socioeconomic class ($p=0.9975$), gender ($p=0.3375$), or race/ethnicity/nationality (0.9789).

Discussion:

Interpretation of this statistic is not trivial. How much does a single question matter? It is worthwhile to parse the arguments for and against the meaningfulness of this result.

Thinking conservatively, this result indicates the correct choice of only a single multiple choice question on each quiz. STEM students answer many thousands of multiple choice questions within their overall education. Previous GPA and pre-testing are better predictors of post-test outcomes, indicating that this effect is subtle. Additionally, multiple choice questions may be answered correctly for a variety of reasons unrelated to deep conceptual knowledge (Stanger-Hall 2012; Darling-Hammond and Adamson 2014)_These interventions did not drastically change student exam outcomes; exams given 2-12 days after the intervention showed no significant different in scores. Indeed, even single interventions did not demonstrate statistically significant increases in performance on a particular pre-post quiz. These arguments reasonably capture doubt as to the ultimate importance of the difference between constructive and interactive teaching.

However, similar arguments hint at a larger importance of this finding. The difference in teaching was subtle: questions, data, and learning pathways were largely identical between version of the activities. Topic and instructor were controlled. While pre/post quizzes were administered in close array around the intervention time, motivation may have been low due to the participation-only nature of scoring for these quizzes [Note: Citation needed for effect of motivation on MC quiz score]. Perhaps most importantly, this real increase in learning was demonstrated for a single conceptually challenging topic. The benefits of interactive learning may drastically scale when used for hundreds of small-scale learning goals for each course. Given the complicated and social nature of human learning (Lave and Wenger 1991), it may be that any small intervention leading to a statistically significant improvement in outcome may hint at the rest of the iceberg of possibilities by making this slight instructional shift in practice. These significant differences in learning were due to a relatively subtle difference in teaching style of a small handful of 50-minute class periods. The use of the repeated measures format may be a effective to boil out the variability inherent in social experiences. If so, then this analytical framework may be one of few that have the power to assess subtle cultural effects appropriately.

Intensely-active strategies at the top of the ICAP spectrum are likely to bring intrinsic benefits not captured within our result. Both constructive and interactive methods give students opportunities to practice collaborative work, which will hopefully lead to improved social skills and experience within this or other fields (Rosenberg et al. 2006; Fine and Harrington 2004). Both methods position students in a growth mindset as apprenticing experts instead of as passive intake automatons as in traditional lecture classrooms (Nasir and Hand 2006). Both are likely to incorporate the same gains seen in a wide variety of STEM teaching environments from even the partial use of active learning strategies (S. R. Freeman et al. 2014; Dauer and Long 2015). Interactive methods have the special characteristic of placing students as 'micro experts', which is likely to improve attitudinal outcomes like confidence and grit (Duckworth et al. 2007). While these likely gains are important for the development of next-generation scientists, all are extra benefits beyond the results described here.

We found no evidence for disparities in learning gains between different instructional strategies in reference to different demographic groups of students. This may result from diversity within registrar-delineated groups, or it may be due to a lack of power in our analytic procedures and statistical analysis. Further research would clearly need to be done to demonstrate equity among intensely active instructional strategies. For now, it is at least encouraging that our close analysis does not demonstrate obvious learning gaps along ethnic, racial, gender, or first-generation-status lines.

Implications:

In light of this improved student outcomes in comparison with high-level constructive activities, interactive teaching strategies are likely to be the superior mode of intensely active instruction. Interactive instruction should be the preferred option for all classrooms in teaching environments with unlimited resources, time, instructor experience, and instructor training or professional development. However, instructors and institutions must negotiate limitations and goals in real education environments. Interactive strategies are likely to require more from instructors both in development time and instructional ability. It may be best practice to limit the use of interactive methods to those topics most easily adapted for student use in this format (for example: subjects for which the learning goals require conceptual understanding across multiple mechanisms). Greater gains overall might be seen in shifts from passive to active instruction, or from active to constructive. More research is needed to better understand the potential gains to students and instructors from interactive or constructive methods. This will help to better guide instructional development decisions at many levels.

Student learning is not a simple scale; the benefits of instructional choices will be mediated by complex characteristics of individual and group learning within a dynamic cultural environment (Bang and Medin 2010). For active learning, these mediations remain incompletely understood. Active classrooms may have some benefit for groups traditionally underserved by more didactic instruction (S. R. Freeman et al. 2014). The extent to which this benefit reaches all students will require deeper cultural research with implementation of intensely active experiences. Our data suggest no overt link between demographically-assigned race, ethnicity, gender and intensely active classroom experiences. This likely means that we do not have the power or the breadth to understand those links yet. This research will be most useful when conducted locally by practitioners to best inform instructional choices [Note: Need citation choice here].

The classrooms used in the experiments described here employed intensely active

learning strategies within an environment in which active learning was a daily norm. The benefits of interactive activities might be predicted to be higher in classrooms where students were already acculturated to their use, or they might be predicted to be lower in situations where the novelty and increased classroom energy would wane after repeated usage. While it is likely that a diversity of instructional methods is best, the extent to which any particular method is useful is unlikely to be precisely described.

These experiments controlled for the identity, talent, and preparation of the instructor. Those variables inform the minute-to-minute practice of teaching in ways that can profoundly impact student learning. Small language choices can invoke stereotype threat or instill a growth mindset (Steele 1997; Duckworth et al. 2007; Dweck 2008). Teachers, even with post-secondary audiences, carry a heavy and delicately balanced burden. To best understand the use of intensely active teaching strategies, the professional development and discipline-based practice habits of instructors must be better understood. Our experimental system controlled for this powerful variable in the simplest incomplete manner possible. Future research into the predictions of frameworks around the use of active learning strategies must necessarily engage directly in the dialogue within and between students and instructors.

Ben Wiggins 9/30/2015 12:43 PM
Formatted: Normal, No bullets or numbering, Tabs:Not at 0.15" + 0.5"

Acknowledgements:

- Grants
- NSF TUES #??
- Help from BERG
- Karen Friesem at CTL
- Grunspan (if not an author outright)
- Students

References

Bang, M., & Medin, D. (2010). Cultural Processes in Science Education: Supporting the Navigation of Multiple Epistemologies. *Science Education, 94*(6), 1008-1026, doi:Doi 10.1002/Sce.20392.

Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist, 49*(4), 219-243, doi:Doi 10.1080/00461520.2014.965823.

Christensen, R. H. B. (2010). ordinal—regression models for ordinal data. *R package version, 22*.

Committee on the Status, C. a. F. D. o. D.-B. E. R. (2012). Discipline-Based Education Rsearch: Understanding and Improving Learning in Undergraduate Science and Engineering. In B. o. S. E. o. t. N. R. Council (Ed.). National Academies Press: National Research Council.

Darling-Hammond, L., & Adamson, F. (2014). *Beyond the bubble test: How performance assessments support 21st century learning*: John Wiley & Sons.

Dauer, J. T., & Long, T. M. (2015). Long‐term conceptual retrieval by college biology majors following model‐based instruction. *Journal of Research in Science Teaching, 52*(8), 1188-1206.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *J Pers Soc Psychol, 92*(6), 1087.

Dweck, C. (2008). Mindsets and math/science achievement. New York: Carnegie Corporation of New York, Institute for Advanced Study, Commission on Mathematics and Science Education.

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).

Fine, G. A., & Harrington, B. (2004). Tiny publics: Small groups and civil society. *Sociological Theory, 22*(3), 341-356.

Freeman, S., Haak, D., & Wenderoth, M. P. (2011). Increased course structure improves performance in introductory biology. *CBE Life Sci Educ, 10*(2), 175-186, doi:10.1187/cbe.10-08-0105.

Freeman, S. R., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., et al. (2014). Active learning increases student performance across STEM disciplines. *Proc Natl Acad Sci U S A*.

Holdren, J. a. L., E (2013). Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. Executive Office of the President: President's Council of Advisors on Science and Technology.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159-174, doi:10.2307/2529310.

Lave, J., & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. New York: Cambridge University Press.

Linton, D. L., Farmer, J. K., & Peterson, E. (2014a). Is Peer Interaction Necessary for Optimal Active Learning? *Cbe-Life Sciences Education, 13*(2), 243-252, doi:10.1187/cbe.13-10-0201.

Linton, D. L., Pangle, W. M., Wyatt, K. H., Powell, K. N., & Sherwood, R. E. (2014b). Identifying Key Features of Effective Active Learning: The Effects of Writing and Peer Discussion. *Cbe-Life Sciences Education, 13*(3), 469-477, doi:10.1187/cbe.13-12-0242.

Nasir, N. S., & Hand, V. M. (2006). Exploring sociocultural perspectives on race, culture, and learning. *Review of Educational Research, 76*(4), 449-475, doi:Doi 10.3102/00346543076004449.

Portney, L., & Watkins, M. (2000). *Foudnations of Clinical Research Applications to Practice*. New Jersey: Prentice Hall Inc.

Rosenberg, J. L., Lorenzo, M., & Mazur, E. (2006). Peer Instruction: Making Science Engaging. *Handbook of College Science Teaching*, 25.

Stanger-Hall, K. F. (2012). Multiple-Choice Exams: An Obstacle for Higher-Level Thinking in Introductory Science Classes. *Cbe-Life Sciences Education, 11*(3), 294-306, doi:Doi 10.1187/Cbe.11-11-0100.

Steele, C. M. (1997). A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance. *American Psychologist, 52*(6), 613-629.

Stockwell, B. R., Stockwell, M. S., Cennamo, M., & Jiang, E. (2015). Blended Learning Improves Science Education. *Cell, 162*(5), 933-936.

Waldrop, M. M. (2015). Why we are teaching science wrong, and how to make it right. *Nature, 523*, 272-274, doi:10.1038/523272a.