ASPECT: A Survey to Assess Student Perspective of Engagement in an Active-Learning Classroom

Benjamin L. Wiggins,[†‡] Sarah L. Eddy,[†§‖] Leah Wener-Fligner,[¶] Karen Freisem,[#] Daniel Z. Grunspan,[@] Elli J. Theobald,[‡] Jerry Timbrook,[**] and Alison J. Crowe[‡*]

[†]Department of Biology, University of Washington, Seattle, WA 98195-1800; [§]Biology Department and [‖]STEM Transformation Institute, Florida International University, Miami, FL 33199; [¶]College of Education, University of Washington, Seattle, WA 98195-3600; [#]Center for Teaching and Learning, University of Washington, Seattle, WA 98195-1265; [@]Center for Evolution and Medicine, Arizona State University, Tempe, AZ 85287-4501; **Department of Sociology, University of Nebraska, Lincoln, NE 68588

ABSTRACT

The primary measure used to determine relative effectiveness of in-class activities has been student performance on pre/posttests. However, in today's active-learning classrooms, learning is a social activity, requiring students to interact and learn from their peers. To develop effective active-learning exercises that engage students, it is important to gain a more holistic view of the student experience in an active-learning classroom. We have taken a mixed-methods approach to iteratively develop and validate a 16-item survey to measure multiple facets of the student experience during active-learning exercises. The instrument, which we call Assessing Student Perspective of Engagement in Class Tool (ASPECT), was administered to a large introductory biology class, and student responses were subjected to exploratory factor analysis. The 16 items loaded onto three factors that cumulatively explained 52% of the variation in student response: 1) value of activity, 2) personal effort, and 3) instructor contribution. ASPECT provides a rapid, easily administered means to measure student perception of engagement in an active-learning classroom. Gaining a better understanding of students' level of engagement will help inform instructor best practices and provide an additional measure for comprehensively assessing the impact of different active-learning strategies.

INTRODUCTION

National reports aimed at improving undergraduate science education have called for a shift away from the traditional "sage on a stage" mode of lecturing toward the use of student-centered, evidence-based instructional approaches (National Research Council, 2003; American Association for the Advancement of Science, 2011; President's Council of Advisors on Science and Technology, 2012). This is due in part to the fact that increasing the amount of active learning in the classroom has been shown to benefit student learning (Freeman *et al.*, 2014). In an active-learning environment, students spend more time coconstructing knowledge with their peers (Chi and Wylie, 2014), which requires the ability to form effective working interactions with friends or peer strangers (Lorenzo *et al.*, 2006). Many factors have been found to influence whether or not students actively engage in small-group work, including English language proficiency, perceived value of the activity, and group composition (Chatman *et al.*, 2008; Dasgupta and Stout, 2014; Grunspan *et al.*, 2014). However, there is little literature on how more social rather than individual learning is impacting students' experience in the classroom (Kurth *et al.*, 2002; Hand, 2006). For example, being the only member of a particular social category (e.g., gender or ethnicity) has the potential

Jennifer Momsen, *Monitoring Editor*

Submitted August 10, 2016; Revised December 21, 2016; Accepted January 6 2017

CBE Life Sci Educ June 1, 2017 16:ar32

DOI:10.1187/cbe.16-08-0244

[†]These authors contributed equally to this work.

*Address correspondence to: Alison J. Crowe (acrowe@uw.edu).

to negatively impact performance of socially disadvantaged or underrepresented groups (Sekaquaptewa *et al.*, 2007; Chatman *et al.*, 2008). As educators, how we structure our active-learning environments could therefore have implications for students' sense of belonging in the classroom and, ultimately, their learning. Better understanding how students perceive their learning environments and why they do or do not choose to engage in an activity will help inform best practices in designing active-learning exercises.

Engagement is a multifaceted concept and includes dimensions ranging from behavioral (being on task) to cognitive (exerting effort) to affective (being invested in a task) (Christenson *et al.*, 2012; Reeve and Lee, 2014). As our interest is in how students engage with active-learning exercises, we will use the term "engagement" here to mean "learning task engagement" as defined by Chapman to encompass "students' cognitive investment, active participation, and emotional engagement with specific learning tasks" (Chapman, 2003, p. 1). Many studies have shown a positive correlation between student engagement and achievement (Dweck, 1986; Wigfield and Eccles, 2000; Hidi and Renninger, 2006; Hulleman *et al.*, 2008; Chi and Wylie, 2014; Reeve and Lee, 2014), leading to the development of a number of different theoretical frameworks to explain this relationship (Dweck, 1986; Wigfield and Eccles, 2000; Chi and Wylie, 2014). Although the underlying motivations driving engagement may vary, it is clear that measuring the extent to which students do or do not engage is important for comprehensive assessment of the effectiveness of active-learning strategies.

There are already several classroom observation tools that can be used to measure overall student participation in the classroom (Sawada *et al.*, 2002; Hora and Ferrare, 2010; Smith *et al.*, 2013; Eddy *et al.*, 2015). An additional observation tool was recently developed to specifically assess student behavioral engagement in large college classrooms (Lane and Harris, 2015). However, these observation tools are limited (by design) to measuring overt behaviors and thus do not capture the internal level of investment or value students are placing on an activity. This can be problematic, according to Pritchard (2008), who documented poor correlation between outward manifestations of traditional "engaged" behaviors (such as sitting upright or looking at the instructor) and student self-reported engagement. This suggests that relying solely on classroom observation may not provide a complete picture of a student's level of involvement. These findings are not surprising, as attentiveness can have many manifestations that fall outside "engaged" behavioral norms, so a student may be deeply engaged in a thought-provoking activity but not exhibit overt signs of engagement (Chi and Wylie, 2014). Ultimately, it is difficult to measure behavioral engagement and even more difficult to measure cognitive and affective engagement through external observation unless student work and attitudes are analyzed (Hart, 1994; Radford *et al.*, 1995).

Alternatively, it is possible to assess student cognitive and affective engagement by asking students to reflect on their own levels of engagement. Several published questionnaires rely on self-report data to provide a more complete view of engagement (Chapman, 2003; Handelsman *et al.*, 2005; Pazos *et al.*, 2010). Although self-report data have the limitation that students may not accurately assess their own levels of engagement (Assor and Connell, 1992), it has the advantage of being able to provide some insight into why students find an activity more engaging, not just whether or not they are visibly engaged. However, of the surveys intended for college students, many are focused on a single aspect of a student's experience such as personal motivation or sense of belonging (Pintrich *et al.*, 1993; Hagerty and Patusky, 1995). Others are geared toward assessing student engagement in a traditional, lecture-based classroom (Handelsman *et al.*, 2005) or are specific for a single type of active-learning strategy such as problem-based learning (Pazos *et al.*, 2010).

Our goal here was to develop a more broadly applicable survey that would enable comparison of the relative effectiveness of different in-class activities at engaging students across the cognitive and affective dimensions of engagement. As we explain later, we have taken a mixed-methods approach to develop a survey that is grounded in the experience of undergraduate biology students and can be used to assess multiple aspects of student self-reported engagement. The survey was designed to be able to capture student engagement for a wide variety of active-learning strategies commonly used in college classrooms. We have chosen to focus on measuring students' self-perception of engagement rather than measuring student behavior in order to capture the cognitive and affective dimensions of engagement. The survey is based on themes that arose during student interviews and focus groups and has been validated in a large introductory biology classroom. The resulting 16-item survey, which we call the Assessing Student Perspective of Engagement in Class Tool (ASPECT), can be used to rapidly obtain quantitative data on student self-reported engagement in an active-learning classroom.

METHODS

Participants

The students who participated in this study were enrolled in one of three quarters of an introductory biology course at a large research university in the Pacific Northwest. The course is the second course of a three-course series, with class size ranging from 370 to 760, depending on the quarter being taught. Different quarters of this course were taught by different instructors, but always included high levels of active learning, including clicker questions, group worksheets, case studies, peer instruction, and whole-class discussions. As described by registrar statistics, students enrolled in this course, over all three quarters, were primarily sophomores (49%) and juniors (40%) and had declared a wide range of majors, typically in the natural sciences. Female students made up on average 60% of the classroom population. In addition, of the students enrolled in the course, 44.2% were Asian Americans, 39.5% were white Americans, 6.3% were international, 5.5% were Latin@s, 1.8% were Black Americans, 1.8% were Hawaiian and Pacific Islander, and 0.8% were Native Americans. Community college transfer students made up 6% of the class, and 46% of the students were first-generation college students.

Overview of Survey Development and Validation

Here, we provide an overview of our survey development process (Figure 1), which follows the process described by Corwin *et al.* (2015) and is consistent with Benson's validation framework (Benson, 1998). We have organized our description of the

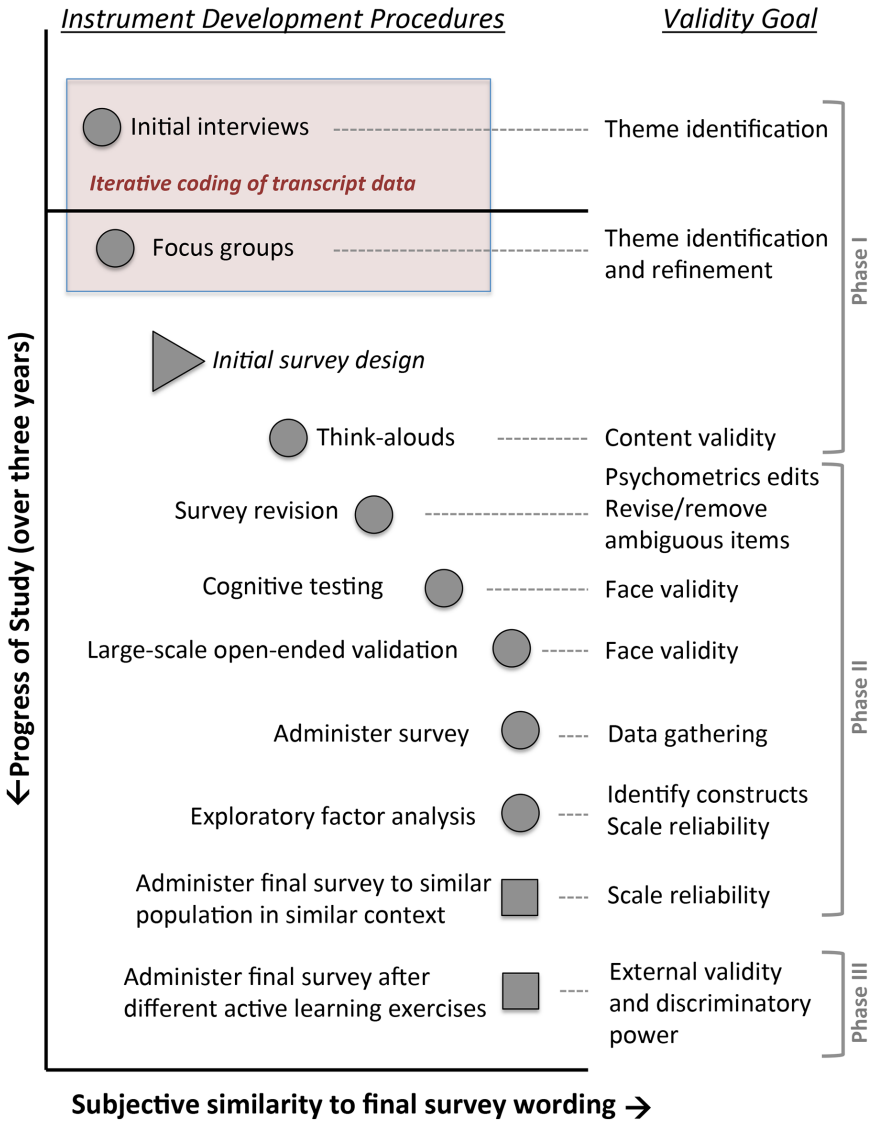

FIGURE 1. Overview of the development process for ASPECT. Final item wording was achieved through an iterative process of development, validation, and revision.

development and validation of ASPECT into three phases that parallel the three stages established by Benson (1998): phase I, in which we 1) develop the constructs to be assessed, 2) design the survey items, and 3) obtain face validity for the survey items; phase II, in which we assess the dimensionality and reliability of the survey; and phase III, in which we gather evidence for the external validity of the survey. Each phase is described in more detail in *Results*.

Phase I: Development and Validation of Constructs Measured in ASPECT

In phase I, we conducted student interviews and focus groups to identify the constructs, or themes, of engagement on which to base the survey. Participants were recruited through blind carbon-copied email forwarded by the instructor to randomly chosen students in the class. The response rate averaged across all groups was 5–7%. In total, 25 participants were recruited into a series of interviews (n = 2) and focus groups (n = 7, ranging from two to five students per focus group) over the course of Fall 2012 and Winter 2013. Participants encompassed qualitatively similar characteristics to the class as a whole in terms of ethnicity, race, gender, and final course grades. In the interviews, we asked students general questions to elicit their thinking about the activity that had taken place in class that day. After transcription and coding, the number of lines of text was used as an (imperfect) approximation of frequency of each code within the transcript. The interview process and student themes arising from the interviews are described in *Results*.

From the themes arising out of the focus groups, we wrote Likert-scale items aimed at determining the overall engagement students experienced in class. Items were edited extensively based on student think-alouds and best practices of survey design (Dillman *et al.*, 2014). The process of question development and revision is illustrated for one question (Figure 2) and included 1) standardizing the number of response alternatives to a six-point Likert scale across all items; 2) separating questions identified as containing two different ideas (i.e., "double-barreled") into two distinct items to ensure that respondents were only asked about one idea per survey item; and 3) revising questions identified as having ambiguous wording to contain more explicit, straightforward language. In addition, several of the original items required students to compare their experiences during an intensive active-learning day with those of a "normal class day." To remove possible confusion or alternative interpretations of a "normal class day," we replaced these items with questions asking students to reflect directly on that day's class (Figure 2). Finally, the survey was shortened to 20 items by removing redundant items (as determined through student think-alouds to be measuring the same general concept). Cognitive testing with a focus group of undergraduate biology students (n = 6) was performed to ensure the survey wording was clear and that students understood the intended meaning of each question. To obtain large-scale face validation of the survey items, we asked students to complete the entire survey online and then explain their thinking in open-ended responses for two to three randomly assigned survey items. This resulted in 30–40 short-answer responses per item. The process of coding these responses is described in *Results*.

Phase II: Validity and Reliability of ASPECT

In phase II, we assessed the dimensionality and reliability of the survey in three steps. First, we used pairwise correlation analysis to determine interitem correlation (Spearman's rank correlation coefficient). Second, we used iterative exploratory factor

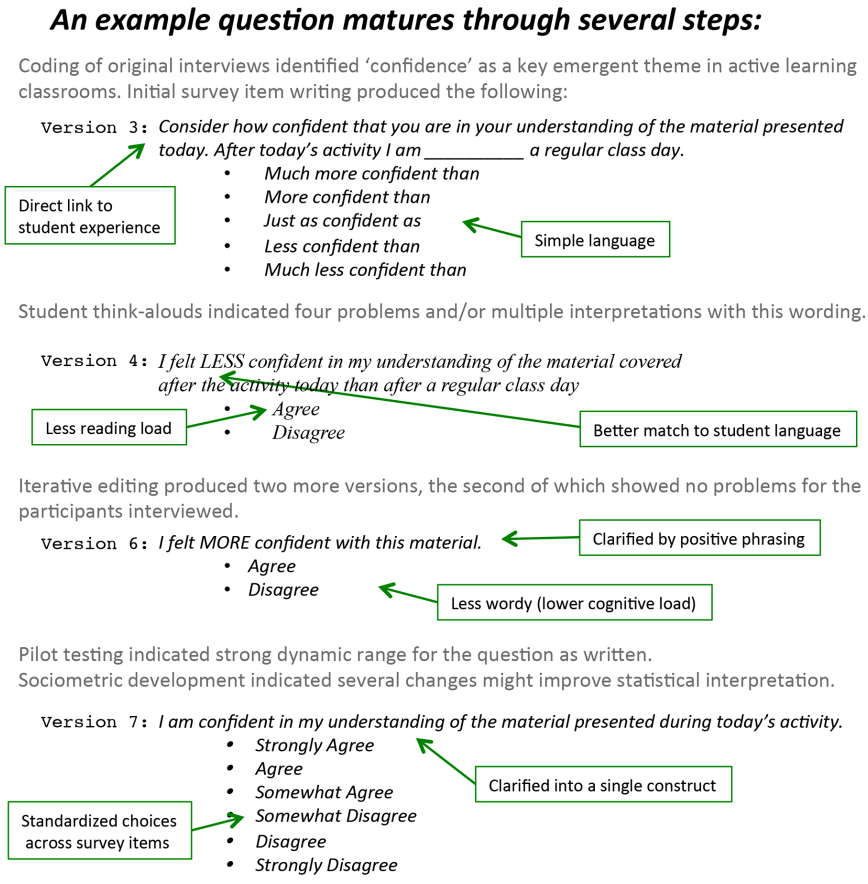

FIGURE 2. Example of development process for one survey item. This question was iteratively improved through the qualitative steps discussed in *Methods*. Examples of specific changes in the development of this question are noted.

analysis (EFA) to determine the dimensionality of the survey and assess the internal consistency of the scales. For this, we used an oblique (promax) rotation, as we hypothesized that different aspects of engagement would be correlated with one another. We performed this EFA analysis on survey responses from online administration of ASPECT in a single quarter of introductory biology ($n = 425$). Students with missing responses ($n = 17$; 4.5% of total) were excluded from this analysis. All EFAs were conducted using the "psych" package in R (Revelle, 2014). Finally, to test the reliability and internal consistency of the scales identified by EFA, we administered ASPECT to a similar population in a subsequent quarter of the same course (an additional $n = 760$ students) and used Cronbach's alpha (Cronbach, 1951) as a measure of the internal consistency of the interrelatedness of the items. In both cases, EFA identified three factors: value of group activity (Value), personal effort (PE), and instructor contribution (IC).

Phase III: External Validity of ASPECT

In phase III, the final stage, we assessed whether ASPECT could discriminate between different activity types and different demographic populations as a measure of external validity of the survey. We compared student responses after completing either 1) a long-activity day in which students worked in groups to complete a worksheet (~30 minutes long) or 2) a short-activity day with a series of clicker-question activities centered around instructor-posed questions. Cronbach's alpha (Cronbach, 1951) was calculated for the short-activity-day responses to measure reliability of the scales. To compare student responses to ASPECT after the two activity types, we summed the Likert-scale score (ranging from 1 to 6) of the questions within a construct (Value = 9 questions, PE = 3 questions, and IC = 4 questions), such that students could indicate Value ranging in score from 9 to 54 points, PE ranging in score from 3 to 18, and IC ranging in score from 4 to 24. We then independently modeled each construct of the survey (Value, PE, and IC) using linear mixed models. Mixed models were necessary, because we had a repeated-measures design in which the same students took ASPECT twice, once after experiencing a short-activity day and once after experiencing a long-activity day. Mixed-effects models can handle the resulting nonindependence of outcomes by including a random effect term for student (Zuur *et al.*, 2009).

Our modeling procedure included three steps. First, we started by fitting a simple model, wherein the outcome was modeled solely as a function of the activity type (and the student random effect). Second, we fitted a complex model, in which the outcome was modeled as a function of the activity type and student demographics, including university grade point average (GPA), gender, first-generation status, and ethnicity (a categorical variable with four levels: white, Asian American, international, and underrepresented minority). After fitting the most complex model within these parameters, we selected the best-fit model by using backward selection, comparing AIC (Akaike's information criterion) from subsequently more simple models to evaluate improvement of model fit; we considered $\Delta AIC < 2$ to be an equivalent fit (Burnham and Anderson, 2002), in which cases we selected the model with the fewest parameters. The third step in our model selection procedure was similar to the second, but we initially fitted a full, saturated model with activity type, student demographics, and all interactions between demographics and activity type. We employed the same backward selection procedure. These models (the simple, complex, and full) test three nested, complementary hypotheses: first, that student engagement is distinguishable by activity type; second, that engagement is distinguishable by student characteristics, controlling for activity type; and third, that engagement is differentially distinguishable by student characteristics on different activity types.

Visual inspection of the residuals (Supplemental Figure S1) revealed that they were unevenly distributed, likely due to the ceiling effect in student responses (Supplemental Figure S2). A ceiling effect occurs in a survey when some respondents who gave the highest response (in our case, 6) *would have* responded at a higher level had they been able to do so. The ceiling effect in our data is an artifact of the Likert-scale nature of student responses (each question was answered on a 1–6 scale, and, as is typical of survey responses, students primarily answered in the upper ranges of this scale); a floor effect is in theory also possible, although our data did not display this pattern (Supplemental Figure S2). To determine whether this ceiling effect influenced our results, we fitted the final nonnull models (selected from the model selection procedure described earlier) as censored regression models (Henningsen, 2011). Censored regression models account for ceiling (and floor) effects by modeling an uncensored latent outcome in place of the censored observed outcome (Henningsen, 2011). The results from the censored regressions indicate qualitatively similar patterns (Supplemental Table S1), indicating that the results from the linear mixed models are not strongly biased.

All models were fitted in R version 3.2.3 (R Core Team, 2015). Mixed-effects models were fitted using the "lme4" package (Bates *et al.,* 2015) and censored regression models were fitted using the "censReg" package (Henningsen, 2016). Code used for fitting models can be found in the Supplemental Material. Owing to institutional review board (IRB) restrictions, data are available only upon request.

RESULTS

Phase I: Development and Validation of Constructs Measured in ASPECT

Coding and Identification of Emergent Themes from Individual Interviews and Focus Groups. We began by recruiting students (n = 2) who had engaged in different active-learning strategies in a large introductory biology classroom for open-ended interviews (Rubin and Rubin, 2011) to answer questions centered around how they perceived the class environment. 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? Follow-up questions were unscripted but were consistently intended to push students to explain their reasoning as deeply as possible. The initial student-generated themes arising from these interviews focused on group dynamics, instructor language, and process-oriented features related to the activity, such as how they were directed to interact with group members.

On the basis of these initial interviews and in hopes of capturing greater depth and breadth of student experiences, we assembled a series of focus groups as described in *Methods*. Focus groups were progressively shifted toward questions and discussions that explored these emergent themes (group dynamics, instructor language, and process-oriented features), using a grounded theory approach (Strauss and Corbin, 1998; Glaser and Strauss, 2009). After each focus group, transcripts were coded independently by two coders (B.L.W. and L.W.-F.). Codes were iteratively revised based on frequent discussion between the coders resulting in unanimous coding at each step; the final consensus codes are shown in Table 1. Through this process, the original themes identified in the initial interviews evolved. The two themes of group dynamics and process-oriented features emerged as a single theme focused on the value of the group activity; the theme of instructor language broadened into the impact of instructor contribution on an activity; and finally, a new theme arose focused on the amount of effort students perceived themselves investing in an activity. This resulted in the three major categories listed in Table 1: 1) utility and intrinsic value of the group activity, 2) personal effort invested during the activity, and 3) instructor contribution to the activity and to student learning.

Initial Survey Item Development and Content Validity. Focusing on the themes that were most prevalent in student talk (Table 1), we developed an initial set of 26 survey items through a short series of research group writing tasks and editing sessions (Dillman *et al.*, 2014). Content validity of the initial questions was provided through seven individual think-alouds (Gubrium and Holstein, 2002). Students read nascent survey items first silently, then aloud, and were then asked to answer the survey items out loud and to justify their reasoning for their answers. Finally, students were asked to explain or identify problematic items and to suggest alternative language if applicable. Items were then edited based on student talk during the think-alouds, with the mutual goals of maintaining coherence of student language and fidelity to the original qualitative emergent themes (Figure 2).

We next revised and refined ASPECT to conform to best practices in survey design (Dillman *et al.*, 2014) as described in *Methods* and illustrated in Figure 2. The revised survey contained 20 items: eight items asking about the value students placed on the activity, seven items asking about student effort and involvement with the material during the activity, and five items asking about the instructor contribution. Three "control" questions were also included at the beginning of the survey to allow us to control for variables we hypothesized might impact student engagement: group size, prior experience with active learning, and having a friend in the group. We refer to this version of ASPECT as "20 + 3" to indicate the 20 engagement items and the three control questions.

Cognitive Testing. Next, to determine whether the language in the revised survey was easily understandable and unambiguous to students, we performed a series of cognitive testing and face validation steps. The goal of cognitive testing was to identify any confusing wording or alternative interpretations of survey items that might lead to students giving the same answer for multiple reasons (Willis, 2004). Participants (n = 6) were randomly recruited to a focus group. Each student first completed the 20 + 3 item survey in paper form, and then the entire focus group worked together to discuss possible interpretations for each item and whether the primary interpretation aligned with the intended interpretation.

Focus group participants unanimously agreed on the primary interpretation of all but one item on ASPECT. For the items agreed upon, the salient interpretations matched the goals and researchers' intentions of the item in each case. The one potentially problematic item (One group member dominated discussion during today's group activity) was interpreted by different members of the focus group as having

TABLE 1. Descriptions and examples of emergent codes from student talk

Category	Code title	Description	Prevalence[a]	Representative quote
Instructor contribution	Instructor effort	Describes student perceptions of the effort spent by instructors both in and outside 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 contribution	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 contribution	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 effort	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 effort	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 effort	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."

[a]Prevalence was determined by counting the lines that were given a particular code title and dividing by the total number of lines of text (3267).

either negative or positive connotations. However, all members agreed that the intent of the question was to ask about group equity. We therefore decided to retain this item in the next step of validation (large-scale face validation, described in the following section). One additional 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 intentionally had large scope, so the inclusion of multiple constructs into "effort" was appropriate, and the item was not changed. The question stems for the final items included in the survey are available in Table 2; all questions were answered on a six-point Likert scale, ranging from 1 = strongly agree to 6 = strongly disagree.

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 in a subsequent quarter of the same course to complete the 20 + 3 item survey online (Supplemental Document S3); students were then asked to provide written explanations for why they answered the way they did for two randomly selected questions on the survey. We had a 96% response rate (n = 383), providing us with 29–40 open-ended responses per item. Student responses were independently coded by three researchers to identify the central themes emerging from student answers. Answers that were too vague to interpret or did not address the question (e.g., "I was sick that day") were removed from analysis.

After independently coding all student responses, three researchers came together to discuss and reach consensus on whether or not students were interpreting items as intended, using an approach similar to that employed by Zimmerman and Bell (2014). Similar to the results described in *Cognitive Testing*,

TABLE 2. Rotated factor loadings for the ASPECT[a]

Survey item		Value of activity	Personal effort	Instructor contribution
VA1[b]	Explaining the material to my group improved my understanding of it.	**0.80**[c]	0.11	−0.13
VA2	Having the material explained to me by my group members improved my understanding of the material.	**0.78**	−0.11	0.00
VA3	Group discussion during the [topic] activity contributed to my understanding of the course material.	**0.79**	0.00	0.04
VA4	I had fun during today's [topic] group activity.	**0.65**	0.04	0.14
VA5	Overall, the other members of my group made valuable contributions during the [topic] activity.	**0.41**	0.05	0.03
VA6	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
VA7	I am confident in my understanding of the material presented during today's [topic] activity.	**0.70**	0.04	−0.04
VA8	The [topic] activity increased my understanding of the course material.	**0.83**	−0.02	0.04
VA9	The [topic] activity stimulated my interest in the course material.	**0.71**	−0.07	0.14
PE1	I made a valuable contribution to my group today.	0.07	**0.73**	−0.04
PE2	I was focused during today's [topic] activity.	0.12	**0.71**	−0.05
PE3	I worked hard during today's [topic] activity.	−0.12	**0.91**	0.07
IC1	The instructor's enthusiasm made me more interested in the [topic] activity.	0.18	−0.7	**0.71**
IC2	The instructor put a good deal of effort into my learning for today's class.	0.02	0.00	**0.75**
IC3	The instructor seemed prepared for the [topic] activity.	−0.11	0.14	**0.72**
IC4	The instructor and TAs were available to answer questions during the group activity.	0.06	0.03	**0.45**
Cronbach's alpha		0.91	0.84	0.78

[a]Questions are reorganized for ease of reading of each factor. Items are considered to be a good fit for loading onto a factor if the loading coefficient is greater than 0.4 and also less than 0.3 on all other factors. Items with factor loadings less than 0.3 were removed. All items had six response items ranging from "strongly agree" to "strongly disagree." VA1 and VA3 had an additional "This did not happen today" response option.
[b]VA refers to a value of group activity scale item; PE to a personal effort scale item, and IC to an instructor contribution scale item.
[c]Factor loadings are bolded in the column pertaining to the factor on which they loaded best.

the question regarding a dominator in the group (One group member dominated discussion during today's group activity) had multiple interpretations but was found to be consistently interpreted as relating to group equity as intended (Supplemental Document S1) and so was retained. However, as described in *Phase II* below, this item was removed from our final EFA analysis due to lack of correlation with other items in the survey. Only one survey item (I engaged in critical thinking during today's group activity) was identified as problematic: although the cognitive testing focus group agreed on a single meaning of this item, the larger-scale analysis of student explanations of this item ($n = 29$) revealed variable interpretations of the term "critical thinking." Interpretations ranged from "the instructions were vague so I had to think critically to understand what the professor wanted" to "this activity evoked critical thinking because I had to think hard to answer the questions." For this reason, and because there is continued debate even among experts as to the definition of critical thinking, we decided to remove this item from the subsequent analysis, resulting in a 19 + 3 item survey. One item (The instructor put a good deal of effort into my learning for today's class) was inadvertently excluded from this large-scale validation process; however, think-alouds and cognitive testing did not reveal any conflicting interpretations of this item. A summary of themes arising from student explanations for their responses is available (Supplemental Document S1).

Phase II: Validity and Reliability of ASPECT

Refinement of Scales. ASPECT was designed to measure three constructs: 1) Value of group activity, 2) PE, and 3) IC. To determine whether survey items would be useful in measuring at least one of these constructs, we performed a pairwise correlational analysis of the 19 items remaining after face validation. Nonuseful items that consistently exhibited low interitem correlations (Spearman's rank correlation coefficient $r < 0.3$ for at least 80% of correlations) were removed (Tabachnick and Fidell, 2007). This resulted in the removal of one item (One group member dominated discussion during today's group activity) that showed no correlation with any other items in the survey, leaving 18 items.

We conducted several iterations of EFA with the remaining 18 items. There was evidence in support of both a three- and four-factor solution. The additional factor that arose in the four-factor solution contained four items, two of which cross-loaded strongly onto other factors. The four items were all related to group function (e.g., Overall, the other members of my group made valuable contributions during the activity; Group discussion during the activity contributed to my understanding of the course material). In the three-factor solution, these items combined with items intentionally constructed to capture "Value of the group activity" to form a single factor, which aligned closely with the Value theme arising from student focus groups. In our discussions with students, the value students placed on an activity was intimately connected to whether or not they perceived their group to be functioning well. Owing to the multiple instances of cross-loading in the four-factor solution and poor support for a fourth distinct construct, we chose to focus on the three-factor solution.

In the three-factor solution, two items (I knew what I was expected to accomplish; I felt comfortable with my group) loaded weakly onto multiple factors (<0.2). These two items were therefore removed from the final factor analysis, resulting in a 16-item survey. The responses to these items may be of particular interest to a researcher or instructor; thus, instead of removing them from the survey, we recommend analyzing them individually along with the third item that was not correlated with the rest of the survey items: "One group member dominated discussion during today's group activity." Both the final 16-item survey and the complete 20 + 3 survey are available (Supplemental Documents S2 and S3).

We conducted a final iteration of the three- factor solution for the EFA with the remaining 16 items. Factor loadings for the final survey (Table 2) were consistently above the suggested minimum cutoff of 0.32 (Tabachnick and Fidell, 2007). Cronbach's alpha, a measure of factor reliability (and therefore scale reliability), was greater than 0.78 for all three factors, providing confidence that the items within each scale are reliably measuring the same construct. Together, these findings provide evidence that the 16-item ASPECT is measuring three distinct constructs (Table 2) and that these constructs are aligned with the themes that emerged from student focus groups in phase I.

The following three factors explained 55% of the variation in student response:

- **Value of group activity**: The first factor consisted of nine items exploring students' 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 alpha for this scale was 0.91. This scale explained 30% of the variation in student response.
- **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; I made a valuable contribution to my group today). Cronbach's alpha for this scale was 0.84. This scale explained 12% of the variation in student response.
- **Instructor contribution**: The final factor included four items and 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; The instructor's enthusiasm made me more interested in the group activity). Cronbach's alpha for this scale was 0.78. This scale explained 13% of the variation in student response.

Scale Reliability. The range of Cronbach's alpha coefficients (Cronbach, 1951) 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 EFA, we administered ASPECT to a similar population of introductory biology students in a consecutive quarter of the same course for which we had performed the EFA. Cronbach's alpha coefficients for each scale ranged from 0.81 to 0.91, again providing evidence for the reliability of the scales (Supplemental Table S2). Histograms of student responses are available (Supplemental Figure S3).

Phase III: External Validity of ASPECT

To be a useful research tool, ASPECT must be sensitive to changing levels of student engagement with different activities. To test its ability to discriminate between activities, we compared ASPECT responses of introductory biology students during two different activity types: 1) a short-activity day with a series of 8–10 clicker-question activities centered around instructor-posed questions, and 2) a long-activity day in which students worked in groups to complete a worksheet (~30 minutes long) followed by clicker questions to check understanding. On the basis of student focus groups and our analysis of student open-ended responses to items on ASPECT, we hypothesized that students would place more value on the short activities compared with the one long activity, because students often voiced frustration regarding infrequent instructor feedback during the long activities. We also hypothesized that students would perceive the instructor putting in more effort on a short-activity day, because the instructor more frequently provides feedback to the entire class than is typical on a class day with a long activity. We did not have an a priori hypothesis about which context would be perceived to elicit more personal effort.

We first tested whether the questions on ASPECT still captured the same three constructs in this new population that had completed a day with short activities by calculating the Cronbach's alpha for each scale (Cronbach, 1951). Because ASPECT was designed to capture student opinion about in-class activities, we reasoned that the same scales should be observed when students reflect on the short instructor-directed activities typical of a regular day, as we found when surveying students after long-activity days. This was supported by our finding that Cronbach's alpha values for each scale on a short-activity day again fell between 0.78 and 0.91 (Supplemental Table S2).

We 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 ASPECT. We found that the ASPECT survey distinguishes between activity types, student populations, and student populations performing different activity types both in the Value students place on the activity and in the IC students perceive, but not the PE students put into the activity (Table 3). Specifically, there is evidence that activity type (Table 3, a–c) and student ethnicity (Table 3, b and c) predict a students' Value of an activity. As we predicted, on average, students value the long activity less than the short activity (Table 3, a–c), Asian-American students value the activity more than white students (Table 3, b and c), and Asian-American students and international students both value the long activity more than white students (Table 3c). Similarly, there is evidence that activity type and student ethnicity predict a student's perception of IC to an activity (Table 3, g–i). As we hypothesized, on average, students perceive less IC on the long activity compared with the short activity (Table 3g-i), Asian-American students perceive more IC than white students (Table 3, h and i), and international students perceive more IC on the long activity than white students (Table 3i). There is no evidence that different groups of students perceive that their PE changes in response to activity type (Table 3, d–f).

TABLE 3. The ASPECT survey is able to discriminate between types of activities (long and short) and types of students (ethnicity) on the Value and IC constructs, but PE was not predictable by student characteristics or activity type

ASPECT construct (outcome)	Intercept	Activity type[1]	Ethnicity[2]		Activity type × ethnicity		ΔAIC[3,4]
a. Value[5]	**43.35**	**−1.09**					4.04
b. Value[6]	**42.57**	**−1.09**	AA	**1.74**			8.7
			Int.	0.51			
			URM	0.83			
c. Value[7]	**43.15**	**−2.19**	AA	0.68	AA:Long	**2.08**	22.61
			Int.	−2.03	Int.:Long	**5.03**	
			URM	1.23	URM:Long	−0.81	
d. PE[5]	**15.08**	0.18					−2.53
e. PE[6]	**15.17**						0
f. PE[7]	**15.17**						0
g. IC[5]	**20.52**	**−1.15**					32.88
h. IC[6]	**20.07**	**−1.15**	AA	**0.96**			35.12
			Int.	0.60			
			URM	0.43			
i. IC[7]	**20.24**	**−1.50**	AA	0.61	AA:Long	0.70	41.71
			Int.	−0.28	Int.:Long	**1.77**	
			URM	0.72	URM:Long	−0.60	

[a]Table shows relationship effect sizes from linear mixed-effects models, in which students were specified as random effects. Superscripts indicate reference groups, starting models, and interpretation notes; boldface coefficients indicate significance to $\alpha < 0.05$. Gray cells indicate variables that were not included in the initial model; the model selection procedure is described in *Methods*.

[1]Reference level: short activity.

[2]Reference level: white; AA stands for Asian American; Int. stands for international; URM stands for underrepresented minority.

[3]Change from null model: outcome ~ 1 + (student random effect).

[4]AIC is used only to compare nested models, in this case, models modeling the same outcome.

[5]Simple model was specified as Outcome ~ Treatment + (student random effect).

[6]Complex model was specified as Outcome ~ Treatment + Demographics + (student random effect). Student demographics included university GPA, ethnicity, first-generation status, and gender.

[7]Full model was specified as Outcome ~ Treatment + Demographics + Treatment × Demographics + (student random effect). Student demographics included university GPA, ethnicity, first-generation status, and gender.

DISCUSSION

We have described here the development of ASPECT, a 16-item survey (Supplemental Document S2) that provides a rapid way to monitor students' perception of engagement. The survey, which takes students on average 6–7 minutes to complete, provides researchers and practitioners a new tool to assess student self-reported engagement in large enrollment active-learning classrooms.

In this mixed-methods study, we triangulated qualitative analysis of students' experience in an active-learning classroom with quantitative analysis of large-scale survey data to gain a richer understanding of student engagement. Based on the themes that emerged from qualitative student interviews and focus groups, ASPECT was intended to elicit student perception of three key constructs of cognitive and affective engagement in the active-learning classroom: 1) utility and intrinsic value of a group activity, 2) personal effort invested during an activity, and 3) instructor contribution to an activity and to student learning. EFA of student responses from a large-enrollment introductory biology class supports the assumption that ASPECT is measuring three discrete factors that align closely with these three constructs. We also provide evidence regarding the reliability of the three scales as measured by Cronbach's alpha coefficient in a similar population under similar conditions. The internal consistency of our findings using this mixed-methods approach provides increased confidence that these three constructs are aspects of the learning experience that affect students' engagement.

Our finding from focus groups and EFA that task value and personal effort are key factors in promoting student engagement has strong support in the sociocognitive literature (Dweck, 1986; Wigfield and Eccles, 2000; Eccles and Wigfield, 2002; Svinicki, 2004; Hidi and Renninger, 2006; Hulleman *et al.*, 2008). Specifically, expectancy value theory predicts that perception of an activity's value will be positively correlated with student interest and engagement (Eccles, 2005). Students place more value on tasks that they see as being either directly connected to their success, such as increasing performance on an exam or having a tangible connection to the world outside the classroom (Eccles and Wigfield, 2002; Hulleman *et al.*, 2008). Motivation to engage in a task is also influenced by how enjoyable the task is perceived to be and whether there is a high expectation of success in completing the task (Eccles and Wigfield, 2002; Eccles, 2005; Svinicki, 2004; Hug *et al.*, 2005). Our qualitative work also identified the importance of instructor contribution to student engagement. Although there is a growing literature on how instructor talk may influence student participation (Myers, 2004; Seidel *et al.*, 2015), future studies will be required to provide additional evidence that perception of instructor effort is positively correlated with student engagement.

When looking at student responses quantitatively, we detected some differences along these three constructs.

Specifically, ASPECT distinguished between activity types in terms of how much value students place on the different activities and how much instructor contribution they perceive but not the amount of personal effort they put forth. Unlike the direct association of motivation with task value (Eccles and Wigfield, 2002), the PE factor measured by ASPECT does not appear to be correlated with task value.

Limitations of ASPECT

ASPECT 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' self-reported involvement during in-class activities. Student perception of engagement is just one measure of engagement and, as with all self-report data, contains inherent biases. To gain a more holistic view of student engagement, one could administer ASPECT in conjunction with other tools designed to measure specific aspects of student behavior such as motivation or sense of belonging (Ryan *et al.*, 1983; Pintrich *et al.*, 1993; Hagerty and Patusky, 1995).

This survey was developed and validated with students in a single introductory biology course at one university and may therefore not be applicable to other course levels in biology, other science, technology, engineering, and mathematics (STEM) majors, or other populations. However, we found ASPECT reliably measures the same constructs in different iterations of the same course taught by different instructors. Furthermore, we have no reason to suspect that ASPECT would be differentially effective in different course levels or in other populations, although further field-testing is necessary to confirm this assumption.

Our intention is for ASPECT to be used to compare different active-learning strategies. Here, we show that ASPECT can discriminate between two different types of active-learning exercises that varied with respect to length and instructor feedback. Further field-testing of the instrument will be important to assess whether ASPECT can differentiate between active-learning exercises that differ in other elements. The best use of this survey will require validation in new classroom environments to ensure that the language and interpretation of ASPECT questions are meaningful for the new population (Lave and Wenger, 1991).

Implications for Research and Teaching

Research. As we elaborate below, ASPECT can be useful for researchers. First, as a research tool that targets student engagement, ASPECT may give insight into the "leaky STEM pipeline"; furthermore, student populations are differentially affected by active-learning activities, and ASPECT could lend insight into the specific differences; and finally, active learning as a research field is moving toward a finer-grained understanding of the most effective aspects of active learning, and one element of efficacy is student engagement.

Defined as the leaky STEM pipeline, only half of the students who enter college in the United States intending to major in a STEM discipline end up completing a science degree (President's Council of Advisors on Science and Technology, 2012). This indicates that far too many talented and interested students are lost along the way (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2007; Drew, 2011; Dasgupta and Stout, 2014). Students who exit STEM come disproportionately from backgrounds historically underrepresented in STEM and report social threats and unwelcoming atmosphere as major factors behind their decisions to leave STEM education (Steele, 1997; Seymour and Hewitt, 1998; London *et al.*, 2012; Graham *et al.*, 2013). The recent shift toward more student-centered learning in STEM classrooms is expected to increase retention (Kvam, 2000; McConnell *et al.*, 2003; Haak *et al.*, 2011), as active learning has the potential to disproportionately benefit underrepresented groups' learning outcomes (Springer *et al.*, 1999; Haak *et al.*, 2011; Eddy and Hogan, 2014). However, an important element to consider in retention of STEM majors is how students engage in classrooms. If paired with retention data, ASPECT could identify areas in which students' value, personal effort, and perceived instructor effort are correlated with attrition.

Furthermore, different student populations are disproportionally affected by active-learning activities (Springer *et al.*, 1999; Haak *et al.*, 2011; Eddy and Hogan, 2014). Underlying cultural factors, including gender spectra, race/ethnicity, and/or socioeconomic backgrounds are thus likely to impact student engagement during active learning. Our results suggest that, not surprisingly, students from different demographic groups perceive the same in-class activity differently. Specifically, Asian-American students saw more value in the group activity than white students. ASPECT could also detect differences in how demographic groups valued different types of activities. For example, the longer activity was more favorably perceived by both international students and Asian students compared with white students. Interestingly, student GPA did not predict whether or not students placed more value on the group activity, suggesting that, all else being equal, students in the top and bottom of the class do not have different perceptions of the activities.

Finally, as the biology education research field moves toward a finer-grained analysis of what makes active learning impactful, considering student engagement via ASPECT may prove beneficial. Because the modes and implementation of active learning vary widely (Andrews *et al.*, 2011; Freeman *et al.*, 2014), it will be important to continue to monitor engagement along with learning outcomes of different demographic populations to determine whether all students are engaging equally. Comparing student responses on ASPECT could enable researchers to assess the impact of different elements of an activity and could provide insight into why one activity is more beneficial than another. For example, by comparing two activities that differ in only one element, researchers could identify pedagogical approaches that influence how much value and personal effort student place on an activity. Additionally, one could compare activities that vary in either their mode of student interaction (e.g., with or without designated group roles) or the method of group assignment (e.g., self-selected, randomly assigned, or instructor assigned). In this way, one could determine, for example, whether students place more value and perceive themselves putting more personal effort into activities with highly structured roles for each individual. Pairing ASPECT data with follow-up student interviews or focus groups could help us to ascertain why students place more value on a particular

activity. Finally, as discussed earlier, ASPECT could help gain perspective on how students' unique characteristics, such as ethnicity, influence their experiences during an active-learning exercise.

Teaching. In addition to being useful to researchers, ASPECT can be helpful for practitioners. Teachers are faced with a myriad of decisions to make daily; when implemented strategically, ASPECT can be one avenue for data-driven decision making. As we enumerate below, by determining levels of student engagement with ASPECT, instructors can inform their decision to continue, modify, or discontinue an activity; and to inform their own teaching practices, instructors can use ASPECT for comparison between activities, student populations, and, potentially, between instructors.

First, as ASPECT is a measure of the level of student engagement on a particular activity, instructors can use these data to inform their decisions to revise or discontinue activities. There are many different and effective active-learning strategies (Tanner, 2013) and a number of different ways to implement active learning (Borrego *et al.*, 2013). After implementing a strategy or teaching an activity, instructors need to decide whether that activity was effective and whether they want to use it again, modify it, or avoid it in the future. ASPECT data can help inform this decision by providing data on the level of student engagement.

Additionally, instructors can use ASPECT data to compare activities, student populations, or, potentially, instructors. Comparing results on ASPECT between two activities can help inform prioritizing one activity over another. Furthermore, as previously discussed, ASPECT can distinguish between student populations, so if an instructor knows that a particular demographic group in a class is struggling academically, they may be able to employ ASPECT to determine whether engagement in this struggling population is also stunted. Finally, ASPECT can be used to determine whether there are certain aspects of instructor behavior that are most effective at engaging students. For example, comparing two activities with different instructor framing techniques could inform how to best set up an activity for students. In this way, ASPECT may also serve as a reminder for novice active-learning instructors that there are many different elements to consider when implementing student-centered strategies, including fostering functional groups and helping students see the value in an activity. Similarly, in a mentoring or coaching relationship, instructors might also be able to compare student engagement between two instructors; paired with additional classroom observation data about instructor habits (for example from a tool like PORTAAL [Eddy *et al.*, 2015] or COPUS [Smith *et al.*, 2013]), instructors may be able to enhance their own soft skills to increase engagement in their own classrooms.

CONCLUSIONS

Our goal with this work was to develop a survey to systematically gather student perception data to compare relative student engagement levels across various active-learning strategies. ASPECT differs from other instruments that have been designed to measure student experiences during active learning (Visschers-Pleijers *et al.*, 2005; Pazos *et al.*, 2010) in that we intentionally designed this survey to be widely applicable for different types of active learning. Our findings suggest that classroom culture, including small-group dynamics and instructor enthusiasm, could influence students' willingness and inspiration to engage in difficult STEM learning tasks. Gathering more information through tools such as ASPECT will help us better understand potential barriers presented by an active-learning environment (Malcom and Feder, 2016) and ideally develop strategies that increase engagement of all students. Our hope is that ASPECT will provide researchers and instructors alike with a tool to rapidly evaluate active-learning strategies from the perspective of the learner. These data can then be used, in conjunction with student performance data, focus group data, and even classroom observation data, to help inform instructional choices in the classroom.

ACKNOWLEDGMENTS

We thank M. P. Wenderoth and J. Doherty for their insightful and detailed comments on earlier versions of this article and the University of Washington Biology Education Research Group as a whole for valuable discussions on this topic. We also thank Kayla Evans and Dr. Linda Martin-Morris for sharing their analysis of student behaviors in the classroom. This research was completed under approved IRB protocol #44438 at the University of Washington and was supported in part by an award from the National Science Foundation (NSF DUE 1244847).

REFERENCES

American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Biology Education: A Call to Action, Washington, DC. http://visionandchange.org/finalreport (accessed 20 June 2016).

Andrews TM, Leonard MJ, Colgrove CA, Kalinowski ST (2011). Active learning not associated with student learning in a random sample of college biology courses. CBE Life Sci Educ 10, 394–405.

Assor A, Connell JP (1992). The validity of students' self-reports as measures of performance affecting self-appraisals. In: Student Perceptions in the Classroom, ed. DH Schunk and J Meece, Hillsdale, NJ: Erlbaum.

Bates D, Mächler M, Bolker B, Walker S (2015). Fitting linear mixed-effects models using lme4. J Stat Softw 67, 1–48.

Benson J (1998). Developing a strong program of construct validation: a test anxiety example. Educ Meas 17, 10–17.

Borrego M, Cutler S, Prince M, Henderson C, Froyd JE (2013). Fidelity of implementation of research-based instructional strategies (RBIS) in engineering science courses. J Eng Educ 102, 394–425.

Burnham KP, Anderson DR (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach, New York: Springer Science.

Chapman E (2003). Alternative approaches to assessing student engagement rates. Pract Assess Res Eval 8, 1–10.

Chatman JA, Boisnier AD, Spataro SE, Anderson CW, Berdahl JL (2008). Being distinctive versus being conspicuous: the effects of numeric status and sex-stereotyped tasks on individual performance in groups. Organ Behav Hum Decis Process 107, 141–160.

Chi MTH, Wylie R (2014). The ICAP Framework: linking cognitive engagement to active learning outcomes. Educ Psychol 49, 219–243.

Christenson SL, Reschly AL, Wylie C (2012). Handbook of Research on Student Engagement, New York: Springer Science.

Corwin LA, Runyon C, Robinson A, Dolan EL (2015). The laboratory course assessment survey: a tool to measure three dimensions of research-course design. CBE Life Sci Educ 14, ar37.

Cronbach LJ (1951). Coefficient alpha and the internal structure of tests. Psychometrika 16, 297–334.

Dasgupta N, Stout JG (2014). Girls and women in science, technology, engineering, and mathematics STEMing the tide and broadening participation in STEM careers. Policy Insights Behav Brain Sci 1, 21–29.

Dillman DA, Smyth JD, Christian LM (2014). Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method, Hoboken, NJ: Wiley.

Drew C (2011, November 6). Why science majors change their minds. New York Times, ED16.

Dweck CS (1986). Motivational processes affecting learning. Am Psychol 41, 1040–1048.

Eccles JS (2005). Subjective task value and the Eccles et al. model of achievement-related choices. In: Handbook of Competence and Motivation, ed. A Elliot and CS Dweck, New York: Guilford, 105–121.

Eccles JS, Wigfield A (2002). Motivational beliefs, values, and goals. Annu Rev Psychol 53, 109–132.

Eddy SL, Converse M, Wenderoth MP (2015). PORTAAL: a classroom observation tool assessing evidence-based teaching practices for active learning in large science, technology, engineering, and mathematics classes. CBE Life Sci Educ 14, ar23.

Eddy SL, Hogan KA (2014). Getting under the hood: how and for whom does increasing course structure work? CBE Life Sci Educ 13, 453–468.

Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, Wenderoth MP (2014). Active learning increases student performance in science, engineering, and mathematics. Proc Natl Acad Sci USA 111, 8410–8415.

Glaser BG, Strauss AL (2009). The Discovery of Grounded Theory: Strategies for Qualitative Research, New Brunswick, NJ: Transaction.

Graham MJ, Frederick J, Byars-Winston A, Hunter AB, Handelsman J (2013). Science education. Increasing persistence of college students in STEM. Science 341, 1455–1456.

Grunspan DZ, Wiggins BL, Goodreau SM (2014). Understanding classrooms through social network analysis: a primer for social network analysis in education research. CBE Life Sci Educ 13, 167–179.

Gubrium JF, Holstein JA (2002). Handbook of Interview Research: Context and Method, Thousand Oaks, CA: Sage.

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.

Hagerty BMK, Patusky K (1995). Developing a measure of sense of belonging. Nurs Res 44, 9–13.

Hand VM (2006). Exploring sociocultural perspectives on race, culture, and learning. Rev Educ Res 76, 449–475.

Handelsman MM, Briggs WL, Sullivan N, Towler A (2005). A measure of college student course engagement. J Educ Res 98, 184–192.

Hart D (1994). Authentic Assessment: A Handbook for Educators, Menlo Park, CA: Addison-Wesley.

Henningsen A (2011). Estimating censored regression models in R using the censReg Package. CRAN. Retrieved from http://cran.r-project.org/web/packages/censReg/vignettes/censReg.pdf.

Henningsen A (2016). Package 'censReg.': 1–10. https://cran.r-project.org/web/packages/censReg/censReg.pdf.

Hidi S, Renninger KA (2006). The four-phase model of interest development. Educ Psychol 41, 111–127.

Hora M, Ferrare J (2010). The Teaching Dimensions Observation Protocol (TDOP), Madison: Wisconsin Center for Education Research, University of Wisconsin–Madison.

Hug B, Krajcik JS, Marx RW (2005). Using innovative learning technologies to promote learning and engagement in an urban science classroom. Urban Educ 40, 446–472.

Hulleman CS, Durik AM, Schweigert SB, Harackiewicz JM (2008). Task values, achievement goals, and interest: an integrative analysis. J Educ Psych 100, 398–416.

Kurth LA, Anderson CW, Palincsar AS (2002). The case of Carla: dilemmas of helping all students to understand science. Sci Educ 86, 287–313.

Kvam PH (2000). The effect of active learning methods on student retention in engineering statistics. Am Stat 54, 136–140.

Lane ES, Harris SE (2015). A new tool for measuring student behavioral engagement in large university classes. J Coll Sci Teach 44, 83–91.

Lave J, Wenger E (1991). Situated Learning: Legitimate Peripheral Participation, Cambridge, UK: Cambridge University Press.

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, 961–979.

Lorenzo M, Crouch CH, Mazur E (2006). Reducing the gender gap in the physics classroom. Am J Phys 74, 118–122.

Malcom S, Feder M (2016). Barriers and Opportunities for 2-Year and 4-Year STEM Degrees: Systemic Change to Support Students' Diverse Pathways, Washington, DC: National Academies Press.

McConnell DA, Steer DN, Owens KD (2003). Assessment and active learning strategies for introductory geology courses. J Geosci Educ 51, 205–216.

Myers SA (2004). The relationship between perceived instructor credibility and college student in-class and out-of-class communication. Commun Rep 17, 129–137.

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine (2007). Rising above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future, Washington, DC: National Academies Press.

National Research Council (2003). BIO2010: Transforming Undergraduate Education for Future Research Biologists, Washington, DC: National Academies Press.

Pazos P, Micari M, Light G (2010). Developing an instrument to characterise peer-led groups in collaborative learning environments: assessing problem-solving approach and group interaction. Assess Eval High Educ 35, 191–208.

Pintrich PR, Smith DA, García T, McKeachie WJ (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). Educ Psychol Meas 53, 801–813.

President's Council of Advisors on Science and Technology (2012). Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering and Mathematics, Washington, DC: U.S. Government Office of Science and Technology. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/fact_sheet_final.pdf (accessed 20 June 2016).

Pritchard GM (2008). Rules of engagement: how students engage with their studies. Newport 1, 45–51.

Radford DL, Ramsey L, Deese W (1995). Demonstration assessment: measuring conceptual understanding and critical thinking with rubrics. Sci Teach 62, 52–55.

R Core Team (2015). R: A Language and Environment for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing. www.R-project.org (accessed 15 July 2016).

Reeve J, Lee W (2014). Students' classroom engagement produces longitudinal changes in classroom motivation. J Educ Psychol 106, 527–540.

Revelle W (2014). Psych: Procedures for Personality and Psychological Research, Evanston, IL: Northwestern University.

Rubin HJ, Rubin IS (2011). Qualitative Interviewing: The Art of Hearing Data, Thousand Oaks, CA: Sage.

Ryan RM, Mims V, Koestner R (1983). Relation of reward contingency and interpersonal context to intrinsic motivation: a review and test using cognitive evaluation theory. J Pers Soc Psychol 45, 736.

Sawada D, Piburn MD, 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 Sci Math 102, 245–253.

Seidel SB, Reggi AL, Schinske JN, Burrus LW, Tanner KD (2015). Beyond the biology: a systematic investigation of noncontent instructor talk in an introductory biology course. CBE Life Sci Educ 14, ar43.

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, 321–327.

Seymour E, Hewitt NM (1998). Talking about Leaving: Why Undergraduates Leave the Sciences, Boulder, CO: Westview.

Smith MK, Jones FH, Gilbert SL, Wieman CE (2013). The Classroom Observation Protocol for Undergraduate STEM (COPUS): a new instrument to

characterize university STEM classroom practices. CBE Life Sci Educ 12, 618–627.

Springer L, Stanne ME, Donovan SS (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: a meta-analysis. Rev Educ Res 69, 21–51.

Steele CM (1997). A threat in the air: how stereotypes shape intellectual identity and performance. Am Psychol 52, 613–629.

Strauss A, Corbin J (1998). Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, Thousand Oaks, CA: Sage.

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

Tabachnick BG, Fidell LS (2007). Using Multivariate Statistics, 5th ed., Needham Heights, MA: Allyn & Bacon.

Tanner KD (2013). Structure matters: twenty-one teaching strategies to promote student engagement and cultivate classroom equity. CBE Life Sci Educ 12, 322–331.

Visschers-Pleijers AJ, Dolmans DH, Wolfhagen IH, van der Vleuten CP (2005). Development and validation of a questionnaire to identify learning-oriented group interactions in PBL. Med Teach 27, 375–381.

Wigfield A, Eccles JS (2000). Expectancy-value theory of achievement motivation. Contemp Educ Psychol 25, 68–81.

Willis GB (2004). Cognitive Interviewing: A Tool for Improving Questionnaire Design, Thousand Oaks, CA: Sage.

Zimmerman HT, Bell P (2014). Where young people see science: everyday activities connected to science. Int J Sci Educ Part B 4, 25–53.

Zuur A, Ieno E, Walker N, Saveliev A, Smith G (2009). Mixed Effects Models and Extensions in Ecology with R, New York: Springer.