Beyond Team Makeup: Diversity in Teams Predicts Valued Outcomes in Computer-Mediated Collaborations

Angela E.B. Stewart
University of Colorado Boulder
Boulder, CO
angela.stewart@colorado.edu

Mary Jean Amon
University of Central Florida
Orlando, FL
mamon@ist.ucf.edu

Nicholas D. Duran
Arizona State University
Glendale, AZ
nduran4@asu.edu

Sidney K. D'Mello
University of Colorado Boulder
Boulder, CO
sidney.dmello@colorado.edu

ABSTRACT
In an increasingly globalized and service-oriented economy, people need to engage in computer-mediated collaborative problem solving (CPS) with diverse teams. However, teams routinely fail to live up to expectations, showcasing the need for technologies that help develop effective collaboration skills. We take a step in this direction by investigating how different dimensions of team diversity (demographic, personality, attitudes towards teamwork, prior domain experience) predict objective (e.g. effective solutions) and subjective (e.g. positive perceptions) collaborative outcomes. We collected data from 96 triads who engaged in a 30-minute CPS task via videoconferencing. We found that demographic diversity and differing attitudes towards teamwork predicted impressions of positive engagement, while personality diversity predicted learning outcomes. Importantly, these relationships were maintained after accounting for team makeup. None of the diversity measures predicted task performance. We discuss how our findings can be incorporated into technologies that aim to help diverse teams develop CPS skills.

Author Keywords
Diversity; team makeup; collaborative problem solving; learning technologies.

CSS Concepts
• **Human-centered computing~Empirical studies in collaborative and social computing**

INTRODUCTION
Consider a team of three college students working collaboratively on a virtual physics lab. The studesnts have not worked together before and must complete a lab problem solving activity on energy transfer in a limited amount of time. All three members of the team come from different backgrounds, bringing their own cultural experiences, academic preparation, and attitudes. In order to effectively complete the task, the students must communicate and coordinate within the heterogeneous team.

This hypothetical situation is broadly referred to as collaborative problem solving (CPS), which occurs when two or more people engage in a coordinated attempt to construct a solution to a problem [34,53]. CPS is considered an important 21st century skill as the workforce becomes increasingly team-based and the nature of work itself becomes increasingly non-routine [53].

Despite its importance, teams often fail to successfully engage in socio-cognitive processes necessary to support effective CPS, such as co-construction of solutions, monitoring progress, and maintaining a positive team dynamic [47,53]. In fact, process loss, where teams fail to achieve performance above theoretical baselines, is a well-documented phenomena in the group work literature [19].

There is also an increased demand for teams to interact in computer-mediated environments as the workforce is increasingly distributed and global [40,53]. Unfortunately, process loss is likely even worse in computer-mediated interactions compared to those that occur face-to-face. Modern computer-interfaces can obscure the transmission of important social signals, like direction of eye gaze or turn-taking [40], and poor bandwidth or other technological limitations further dampen communication.

Accordingly, modern educators have emphasized the need for students to develop CPS skills, especially in the context of computer-mediated communication [53]. We envision a 21st century solution for this challenge, where next-generation collaborative learning technologies can facilitate customized experiences to foster development of CPS skills. Such technologies should personalize learning content, goals, and feedback for the team at hand, taking into account the background, skills, and attitudes of the teammates. We take a step in this direction by focusing on one aspect of successful CPS – effective collaboration for diverse teams.

We choose to focus on diversity because the modern workforce requires individuals that are able to successfully collaborate on diverse teams [37,53]. Teams are rapidly becoming more global, requiring people with different

CHI 2020, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04...$15.00.
DOI: https://doi.org/10.1145/3313831.3376279

cultures and beliefs to work together [52,53]. Further, women and millennials have changed workplace demographics, bringing along with them vital new attitudes and values [45,52]. Additionally, the growing "gig" economy has given rise to temporary workers in spontaneously formed teams [18,23]. In these short-term teams, effective CPS processes must be quickly established, no matter the background of individual members of the team.

There is limited work that explicitly aims to develop CPS skills with diversity of the team at the forefront (e.g. [36]). The most closely related work focuses on assessing CPS skills, rather than developing it [6,24,36,46]. Thus, our long-term goal is to develop computer interfaces that teach CPS skills in a manner that is sensitive to team diversity. In particular, such an interface could assess team diversity along many pertinent dimensions (e.g. differences in personality or attitudes towards the team) and customize the learning experience accordingly. For example, information on team diversity could be used to select training goals, or provide feedback.

An important first step towards that end includes a deeper understanding of how team diversity is associated with team performance. Accordingly, the present paper aims to uncover what dimensions of team diversity predict objective and subjective outcomes during computer-mediated CPS activities among triads. We expand work on traditional measures of visible (demographic) diversity by quantifying a variety of dimensions of diversity, such as personality, attitudes towards teamwork, and prior domain experience. We include demographic measures for study (e.g. race, gender, age, first language) as they have been linked to group performance [16], perception by others [51], attitudes towards teamwork [1], cultural work norms [45], and effective verbal contributions [49]. We expand to include personality as it has been associated with CPS performance, self-reported perception of the collaboration quality [44], and team communication [48]. Further, survey-based personality measures can be considered a measure of self-reported identity, and thus important for inclusion. Attitudes towards teamwork (e.g. assessment of leadership, or teamwork self-efficacy), have been extensively shown to influence team decision making [2,17] and have been linked with perception of collaborative interactions [20]. Finally, prior domain experience is expected to be associated with CPS task performance (e.g., [30]), making it key for inclusion as well. We find that diversity in several dimensions predicts CPS outcomes after accounting for pertinent covariates.

RELATED WORKS

The literature on the effects of group composition is vast and covers factors such as gender [11,39], ethnicity [50], teammate familiarity [14], team member ability [50], and personality [9,43], to name a few. We focus our review on the relationship between diversity and collaborative outcomes, as it is most relevant to our work.

Traditionally, team diversity has been studied in terms of demographic and psychological factors. Team diversity has been quantified in terms of standard deviation, relative standard deviation, or Euclidean distance of a measure across teammates [26,31,44]. Demographic features alone cannot capture the intricacies of diversity and as such are often referred to as surface-level diversity [15]. That said, demographic measures still heavily influence collaborative experiences like sense of belonging [10], perception by teammates [51], and attitudes towards group work [1]. The influence of demographic diversity has been described as a "double-edged sword" that leads to both positive and negative team outcomes [28]. For example, diversity in nationality within teams is associated with increased collective knowledge [8] and better performance [29], but a decreased amount of overall interaction [32]. Effects of demographic diversity can also vary over time. Though team members may initially judge one another based on demographic characteristics, team differences may become less salient over time as the group reaches a consensus on team values and beliefs [5,15,25]. Thus, understanding the effects of team diversity requires examining nuanced relationships with temporal and contextual factors as well.

Research has gone beyond studying visible demographic characteristics to quantifying diversity in terms of psychological characteristics, referred to as deep-level diversity [15]. Cognitive diversity is one such aspect and is defined by differences in beliefs, knowledge, skills, thinking styles, or values [41]. Cognitive diversity enhances performance when creative solutions are needed and minority views can lead teams to explore alternative solutions [7,26,38]. For example, a study with third-year engineering students working together during a 16-week semester found that cognitive diversity was associated with project design outcomes, including expert-rated value, user satisfaction, and effectiveness [26]. As with demographic diversity, more nuanced contextual effects may moderate the relationship between cognitive diversity and team performance. For example, Park et. al. [31] report that task knowledge diversity is positively associated with team creativity, but this relationship is negatively moderated by status inequality among team members.

Although research on deep diversity often focuses on teams' knowledge and cognitive diversity, a limited number of research studies have explored additional types of diversity such as personality differences. Pieterse et. al. [33] examined the performance of short-lived student software engineering teams and personality diversity, as indicated by differences in their Myers-Briggs Type Indicators. They found no differences in degree of collaboration or quality of product based on personality diversity.

Most similar to our work is a study that examined the relationship between group diversity, objective outcomes (task performance and posttest score), and subjective outcomes (teammate ratings of performance and

collaboration quality), all during a computer programming CPS task [44]. This study found that gender diversity did not predict outcomes, but personality diversity negatively predicted task performance and teammate ratings of the collaboration quality. Further, they found that diversity in teammate rating of performance negatively predicted learning, hypothesizing that this was related to lack of shared task alignment amongst the group.

NOVELTY

Our work is novel in several respects. First, we study diversity in short-term, spontaneously formed teams. Most of the work on diversity examines organizational teams (e.g. [5,7,25,29,31,41]) and other types of long-term teams (e.g. [8,26]), which may be able to establish familiarity and group work norms over the course of multiple interactions. In contrast, short-term teams face the challenge of quickly engaging each other in CPS processes, such as goal definition or co-construction of solutions.

Second, the limited work on short-term teams only considers one or two dimensions of diversity (e.g. gender and personality [44], personality only, [33], culture [32], and cognition [38]). Thus, it is unclear as to which dimensions should be included in learning technologies that customize CPS skill content for the current team. Accordingly, we study diversity across a variety of dimensions, including demographics, personality, attitudes towards teamwork, and prior domain experience.

Finally, we predict CPS outcomes from diversity after controlling for the general team makeup. We distinguish between the two in that diversity quantifies how different a team is along individual difference measures. Conversely, team makeup refers to the overall composition of a team on those individual difference measures. For example, when examining prior knowledge, diversity quantifies the extent of variation among prior knowledge of teammates, whereas makeup measures the average prior knowledge possessed by the team. There is research that examines either diversity (see Related Works) or general team makeup (e.g. [9,11,14,39,43]); however, the former fails to account for the latter, which raises the possibility that diversity might not predict outcomes above team makeup. Accordingly, we address the question as to whether diversity has incremental predictive power after accounting for team makeup.

DATA COLLECTION

Data collection protocols were approved by our designated Institutional Review Boards and all participants provided consent prior to any data collection.

Participants

Participants were 288 students from two large public universities in the United States (111 from School 1, 177 from School 2). Students were assigned to 96 triads based on scheduling constraints. Forty-six students from 25 teams (26%) indicated they knew at least one person from their team prior to participation. Participants were compensated either with a $50 Amazon gift card (95.8%) or with 3.5 hours of course credit (4.2%) for the two-part study that included an at-home survey and an in-lab session.

Problem Solving Environment (Physics Playground)

Students participated in a CPS task using Physics Playground (Figure 1), which is a highly engaging, two-dimensional educational game that aims to teach students basic Newtonian physics concepts (e.g., Newton's laws, energy transfer, and properties of torque) [4,27]. In Physics Playground, students complete game levels by using the mouse to draw simple machines (i.e., ramps, levers, pendulums, and springboards) that navigate a green ball to a red balloon. All objects that students draw, as well as pre-existing agents in the levels, obey the laws of physics. Figure 1 depicts a team using a lever (pre-existing agent in game) and a weight (drawn by team) to roll the green ball towards

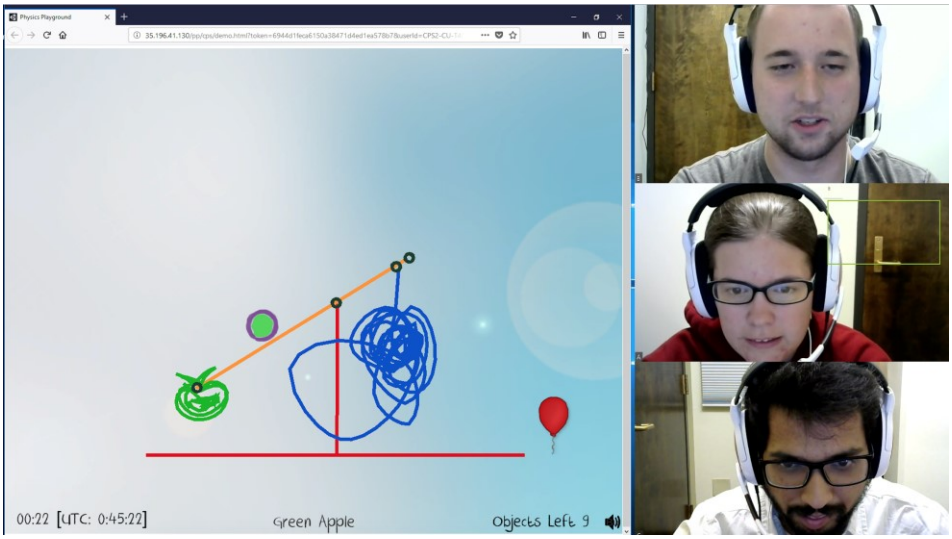


Figure 1. A team playing Physics Playground. The triad used a lever and weight to navigate the green ball to the red balloon.

the red balloon. A team earns a gold trophy when they successfully navigate the green ball to the red balloon using few objects. If more objects are used, then a successful solution earns a silver trophy. Students could restart, exit, or change levels at any time during gameplay. There were no hints or support mechanisms in the game with the exception of a tutorial on game mechanics that teams could optionally view at any time during the collaboration. Each game level had an expert-rated difficulty score (easy, medium, or hard) based on physics knowledge required to solve the level and difficulty of the game mechanics.

At-Home Surveys

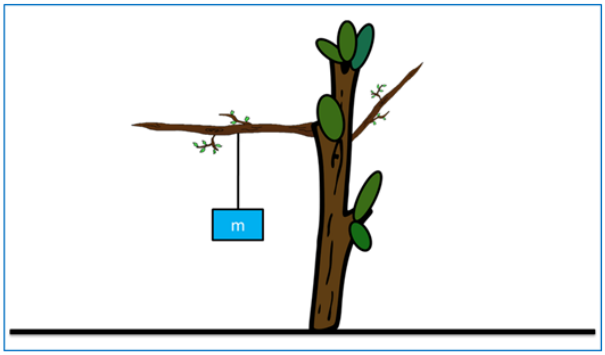
Students were emailed a survey at least 24 hours prior to their scheduled lab session. The purpose of the survey was to assess individual difference measures, such as demographics, personality, attitudes towards leadership and teamwork, and physics competency. The demographic questionnaire assessed the student's gender, race, age, first language, and formal physics coursework. We used the validated short version of the Big Five inventory [13] to assess personality in the following dimensions: extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience (Cronbach's alpha = .75, .09, .55, .59, .31, respectively – see Results for discussion on these reliabilities). We assessed leadership self-efficacy, which is belief in one's leadership capability, with the Leadership Domain Identification Measure (Cronbach's alpha = .82) [17]. Collectivism (willingness to work in teams) and teamwork self-efficacy (personal perception of one's ability to work in teams) were measured using the Individual Satisfaction with the Team Scale (Cronbach's alpha = .79 and .71, respectively) [20]. Finally, we used a validated survey to measure physics self-efficacy (Cronbach's alpha = .85) [22]. Example survey items are shown in Table 1.

Table 1. Examples of validated individual difference and outcome measures are shown.

Measure	Survey Question
Extraversion	I see myself as extraverted, enthusiastic.
Agreeableness	I see myself as sympathetic, warm.
Conscientiousness	I see myself as dependable, self-disciplined.
Emotional Stability	I see myself as calm, emotionally stable.
Openness to Experience	I see myself as open to new experiences, complex.
Leadership	I am a good leader.
Teamwork Self-Efficacy	I can work very effectively in a group setting.
Collectivism	I prefer to work with others in a group than to work alone.
Physics Self-Efficacy	I know I can stick to my aims and accomplish my goals in physics.
CPS Quality	Our team responded to others' questions and ideas thoughtfully.
Inclusiveness and Team Norms	Everyone on the team worked to reach our performance goals.

Students also completed an expert-created ten-item physics pretest that assessed knowledge of energy transfer and properties of torque, which corresponded to the Physics Playground levels selected for the CPS activity (see In-Lab Session). This was a parallel-form version test (versions A and B), that was counterbalanced across participants. An example test question is shown in Figure 2.

After completing the physics pretest, students learned how to use the Physics Playground environment with a short tutorial that taught them how to draw simple machines such as ramps and springboards. After completing the tutorial, students were given 15 minutes to complete five easy levels to familiarize themselves with the game. They then completed other activities not relevant to the present study.



An object is hanging on a tree branch. What would make the branch *less likely* to break?
A. By making the object heavier
B. By moving the object farther from the tree trunk
C. By moving the object closer to the tree trunk
D. Moving the object won't make a difference

Figure 2. An example of a properties of torque pre/posttest question is shown. The correct answer for this question is C.

In-Lab Session

The in-lab session commenced at least 24 hours after the at-home surveys were administered. Students were each assigned to one of three computer-enabled workstations that were either partitioned in the same room using dividers or were located in different rooms (depending on the school where data was collected). They all had video conferencing capabilities and screen sharing through Zoom (https://zoom.us). Each computer was equipped with a webcam and headset microphone so students could see and hear each other. Additional sensors not relevant to the current study were also included.

Teams interacted with Physics Playground for one warmup block and two experimental blocks. One randomly assigned team member controlled the mouse during a block, and this student's screen was shared. A different team member was given control of the mouse during each block such that each student controlled the interaction for one block.

Teams first completed a 15-minute warmup. They were instructed verbally and with on-screen instructions to use the time to familiarize themselves with their teammates and play a few levels together. They were given five easy-to-medium levels corresponding to energy transfer and properties of torque physics concepts. Teams were shown an on-screen warning when ten and five minutes were left in the block.

After the warmup, screen sharing was disabled and students individually rated their emotional valence (1 = *very negative*, 5 = *very positive*) and arousal (1 = *very sleepy*, 5 = *very active*). Students also completed a validated six-item Likert-style (1 = *disagree strongly*, 7 = *agree strongly*) questionnaire assessing perceived CPS quality along the following dimensions: sharing understanding of problems and solutions, establishing common ground, responding to others' questions/ideas, monitoring execution, fulfilling roles on the team, and taking initiative to advance the collaboration (Cronbach's alpha = .90) [47]. This was followed by a three-item inclusiveness and team norms questionnaire (using the same 7-point scale) that assessed how inclusive the team was and whether they worked towards task performance or socially-oriented goals (Cronbach's alpha = .77) [12].

Teams then collaborated for two 15-minute experimental blocks where each block had a different CPS goal. In one goal manipulation, teams were instructed to "solve as many levels as possible." The purpose of this manipulation was to prioritize solution quantity. In the other manipulation, teams were instructed to "get as many gold trophies as possible." The purpose here was to focus teams on quality solutions, and teams were reminded that gold trophies are earned by using fewer objects in the solution. Instructions for each experimental block were provided verbally and on screen.

There was also a physics concept manipulation where teams were either presented with seven energy transfer levels or six properties of torque levels. All levels were of medium-to-hard difficulty. Goal and physics concept were counterbalanced across teams in a 2 × 2 within-subject design. For example, a team could be assigned the golds manipulation and energy transfer levels in the first experimental block followed by the levels manipulation and properties of torque levels in the second experimental block.

Teams were shown the same on-screen warnings as the warmup when they had ten and five minutes left in the block. However, they were also reminded of their goal condition (levels or golds) along with the warning. After each experimental block, screen sharing was again disabled and students individually completed the same surveys that they completed after the warmup. After both blocks, students individually completed a physics posttest, which was a parallel-form version of the pretest. Assignment of test version (A or B) as pre- or posttest was counterbalanced across students. Teams also completed an unrelated task not analyzed here.

DATA ANALYSIS

We investigated associations between team makeup/diversity across four dimensions (demographic, personality, attitudes towards teamwork, and prior domain experience) and five outcome variables (task score, posttest score, valence, arousal, perception of collaboration).

Individual Makeup Measures

We computed a vector for each student, in each of the four dimensions (demographic, personality, attitudes towards teamwork, prior domain experience). For demographics, we computed a four element vector from self-reported age, first language (English or not), gender (female or not), and race. Race distributions varied widely for the two schools (Figure 3). Thus, we contextualized each student's race compared their peers. To do this, we coded each student's race as the proportion of students in their school that identified as that particular race. Thus, a higher value represents a student identifying as a racial majority group. For example, at School 1, 7% of the students reported being Hispanic, and were thus a minority group; their race value was coded as .07. At School 2, Hispanic students were a larger population (39%), thus we coded those students' race value as .39. In a supplemental analysis (see Results), we examined race on its own. Each student was then represented as a seven-dimensional dummy-coded race vector. We scaled the age and race elements from 0 to 1 to ensure that all elements had the same range. This was required to ensuring our diversity metric equally weighted each element (see Diversity and Team Makeup).

There were two survey items per Big-Five personality dimension which we summed [13] to yield a five-element personality vector (extraversion, agreeableness, conscientiousness, emotional stability, and openness to new experience).

We constructed an attitudes towards teamwork vector from the leadership self-efficacy, teamwork self-efficacy, and collectivism scales. We first computed the mean across the items in each scale to yield a leadership, teamwork, and collectivism score. The first two measures were correlated (Pearson's $r = .57$), so we combined them by z-scoring each, and then averaging the z-scores. This measure was scaled from 1 to 7 to match the range of the collectivism scale. Thus, the attitudes towards teamwork vector had two elements: (1) teamwork and leadership self-efficacy and (2) collectivism.

We constructed a prior domain experience vector from the formal physics coursework (scored as 0 for no courses, 1 for high school, 2 for introductory college, and 3 for multiple college courses), physics self-efficacy (average of individual items), and physics pretest score (number of items correct) measures. The other two measures were scaled from 1 to 7 to match the range of the physics self-efficacy measure. Descriptive statistics for our measures are shown in Table 2.

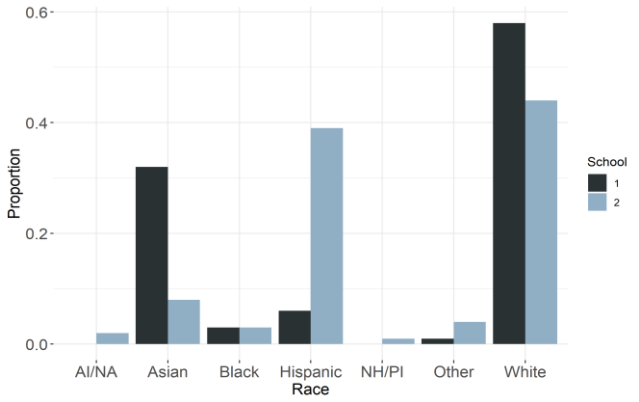


Figure 3. Reported race of students by school. Note, AI/NA refers to American Indian/Native Alaskan; NH/PI refers to Native Hawaiian/Pacific Islander.

Table 2. Mean (*M*), standard deviation (*SD*), and range (*R*) for elements (unscaled) in team makeup vectors are shown. Note, race is omitted, as it is shown in Figure 3.

Elements	*M*	*SD*	*R*
Demographics			
Gender Female	.56	.50	0 – 1
Age	21.72	4.60	18 – 49
English First Language	.74	.44	0 – 1
Personality			
Extraversion	8.01	3.06	2 – 14
Agreeableness	9.91	2.07	2 – 14
Conscientiousness	11.13	2.15	2 – 14
Emotional Stability	9.31	2.54	2 – 14
Openness	10.58	1.95	2 – 14
Attitudes Towards Teamwork			
Leadership/Teamwork Self-Efficacy	5.27	.84	1 – 7
Collectivism	3.89	1.23	1 – 7
Prior Domain Experience			
Prior Physics Courses	1.07	1.07	0 – 3
Physics Self-Efficacy	4.66	1.28	1 – 7
Pretest Score	6.49	1.80	1 – 10

Diversity and Team Makeup

We condensed the elements in each of our focal dimensions (demographic, personality, attitudes towards teamwork, and prior domain experience) to a single, interpretable team-level diversity metric that captures how different a team is along a particular dimension. For a given dimension, our diversity metric was computed as the mean of the pairwise Euclidean distances (*d*) between the three vectors (A, B, C) representing the three students in a team: diversity = mean [d(A, B), d(A, C), d(B , C)]. For example, we compute personality diversity as the mean pairwise Euclidean distances between the five-dimensional personality vectors of each student. Thus, a higher score corresponds to a more diverse (dissimilar) team. Distributions of our diversity measure are shown in Figure 4. Based on these histograms, we note that there are indeed team-level differences in diversity that our metric is able to capture.Team makeup quantifies the team's mean level of the component elements of a dimension. For example, for personality, we average extraversion across team members as a measure of how extraverted the team is overall. This was repeated for the agreeableness, conscientiousness, emotional stability, and openness elements in the personality vector. In all, each team had four diversity scores (demographic, personality, attitudes towards teamwork, and prior domain experience), and 14 team makeup scores (four demographic, five personality, two attitudes towards teamwork, and three prior domain experience).

Outcome Measures

We computed objective and subjective CPS outcomes for the two experimental blocks. The warmup block was not used in analysis as its purpose was to familiarize the team with each other and the CPS environment. Block-level measures were combined to obtain a single team-level measure since the team is the unit of analysis (i.e. team-level diversity).

Recall that teams were assigned the energy transfer concept for one block and the properties of torque concept for the other. We found that task score (computed as the proportion of trophies a team earned) varied significantly across the two concepts with a mean of .19 for energy transfer and .63 for properties of torque ($p < .001$ on a paired-samples *t*-test). Note, task score did not vary significantly across block number (first or second) or goal manipulation (levels or golds). Accordingly, to combine outcomes across blocks, we first z-scored each outcome measure by concept (to remove concept-related variability) and then averaged the two scores to yield a single team-level outcome variable. We adopted this approach for the two objective outcomes: task score and posttest score (number of items pertaining to each concept correctly answered on the physics posttest). We did the same for our three subjective outcomes: self-reported valence, arousal, and a collaboration perception measure, obtained by aggregating the CPS quality and inclusiveness and team norms measures. To aggregate, we separately averaged the six items in the CPS quality measure and the three items in the inclusiveness and team norms measure. These averages were highly correlated (Pearson's $r = .79$), so we combined them by *z*-scoring each and taking the mean. Table 3 show descriptive statistics before outcome metric aggregation.

Table 3. Mean (M), standard deviation (SD), and range (R) of outcome metrics before aggregation are shown.

Elements	M	SD	R
Energy Transfer Trophies	.63	.19	0 – 1
Properties of Torque Trophies	.19	.28	0 – 1
Posttest Score	6.85	1.95	0 – 10
Valence	3.73	1.05	1 – 5
Arousal	3.44	1.16	1 – 5
CPS Quality	6.23	.85	1 – 7
Inclusiveness/Team Norms	6.39	.82	1 - 7

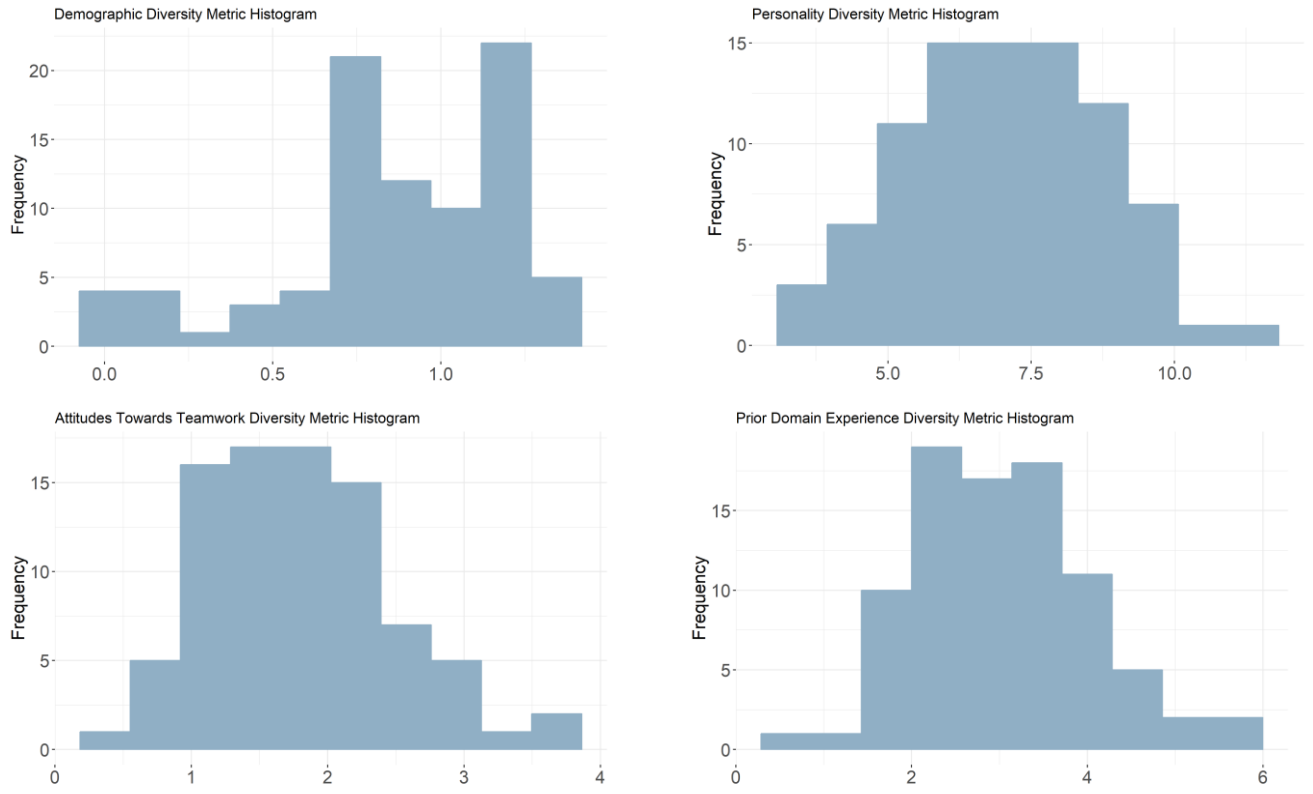


Figure 4. Histograms of the team-level diversity dimensions are shown.

Data Recovery and Removal

Occasionally, students did not complete the at-home measures before completing the in-lab portion (31 students from 21 teams). We emailed a follow-up survey to these students which included the demographic, personality, attitudes towards teamwork, and prior domain experience items. A total of 21 students from 15 teams completed the follow up survey.

We did not include the pretest in the follow-up survey as students had already completed the in-lab portion of the study when the survey was administered. Instead, we replaced missing pretest scores for these individuals with the mean pretest score for their school (7.49 for School 1 and 5.86 for School 2). This yielded a total of 86 teams (from 96 possible triads) where all three teammates reported the individual difference measures needed to compute the diversity and makeup measures. Of these, due to technical issues, one team did not complete the physics task, so they were excluded, leaving 85 teams.

Five teams did not complete the second experimental block due to technical errors, so their outcome measures were only computed for the first experimental block. Teams occasionally did not complete one or more of the measures for a variety of reasons. To maximize sample size, we included teams that had at least one of the outcome variables, so number of teams in the subsequent analyses varies slightly by outcome variable under consideration: 85 for task score, 82 for posttest, and 84 for valence, arousal, and collaboration perception.

RESULTS

We individually regressed each outcome (task score, posttest score, valence, arousal, and collaboration perception), on the four team-level diversity metrics (demographic, personality, attitudes towards teamwork, prior domain experience). Each diversity measure was included in a separate model. The mean value of the component elements of the diversity vectors (makeup scores) were included as covariates in the models to assess the influence of diversity above team makeup. For example, we include mean extraversion, agreeableness, conscientiousness, stability, and openness as covariates in the models that used personality diversity as the primary predictor. We also included school (School 1 or School 2) as a covariate to account for school-level differences in outcomes. Additionally, when predicting posttest score, we included pretest score as a covariate to isolate learning gains beyond prior knowledge (pre-post Pearson's r = .67). In total, we computed 20 regression models (five outcome variables × four team diversity dimensions) as shown in Table 4.

Main Results

We found that demographic diversity positively predicted valence (marginal effect) and arousal beyond the team makeup variables, suggesting there is not a specific team

demographic makeup profile, but diversity, that influences valence and arousal. We also found that having English as a first language predicted posttest score, ostensibly because the task was primarily verbal and conducted in English. Interestingly, older students achieved better learning gains, but were less positive about the outcomes of the collaboration. Due to limitations in our method of coding race, we investigated racial diversity in a follow-up analysis.

Diversity in personality did not predict objective task performance (task score), but the more agreeable teams had lower task scores. In this case, it is possible that agreeable teams focused on socially-oriented group work norms (i.e. avoiding conflict), rather than pursuing successful task completion [21]. Importantly, personality diversity did predict posttest score and did so above the team's personality makeup variables, which were non-significant predictors.

We found that neither personality diversity nor the individual personality facets predicted students' perception of the collaboration. However, extraverted teams reported more positive affect after the task, which is reasonable given that the task is inherently social. Interestingly, teams higher in emotional stability and agreeableness reported more arousal, in contrast to those higher in openness. Overall, personality makeup appears to be more related to arousal than any of the other outcome variables. We further investigated personality in a follow-up analyses, due to reliability limitations.

Diversity in attitudes towards teamwork did not predict any of the objective task outcomes. However it was negatively associated with all three subjective outcomes, although it was only significant for arousal. Thus, teams with more disparate attitudes towards teamwork reported less emotional activation, possibly signaling differing task expectations. With respect to the individual makeup variables, leadership and teamwork self-efficacy predicted positive perceptions of the collaboration, which is an expected finding. Somewhat unexpected was that collectivism negatively predicted posttest score. Since the posttest is an individual measure of learning, this suggests that attitudes that promote effective collaboration (a group outcome) might not be beneficial for individual learning. Some initial evidence may be obtained by the positive (though non-significant) association between diverse attitudes towards teamwork and posttest scores.

In general, the prior domain experience variables did not significantly predict any of the outcomes other than expected associations with pretest scores. That said, the strongest, albeit non-significant, predictor of task score was diversity

Table 4. Standardized beta coefficients are shown. p-values less than .10 are bolded.

	Objective Outcomes		Subjective Outcomes		
Predictors	**Task Score**	**Posttest Score**	**Valence**	**Arousal**	**Collaboration Perception**
Demographics					
Demographic Diversity	-.01 (.93)	-.04 (.66)	**.21 (.10)**	**.27 (.03)**	.00 (.98)
English First Language	.14 (.25)	**.18 (.04)**	-.04 (.74)	-.05 (.68)	.01 (.93)
Female	-.16 (.16)	.00 (.96)	-.07 (.52)	-.13 (.23)	-.05 (.70)
Race	.06 (.63)	.06 (.55)	.09 (.48)	-.10 (.42)	-.05 (.71)
Age	-.11 (.35)	**.18 (.04)**	-.16 (.18)	-.02 (.86)	**-.21 (.10)**
Pretest Score		**.46 (<.01)**			
School (School 1)	**.26 (.04)**	**.21 (.06)**	.06 (.64)	-.03 (.80)	.08 (.54)
Personality					
BFI Diversity	.03 (.77)	**.16 (.08)**	.15 (.20)	-.02 (.83)	.12 (.32)
Extraversion	.07 (.52)	-.02 (.85)	**.25 (.04)**	-.03 (.78)	.08 (.49)
Agreeableness	**-.22 (.08)**	.04 (.68)	-.11 (.39)	**.22 (.07)**	.15 (.27)
Conscientiousness	.06 (.57)	.00 (.96)	-.05 (.64)	-.15 (.19)	-.03 (.83)
Emotional Stability	.04 (.72)	.00 (.98)	.17 (.16)	**.40 (<.01)**	.06 (.61)
Openness	.09 (.49)	.03 (.74)	-.03 (.79)	**-.29 (.02)**	-.06 (.63)
Pretest Score		**.46 (<.01)**			
School (School 1)	**.32 (.01)**	**.26 (.03)**	.09 (.41)	-.10 (.35)	-.04 (.73)
Attitudes Towards Teamwork					
Attitudes Diversity	.00 (.98)	.11 (.22)	-.12 (.33)	**-.26 (.03)**	-.17 (.13)
Leader/Teamwork Self-Efficacy	.05 (.64)	.06 (.50)	.06 (.63)	-.15 (.20)	**.23 (.05)**
Collectivism	.12 (.26)	**-.16 (.05)**	.03 (.79)	.16 (.14)	.10 (.36)
Pretest Score		**.48 (<.01)**			
School (School 1)	**.30 (<.01)**	**.26 (.01)**	.10 (.37)	.01 (.92)	.03 (.80)
Prior Domain Experience					
Prior Domain Exp. Diversity	-.16 (.17)	-.06 (.50)	.02 (.84)	-.07 (.59)	-.11 (.39)
Prior Physics Courses	.18 (.26)	.20 (.11)	.01 (.97)	.11 (.53)	-.07 (.71)
Physics Self-Efficacy	.05 (.65)	.14 (.12)	.05 (.68)	.01 (.95)	.03 (.83)
Pretest Score	.17 (.25)	**.34 (<.01)**	**.27 (.10)**	.09 (.57)	.02 (.90)
School (School 1)	.10 (.53)	**.23 (.07)**	-.10 (.58)	-.13 (.47)	.08 (.65)

in prior domain experience. It is possible that teams with differing prior domain experience had more difficulty coordinating their domain knowledge and skills to earn a high task score, though this finding needs further consideration since it is non-significant.

We also found that school affiliation was a consistent predictor of both objective outcomes (task score and posttest score) but not of the subjective outcomes. This might be explained by educational achievement difference between the schools (average ACT scores, which index scholastic achievement, was 30.6 and 25.0 for Schools 1 and 2, respectively). Future work should consider including scholastic achievement as an additional dimension of diversity and makeup.

Follow-up Analyses

We conducted follow-up analyses to supplement these main results. Because all of our team diversity and makeup measures were obtained via self-reports, the reliability and validity of these measures is of importance. Fortunately, with the exception of personality, reliability of all of our self-report measures exceeded a Cronbach's alpha of .70.

Reliability for the short-BFI has traditionally been a concern, [3], and indeed is low in our study (see At-Home Surveys). Therefore, we conducted an additional analysis by eliminating BFI facets with very low reliability (agreeableness and openness). Specifically, we constructed a personality diversity vector with extraversion, conscientiousness, and emotional stability as these had alphas above 0.50.

Specifically, when using extraversion, conscientiousness, and emotional stability, personality diversity was positively related to posttest (β = .05, p = .26), which was also the case when using all five facets BFI (β = .16, p = .08). That said, the relationship was weak in both cases, suggesting personality measures alone are not predictive of objective outcomes. Further, extraversion was a marginal predictor of valence when using extraversion, conscientiousness, and emotional stability (β = .08, p = .06). This relationship was stronger when using all BFI factors (β= .25, p = .04), but in the same direction. Overall, the relationships were in a similar direction, yet a little weaker when using only the reliable BFI measures.

We also conducted a follow-up analysis on racial diversity, given that our method of coding race does not account for the inherent complexities of race, which may be intertwined with access to educational resources and perception by teammates. These factors in turn might influence CPS. We encoded a student's race directly rather than considering it with respect to the majority at their school. Each student's race was represented as a seven-element binary vector and diversity and makeup variables were computed on race alone. Race diversity did not predict any outcomes. There was a significantly positive relationship between a team's makeup and posttest score (β = 2.47, p = .02 for black, β = 1.34, p = .02 for white, β = 1.37, p = .04 for Asian), potentially signaling that teams with more similar racial backgrounds had more common ground to achieve higher learning gains. There was also a significant relationship between race and arousal in that teams with more American Indian/Native Alaskan students reported significantly lower arousal (β = -3.52, p = .04). These students were a minority in our dataset, and these results suggest that such students would benefit from personalized CPS support.

DISCUSSION

Our main goal was to investigate how diversity in team composition is associated with team-level outcomes in a STEM-based collaborative problem solving (CPS) task. We developed a novel approach for reducing the complexity in characterizing team diversity along four dimensions: demographic, personality, attitudes towards teamwork, and prior domain experience. We investigated associations between team diversity and collaborative outcomes, specifically, objective measures of team performance and subjective judgements of the collaboration. In what follows, we summarize main findings, discuss applications, limitations, and ideas for future work.

Main Findings

We identified robust links between team diversity and team-level outcomes, specifically objective measures of learning and subjective judgments of the group's interaction. Teams that were more demographically diverse (e.g., race, age, gender, first language), reported more positive affect (valence) and higher energy (arousal) after the collaborative interaction. Teams that had more varied personalities learned more as a result of the collaboration, and teams with disparate attitudes towards teamwork reported lower energy (arousal) after the interaction.

Importantly, associations between team diversity and outcomes were obtained after controlling for the overall makeup of the team. Team makeup has been shown to be predictive of outcomes in previous studies [9,43,44], which was also the case in our work. However, our results point to diversity as the key characteristic of the team that influences outcomes beyond mere makeup. We found cases where diversity was a significant predictor of CPS outcomes, but makeup was not; the reverse was also true, and there were no cases where both were predictive. Thus, it is imperative that future work investigates these dimensions in tandem instead of focusing on one or the other as is currently done (e.g. [26,43,44]).

In contrast to prior work on team diversity, we studied short-term collaborations (two 15-minute interactions) for teams that generally did not know each other prior to the collaboration. Previous literature suggests that demographic factors play a bigger role in collaborative outcomes in the short-term, but non-visible aspects of diversity become more salient as the team becomes more familiar with each other over long-term interactions [5,15,25]. We found that both demographic diversity as well as non-visible aspects of

diversity (personality and attitudes towards teamwork) predicted CPS outcomes (learning gains and arousal). Therefore, we show that even in short-term interactions, non-visible diversity still influences CPS.

There were also some unexpected null effects. Most importantly, we were largely unable to predict CPS task performance with any of diversity or makeup measures. The exception was the BFI personality dimension of agreeableness that negatively predicted task performance. It is possible that more agreeable teams were focused on conflict minimization and achieving positive social interactions rather than task completion. That said, the difficulty in predicting task performance suggests that it might have to more to do with what teams do over the course of the interaction rather than what they bring to the interaction. As such, the destiny of the team is not predetermined, but depends on the team's behaviors.

Applications

We envision that our work can be applied to collaborative-learning interfaces that personalize CPS tasks and goals based on different dimensions of diversity. We are in the very early stages of research and more work is needed before our findings are actionable; however, we can illustrate some preliminary ideas at this time. In particular, we found that teams with less diverse personalities learned less from the collaboration. Accordingly, the system could target learning outcomes for such teams, for example, by suggesting they demonstrate their reasoning for implementing a solution [35]. Similarly, teams with disparate attitudes towards teamwork experienced low arousal, which could signal task disengagement [42]. Accordingly, throughout the task, the system could suggest engaging activities that simultaneously support productive CPS. For example, the system could suggest the team spend a few minutes generating new ideas, refining old ones, or reflecting on prior results [53]. This would serve as a reminder that these tasks are essential to productive CPS.

Limitations and Future Work

Like all studies, ours has limitations that should be addressed in future research. First, we focused on individual dimensions of diversity but did not examine them in concert. In future work, analyses could be expanded to include how different aspects of diversity interact to predict outcomes.

Second, our data was collected in a highly controlled lab setting, so generalizability to a more authentic context is limited. We also considered only one task, so future work should examine the relationship between team diversity and CPS outcomes for different CPS tasks and task types, such as creative versus analytical tasks.

Third, we did not focus on measures of what occurred in the collaboration itself (i.e., how the team spoke to each other, kinds of ideas they generated, whether they payed attention to on screen content). It is highly likely that behaviors during the collaboration can be used to supplement our findings. Accordingly, future work could consider a collaboration "timeline" where factors in place before the collaboration (e.g. diversity and team makeup) are combined with behaviors exhibited during the collaboration.

Fourth, some of our measures did not meet adequate standards for reliability. Specifically, our short measure of personality (short BFI) exhibited low reliability, so the findings pertaining to personality warrant replication with the full version of the BFI.

Further, a key limitation in our work is that our diversity measures did not significantly predict task performance. It might be the case that diversity itself is not pertinent when it comes to objective task performance. More likely, however, diversity in dimensions other than those investigated here (e.g., measures of cognition or problem-solving ability) might be predictive.

Finally, it is impossible to infer causation in our work, particularly for subjective outcomes. We do not know if team-level differences influenced collaborative behaviors which influenced attitudes, or if team-level differences influenced attitudes towards the collaboration.

CONCLUSION

Our work is an initial step in building interfaces for computer-mediated collaboration that teach CPS skills tailored to the diversity of the team. We found show that several diversity dimensions predict CPS outcomes even after controlling for the overall makeup of the team. With subsequent research, next-generation collaborative interfaces can leverage these findings to support effective collaboration among diverse teams.

ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation (NSF DUE 1745442) and the Institute of Educational Sciences (IES R305A170432). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] Kacey Beddoes and Grace Panther. 2018. Gender and teamwork: an analysis of professors' perspectives and practices. *Eur. J. Eng. Educ.* 43, 3 (2018), 330–343. DOI:https://doi.org/10.1080/03043797.2017.1367759

[2] Cigdem Beyan, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. 2016. Identification of emergent leaders in a meeting scenario using multiple kernel learning. In *Proceedings of the 2nd Workshop on Advancements in Social Signal Processing for Multimodal Interaction*, 3–10.

[3] Jack Block. 1995. A contrarian view of the five-factor approach to personality description. *Psychological Bulletin 117*, 187–215. DOI:https://doi.org/10.1037/0033-2909.117.2.187

[4] Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2015. Automatic detection of learning-centered affective states in the wild. In *Proceedings of the 20th international conference on intelligent user interfaces*, 379–388.

[5] Jennifer A Chatman and Francis J Flynn. 2001. The influence of demographic heterogeneity on the emergence and consequences of cooperative norms in work teams. *Acad. Manag. J.* 44, 5 (2001), 956–974. DOI:https://doi.org/10.2307/3069440

[6] Pravin Chopade, Kristin Stoeffler, Saad M Khan, Yigal Rosen, Spencer Swartz, and Alina von Davier. 2018. Human-Agent Assessment: Interaction and Sub-skills Scoring for Collaborative Problem Solving. In *Artificial Intelligence in Education*, 52–57.

[7] Taylor H Cox and Stacy Blake. 1991. Managing Cultural Diversity: Implications for Organizational Competitiveness. *Exec.* 5, 3 (1991), 45–56. Retrieved from http://www.jstor.org/stable/4165021

[8] Petru L Curşeu and Helen Pluut. 2013. Student groups as learning entities: The effect of group diversity and teamwork quality on groups' cognitive complexity. *Stud. High. Educ.* 38, 1 (2013), 87–103. DOI:https://doi.org/10.1080/03075079.2011.565122

[9] Petru Lucian Curşeu, Remus Ilies, Delia Vîrgă, Laurenţiu Maricuţoiu, and Florin A Sava. 2019. Personality characteristics that are valued in teams: Not always "more is better"? *Int. J. Psychol.* 54, 5 (2019), 638–649. DOI:https://doi.org/10.1002/ijop.12511

[10] Alberto Esquinca and Lidia Herrera-Rocha. 2019. Board 90: Latinx Persistence In and Beyond the Degree: Intersections of Gender and Ethnicity (Research). In *2019 ASEE Annual Conference & Exposition*.

[11] Graham D. Fenwick and Derrick J. Neal. 2001. Effect of gender composition on group performance. *Gender, Work Organ.* 8, 2 (2001), 205–225. DOI:https://doi.org/10.1111/1468-0432.00129

[12] Donald G. Gardner and Jon L. Pierce. 2016. Organization-based self-esteem in work teams. *Gr. Process. Intergr. Relations* 19, 3 (May 2016), 394–408. DOI:https://doi.org/10.1177/1368430215590491

[13] Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann. 2003. A very brief measure of the Big-Five personality domains. *J. Res. Pers.* 37, 6 (2003), 504–528. DOI:https://doi.org/10.1016/S0092-6566(03)00046-1

[14] Deborah H. Gruenfeld, Elizabeth A. Mannix, Katherine Y. Williams, and Margaret A. Neale. 1996. Group composition and decision making: How member familiarity and information distribution affect process and performance. *Organ. Behav. Hum. Decis. Process.* 67, 1 (1996), 1–15. DOI:https://doi.org/10.1006/obhd.1996.0061

[15] David A Harrison, Kenneth H Price, Joanne H Gavin, and Anna T Florey. 2002. Time, teams, and task performance: Changing effects of surface- and deep-level diversity on group functioning. *Acad. Manag. J.* 45, 5 (2002), 1029–1045.

[16] Cedric Herring. 2009. Does Diversity Pay?: Race, Gender, and the Business Case for Diversity. *Am. Sociol. Rev.* 74, 2 (2009), 208–224. DOI:https://doi.org/10.1177/000312240907400203

[17] Crystal L Hoyt and Jim Blascovich. 2010. The role of leadership self-efficacy and stereotype activation on cardiovascular, behavioral and self-report responses in the leadership domain. *Leadersh. Q.* 21, 1 (2010), 89–103. DOI:https://doi.org/https://doi.org/10.1016/j.leaqua.2009.10.007

[18] Otto Kässi and Vili Lehdonvirta. 2018. Online labour index: Measuring the online gig economy for policy and research. *Technol. Forecast. Soc. Change* 137, (2018), 241–248. DOI:https://doi.org/https://doi.org/10.1016/j.techfore.2018.07.056

[19] Norbert L. Kerr and R. Scott Tindale. 2004. Group performance and decision making. *Annu. Rev. Psychol.* 55, 1 (February 2004), 623–655. DOI:https://doi.org/10.1146/annurev.psych.55.090902.142009

[20] José M de la Torre-Ruiz, Vera Ferrón-Vílchez, and Natalia Ortiz-de-Mandojana. 2014. Team Decision Making and Individual Satisfaction With the Team. *Small Gr. Res.* 45, 2 (2014), 198–216. DOI:https://doi.org/10.1177/1046496414525478

[21] Claus W. Langfred. 1998. Is group cohesiveness a double-edged sword? *Small Gr. Res.* 29, 1 (February 1998), 124–143. DOI:https://doi.org/10.1177/1046496498291005

[22] Christine Lindstrøm and Manjula D Sharma. 2011. Self-efficacy of first year university physics students: Do gender and prior formal instruction in physics matter? *Int. J. Innov. Sci. Math. Educ. (formerly CAL-laborate Int.* 19, 2 (2011), 1–19.

[23] James Manyika, Susan Lund, Jacques Bughin, Kelsey Robinson, Jane Mischke, and Deepa Mahajan. 2016. Independent work: Choice, necessity, and the gig economy (McKinsey global institute report).

[24] Margaret M. McManus and Robert M. Aiken. 2016. Supporting Effective Collaboration: Using a Rearview Mirror to Look Forward. *Int. J. Artif. Intell. Educ.* 26, 1 (2016), 365–377. DOI:https://doi.org/10.1007/s40593-

015-0068-6

[25] Philip Meissner, Malte Schubert, and Torsten Wulf. 2018. Determinants of group-level overconfidence in teams: A quasi-experimental investigation of diversity and tenure. *Long Range Plann.* 51, 6 (2018), 927–936. DOI:https://doi.org/https://doi.org/10.1016/j.lrp.2017.11.002

[26] Jessica Menold and Kathryn Jablokow. 2019. Exploring the effects of cognitive style diversity and self-efficacy beliefs on final design attributes in student design teams. *Des. Stud.* 60, (2019), 71–102. DOI:https://doi.org/https://doi.org/10.1016/j.destud.2018.08.001

[27] Juan Miguel, L Andres, Ma Mercedes, T Rodrigo, and Jessica O Sugay. 2014. An exploratory analysis of confusion among students using Newton's Playground. *Proc. 22nd Int. Conf. Comput. Educ.* (2014).

[28] Frances J Milliken and Luis L Martins. 1996. Searching for common threads: Understanding the multiple effects of diversity in organizational groups. *Acad. Manag. Rev.* 21, 2 (1996), 402–433.

[29] Bo Bernhard Nielsen and Sabina Nielsen. 2013. Top management team nationality diversity and firm performance: A multilevel study. *Strateg. Manag. J.* 34, 3 (2013), 373–382. DOI:https://doi.org/10.1002/smj.2021

[30] Timothy J Nokes-Malach, Michelle L Meade, and Daniel G Morrow. 2012. The effect of expertise on collaborative problem solving. *Think. Reason.* 18, 1 (2012), 32–58. DOI:https://doi.org/10.1080/13546783.2011.642206

[31] Won-Woo Park, Joon Yeol Lew, and Eun Kyung Lee. 2018. Team knowledge diversity and team creativity: The moderating role of status inequality. *Soc. Behav. Personal. an Int. J.* 46, 10 (2018), 1611–1622.

[32] S Paul and S Ray. 2013. Cultural Diversity, Group Interaction, Communication Convergence, and Intra-group Conflict in Global Virtual Teams: Findings from a Laboratory Experiment. In *2013 46th Hawaii International Conference on System Sciences*, 344–352. DOI:https://doi.org/10.1109/HICSS.2013.156

[33] V Pieterse, M Leeu, and M van Eekelen. 2018. How personality diversity influences team performance in student software engineering teams. In *2018 Conference on Information Communications Technology and Society (ICTAS)*, 1–6. DOI:https://doi.org/10.1109/ICTAS.2018.8368749

[34] Jeremy Roschelle and Stephanie D Teasley. 1995. The Construction of Shared Knowledge in Collaborative Problem Solving. In *Computer Supported Collaborative Learning*, 69–97.

[35] Carolyn Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *Int. J. Comput. Collab. Learn.* 3, 3 (September 2008), 237–271. DOI:https://doi.org/10.1007/s11412-007-9034-0

[36] Yigal Rosen. 2015. Computer-based assessment of collaborative problem solving: Exploring the feasibility of human-to-agent approach. *Int. J. Artif. Intell. Educ.* 25, 3 (2015), 380–406. DOI:https://doi.org/10.1007/s40593-015-0042-3

[37] Eduardo Salas, Nancy J Cooke, and Michael A Rosen. 2008. On Teams, Teamwork, and Team Performance: Discoveries and Developments. *Hum. Factors* 50, 3 (2008), 540–547. DOI:https://doi.org/10.1518/001872008X288457

[38] Jürgen Sauer, Tobias Felsing, Holger Franke, and Bruno Rüttinger. 2006. Cognitive diversity and team performance in a complex multiple task environment. *Ergonomics* 49, 10 (2006), 934–954. DOI:https://doi.org/10.1080/00140130600577502

[39] Victor Savicki, Merle Kelley, and Dawn Lingenfelter. 1996. Gender and group composition in small task groups using computer-mediated communication. *Comput. Human Behav.* 12, 2 (1996), 209–224. DOI:https://doi.org/10.1016/S0747-5632(96)00024-6

[40] Julian Schulze and Stefan Krumm. 2017. The "virtual team player": A review and initial model of knowledge, skills, abilities, and other characteristics for virtual collaboration. *Organ. Psychol. Rev.* 7, 1 (February 2017), 66–95. DOI:https://doi.org/10.1177/2041386616675522

[41] Shung J Shin, Tae-Yeol Kim, Jeong-Yeon Lee, and Lin Bian. 2012. Cognitive team diversity and individual team member creativity: A cross-level interaction. *Acad. Manag. J.* 55, 1 (2012), 197–212.

[42] Gale M Sinatra, Benjamin C Heddy, and Doug Lombardi. 2015. The Challenges of Defining and Measuring Student Engagement in Science. *Educ. Psychol.* 50, 1 (2015), 1–13. DOI:https://doi.org/10.1080/00461520.2014.1002924

[43] Arjumand Bano Soomro, Norsaremah Salleh, and Azlin Nordin. 2015. How personality traits are interrelated with team climate and team performance in software engineering? A preliminary study. In *2015 9th Malaysian Software Engineering Conference (MySEC)*, 259–265.

[44] Angela E. B. Stewart and Sidney K D'Mello. 2018. Connecting the Dots Towards Collaborative AIED: Linking Group Makeup to Process to Learning. In

International Conference on Artificial Intelligence in Education, 545–556.

[45] Jeanine S Stewart, Elizabeth Goad Oliver, Karen S Cravens, and Shigehiro Oishi. 2017. Managing millennials: Embracing generational differences. *Bus. Horiz.* 60, 1 (2017), 45–54. DOI:https://doi.org/https://doi.org/10.1016/j.bushor.2016.08.011

[46] Kristin Stoeffler, Yigal Rosen, Maria Bolsinova, and Alina von Davier. 2018. Gamified Assessment of Collaborative Skills with Chatbots. . 343–347. DOI:https://doi.org/10.1007/978-3-319-93846-2_64

[47] Chen Sun, Valerie J Shute, Angela E.B. Stewart, Jade Yonehiro, Nicholas Duran, Sidney K. D'Mello, Sidney D'Mello, and Sidney K. D'Mello. 2020. Towards a generalized competency model of collaborative problem solving. *Comput. Educ.* 143, (2020), 103672. DOI:https://doi.org/https://doi.org/10.1016/j.compedu.2019.103672

[48] T Walle and J E Hannay. 2009. Personality and the nature of collaboration in pair programming. In *2009 3rd International Symposium on Empirical Software Engineering and Measurement*, 203–213. DOI:https://doi.org/10.1109/ESEM.2009.5315996

[49] Yuko Watanabe and Merrill Swain. 2007. Effects of proficiency differences and patterns of pair interaction on second language learning: collaborative dialogue between adult ESL learners. *Lang. Teach. Res.* 11, 2 (2007), 121–142. DOI:https://doi.org/10.1177/136216880607074599

[50] Ian A.G. G Wilkinson and Irene Y.Y. Y Fung. 2002. Small-group composition and peer effects. *Int. J. Educ. Res.* 37, 5 (2002), 425–447. DOI:https://doi.org/10.1016/S0883-0355(03)00014-4

[51] Joanna Wolfe and Elizabeth Powell. 2009. Biases in Interpersonal Communication: How Engineering Students Perceive Gender Typical Speech Acts in Teamwork. *J. Eng. Educ.* 98, 1 (2009), 5–16. DOI:https://doi.org/10.1002/j.2168-9830.2009.tb01001.x

[52] Intuit 2020 Report: Twenty Trends that will Shape the Next Decade.

[53] 2015. *PISA 2015 Collaborative Problem Solving Framework*.