# Computational Methods for Assessing and Improving Quality of Study Group Formation

*Ana Tudor*

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2022-163
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-163.html

May 20, 2022

Acknowledgement

I would like to thank Professor Gireeja Ranade for supporting the development of this project, encouraging my ideas, and always challenging me to grow. I would like to thank Sumer Kohli and Neelesh Ramachandran for their tireless efforts in developing the study group formation system that allows students to receive groups every year. I would like to thank Gloria Tumushabe for her vision and work in shaping this project towards equitable opportunities for all students. I would like to thank UC Berkeley Undergraduate and Graduate Student Instructors for collaborating with us to bring our vision to classrooms, and for giving their all towards improving the learning experience of students.

# Computational Methods for Assessing and Improving Quality of Study Group Formation

by: Ana Tudor

May 2022

**Research Project**

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

**Committee:**

Professor Gireeja Ranade
Research Advisor

* * * * * * *

Professor Armando Fox
Second Reader

**Acknowledgements**

I would like to thank Professor Gireeja Ranade for supporting the development of this project, encouraging my ideas, and always challenging me to grow. I would like to thank Sumer Kohli and Neelesh Ramachandran for their tireless efforts in developing the study group formation system that allows students to receive groups every year. I would like to thank Gloria Tumushabe for her vision and work in shaping this project towards equitable opportunities for all students. I would like to thank UC Berkeley Undergraduate and Graduate Student Instructors for collaborating with us to bring our vision to classrooms, and for giving their all towards improving the learning experience of students. Finally, I would like to thank my family and friends for being my own community, and for being the reason I am able to do this work.

This project is dedicated to all the students who have dreamed of a community to learn and grow with, especially those who haven't seen it come to fruition.

# Contents

# Abstract

The student peer group has been established as one of the most important influences on student development. As such, ensuring students have access to high quality classroom peer groups, referred to as study groups, is beneficial to student learning. However, both in instructor-assigned and self-formed groups, students may encounter less than positive experiences. Notably, students from underrepresented communities often face challenges in finding social support for their education when compared with those from majority groups. Several algorithmic systems have been developed to allow instructors to form study groups informed by student preferences and needs. For existing systems, addressing student feedback can help ensure that students of all demographics receive acceptable peer group support.

This focus of this project concerns incorporating student feedback to improve algorithmic study group formation. The project considers three aspects of this problem: devising a survey to collect student feedback, analyzing impact based on feedback data, and investigating a computational method to improve groups based on student feedback. This work is applied in the context of an existing Scalable, Inclusive Matching of Groups (SIM-G) system, which inclusively generates preference-based study groups for student. SIM-G currently operates in large introductory classrooms at UC Berkeley.

First we focus on developing and validating a survey which assesses the quality of study groups. This is based on a construct of group quality which includes reliability and availability of the peer group, effectiveness in aiding course learning, and student psychological safety. The survey is demonstrated to provide a valid and reliably consistent measure of group quality, with suggested deployment after slight revision.

Next, the project conducts analysis of the impact of study group formation in large EECS classrooms at UC Berkeley, based on datasets generated from group matching in courses using the SIM-G system. It is found that study groups matched by SIM-G have roughly equitable outcomes across demographics, and that students from under-represented demographics preferentially choose software-matched groups over self-formed groups. The analysis also reveals opportunities for improvement in providing study groups in classrooms, namely in facilitating student meetup and communication. Finally, positive performance in assessment grades is correlated with a combination of measures of student comfort in the group, and frequency of group interaction.

Finally, the project explores a method for computationally forming and improving study groups, via a Reinforcement Learning model. A system is outlined for regressing on study group quality based on preference and demographic features of each student. The project also attempts the incorporation of a clustering model towards group formation, ultimately rejecting it as appropriate for this application. Ultimately, this modeling is used to approach iterative improvement of study groups within a Reinforcement Learning Actor-Critic framework, demonstrating its potential feasibility given a more appropriate group formation model.

# Chapter 1: Background

## 1.1 Prior Work

The valuable positive impact of instructor-provided group work on student development has been established [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. This positive impact is reinforced by other works finding learning can be viewed as a social experience [14, 15], and collaborative social networks are found to correlate with positive individual success [16, 17, 18, 19, 20, 21]. These effects extend to the context of engineering classrooms, where student learning often takes the form of collaborative work on assignments and projects.

However, many students do not necessarily find such peer groups to be equitably accessible to them, and even when they do, may feel excluded within the context of their group. Solo members of racial minorities in social groups often perceive higher rates of discrimination [22], women face social isolation and pressure in peer academic groups [23, 24, 25, 26, 27], and both women and racial minority group students have documented trends of experiencing social exclusion in classroom contexts [28, 29]. Additionally, trans or gender non-conforming (GNC) students face low retention rates in STEM, with work positing this is due to cultural hostility and social exclusion [30]. These effects may be associated with correlations found between less positive classroom climate perception for GNC students, and significant associations between positive class climate perception and institutional support/resource use [31].

These trends all surface at UC Berkeley, where introductory classes often have thousands of students, and with disproportionate under-representation of students from certain minority groups (e.g. fewer than ten Black students in a class of 1000). Such under-representation is present in the classroom datasets used in this project.

### 1.1.1 Group Matching Software

To target these issues and promote the accessibility of study groups in introductory Electrical Engineering and Computer Science (EECS) classrooms at UC Berkeley, a formation system for Scalable, Inclusively-Matched Groups (SIM-G) was developed. The foundation of the project lies in work by Gloria Tumushabe [32], Gireeja Ranade, Sumer Kohli, and Neelesh Ramachandran [33]. These efforts established the viability of a scalable matching system to promote peer academic networks, and asserted the positive and equitable impact such formation can have.

In the SIM-G system, study groups are formed for students opting in to use the study-group matching system, which was completely optional with no grade incentives. Students were also allowed to form their own groups and simply indicate that. Matching surveys are sent out at the very beginning of the semester as part of the first homework assignment (see Appendix C), and groups were released to students online a week later, by course staff. Feedback on the study groups is collected through a mid-semester survey, and a final evaluation survey (see [34]).

If students found that their assigned or self-formed study groups were not working well for them, they could request a different study group during the mid-semester survey. This enables a "reassignment round", in which the same matching algorithm is executed within the group of students who requested reassignment, and new study groups are formed from those students. Since very few reassignments were requested in the third round of reassignment, we did not conduct further rounds.

## 1.2 Problem Approached in this Thesis

Given the established benefits of stable peer academic groups (study groups), and given this existing infrastructure for organizing study groups on a large scale, it becomes imperative to assert standards of quality for provided study groups. Diakopoulos writes that for technological development, incorporation of periodic feedback, and actionable change based on that feedback, is ethically imperative when engaging in product design [35]. For all intents and purposes, a study group formation system brought to scale should be considered a technological service product provided to students. Given that individuals of certain demographics have historically faced disproportionate cultural exclusion within the classroom, and given that individual students do not always encounter a compatible group, periodic check-ins should be implemented by default to ensure groups are meeting student needs. Additionally, actions should be taken on feedback and needs voiced by students.

With this perspective, group reassignment based on feedback from students is a key feature of the system, and this distinguishes SIM-G from other group-formation approaches [36, 37]. However, the initial surveys implemented to assess student feedback were rudimentary, and addressing student dissatisfaction initially took the form of simply reassigning students via the same system, within a smaller subset of the classroom.

In order to better take into account individualized student feedback to form study groups, two primary problems emerge:

1. How does one go about assessing whether a study group is functioning well for a student?
2. How does one go about improving a student's experience based on their communicated impression of group quality?

To approach problem 1, this thesis first establishes an improved survey instrument with which to measure high quality study group function (Ch. 2), and then establishes methods for interpreting the results of deploying such an instrument (Ch. 3), while reflecting on perceived impact of the SIM-G matcher in past classroom contexts. To approach problem 2, this thesis approaches the computational problem of iterative improvement of study groups over a semester as a Reinforcement Learning problem, training on group-combined student information about matching preferences and resulting group quality (Ch. 4). In presenting viable methods with which to address both questions, this thesis enables the improvement of future study group generation via computational methods.

# Chapter 2: Measuring the Quality of Social & Academic Support Provided by Study Groups

## 2.1 Introduction

Improving the quality of a study group should include an informed understanding of what this quality entails, and an appropriate survey to measure those key aspects in existing groups. Once a valid and reliable measure of group quality is created, it can justify an analysis of which group formation measures best improve groups for students. Such a measure also enables an opportunity for adapting machine learning models towards the prediction of study group quality, for anticipated groups of students who have not yet interacted, but whose matching preferences and demographic characteristics are already known. However, relatively few inquiries have been conducted into the comprehensive definition of, and measurement of, the level of quality at which a study group provides social and academic support for a student.

This chapter engages in the following process to ensure a valid measurement of study group quality:

- Development of a definition of a student's study group quality. This requires the identification of tangible characteristics of group interactions, and student experience, that ultimately contribute to positive academic and social outcomes.
- Designing a survey, or instrument, to accurately measure the occurrence of these characteristics.
- Analyzing the instrument's reliability, or consistency in associating responses which should theoretically be often observed in conjunction with each other, without being susceptible to inconsistency in student responses.
- Analyzing the instrument's validity, or ability to measure the developed definition of quality.

The survey construction and modeling of resulting responses are thoroughly discussed. The Rasch model is employed to estimate a student's study group quality level based on their responses [38], and the model results are validated via a number of techniques derived from Item Response Theory, which is heavily utilized in research in education. All modeling and validation techniques are chosen as per recommendations from Professor Mark Wilson, detailed in [39], and all methods will be thoroughly described in Chapter 2.4.

In summary of results, the instrument is found to have acceptable reliability, based on an internal consistency coefficient value of 0.915, and a Spearman-Brown alternate forms reliability coefficient of 0.84 (Sec.2.5). A measure of construct validity, via Wright Map interpretation, demonstrated consistent predicted group quality levels among each of two subsets of questions, indicating these subsets measure slightly independent characteristics of group quality (Sec. 2.6.1). Internal validity was evaluated as acceptably high, based on a Spearman's Rho value of 0.893 (Sec. 2.6.2. Two

questions are suggested to be amended based on poor item fit (Sec.2.6.3), specifically those with poor item fit due to question wording ambiguity. The author suggests the slight revision and re-validation of the survey, and subsequent deployment for study group improvement after validation.

### 2.1.1 Defining Quality of Study Groups

The task of identifying key characteristics of study groups, study groups, which can contribute to positive social and academic outcomes, has been approached by many in the education research space. One study which attempted this task [40] focused on assessing the success of groups based on a student's experience of engaging in group work. Conducted via open-ended survey questions, the authors found that key features of self-reported successful groups included how well a group facilitated learning, how well the group functioned as a dual between studying and socialization, and how well the group was organized. Some studies focus on other aspects, such as: teacher's impressions that successful cooperative learning involves overall social balancing/composition of groups [41], how social media use can positively supplement group function [42], and how group sizes and stability of consistent interactions with the same members can impact the social/academic experience of group work [43]. Common themes in these works center around a balance between social function and academic efficacy of a group, in combination with accessibility and frequency of interaction. These four themes were major contributors to the final group quality characterizations identified.

## 2.2 Construct Map

To arrange these findings into a theory of what might constitute differing levels of study group quality, we employ an Item Response Theory (IRT) model [39, 44]. IRT models that a set of items, or measurement tools constituted by survey questions in our case, attempts to measure some real world target phenomenon relating to a human subject. The phenomenon is assumed to operate at differing levels of magnitude. The human subject experiencing the phenomenon is denoted as the "respondent" to the set of phenomenon-measuring items.

The items must relate to the target phenomenon in a theoretical way. This may entail items directly asking about characteristics at differing magnitude levels of the phenomenon. This may also entail items attempting to measure characteristics of adjacent phenomena, that the target is dependent on.

To represent this theoretical structure, a **construct** is defined as a sequence of magnitude levels for the target phenomenon, along with characteristics found at each level. A **subconstruct** is defined as any additional phenomena which a respondent may experience at varying degrees, and which contribute to the target construct.

A **construct map** structurally defines the overarching target phenomenon, the levels at which it may be expressed, and the characterization of each of these levels. Further work attempts to build and operate within this construct map framework. The mathematical modeling methods, performed after the design of the construct map and the item set, is discussed in Section 2.4

### 2.2.1 Construct Map Definition

Within this construct map framework, the corresponding target phenomenon is **quality of social and academic peer support provided by a study group**. A student respondent may experience this quality of study group support to varying degrees. Four levels of study group quality are defined going forward, namely:

- High Quality Group Support, corresponding to the best case study group scenario, where a student experiences a highly socially comfortable and academically effective study group.
- Acceptable Group Support, corresponding to a student experiencing some social and academic benefits from a study group.
- Low Group Support, corresponding to a student experiencing few social and academic benefits from a study group.
- Negligible Group Support, corresponding to a student experiencing no benefits.

In designing a set of exhaustive characteristics of each of these quality levels, all existing definitions of social academic support reviewed in Section 2.1.1 were incorporated, along with feedback from existing students. Four separate contributing subconstructs, or phenomena contributing to group quality, were identified as follows, with their corresponding literature-reviewed themes denoted alongside:

- **Reliability of group engagement and presence**, which targets both the theme of social contribution of the group, and consistency of group interactions.
- **Availability of the group**, with a focus on frequency of interactions, targets the necessity of frequent group interactions (whether whether via in-person meetings or digital communications).
- **Effectiveness of the group in providing course learning** for the student, which targets the theme of academic benefits and activities occurring.
- A **student's social comfort within the group**, which targets the themes of social contribution and psychological safety in the group.

| Sub-construct | Respondent Experience | Perceptible behavior characteristic |
|---|---|---|
| Reliability of support network | Feeling of security and/or confidence in accessing group members | Consistency of responses, amount of group that interacts with the student |
| Availability of support network | Groups meets at frequency desired by the student | Overall high frequency of interactions |
| Effectiveness of group in course learning | Feeling that group members are useful to studying | How often a student studies and/or does assignments with the group |
| Comfort with the group | Feeling comfortable sharing ideas/questions, and with their role | Student is included in the group, and initiates interactions frequently |

Figure 1: Sub-constructs for factors contributing to group quality, and the corresponding characteristics a respondent might experience in that factor with a high quality, or a perceptible behavior in that factor for a high quality group

The four sub-constructs outlined above were integrated as dimensions of group quality. Experiences along these dimensions are taken in combination to contribute to an overall measure of study

group quality. Additionally, effort was made to identify subconstruct characteristics of groups along two dimensions: both in student's emotional perception of interactions with their group (denoted as Respondent experience), and more externally measurable characteristics of group functioning (denoted as Perceptible Behavior Characteristics). The sub-constructs, along with corresponding internal respondent experiences and externally perceptible characteristics are detailed in Figure 1.

**The complete construct map is available in Figure 2**, incorporating criteria at differing levels of these subconstructs. It is meant to be read in full, as it informs all subsequent formulation of questions.

## 2.3 Item Panel Format

Working within the Item Response theory design model, the group quality measurement survey is referred to below as the "instrument", and all questions designed for this survey are referred to as "items".

Items were developed to cover each of the sub-construct sections detailed above in Figure 1. In designing these items, importance is placed on both respondent internal experience and external group behavior, addressing degrees of these external characteristics and internal experiences outlined throughout the differing levels on the Construct Map in Figure 2. For this reason, items within each sub-construct were constructed to address both internal experience and external behavior.

Item response options were developed at scales that correspond to one of the four overarching construct levels, given they reflect varying levels of their sub-constructs. Options for all items are polytomous (offering more than two choices), based on a 4-level Guttman-scale model [45]. Each item option is carefully worded to indicate clear meaning to the subject taking the survey, such that the responses to any option could be interpreted to reflect very similar experiences across any respondent. Occasionally, multiple item response options within the same question reflect the same construct map level. The questions are presented to the students in matrix formats, with questions grouped by the response option types, to allow the survey to appear of a shorter length to the students. Please refer to Appendix B for a copy of the survey as viewed by respondents.

The resulting 12 items are listed in Figure 3 below, and they are grouped by the corresponding sub-construct.

### 2.3.1 Alternate Forms Format

In order to provide an additional measure of internal reliability, the item set was designed to be divided into two sets of questions, Forms A and B. This measure of reliability is based on ensuring consistency of responses between two different sets of items, which still target the same theoretical phenomena. The larger item set was therefore split to ensure there were pairs of questions of which both would target the same subconstruct. In the case of sub-constructs with three questions assigned to each, namely "Effectiveness of Learning" and "Comfort with the Group", a pair of questions was chosen from each to corresponded to respondent's subjective experience rather than a measurable phenomenon, and which hit the related concept of motivation towards academic

| Internal experience characteristics of a supported student (respondent) | Quality of peer support in study groups | External/perceptible characteristics of a supported student (item responses) |
|---|---|---|
| 1. Students feel very secure in the availability of their group<br>2. **Students feel the group meets at the frequency they desire**<br>3. Students feel their **teammates are consistently conducive to learning**<br>**4. Students feel very comfortable sharing ideas or asking questions** with their support network- they feel comfortable in group contexts | High Quality Group Support | 1. Students receive a **high consistency of responses, and group frequently initiates interactions**<br>2. Students collaborate with their group **at high frequency, 1+ times a week**<br>3. Students use their support network for assignments/course studying, and perform better on these activities when they do so<br>4. Students frequently initiates interactions/ participates in group settings |
| 1. Students feel secure in the availability of their group<br>2. Students feel the group meets at the frequency they desire<br>3. Students feel their teammates are conducive to their learning<br>4. Students feel **somewhat comfortable in group contexts** | Acceptable Group Support (Average working study group) | 1. Students collaborate with their group **somewhat consistently (every 1-2 weeks), and group initiates interactions**<br>2. Students reach out to teammates as frequently as in other classes<br>3. Students sometimes use their support network for assignments/course studying, and perform better on these activities when they do so<br>4. Students occasionally initiates interactions/ participates in group settings |
| **1. Students don't feel secure in the stability of their support network - but there are a few peers they can consistently reach out to**<br>2. Students feel the **group does not meet at the frequency they desire**<br>3. Students feel their teammates are conducive to learning<br>4. Students may not feel comfortable in group contexts | Low Group Support (Group breakdown) | 1. Students **collaborate with only a few peers somewhat consistently** (every 2-3 weeks), and those peers may **initiate interactions**<br>2. Students sometimes use their support network for assignments/course studying, but **may not perform better on these activities**<br>3. **Occasional replies when reaching out** in group modes of communication - **consistent replies from 1+ people**<br>4. Students occasionally initiates interactions/ participates in group settings |
| 1. Students don't feel secure in the stability of their support network<br>2. Students feel the group does not meet at the frequency they desire<br>3. **Students do not feel their teammates are conducive to their learning**<br>**4. Students do not feel comfortable in group contexts** | Negligible Group Support | 1. Students **only collaborate with peers a few times, to no times, during a course**<br>**2. Students do not use their support network for assignments/course studying, and/or do not perform better when they do**<br>3. **Low to no replies when reaching out** in group modes of communication - may have occasional replies from at least one person<br>4. **Students infrequently reach out to teammates, or feel teammates do not reach out to them.** Students infrequently/ never participate in group or peer settings |

Figure 2: Construct Map. Each level of study group support quality is provided in the central column, with corresponding internal experience characteristics and external characteristics provided in the leftmost and rightmost columns, respectively. Characteristics are listed per subconstruct level, as specified in the following key:

Key: 1 - Group reliability

— 2 - Group availability

— 3 - Effectiveness of learning in the group

— 4 - Comfort with the group

(a) Group Reliability items

| Question | Question type | Construct map levels, and options corresponding to them | | | |
|---|---|---|---|---|---|
| | | Negligible | Low | Acceptable | High quality |
| I initiate interactions with my group __ | Perceptible characteristic | Never | A few times a month | Once a week | More than once a week |
| This percentage of the group regularly participates in study group activities | Perceptible characteristic | None | Few | Many | Most/All |
| Group members respond to ___ of group interactions or study activity initiation | Perceptible characteristic | None | Few | Many | Most/All |
| Other group members initiate interactions ___ | Perceptible characteristic | Never | A few times a month | Once a week | More than once a week |

(b) Group Availability items

| Question | Question type | Construct map levels, and options corresponding to them | | | |
|---|---|---|---|---|---|
| | | Negligible | Low | Acceptable | High Quality |
| I wish I could have interacted with my group more frequently. | Respondent experience | Strongly agree | Agree | Disagree | Strongly disagree |
| I interact with my study group ___ | Perceptible characteristic | Never | A few times a month | Once a week | More than once a week |

(c) Effectiveness in Course Learning items

| Question | Question type | Construct map levels, and options corresponding to them | | | | |
|---|---|---|---|---|---|---|
| | | Negligible | Low | Acceptable | | High quality |
| I perform better on assignments when I collaborate with group members. | Respondent experience | Strongly disagree | Disagree | Agree | | Strongly agree |
| I collaborate with members of the group on this percentage of assignments* | Perceptible characteristic | None | | Few | Many | Most/All |
| I collaborate with members of the group on studying for this percentage of exams* | Perceptible characteristic | None | | Few | Many | Most/All |

(d) Comfort in Group items

| Question | Question Type | Construct map levels, and options corresponding to them | | | |
|---|---|---|---|---|---|
| | | Negligible | Low | Acceptable | High quality |
| I would like to work again with some or all of the people I met in my group, if they take the same future courses. | Respondent experience | Strongly disagree | Disagree | Agree | Strongly agree |
| I feel comfortable asking questions in the group. | Respondent experience | Strongly disagree | Disagree | Agree | Strongly agree |
| I feel comfortable with the role and contributions I make in this group. | Respondent experience | Strongly disagree | Disagree | Agree | Strongly agree |

Figure 3: Items, organized in subfigures corresponding to subconstructs. Item wordings are provided in the "Question" column, with a description in the "Question type" column of whether they are meant to measure a respondent experience or perceptible characteristic. The "Construct map levels, and options corresponding to them" section in each subfigure provides the wording of options per each question, listed under its corresponding study group quality level.
* Here, the options "Few" and "Many" both correspond to the construct level of "Acceptable". This is because acceptable quality groups may have varying degrees of collaboration on assignments and/or exam studying.

| Form A | Form B |
|---|---|
| (1) I initiate interactions with my group ___ | (1) Other group members initiate interactions ___ |
| (1)___ of the group regularly participates in study group activities | (1)Group members respond to ___ of group interactions or study activity initiation |
| (2) I wish I could have interacted with my group more frequently. | (2) I interact with my study group ___ |
| (3) I collaborate with my group on studying for _____ of the homeworks. | (3) I collaborate with my group on studying for _____ of the exams. |
| (4) I feel comfortable with the role and contributions I make in this group. | (3) I perform better on assignments and/or exams when I collaborate with group members. |
| (4) I feel comfortable asking questions in the group. | (4) I would like to work again with some or all of the people I met in my group, if they take the same future courses. |

Figure 4: Items as divided between Alternate Forms A (left column) and B (right column). Numerical descriptors are provided in parentheses prior to each item, representing their corresponding subcontruct.
Key: 1 - Group reliability
— 2 - Group availability
— 3 - Effectiveness of learning in the group
— 4 - Comfort with the group

success. This pair can be viewed in the 5th row of Figure 4. The concept of benefiting from collaborating with the group would likely correlate with feeling comfortable with one's role and contributions. The final split, reflecting these subconstruct targets, is available in Figure 4.

## 2.4 Instrument Modeling and Dataset

This section is primarily concerned with an analysis of this survey's reliability and validity, as defined per standards developed by experts in Education Measurement, via methods recommended in [39].

Specifically, once survey responses are collected, an Item Response model may be applied to best ascertain the level of study group quality which the respondents have experienced. The model applied for this instrument is the Polytomous Rasch model [46, 38], which in this case best approximates the probability of a student choosing a certain option per item with multiple options, given their overall study group quality.

### 2.4.1 Variable and Methods Definitions

Per the Polytomous Rasch model, each item option corresponds to some level of group quality. The group quality value of option $k$ for item $i$ corresponds to the option's group quality level, and is denoted as $\delta_{i,k} \in (-\infty, \infty)$. The two extremes of the scale represent negligible group quality to high group quality.

Per the model, each student respondent also falls somewhere on the group quality level numerical scale. For student $n$, their numerical study group quality level is denoted as $\theta_n$, on a scale of $\theta_n \in (-\infty, \infty)$. Group quality levels of both students and item options are estimated on the same scale.

A student $n$'s response to item $i$ is represented as $X_{i,n}$. Per the model, the probability of $X_{i,n}$ being a certain option $x$ is mathematically modeled as:

$$Pr(X_{i,n} = x|\theta_n, \delta_{i,x}...\delta_{i,0}) = \frac{\exp \sum_{k=0}^{x} (\theta_n - \delta_{i,k})}{1 + \sum_{j=0}^{x_{\max}} \left( \exp \sum_{k=0}^{j} (\theta_n - \delta_{i,k}) \right)}$$

This probability model is not exact, as it does not model the dropoff in probability of choosing a certain option once a student's level $\theta_n$ is far above a certain item. Rather, the model ensures that the probability of choosing a higher level option, rises to a greater probability, once a student's level $\theta_n$ surpasses the level of the higher option. Mathematically, this is expressed as:

$$\theta_n > \delta_{i,x} \implies Pr(X_{i,n} = x|\theta_n, \delta_{i,x}...\delta_{i,0}) > Pr(X_{i,n} = x-1|\theta_n, \delta_{i,x-1}...\delta_{i,0})$$

and

$$\theta_n = \delta_{i,x} \implies Pr(X_{i,n} = x|\theta_n, \delta_{i,x}...\delta_{i,0}) = Pr(X_{i,n} = x-1|\theta_n, \delta_{i,x-1}...\delta_{i,0})$$

This model assumes an ordering of options exists such that $\delta_{i,x} > \delta_{i,x-1}, \quad \forall x \in (0, ..., K)$.

In further analyses performed in this chapter, the following terms will be used and defined as such:

- MLE: the maximum likelihood estimate of a parameter, such as $\theta_n$ for a student, or $\delta_{i,x}$ for a given item option.

- Logit: this is the unit for the scale of estimated group quality. The term derives from modeled parameters for group quality level falling on a log probability scale. For example the logarithm of a ratio between different modeled probabilities is denoted along this scale as follows: $\text{logit}(k : k-1) = \log\left(\frac{Pr(X_{i,n}=k)}{Pr(X_{i,n}=k-1)}\right) = \sum_{k=0}^{k}(\theta_n - \delta_{i,k}) - \sum_{k=0}^{k-1}(\theta_n - \delta_{i,k}) = \theta_n - \delta_{i,k}$
- Thurstone thresholds [47]: also referred to as option thresholds, these thresholds represent the group quality logit level at which a student's probability of selecting a sequentially higher option surpasses the probability of selecting the preceding option. We note a special modeling case based on this example: when the probabilities of a student responding to two sequential item options is equal, we denote $Pr(X_{i,n} = k) = Pr(X_{i,n} = k-1)$. In this case, $\log\left(\frac{Pr(X_{i,n}=k)}{Pr(X_{i,n}=k-1)}\right) = 0 = \theta_n - \delta_{i,k}$, and therefore $\theta_n = \delta_{i,a}$. This implies that when student group quality level is estimated to be equal to the level represented by a higher item option, we model that student is equally likely to select the immediately preceding option. As soon as a student is slightly above an option's modeled group quality, or has passed the option's Thurstone threshold, $\delta_{i,k}$, the student is more likely to select that option than any other.

Now that the Polytomous Rasch model has been established, numerous methods exist for approximating group quality parameters based on data of student responses to surveys. In this case, the process of fitting parameters $\theta_n$ and $\delta_{i,k} \forall i, k$ is approached via a Maximum Likelihood Estimation method, performed via the BASS software [48]. We intend to validate this instrument using this software package (Sec. 2.5. Sec. 2.6). We are not experts in parameter estimation for survey validation, and as such take this software package as given. If student group quality levels are to be estimated using the Polytomous Rasch model described above, and used in concrete classroom settings, the author recommends the validation of a high quality estimator for these parameters.

### 2.4.2 Data Collection

The pilot data for this survey consists of 85 survey respondents from the Fall 2021 offering EECS 16B. All students represented in this data provided consent to allowing their anonymized data to be used for research purposes, under the project IRB with Protocol ID is 2020-08-13526, for which the PI is Prof. Gireeja Ranade. This group of students consists of the counts across demographic categories denoted in the following table, with students selecting "Prefer not to answer" denoted as "PNA":

| Racial Demographic | Asian/ Asian American | Black/ African American | Hispanic | White | Mixed Race | PNA |
|---|---|---|---|---|---|---|
| | 50 | 1 | 8 | 8 | 13 | 5 |
| Gender Demographic | Female | Male | Gender non-conforming | | | PNA |
| | 24 | 55 | 3 | | | 3 |
| Year Demographic | Sophomore | Junior | Senior | Senior Tranfer | | PNA |
| | 38 | 28 | 15 | 1 | | 3 |

## 2.5 Analysis of Reliability

Any instrument that takes the form of a survey is subject to variability, or error, in the student's replies. As described by Wilson in Sec 7.1 of *Constructing Measures: An Item Response Modeling Approach* [39], this may take one of four forms:

1. Variability in answers due to a student's interest, mood, or health;
2. Variability in answers based on the conditions or time in which the student is taking a survey;
3. Measurement inaccuracy due to item wording, or style of presentation of the instrument;
4. Measurement inaccuracy due to inconsistency of scoring the items. Due to the multiple choice form of the item set in consideration, no human scoring is required, so inaccuracy in this area will not be taken into consideration.

The intent of measuring the instrument's reliability lies in assessing the variability of the measurement stemming from any of these categories. High instrument reliability implies lower probable error of measurement, and can be ensured via internal consistency of responses across questions. In effect, if students responses between different sets of items correlate highly with each other, this provides assurance that the instrument is able to draw signal about the construct being measured, with fallbacks even in case of some kind of variability or inaccuracy.

### 2.5.1 Internal Consistency Coefficient

The internal consistency coefficient, $r \in [0, 1]$, is defined in Wright & Masters, 1982 [49], and is also known as the separation reliability coefficient. This coefficient measures the amount of variance in data that is captured by parameter estimates according to the Rasch model, and not explained by error in Rasch model predictions. $r$ can be viewed as a measure of consistency of the model in predicting variation in student responses. It can be calculated as follows:

$$\hat{\theta} = \frac{1}{N} \sum_{n=1}^{N} \theta_n$$

Observed total variance of estimated student parameters: $\quad Var(\hat{\theta}) = \frac{1}{N-1} \sum_{n=1}^{N} (\hat{\theta}_n - \bar{\theta})^2)$

$$\text{Expected response to an item } i: \quad E_{i,n} = \sum_{k=1}^{K_i} kPr(X_{i,n} = k|\theta_n, \delta_i)$$

Mean-square error(MSE) of expected responses compared to actual responses:

$$MSE(\theta) = \frac{1}{N}\sum_{n=1}^{N}(X_{i,n} - E_{i,n})^2$$

Variance accounted for by the model = Observed total variance - MSE:

$$Var(\theta) = Var(\hat{\theta}) - MSE(\theta)$$

$$\text{Proportion of variance accounted for by the model:} \quad r = \frac{Var(\theta)}{Var(\hat{\theta})}$$

The internal consistency coefficient for this instrument is reported at 0.915. Expert estimates rate coefficient values between 0.9 and 0.94 to indicate functional internal consistency [50]. This result indicates reasonable internal consistency of model predictions at similar construct levels, across questions in the instrument.

### 2.5.2 Alternate Forms Reliability

An analysis of alternate forms reliability was conducted using a correlation between MLE parameters per student, and a Spearman-Brown measure of reliability. Each student responds to the questions on both forms A and B, outlined in Figure. 4. The hypothesized effect would be that each student's group quality would be estimated at the same level, based on their responses even to the different forms. A high correlation between estimated group quality per student, modeled from the item sets of each separate form, would indicate that the measurement of group quality can be well replicated by using just one of the item sets. This ultimately serves as assurance that, given some inaccuracy in response to one question in Form A, for example, its noisy effect on final group quality estimate may be minimized by accurate responses to Form B.

Performing the correlation in student MLE between the forms, produced an $R^2$ value of 0.7245, and a Spearman-Brown reliability value of the overall instrument at 0.840, indicating an moderately high positive correlation between the form responses. The correlation is visualized in Figure 5. The moderately high values indicate that there is reasonable internal consistency ensured by the parallel nature of these two forms. However, the values being less than 0.9 indicate imperfect correlation in responses to the forms, likely because they were not developed with perfectly duplicated question topics, and therefore measure some slightly independent concepts from each other.

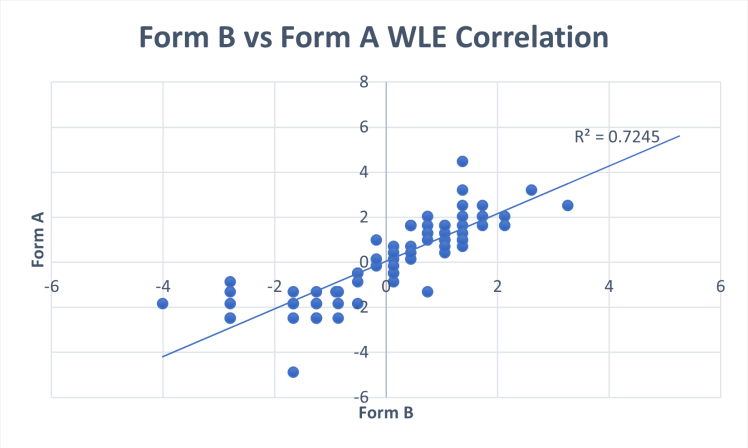



Figure 5: Comparison of estimates of student group quality level, based on item sets separated into alternate Forms A and B. The y-axis represents student parameter estimates $\theta_n$ for all students $n$ responding to form A. The x-axis similarly represents student parameter estimates in form B.

## 2.6 Analysis of Validity

The validity of an instrument can be defined as the alignment of modeled results with the theoretical construct backing. This section presents multiple possible measures of instrument validity, as recommended in Chapter 8 of *Constructing Measures: An Item Response Modeling Approach* [39].

### 2.6.1 Construct Validity

A primary form of instrument validity is the valid relationship between the construct / phenomenon intending to be measured, the content of the item set, and the resulting responses to the item set. The relationship between the construct and the item set has been detailed and justified thoroughly in Section 2.2.

A Wright Map is provided as an additional measure of construct validity [49]. A Wright map is intended to visualize, for each item, the Rasch-model based Thurstone thresholds at which a student's probability of selecting a sequentially higher option surpasses the probability of selecting the former option. The calculation of these Thurstone thresholds is described in Section 2.4.1, but the thresholds essentially represent the item option parameter values $\delta_{i,k}$, for a given response $k$ to item $i$. A Wright Map with high validity would demonstrate similar values in thresholds for item options representing the same construct levels. This sort of close alignment in corresponding thresholds is referred to as "banding".

Available in Figure 6, the Wright Map shows a strong band of very similar question thresholds indicating *High Quality Group Support*. *Acceptable Group Support* bands can be somewhat divided into two groups of questions, which each demonstrate similar threshold likelihoods of moving to the next construct level. The first four questions in the Wright Map, of which three relate to the Group Comfort subconstruct, and which all ask about subjective "Respondent Experience" as described in the Construct Map Section 2.2 and Figure 3, show relatively low thresholds for

Figure 6: Wright Map of Items on Study Group Quality Survey. Histogram of estimated student group quality along the log-probability scale is available at the top. Thurstone thresholds per item, on the logit scale, are represented along the x axis, with items along the y axis. The histogram at the top is a distribution of student group quality level estimates, $\theta_n \quad \forall n$. We see the first four items demonstrate very similar threshold values corresponding to levels of group quality, and the last seven items also demonstrate very similar threshold values corresponding to levels of group quality.

answering at options regarding *Low Group Support* or *Acceptable Group Support*. Additionally, all these questions' options were worded from "Strongly disagree", "Disagree", "Agree", and "Strongly agree". This may describe that respondents found generally it easier to "Disagree" than to "Strongly disagree" when asked if a particular quality of their group was good, and generally found it much easier to "Agree" than to "Strongly agree".

The only question for which these subjective quality thresholds are an exception, and the threshold from *Low Group Support* to *Acceptable Group Support* is very high, is worded as follows:

*"I wishcould have interacted with my group more frequently."*

The intent of this question was to gauge whether the group was meeting at the frequency the student desired, and so if the group quality would be high. Therefore, if the student agreed that their group should interact more frequently, the response was scored at Low group quality, and if they disagreed with the idea their group should interact more frequently, the response was scored as Acceptable. The high threshold values indicate that the probability of answering "Agree" (at Low Quality) was relatively high, in comparison to answering either "Disagree" or "Strongly Disagree". This points out ambiguity in the question, and a tendency for students with high quality groups to answer "Agree" - since a student may well want to interact more with their group if they like it! Outside of changing the wording of the question, the scoring may be more accurate at Acceptable group quality (rather than Low) if a student agrees they want to interact with their group more often.

The last seven questions on the Wright Map all show consistent banding around all of the thresholds. These questions all generally measure perceptible characteristics in group quality, regarding frequency of types of interaction or initiation.

One option to increase consistency of these questions with the construct, and with the student experience questions, would be to simply lower the frequencies listed in the last seven question options, and score slightly lower frequencies as being at Acceptable. This fix, however, would be performed based on the assumption that there is high correlation in level of responses, with some offset of level corresponding to the options.

To investigate this assumption, two different model parameterizations were performed, based on the separation of the item set into the five "experience" questions, and the seven "frequency" questions. A measure of correlation between MLE of student group quality was performed, wherein a high correlation with non-zero offset would demonstrate a need for shifting over of construct levels corresponding to the options in one of these two sets of questions.

The correlated Likelihood Estimates of each student's overall score, estimated via two separate item sets, is visualized in Figure 7, as measured within each separate subset of questions.

The resulting positive correlation is quite low, with an $R^2$ value of 0.46. The overall distribution is shifted a little higher for the frequency option questions, but for individual students the actual combinations of responses do not necessarily correlate. This indicates the hypothesis of correlation with offset between these question subsets is likely incorrect, indicating that perhaps internal experiences that a student feels with their group do not always correlate with the other components

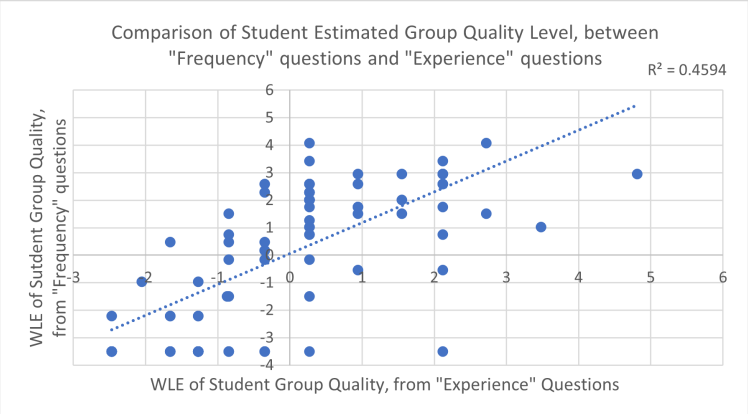



Figure 7: MLE of student group quality level, estimated from "Frequency" items, vs MLE of student group quality estimated from "Experience" items.

of the group being helpful to their academic experience. Instead, it appears that many students may feel comfortable with a group, while still not drawing many interactions or high academic value from the group. Overall, the author would not change the construct map in light of these results, as it still makes sense to measure the final group quality as a combination of different possibly independent sub-constructs. However, the author would consider amending the options for "Experience" questions to a set that captures a more complex set of internal processes, and leads to fewer default responses to "Agree" when asked about comfort.

### 2.6.2 Internal Structure Validity - Spearman's Rho

A measure of the internal structure validity was performed using Spearman's Rank-Order correlation test, a widely-used statistical test for measuring the strength in association between rank of empirical measurements with some theoretical ordering of rank. In this case, empirical rank is assigned to the rank of overall estimated parameters for item options, $\delta_{i,k}$, and predicted rank is assigned based on their corresponding construct level (Fig. 8). A Spearman's Rho value of 0.893 was obtained, which is acceptably high and indicates few item options that violated the predicted rank order.

### 2.6.3 Internal Structure Validity - Item Fit

A final measure of internal structure validity may be analyzed via the "Item Fit", or INFIT, of each response option per item, detailed in Section 6.2.2 of [39]. Item Fit represents the ratio of error in prediction from item parameters, versus expected variance from students. Item Fit is calculated via a mean-square fit statistic per item option [51], as follows:

$$\text{Expected score:} \quad E_{i,n} = \sum_{k=1}^{K_i} kPr(X_{i,n} = k|\theta_n, \delta_i)$$

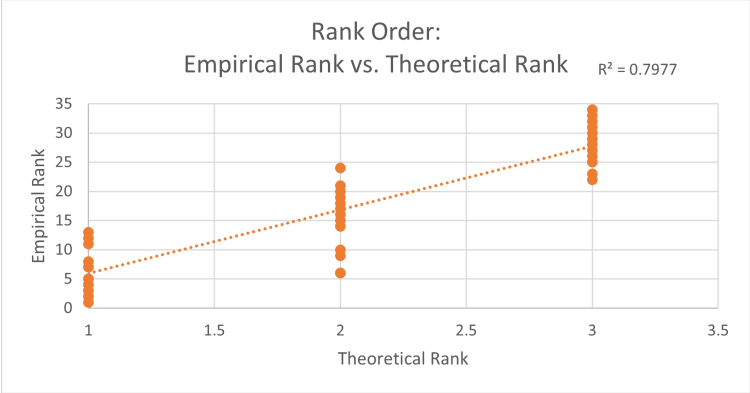



Figure 8: Spearman's Rho Visualization: Empirical rank of item option parameter estimates ($\delta_{i,k}$), vs their theoretical rank given the construct map.

$$\text{Residual of score:} \quad Y_{i,n} = X_{i,n} - E_{i,n}$$

$$\text{Expected squared residual:} \quad W_{i,n} = \sum_{k=1}^{K_i} (k - E_{i,n})^2 Pr(X_{i,n} = k | \theta_n, \delta_i)$$

$$\text{Mean-square expected residual over respondents:} \quad \sum_{n=1}^{N} W_{i,n}/N$$

$$\text{Mean-square observed residual over respondents:} \quad \sum_{n=1}^{N} Y_{i,n}^2/N$$

$$\text{Item Fit (Mean-square fit statistic):} \quad MX_{i,n} = \sum_{n=1}^{N} Y_{i,n}^2 / \sum_{n=1}^{N} W_{i,n}$$

Item fit was measured via this mean square fit statistic (INFIT) per each item response option, visualized in Figure 9. Most all item option INFIT values fell between 0.75 and 1.3 mean square fit, meaning generally reasonable variability in their scores in quality of study groups, without too much interdependence. Two questions, numbers 4 and 5, had options with mean square fit falling above 1.3.

Question 4 was worded as follows:

*"I perform better on assignments and/or exams when I collaborate with group members."*

The variability in answering the "Strongly Agree" indicates that the question should likely be reworded to separate performance on assignments and performance on exams, since these may be independent, and a student's study group quality may not be accurately reflected performing better in both these categories.

Question 5 was worded as follows:

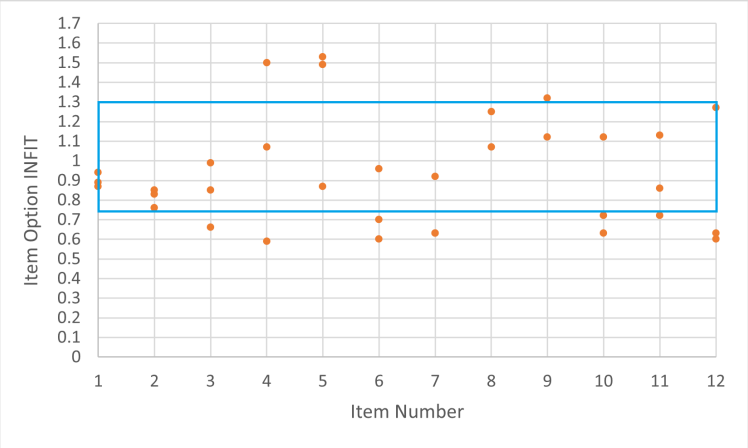



Figure 9: Spearman's Rho Visualization: Rank of item options, based on estimated $\delta_{i,k}$, vs their theoretical rank given the construct map. Questions 4 and 5 demonstrate item options chosen with variance falling above the standard 1.3 acceptable threshold.

*"I wish I could have interacted with my group more frequently."*

This question has been addressed in Section **??** and the high variability in responses may very well be due to the ambiguity of the wording.

### 2.6.4 Response Processes

Response process validity aims to demonstrate that a respondent's experience of interacting with the instrument matches the intended experience. It also aims to ensure that the instrument items collect the intended information from respondents, and that respondents were able to communicate all aspects relating to their experience which the instrument intended to measure. Response process interviews were conducted with personal acquaintances of the author, after they took the survey. These acquaintances were not in any class impacted by the outcomes of this project's study group formation. Responses to these interviews should not be considered as rigorous, unbiased indicators of quality.

In order to analyze response process validity, the following key pieces of feedback were collected from optional post-survey questions, and three interviews. These responses either inform design changes to be made, or validate design choices already made - a design decision is described at the end of each question.

**Question to interviewee: Were any questions confusing, difficult, or uncomfortable to answer?**

*Response: 'I think that the question of if people work better in groups should be split into exam and assignments, as I feel that groups are less effective for me for tests but more effective for assignments.'*

Design decision: change the wording of this question (question 4 in Fig. 5) by splitting it into two questions, as suggested in this interviewee response.

*Response: 'At first I was a little confused on why the answer choices for the 2nd page used options such as "few" and "many" as opposed to language such as "not often" and "often" - but I got it after a little bit. '*

Design decision: do not change the option choice style, as it still took the respondent only three minutes to complete the survey, and the overall item analysis for these questions are very reasonable. Additionally, inducing some pause for thought is not undesirable - it may result in respondents thinking more carefully about their answers.

*Response: 'Overall it was hard to answer some of the questions since my original group only had two people.'*

Design decision: In the future, include a question asking about number of students they ended up working with, and perform external variable analyses on this information to analyze the impact that different-sized groups have on overall responses.

**Question to interviewee: Are there any other aspects to how you interacted with your study groups that you believe we missed?**

Summary: The responses mentioned that it might be fruitful to ask about other social aspects of the group besides comfort or desire to interact with them, or to ask what benefits students take from study groups. Some examples include how often the group interacts in informal settings, or number of shared with the group.

Design decision: These questions might be interesting to analyze in the context of how a study group impacts a student's social life, and some extensive time should be spent designing what such items would look like. It is unclear whether they would provide additional independent information, on top of the existing items surrounding student social comfort.

**Question to interviewee: How long did it take to complete the survey?**

Summary: All responses fell in the range of 2-3 minutes.

Design decision: This is a short amount of time, which validates the legibility and short length choices which went into designing the instrument, and which ideally will function to maximize response rate.

As a note, there was no mention of discomfort or emotional difficulty in answering any of the questions.

### 2.6.5 External Variables

A measure of construct validity may also be investigated through the comparison of distributions of MLE group scores of students in different demographic groups. Finding significant differences in distributions through this method, however, does not always indicate issues with the construct, as some groups may simply have different experiences overall. However, such an analysis may point out areas of interest for whether the instrument itself may produce a differential effect.

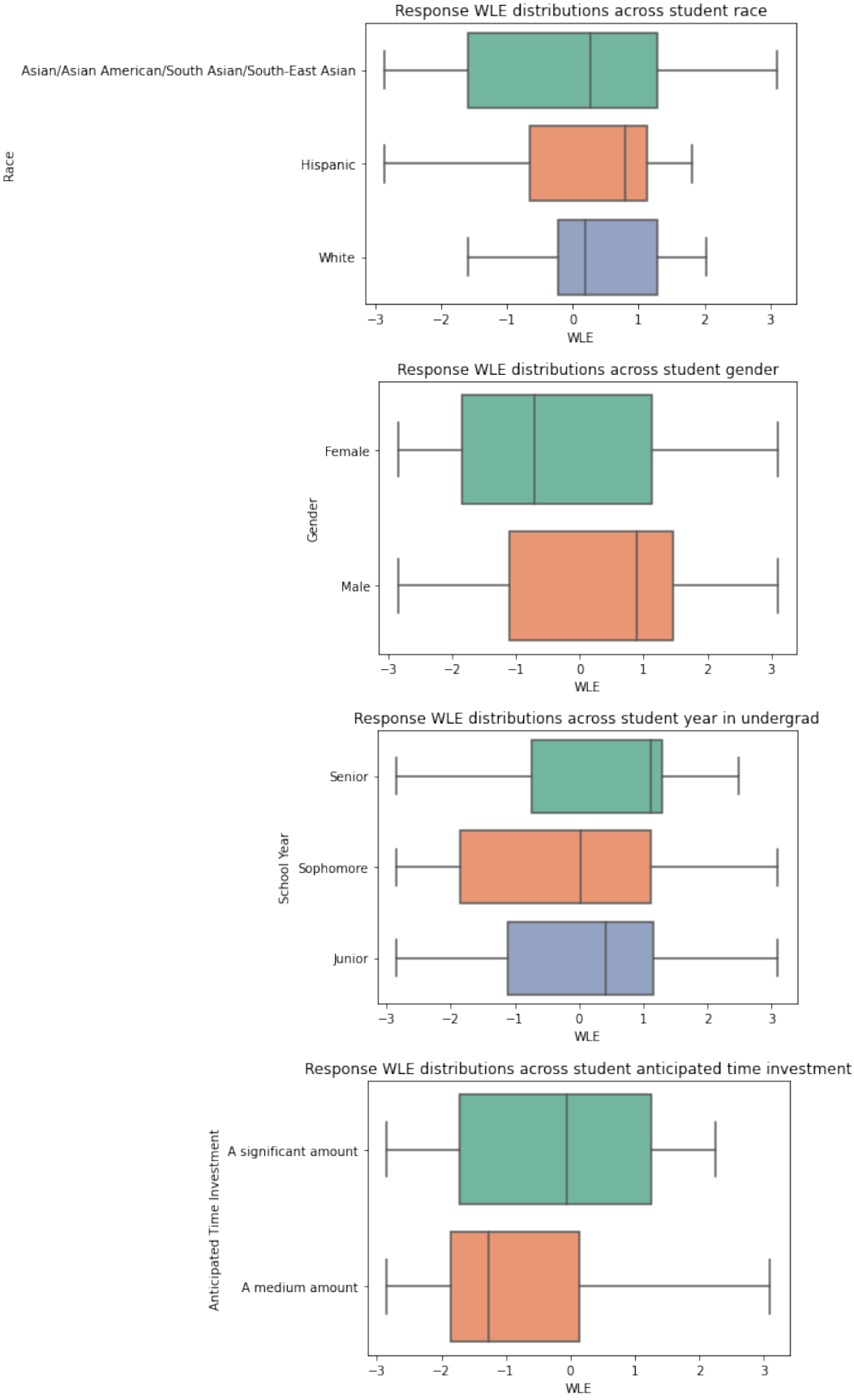



Figure 10: Bar plots of WLE student parameter distributions, as estimated from all responses of students in the class. WLE student parameter distributions are compared between demographics of students, across Race, Gender, School Year. Distributions are also compared between students with different answers to the question "How much time are you hoping to put into this class?", with answers given at the beginning of the semester.

Analysis was conducted on whether MLE score distributions of overall student answers differed significantly across a number of variables, including race, gender, year in undergrad, and the time students stated they hoped to put into the course at the beginning of the year (Fig.10. Although some MLE group score medians differed across these categories, none differed significantly outside 95% confidence intervals. A higher median MLE was observed for men vs women students, with the fringe distributions not varying too greatly, possibly indicating slightly higher study group quality experienced by men students than women students.also found a higher median MLE for students stating they hoped to put a significant amount of effort into the course, which makes sense given they likely invest more time in and thus experience higher quality study groups than counterparts who wish to invest less time in their study groups. No differences worth mentioning were found across race and year in undergrad.

### 2.6.6 Consequential Validity

Consequential validity is assured via analysis of whether future uses of this instrument use any measurements appropriately, and employ an accurate interpretation of what the instrument attempts to measure.

Future versions of this survey will be heavily used for improving/devising study groups in the future, for many students. ML models are intended to be applied for the formation of study groups, using periodic responses from students to this survey to dynamically improve the quality of groups. Although research is currently being iterated on this instrument, actual study groups will not be delivered to students as a product of these survey responses until it has been further iterated on and validated.

## 2.7 Conclusion

The internal consistency coefficient of 0.915 and alternate forms reliability of 0.84 demonstrate good reliability of the instrument to measure the desired combination of characteristics of high quality study groups. A measure of construct validity, via Wright Map interpretation, demonstrated consistent predicted group quality levels among each of two subsets of questions, indicating these subsets measure slightly independent characteristics of group quality. Internal validity was evaluated as acceptably high, based on a Spearman's Rho value of 0.893. The external variables analysis of responses across demographics demonstrates no significant differences in responses between demographic groups.

The author primarily suggests modification of two items of the instrument, Questions 4 and 5, as outlined in the Internal Structure Validity sections (Sec. 2.6.3). Otherwise, the item set demonstrated validity and ability to be modeled to predict student responses, across the rich multidimensional expression of student study group experience and demographics. A slightly revised instrument should be validated with a similar process in subsequent offerings of the study group formation process.

# Chapter 3: Impact Analysis of Study Group Formation

## 3.1 Introduction

This research project, and its effort to make study groups more available in the larger Electrical Engineering and Computer Science (EECS) community at UC Berkeley, has been running since 2020. Over the semesters of formalized partnering with large introductory EECS courses to provide software-matched study groups, as well as allow for self-matched groups, a large amount of data on study group preferences and feedback has been collected from students. This chapter aims to provide a thorough analysis of the experiences of students in study groups, whether these groups were software-assigned or self-formed.

This analysis provides a unique opportunity for insight in the social impact of study groups in a remote classroom setting, as software-assigned study groups in the semester of Fall 2020 were conducted in a remote environment during the COVID-19 period. It further extends a contrast to how study groups functioned in a non-remote setting, in the semester of Fall 2021. Ultimately, the insight from this chapter points towards the benefits of integrating study groups into any large classroom, particularly in the context of remote learning.

### 3.1.1 Methods Summary

The software used to generate groups is discussed in the introductory Chapter 1. In analyzing whether positive experiences are present for students in software-assigned or self-assigned study groups, one method would involve comparison of course software- or self-assigned datasets against a number of control, randomly assigned groups. However, due to the desire to maximize group quality across courses where study group matching was running, such a randomized control group was not available. Indeed, given the assumption that randomly-assigned groups do not address student needs or wishes, it may not be ethical to run such control groups.

Instead, we identify evidence of positive impact within a given group of students by testing the hypothesis that more than 50% of that group of students had a positive experience with a particular indicator. The following research questions are analyzed accordingly:

- Whether study group experiences were positive across multiple aspects of group quality
- Whether differing demographics experienced study groups similarly
- Whether study group experiences changed across semesters
- How the experiences of students in matched groups compare to those in self-formed groups
- Whether differences in study group experience correlated with differences in exam grades

Of particular interest was analyzing whether any under-represented demographic groups faced any difference in study group experience in comparison to peers who were not members of their demographic group. It is discussed extensively in the introductory Chapter 1 how study groups

provide fundamental formative educational opportunities, and how under-represented demographics groups have historically not seen the same inclusion in this social academic contexts. Therefore we analyze across all demographics if disproportionate differences in reported experiences exist.

The primary datasets available for analysis were collected in the semesters of Fall 2020 and Fall 2021. Each dataset includes include the following information, for a given student participating in a study group:

- A variety of matching preferences, such as schedule availability, previous exposure to content, etc.
- Demographic information about their gender, year, and race
- Whether they requested reassignment
- Study group feedback during a semester if they requested reassignment, as many times as they requested reassignment
- Final feedback at the end of the semester. This feedback contained a variety of indicators, or questions addressing factors contributing to the quality of a study group.

Regarding statistical analysis methods to answer research questions, significant results are determined as follows:

- Fisher's exact test for proportion differences in two categories, at a significance level of $\alpha = 0.05$, was used for comparing rates of positive responses to certain study group quality indicators, between different subgroups. For example, it would be used to test for significant difference in proportion of women students reporting high rates of interaction with their group, versus the proportion of men students reporting high rates of interaction with their group.

- Student's t-test, at a significance level of $\alpha = 0.05$, was used for hypothesis testing in subgroup continuous variable differences. For example, it would be used in testing differences in mean Midterm scores between self-formed groups and software-assigned groups.

All analysis presented in this chapter is drawn from students consenting to anonymized analysis of their information and study group experiences, intended for research purposes alone. This analysis is enabled under an IRB-approved study, for which Prof. Gireeja Ranade is the PI, and for which the Protocol ID is 2020-08-13526.

### 3.1.2 Results Summary

Main analytical findings include:

- In a remote semester, when study groups were primary methods of students finding social and academic peer networks, study group quality for software-matched and self-matched groups was overwhelmingly positive. No significant differences were found in study group experiences between majority and non-majority groups for software-matched students. For groups large

enough to analyze with significance, $> 50\%$ of each group reported positive experiences with software-matched groups.

- Study group quality decreased significantly for both software-matched and self-matched students in Fall 2021. This is possibly due to external circumstances. Anecdotally, the most common cited reason was lack of time to meet up.

- Students in self-matched groups report consistently more positive experiences, on average, than students in software-matched groups. However, students from under-represented demographics were less proportionally represented in self-matched groups, in comparison to software-matched groups. Additionally, students from under-represented gender and student year demographics occasionally reported significantly worse experiences in comparison to majority groups.

- Very few significant differences were found between group experiences of majority and non-majority groups in software matched groups. Specifically, no differences were found in experiences across racial or gender demographics.

- Hispanic students requested reassignments significantly more often than non-Hispanic students, and White students also requested reassignments significantly more often than non-White students. These results support the decision to offer at least one reassignment along the course of a course term, as otherwise there may be students, occasionally disproportionately of underrepresented demographics, left with lower-quality groups from the outset.

- Feeling comfortable asking questions, and feeling comfortable sharing ideas, were found to correlate significantly with higher exam scores, for all software-matched students.

## 3.2 Analysis and Evaluation of 16A Fall 2020

The first available comprehensive dataset is drawn from the Fall 2020 offering of EECS 16A at UC Berkeley. In this semester, group reassignments were offered twice, and surveys were incorporated into homeworks and heavily promoted by course staff. Additionally, the remote nature of the semester possibly contributed to students relying on study groups to build their social sphere. In combination, these factors led to a very large dataset of consenting student data. The analysis of Fall 2020 data in this section was performed as part of [33], and builds off of analysis performed by Gloria Tumushabe in Spring 2021 [32].

### 3.2.1 Overview of student impact

Analysis of students participating in study groups across EECS16A Fall 2020 is performed over a total of 477 consenting students. 143 students reported having self-formed groups, and 334 students asked for software-matched groups. The demographic distribution across categories can be seen in column A of Fig. 11, with majority student year being Freshmen, majority student gender being

| Demographic group | (A) | (B) | (C) | (D) |
|---|---|---|---|---|
| Women | 139 | 103 | 74.1% | (66.8%, 81.4%) |
| Men | 326 | 221 | 67.8% | (62.7%, 72.9%) |
| Gender non-conforming/ Genderqueer | 2 | 2 | 100% | - |
| Other/Prefer not to answer | 10 | 10 | 100% | - |
| Black/ African American | 7 | 6 | 85.7% | (59.8%, 100%) |
| Hispanic | 39 | 28 | 71.8% | (57.7%, 85.9%) |
| Native American/ Alaska Native/ Hawaiian Native | 9 | 6 | 66.7% | (35.9%, 97.5%) |
| White | 86 | 66 | 76.7% | (67.8%, 85.7%) |
| Asian | 345 | 233 | 67.5% | (62.6%, 72.4%) |
| Other/Prefer not to answer | 27 | 20 | 74.1% | (57.5%, 90.6%) |
| Freshman | 323 | 214 | 66.2% | (61.1%, 71.4%) |
| Junior or Senior Transfer | 66 | 53 | 80.3% | (70.7%, 89.9%) |

Figure 11: Demographic distribution of students in study groups in EECS16A Fall 2020. Column (A) shows the total counts of students in each demographic subgroup who either self-formed or requested software-assigned groups. Column (B) shows the count of students of each demographic subgroup who requested software-assigned groups, and column (C) shows percentages of students in subgroup who requested software-assigned study groups. Column (D) shows population-level confidence intervals on percentages in column (C).

Male, and majority student racial groups being Asian and White. All other demographic subgroups are considered non-majority groups. The 16A Fall 2020 final evaluation survey (see Appendix A) had five questions related to the quality of the study group experience for each student: (1) the frequency of interaction of the study group, (2) the number of students in the study group which regularly participated in interactions, (3) the comfort of the student in sharing ideas with their group, (4) the comfort of the student in asking questions in their group, (5) whether the student wants to take future courses with their group. As an additional factor, we were also interested in understanding associations with how many times a student requested reassignment to a new group.

When analyzing responses in these factors, we classify "positive" responses, i.e. indicators that the student had a good experience in their study group, in Figure 12 below.

| Indicator of group quality | Positive response definition |
|---|---|
| Future courses | Students state they hope they can, or definitely will, take future courses with their group |
| Group interaction | Students interacted with their group once a week or more |
| Group participation | Some, most, or all members participated in group interactions |
| Comfort asking questions | Students agreed or strongly agreed that they feel comfortable asking questions in their group |
| Comfort sharing ideas | Students agreed or strongly agreed that they feel comfortables haring ideas in their group |

Figure 12: Definitions of positive responses to differing question indicators of study group quality. Students are considered to have responded positively to an indicator if they meet the conditions in the "Positive response definition" column.

| Group Quality Indicator | Software-Assigned | Self-Formed |
|---|---|---|
| Future courses | 52.4% (47.0%, 57.8%) | 97.2% (94.5%, 99.9%) |
| Group interaction | 56.6% (51.3%, 61.9%) | 95.8% (92.5%, 99.1%) |
| Group participation | 62.3% (57.1%, 67.5%) | 95.1% (91.6%, 98.6%) |
| Comfort asking questions | 74.2% (69.6%, 78.9%) | 93.7% (89.7%, 97.7%) |
| Comfort sharing ideas | 78.4% (74.0%, 82.9%) | 95.1% (91.7%, 98.6%) |

Figure 13: Percentages (with confidence intervals) of students who reported positive responses to the five group quality indicators. First column: Students in software-matched groups. Second Column: Students in self-formed groups.

When looking at all the students who received software-matched groups, we see evidence of overall positive group experiences in this dataset (see Fig. 13 for response percentages and confidence intervals). For example, 74% of students report feeling comfortable asking questions in their groups, 78% of students report feeling comfortable sharing ideas. We performed significance tests in comparison to the null hypothesis that 50% of students would have positive study group experiences and 50% of students would have negative study group experiences. For all quality indicators, over 50% of our sample had positive experiences. Performing analyses with proportion z-tests at $\alpha = 0.05$, over 50% of all software-matched students had positive experiences with group interaction, group activity, and comfort asking questions and sharing ideas, as can be noted in the second column of Fig. 13.

In the 16A Fall 2020 dataset, self-formed groups reported extremely positive results across all the metrics, as can be observed in the rightmost column of Fig. 13. One perspective is that self-formed groups can be viewed as a gold standard, where students already know they feel comfortable and productive when working with students they choose. However, as will be later discussed in a detailed comparison of self-formed groups to software-matched groups Section 3.4, demographic subgroups of students in self-formed groups do not all experience this gold standard phenomenon.

**Correlations in group quality indicators**

To preface the upcoming analysis, we note that positive responses to certain study group quality indicators tended to correlate with some more than others, using the Pearson correlation coefficient (see Fig. 14). Specifically, positive responses to the group interaction and group activity questions correlated highly with each other, and at moderate levels with the other three indicators. Additionally, the comfort indicators correlated highly with each other.

Whether a student indicated a desire to take future courses with their group did not strongly correlate with a student's comfort asking questions and sharing ideas in the group. This indicates that the comfort in a study group may not be fully aligned with whether a student wants to take future courses with a group. This may be because future academic logistics impact the desire to take future courses, which is independent of the quality of the study group.

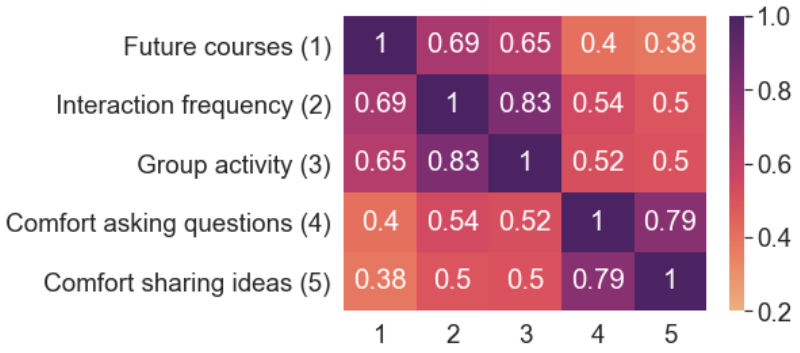



Figure 14: Pearson correlation matrix of positive responses to EECS16A Fall 2020 study group quality indicators. Group quality indicators 1, 2, and 3 share high correlation amongst each other, and indicators 4 and 5 share high correlation amongst each other. Low to moderate correlation is shared in indicators across these subsets.

In general, we consider these differing correlations to point towards different forms of well-functioning study groups, and do not define any particular combination to be an ideal study group. We therefore will conduct analysis of study group quality across individual indicators.

**Requesting reassignments**

A small but sizeable proportion of students requested reassignment to new groups at least once, with 27% of students requesting only one reassignment, 2% of students requesting two reassignments, and $< 0.5\%$ of students requesting three reassignments. Using 2-sample proportion z tests, Hispanic students (and White students) requested reassignments significantly more often than non-Hispanic (non-White) students, respectively indicating that initial group matches did not work out as well for Hispanic and White students. Asian students requested reassignment significantly less often than non-Asian and non-White students. These results support the decision to offer at least one reassignment along the course of a course term, as otherwise there may be students, occasionally disproportionately of underrepresented demographics, left with lower-quality groups from the outset. We also note in Section 3.2.3 that never requesting reassignments can be associated with higher performance on exams. There may be different reasons for this and we do not have any insight into possibly hidden variables, so it is hard to establish any causality. However, we believe it is important to have a quick turnaround for reassigning groups to dissatisfied students.

### 3.2.2 Student impact within demographic groups

To evaluate if our system worked well for students from underrepresented groups, we compare student responses across demographic groups — column A in Fig. 11 provides the count breakdown across these demographic groups. Across both software-matched and self-formed groups, Men are the majority gender group, Asian and White students are majority racial groups, and Freshmen are the majority year group.

In general, students who identify as being from an underrepresented racial or gender demo-

graphic did not demonstrate significant differences in study group quality compared to majority demographic groups, with a few exceptions. We interpret this positively — that with high confidence, the process of forming groups without racial or gender singletons enables underrepresented groups to have study group experiences on par with other groups, which was a major goal for us. We did, however, notice some differences in study group outcomes based on student year; there are likely many reasons for this.

We were not able to control against courses forming study groups through other methods, so we focus our analysis on normative statements to be made about experiences in our pilot group. We additionally do not draw any gender or race comparison conclusions about students who preferred not to state their gender or racial identity.

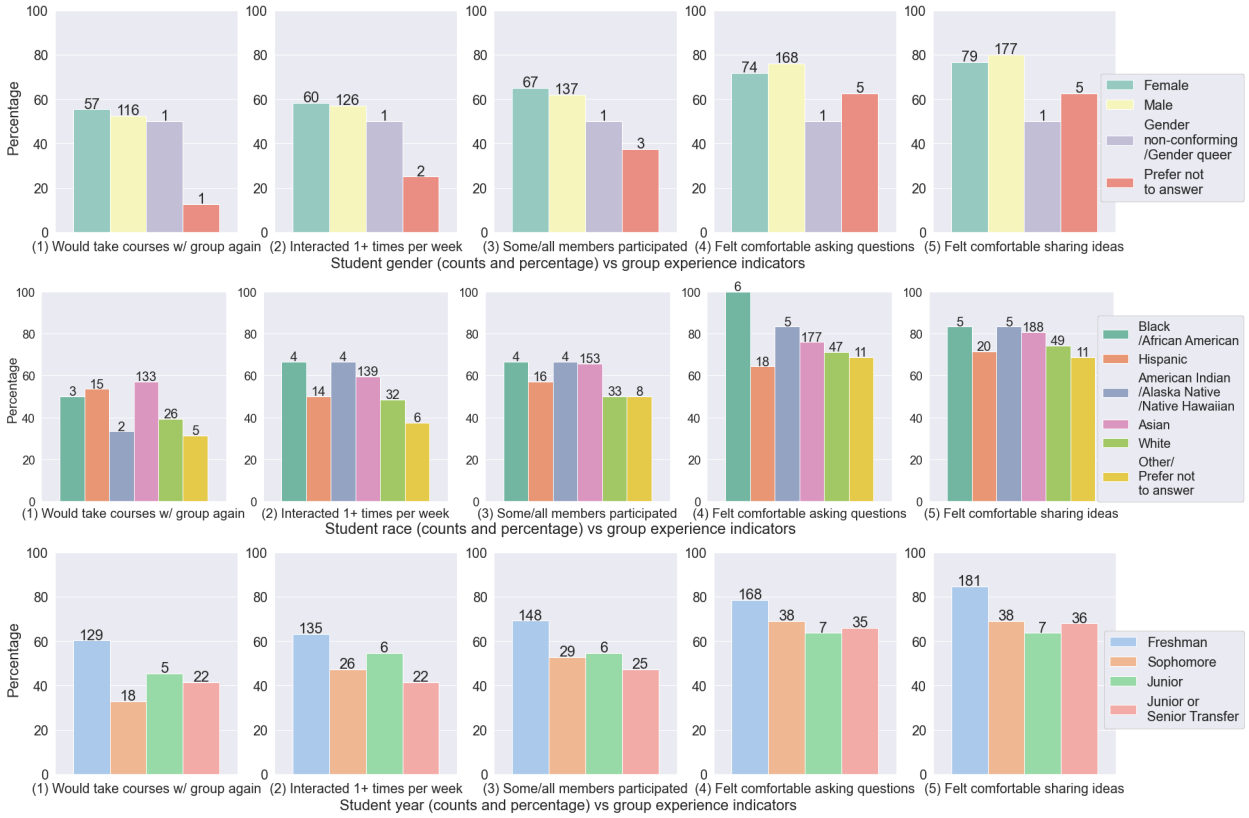



Figure 15: Drawn from dataset of EECS16A Fall 2020 course offering. Comparisons between students of differing demographics, in percentages ($y$-axis) answering positively to key study group quality questions, and counts (displayed on bars) answering positively to key study group quality questions. For student race, students of mixed race were counted in each of the racial categories they indicated they identified with.

### Group quality vs student gender

Using 2-sample proportion z-tests, we found no significant difference between students in the majority group (men) and those who identified as any other gender. This can be observed in Fig. 15, with the yellow (men) and green (women) bars being of similar heights. The small numbers

of students represented in the "Gender non-conforming / Genderqueer" (GNC) and "Prefer not to answer" bars do not significantly skew the non-male sample group towards negative experience proportions. Additionally, confidently more than half of all women and men students in software-assigned groups responded positively to 4 of 5 group quality indicators, aligned with our general population results (see plots 2-5 of group quality indicators in the student gender section, in Fig. 15). One student identifying as GNC had positive group experiences in all categories, and one other GNC-identifying student had negative group experiences in all categories. Due to the very small count of GNC-identifying students, we are not able to generalize these results.

**Group quality vs student race**

Within differing racial subgroups, we find some slight differences in student group experiences. Most notably, no significant differences were found between all students of majority racial groups (White or Asian) and all students of underrepresented racial groups (non-White and non-Asian). All analyses below on individual racial subgroups support this finding.

**Hispanic-identifying students** did not have statistically significant differences in their responses, indicating that final group assignments were comparable in quality to other demographic subgroups. We also note that we can say with 95% confidence that the majority of Hispanic students felt comfortable asking questions and sharing ideas in their final groups. These findings can be observed in the second row of Fig. 15, when qualitatively comparing the height of the percentage bars of Hispanic students to other subgroup bars.

**Students identifying as Black / African-American (AA)** did not show statistically significant differences in any study group indicators, in comparison to non-Black/AA students. For the six Black/AA students who participated in software-matched study groups, there were very positive experiences across all indicators. All six students felt comfortable asking questions, and five of six felt comfortable sharing ideas in their groups (Fig. 15). The small sample size limits our interpretations here, but this is promising.

**Students identifying as Native American, Alaska Native, or Native Hawaiian** (referred to shorthand as Native American), we observed four of the six students indicating positive responses across all study group indicators, and five of the six feeling comfortable asking questions and sharing ideas in their groups (Fig. 15). Again, due to small sample sizes, generalizing statements about their experiences or comparisons to other groups cannot be made with confidence.

**White-identifying students** had significantly less positive study group experiences than non-White students in a few categories, namely: wanting to take courses with their study groups less often, number of students who participated in their groups (seen in the second section of Fig. 15).

**Asian-identifying students** had significantly more positive study group experiences than non-Asian students (Fig. 15). However, when performing tests of significant difference between Asian students and all other non-White or non-Asian students, we found no significant differences in study group quality. Similarly, we found no significant differences in study group experiences between White students and all other non-White and non-Asian students. This may indicate that there are different social and cultural experiences in study groups between two demographics that

may be considered majority groups in engineering classrooms at UC Berkeley, but non-majority groups do not face disproportionately different study group experiences in comparison to either.

**Comparisons in group quality vs student year**

When considering a student's year in school at the time of the course, it became clear there are extremely different study group outcomes depending on a student's year. In many ways, our group matching software was geared towards first-year students, who are the least connected socially. As such, it is promising to note that Freshman students had overwhelmingly positive study group experiences, with significantly higher proportions of positive responses to all five study group indicators in comparison to non-freshmen (Fig. 15).

Transfer students, however, had significantly less positive responses in the indicators of: group group interaction, group activity, and comfort sharing ideas in their group (Fig. 15). This is an additional possible indicator that post-processing matching to ensure non-singletons of certain groups may actually be quite important, our software matcher in Fall 2020 did not fully ensure against student year singletons. Although we partition on student year to ensure groups predominantly comprised of the same year, transfer students may need to be considered an underrepresented group in themselves, and other as of yet unaddressed factors may impact their study group success.

### 3.2.3 Association of high quality study groups and student grades

Although there are clear self-contained benefits to having a high-quality study group, we conducted analyses to verify whether indicators of higher-quality study groups might independently correlate with higher test scores. We find that feeling comfortable asking questions and sharing ideas in a study group is a key indicator of student success, and that study group environments which are likely to encourage these feelings are also highly effective academic supports.

At the classroom-level, all software-matched students demonstrated significant associations between feeling comfortable sharing ideas in their groups, and increases in both midterm and final exam scores, as seen in the top section of Fig. 16. For example, in the "Final score" column of the second row in Fig. 16, we observe that students who felt comfortable asking questions averaged 72.22 on the final, versus students who didn't feel comfortable asking questions averaging near 66.08, with a confidence of 98.9% of this being a significant difference reflected in the wider classroom. Similar results were seen around comfort sharing ideas, as well as not requesting reassignments. All software-matched students also demonstrated significant associations between feeling comfortable sharing ideas in their groups, and increases in Midterm 1 (MT1), Midterm 2 (MT2), and Final scores.

These findings might be explained by the external factors of a student's academic proactiveness and general social comfort to be associated with their better performance as a student, and we cannot draw any firm conclusions. However, given many students feel comfortable in groups and are not requesting reassignments, it could also be inferred that groups afford them the socio-academic support to exercise their academic proactiveness.

Specializing to freshmen, our largest but also one of our target demographics, we find even larger

| Student group | Indicated positive response / indicated negative response | MT1 Score (Max 50) | MT2 Score (Max 50) | Final Score (Max 100) | Misc. diff |
|---|---|---|---|---|---|
| **All software-matched students** | Did not request reassignment / did request reassignment 1-3 times | 40.27 / 37.71 ( p=0.003) | 38.72 / 36.42 (p=0.025) | 72.01 / 67.35 (p=0.046) | - |
| | Felt comfortable asking questions / did not | - | 38.68 / 36.22 (p=0.022) | 72.22 / 66.08 (p=0.011) | - |
| | Felt comfortable sharing ideas / did not | 40.04 / 37.60 (p=0.011) | 38.81 / 35.27 (p=0.002) | 71.96 / 65.86 (p=0.046) | - |
| **Freshmen** | Did not request reassignment / did request reassignment 1-3 times | 41.40 / 37.60 (p=0.0006) | 40.44 / 37.17 (p=0.007) | 74.56 / 68.55 (p=0.046) | - |
| | Felt comfortable sharing ideas / did not | 40.96 / 37.41 (p=0.009) | - | - | - |
| **Transfer** | Would take future courses with group / would not | - | 32.00 / 39.57 (p=0.001) | 61.02 / 73.71 (p=0.012) | MT1-MT2: -0.075 / 0.012 (p=0.018) |
| | Any group interaction / no interactions with group | 34.96 / 39.43 (p=0.013) | 32.45 / 38.98 (p=0.005 | 61.65 / 72.81 (p=0.030) | |
| | Some to most members participated / no members participated | 35.64 / 39.31 (p=0.024) | 33.08 / 39.16 (p=0.009) | - | - |
| **Students with B on MT1** | Any group interaction / no interactions with group | - | - | 72.71 / 65.78 (p=0.008) | MT1-Final: -0.075 / -0.14 (p=0.003) |
| | Some to most members participated / no members participated | - | - | - | -0.084 / -0.130 (p=0.047) |
| | Felt comfortable asking questions / did not | - | 38.99 / 36.53 (p=0.044) | 72.18 / 63.91 (p=0.004) | -0.082 / -0.154 (p=0.004) |

Figure 16: Student's t-tests on difference in sample means, between exam grades within demographic groups, split on positive/negative responses to group experience indicators in the group survey. Only cells with significant differences in sample means are highlighted, with p-values provided. The Miscellaneous Differences column contains percentage-score changes for a student group, either from Midterm 1 to the Final, or from Midterm 2 to the Final, as indicated. Final scores were overall lower than midterm scores, so lower decreases in percent scores are interpreted as positive results.

gains on exam scores for those who did not request reassignments. These results are especially promising given that over 75% of freshmen never requested regroup assignments.

Finally, we explored the impact of study groups on students in different grade ranges. The most impact was seen in the B-range students. Students with B-range MT1 scores (between 68 and 89 percent, based on the class grading scale) saw significantly higher Final exam scores associated with high group interaction groups, and with feeling comfortable asking questions in the group. Additionally, knowing that the general distribution of Final scores was much lower than MT1 scores, students at the B range demonstrated significantly lower decreases from MT1 to the Final associated with several factors: frequently interacting groups, multiple group members participating, and feeling comfortable asking questions in their groups (Fig. 16). These results suggest that for students who entered the class performing at a mid-level, being part of active and comfortable study groups benefited their grade. This range of grades constitutes a significant portion of the student population.

We also considered students who received A, and C to below C-level, grades on MT1, but found

no significant associations between their scores and group quality indicators. This may be because students scoring at an A-level are well-prepared to succeed independently, and students with C-level grades may have external factors affecting performance and their ability to engage effectively with groups.

However, our findings above differed significantly when considering transfer students — this demographic group had generally lower exam performance in general. In addition, students desiring to take courses with their group again, having higher group interaction in group, and having higher group member activity, tended to be associated with having lower exam scores. It is difficult to reason about this phenomenon, due to the large variability of transfer student backgrounds and personal situations.

## 3.3 Analysis and Evaluation of Fall 2021 Course Runs

Extending into the Spring 2020 and Fall 2021 semesters, the research group collected further information about how study groups were operating, and aimed to integrate this feedback into improving the matching process in future iterations. Datasets from classes in Spring 2021, although available for analysis, were not analyzed in this report. In Fall 2021, the information of around 300 students was made available for analysis, across introductory classes EECS 16A and EECS 16B at UC Berkeley. This dataset is comparably smaller to the Fall 2020 dataset. In general, fewer students opted for participating in the study group process, and feedback surveys received lower rates of response. These phenomena are possibly due to the fully in-person nature of the Fall 2021 semester.

### 3.3.1 Survey Formats

Due to the iteration in feedback surveys discussed in Chapter 2, only a subset of questions are comparable between the Fall 2020 and Fall 2021 semesters. Additionally, surveys for running study group formation were not fully standardized across courses in Fall 2021. The sources of feedback available for analysis in Fall 2021 are differentiated as follows:

- From the Fall 2021 offering of EECS 16A, a final survey was conducted asking students about the quality of the study group they primarily interacted with during that semester. This survey was identical to the final survey conducted in the Fall 2021 offering of EECS 16B, and is denoted as Feedback Survey Version 2 in Chapter 2. Due to this survey being included on an optional homework, as an optional question, fewer than 20 students responded to this survey.

- From the Fall 2021 offering of EECS 16A, a midsemester survey was conducted asking students about the quality of their ongoing study group, as well as offering reassignment for students. Around 150 students replied to this survey and also provided consent for analysis of their information. Several feedback questions were provided in this midsemester survey, but the three main questions used for analysis are detailed in Figure 17. All three questions

share common wording with questions in the EECS 16B Fall 2021 survey, and the "Personal interaction" and "Group participation" questions share common wording with questions in the EECS 16A Fall 2020 survey.

- From the Fall 2021 offering of EECS 16B, a final survey was conducted asking students about the quality of the study group they primarily interacted with during that semester. This survey is denoted as Feedback Survey Version 2 in Chapter 2, containing the 11 group quality indicators in Figure 18.

| Indicator | Positive response definition |
|---|---|
| Personal interaction | The student interacted with their group once a week or more |
| HW collaboration | The student collaborated with group members on few, many, or all homework assignments |
| Group participation | Some, most, or all members participated in group interactions |

Figure 17: Definitions of positive responses to differing question indicators of study group quality, in the Summary Indicators Fall 2021 Dataset. Students are considered to have responded positively to an indicator if they meet the conditions in the "Positive response definition" column.

| Indicator | Positive response definition |
|---|---|
| Personal interaction | The student interacted with their group once a week or more |
| Personal initiation | The student initiated interactions with their group once a week or more |
| Group initiation | Group members besides the student initiated interactions once a week or more |
| HW collaboration | The student collaborated with group members on few, many, or all homework assignments |
| Exam collaboration | The student collaborated with group members on studying for few, many, or all exams |
| Group participation | Some, most, or all members participated in group interactions |
| Response rate | Some, most, or all interaction initiations were responded to by group members |
| Interaction wish | Students agreed or strongly agreed that they wish they could have interacted more with their group |
| Comfort asking questions | Students agreed or strongly agreed that they feel comfortable asking questions in their group |
| Comfort sharing ideas | Students agreed or strongly agreed that they feel comfortable sharing ideas in their group |
| Desire to collaborate again | Students agreed or strongly agreed that they hope to collaborate with their group members again |

Figure 18: Definitions of positive responses to differing question indicators of study group quality, in the Detailed Indicators Fall 2021 Dataset. Students are considered to have responded positively to an indicator if they meet the conditions in the "Positive response definition" column.

Given these different data sources, we wished to both summarize feedback in a way that allowed comparison across the Fall 2020 and Fall 2021 semesters, and also analyze the detailed group quality feedback provided by Feedback Survey Version 2. Combined analysis was therefore conducted based on two differing, but overlapping, datasets:

- Summary Indicators Fall 2021 Dataset: contains three study group quality indicators, over the common questions between the mid-semester survey from EECS 16A and the final survey from 16B Fall 2021. This dataset is collected from 236 consenting students, of which 167 students were in software-matched groups and 69 students were in self-formed groups. Demographic breakdowns for this dataset are provided in Figure 19a, where we note that this dataset is composed of predominantly White and Asian students, and Male students, with 16 Hispanic students. $< 10$ students were represented from each of the Black/African American, MENA, and Native American/Alaska Native/Hawaiian Native racial categories.
- Detailed Indicators Fall 2021 Dataset: contains 11 study group quality indicators, over the final response surveys from EECS 16A and 16B Fall 2021, collected from 95 consenting students, 53 students in software-matched groups and 42 students in self-formed groups. This dataset will be denoted as the . Demographic breakdowns for this dataset are provided in Figure 19b, where we note that this dataset is composed of predominantly White and Asian students, and Male students, with $< 10$ students from under-represented racial minorities represented in each racial category. These low numbers mean very few significant conclusions can be drawn from analysis of students from under-represented demographics in this dataset. As such, the Summary Indicators Fall 2021 Dataset will be the only one used in this chapter to test for significant differences in study group experiences between demographic categories.

| Demographic group | (A) | (B) | (C) |
|---|---|---|---|
| Women | 76 | 58 | 76.3%(66.8%,85.9%) |
| Men | 154 | 105 | 68.2%(60.8%,75.5%) |
| GNC | 1 | 1 | 100% |
| Black/ African American | 3 | 1 | 33.3%(0.0%,86.7%) |
| Hispanic | 24 | 16 | 71.8 (47.8%,85.5%) |
| MENA | 4 | 4 | 100.0% |
| Native American/ Alaska Native/ Hawaiian Native | 4 | 4 | 100.0% |
| White | 38 | 27 | 71.1%(56.6%,85.5%) |
| Asian | 165 | 114 | 69.1%(62.0%,76.1%) |
| Freshman | 87 | 64 | 73.6%(64.3%,82.8%) |
| Sophomore | 74 | 52 | 70.3%(59.9%,80.7%) |
| Junior or Senior Transfer | 21 | 17 | 81.0%(64.2%,97.7%) |

(a) Demographic distribution of students in the Summary Indicators Fall 2021 Dataset.

| Demographic group | (A) | (B) | (C) |
|---|---|---|---|
| Women | 27 | 18 | 66.7%(48.9%,84.4%) |
| Men | 64 | 33 | 51.6%(39.3%,63.8%) |
| GNC | 1 | 1 | 100% |
| Black/ African American | 1 | 1 | 50.0%(0.0%,100%) |
| Hispanic | 12 | 6 | 50.0%(21.7%,78.3%) |
| MENA | 3 | 3 | 100.0% |
| Native American/ Alaska Native/ Hawaiian Native | 0 | 0 | 0% |
| White | 14 | 7 | 50.0%(23.8%,76.2%) |
| Asian | 64 | 35 | 54.7%(42.5%,66.9%) |
| Freshman | 8 | 5 | 62.5%(29.0%,96.0%) |
| Sophomore | 41 | 23 | 56.1%(40.9%,71.3%) |
| Junior or Senior Transfer | 2 | 1 | 50.0%(0.0%,100.0%) |

(b) Demographic distribution of students in the Detailed Indicators Fall 2021 Dataset.

Figure 19: Demographic breakdowns in the Summary Indicators Fall 2021 dataset, and Detailed Indicators Fall 2021 dataset. Column (A) shows the total counts of students in each demographic subgroup who either self-formed or requested software-assigned groups. Column (B) shows the count of students of each demographic subgroup who requested software-assigned groups, and column (C) shows percentages of students in subgroup who requested software-assigned study groups. Adjacent in Column (C) are population-level confidence intervals on percentages in column (C).

### 3.3.2 Positive response definitions

When performing analyses across the factors in the Summary Indicators Fall 2021 Dataset, we define positive responses corresponding to each indicator feedback question, in Figure 17.

Similarly, when performing analyses across the factors in the Detailed Indicators Fall 2021 Dataset, we define positive responses corresponding to each indicator feedback question in Figure 18.

### 3.3.3 Correlations analysis

Regarding the Detailed Indicators Fall 2021 dataset, validation of the Study Group Survey Version 2 used in this dataset is provided in Chapter 2, based on the pilot results from EECS 16B Fall 2021. Analysis of Pearson correlations, between the 11 group quality indicator variables in this dataset, further validates the findings of this pilot analysis. Specifically, we find that questions designed to target specific subconstructs of study group quality share high correlations (0.7 and above) with other questions within that subconstruct, and moderate correlations (0.4 and above) with other questions (Fig. 20. All subconstructs are detailed in Chapter 2.

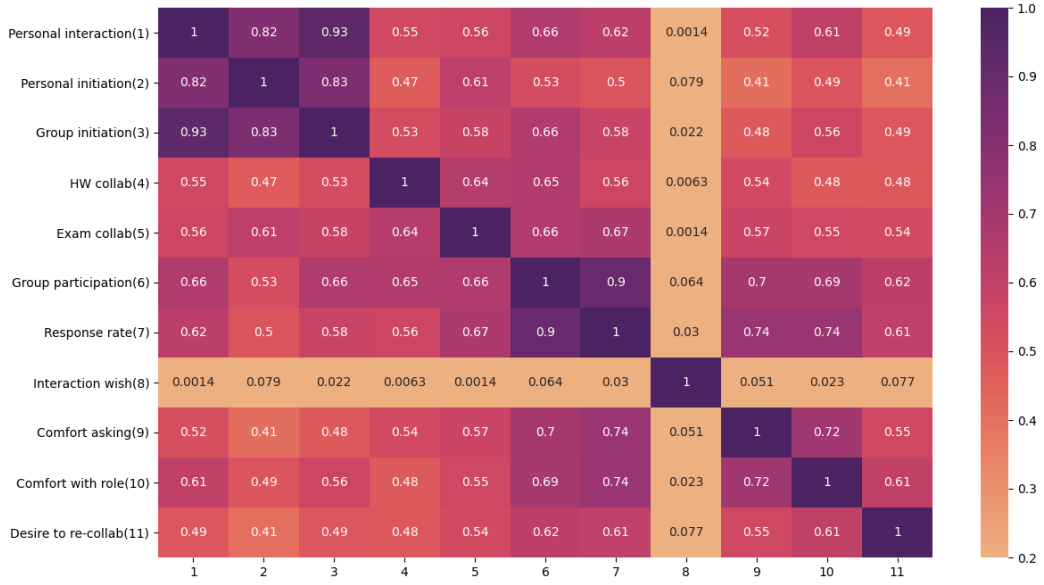



Figure 20: Pearson correlation matrix of positive responses to study group Detailed Indicators, from EECS16B and EECS16A Fall 2021 final surveys. Group quality indicators 1, 2, and 3 share high correlation amongst each other, and indicators 6, 7, 9, and 10 share high correlation amongst each other. Questions 4, 5, and 11 share moderate correlations with all other indicators besides question 8. Low to moderate correlation is shared in indicators across these subsets. Question 8 shares no correlation with all other questions.

The standout question is question 8, denoted "Interaction wish(8)" in the Pearson correlation matrix in Figure 20, which shares near-zero correlation with all other questions. This indicates responses which do not correlate with any other measure of study group quality, likely due to poor question design, and the question has been removed from subsequent iterations of this survey. This removal is based on both the lack of correlation with any other indicator of quality, and also based on evidence from Chapter 2 that responses to this question did not align well with other similar responses to group quality.

This correlation analysis of the Detailed Indicators Fall 2021 Dataset enables an understanding that study group quality can be comprehensively viewed as a combination of these indicators. Positive responses to detailed dimensions of group quality correlate slightly with each other, but capture slightly independent qualities as well, and thus no single indicator is a perfect predictor of other dimensions of study group experience.

Pearson correlations between positive responses in the Summary Indicators Fall 2021 Dataset are provided in Figure 21. The analysis of correlations between the three summary indicators reveals high correlation between positive responses to "Personal interaction" and "Group participation", and moderate correlation between "Homework collaboration" and the other two indicators. Although many dimensions of study group experience are not captured in these three indicators, especially comfort metrics, these metrics are still sufficiently varied and moderately correlating to be said to capture related but somewhat independent aspects of study group quality. We therefore will conduct analysis of study group quality across all individual indicators.

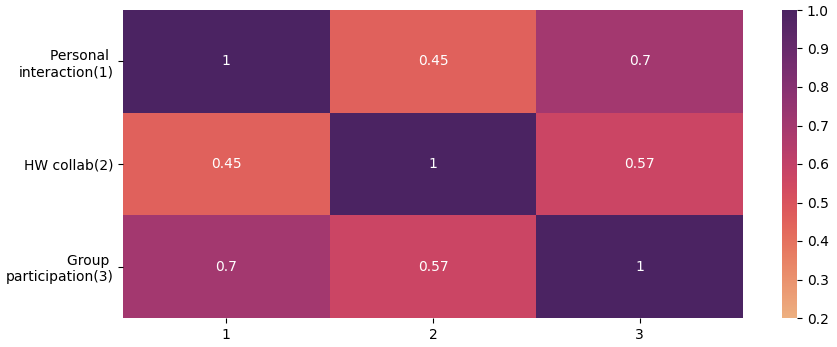



Figure 21: Pearson correlation matrix of positive responses to Summary Indicators, from EECS16B Fall 2021 final surveys and EECS16A Fall 2021 mid-semester survey. All three indicators share moderate to high correlations with each other.

### 3.3.4 Overview of student impact in Fall 2021

Overall, rates of positive responses in group quality for software-matched students were much lower for Fall 2021 datasets, in comparison to Fall 2020 datasets.

In the Summary Indicators Fall 2021 dataset, across all group quality indicators, software-assigned students did not report positive experience in proportions $> 50\%$. Overall percentages of

positive response are reported in Figure 22, with confidence intervals positive response proportions across "Interaction with group", "HW Collaboration", "Group participation" reported to fall above 20%. For self-matched study groups, rates of positive response were consistently above 50%, but analysis of a decrease in rates for software-matched students across semesters as well is documented in Section 3.3.5.

| Group Quality Indicator | Software-Assigned | Self-Formed |
|---|---|---|
| Personal interaction | 34.7% (27.5%,42.0%) | 63.8% (52.4%,75.1%) |
| HW collaboration | 30.5% (23.6%,37.5%) | 63.8% (52.4%,75.1%) |
| Group participation | 45.5% (38.0%,53.1%) | 78.3% (68.5%,88.0%) |

Figure 22: Percentages (with confidence intervals) of students who reported positive responses to indicators of the Summary Indicators Fall 2021 Dataset. First column: Students in software-matched groups. Second Column: Students in self-formed groups.

In the Detailed Indicators Fall 2021 dataset, across all group quality indicators, matched students did not report positive experience in proportions $> 50\%$. Rates of positive response are reported in Figure 23, with all observed positive response proportions reported to fall below 50% for students in software-matched groups. For students in self-formed groups, overall proportions of positive responses fell significantly $> 50\%$.

| Group Quality Indicator | Software-Assigned | Self-Formed |
|---|---|---|
| Personal interaction | 52.8% (39.4%,66.3%) | 88.1% (78.3%,97.9%) |
| Personal initiation | 54.7% (41.3%,68.1%) | 88.1% (78.3%,97.9%) |
| Group initiation | 45.3% (31.9%,58.7%) | 83.3% (72.1%,94.6%) |
| HW collaboration | 30.2% (17.8%,42.5%) | 76.2% (63.3%,89.1%) |
| Exam collaboration | 18.9% (8.3%,29.4%) | 76.2% (63.3%,89.1%) |
| Group participation | 41.5% (28.2%,54.8%) | 85.7% (75.1%,96.3%) |
| Response rate | 41.5% (28.2%,54.8%) | 88.1% (78.3%,97.9%) |
| Interaction wish | 71.7% (59.6%,83.8%) | 83.3% (72.1%,94.6%) |
| Comfort asking questions | 49.1% (35.6%,62.5%) | 92.9% (85.1%,100.0%) |
| Comfort sharing ideas | 45.3% (31.9%,58.7%) | 92.9% (85.1%,100.0%) |
| Desire to collaborate again | 37.7% (24.7%,50.8%) | 90.5% (81.6%,99.4%) |

Figure 23: Percentages (with confidence intervals) of students who reported positive responses to indicators of the Detailed Indicators Fall 2021 Dataset. First column: Students in software-matched groups. Second Column: Students in self-formed groups.

**Correlations of Study Group Quality with Grades EECS 16B, Fall 2021**

Grade data for Fall 2021 was available in the form of midterm scores, final exam scores, lab grade, participation grade, and overall grade in the course. A brief analysis of student grade data in correlation with study group quality indicators, analyzed across students, revealed no significant differences were observed in grades in correlation with different responses to study group quality indicators.

### 3.3.5 Comparison between Fall 2020 and Fall 2021

The non-positive overall study group experiences of Fall 2021 fall in contrast to the results discussed in the EECS16A Fall 2021 dataset in Section 3.2.1. This decrease in overall quality of study group experiences is reflected in self-matched study groups as well. This phenomenon is illustrated in Figure 24, where the two indicators from the Summary Indicators Fall 2021 Dataset which share wording with the 16A Fall 2021 dataset are compared in positive response rates across the semesters. A consistent decrease in positive responses from Fall 2020 to Fall 2021 is observed in Figure 24, for both software-grouped and self-matched student groups. This possibly indicates a phenomenon not in decrease of quality of software-provided groups, but rather in external factors affecting the motivations of students to interact with or make time for their groups. This may also indicate an association with different course offerings and course staff between semesters, and the degree to which students engage with their groups. The primary circumstantial difference between these two semesters are differing course instructors, the remote learning nature of courses in Fall 2020, and the return to in person in Fall 2021. The next section discusses an array of free response feedback, which builds a larger picture of the reasons students found groups to be less than successful.

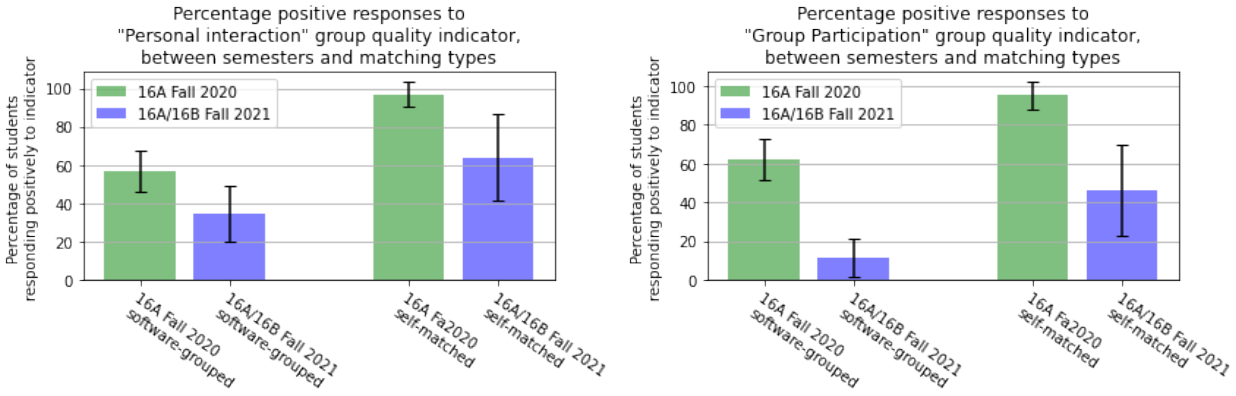



Figure 24: Comparisons are shown between students answering positively, in percentages ($y$-axis), to group quality questions corresponding to "Personal interaction" and "Group participation" indicators, with question wording and positive response bucketing described in 3.3.2. These are the group quality indicators with common wording shared between semesters. The figure denotes a consistent decrease in positive responses from Fall 2020 and Fall 2021, for both software-grouped and self-matched student groups.

**Anecdotal analysis from unsuccessful groups**

In aggregating feedback from both successful and unsuccessful groups, we were unable to conduct systematic student interviews to understand the student experience beyond the surveys. However, questions were included in the feedback surveys allowing students to provide open-ended responses about what went well or poorly. Examples capturing the general trends in feedback from unsuccessful groups are as follows:

- *"The two study groups that I was assigned completely fell through. No one talked after I initiated the conversation multiple times in both of them."*
- *"Most of the students I was matched with didn't show up to the group. Some didn't respond to any of our emails. Besides me, just one other person went to our meetings. I was the only one willing to keep up communications. Our last meeting was the 4th week of the semester."*
- *"I went through three rounds of study group pairing and only in the last round did I get responses from the people I was matched with. Even then, we weren't able to ever meet."*
- *"No negative experiences, but they never responded to my requests to meet up"*
- *"We literally never met"*
- *"I think that being paired with some random study groups, there always needs to be one person willing to initiate. We had a couple messages in the beginning and then no one messaged the group chat anymore so the study group kind of died out."*
- *"There should be consequences for students who don't reply at all. I requested re-match once, and out of the other 6 people that I was paired with, only one of them kept in contact for longer than a few weeks."*
- *"I was assigned to two study groups via the form (the first one fell through due to lack of communication, including on my part). The second time no one reached out either and I didn't feel comfortable initiating that conversation. I wish that these groups could have been formed more organically through discussion or lab section."*

In summary, the vast majority of students whose matched groups did not work well had group members either never reply to their attempts to initiate, or were not able to meet due to scheduling conflicts. This corresponds to the overall proportions of positive responses of software-matched groups being lowest in the "Group participation" and "Response rate", as well as the "Homework collaboration", "Exam collaboration", and "Desire to collaborate again" which are somewhat contingent on the group already having met. Overall, this feedback indicates that the biggest problem to solve, in the face of many students not responding to initiation and/or not meeting up with their groups at all, would be to:

- Facilitate formal avenues for communication
- Better ensure common available meeting times for all group members
- Make students aware of those common meeting times

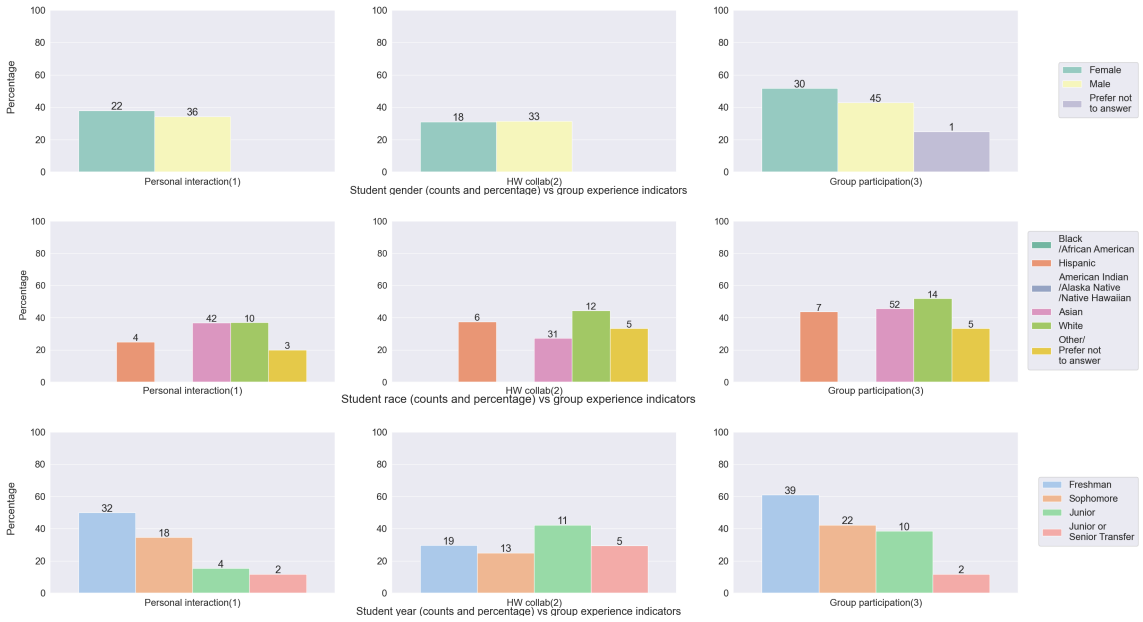



Figure 25: Drawn from software-grouped students in the Summary Indicators Fall 2021 dataset. Comparisons are shown between students of differing demographics, in percentages ($y$-axis) answering positively to key study group quality questions, and counts (displayed on bars) answering positively to key study group quality questions. For student race, students of mixed race were counted in each of the racial categories they indicated they identified with.

### 3.3.6 Student impact between demographic groups

To analyze if certain demographic subgroups experienced better or worse study groups through our Fall 2021 matching process, impact was analyzed using the Summary Indicators Fall 2021 Dataset. Positive response rates to each group quality indicator, for members of one subgroup, were tested for significant difference in comparison to all non-members of that subgroup. Positive experiences for each indicator are defined in Figure 17, in "Personal interaction", "HW collaboration", and "Group participation".

Very few significant differences were found between group experiences of majority and non-majority groups. Specifically, no differences were found in experiences across racial or gender demographics. Some significant differences were seen in experiences of subgroups of different years, specifically:

- Freshmen students saw significantly higher rates of group participation and interaction. An estimated 60% of Freshmen students had positive rates of group participation, and an estimated 50% having positive rates of interaction with their group (Figure 25).

- Juniors experienced significantly lower rates of group participation and interaction with their group than non-Juniors (Figure 25).

- Transfer students experienced significantly lower rates of group participation and interaction

with their group than non-Juniors (Figure 25).

Freshmen students are a key group for which we are interested in providing high-quality groups, in particular because very few freshmen know students coming in to the course, and even if they do, we see in Section 3.4.2 that self-formed groups for freshmen are not significantly higher quality that software-formed groups. Therefore, it is positive that they saw overall positive experiences. However, it seems a number of upperclassmen, and especially transfer students, have not found high quality groups through the software matching process. This analysis further reinforces the idea that concentrated effort should be made in the study group process to enable more convenient meeting times for all students. Given that student year is the only demographic group where students of different demographics have consistently different experiences, the author advises performing singleton prevention for student year in future iterations of running the study group matching process, in a similar manner to the prevention of racial and gender singletons.

## 3.4 Comparison to self-formed groups

As previously discussed, in the 16A Fall 2020 dataset, self-formed groups reported extremely positive results across all the metrics, as can be observed in the rightmost column of Fig. 13. However, not all students of demographic subgroups in self-formed groups experience this gold-standard phenomenon of significantly more positive group experience. This section engages in a detailed comparison of demographic outcomes when compared between software-matched and self-formed groups.

### 3.4.1 Demographic balance between software-assigned and self-formed groups

We notice that students of underrepresented demographics in EECS 16A Fall 2020 were more likely than majority students to request study groups. For example we observe, in the second section of column C of Fig. 11, that Black and Hispanic student proportions of requesting software-matched study groups are higher than Asian student proportions. Similarly, proportions of women and gender non-conforming / genderqueer students requesting software-matched study groups are higher than proportions of men, observable in the first three rows of Fig. 11. Therefore, even though self-matched groups seem to have more positive responses than software-matched groups, it is useful to a large population of students, and especially traditionally underrepresented students, to provide inclusive study group options.

Under-represented demographics in Fall 2021 continue to be better-represented in requesting/receiving software-matched study groups. Specifically, in the Summary Indicators Fall 2021 Dataset, higher rates of women requested study groups than men (see Column C compared between Women and Men in Figure 19a). Higher rates of Hipanic students requested study groups than White and Asian students, and all MENA students and Native American/Alaska Native/Native Hawaiian students requested study groups (see Column C compared between racial subgroups in Figure 19a). One of three Black students requested study groups, but due to the small sample size, no strong

conclusions can be drawn from these trends.

We do not use the Detailed Indicators Fall 2021 dataset to draw any significant conclusions about demographic differences in experience.

### 3.4.2 Demographic difference tests in self-formed groups

In significant ways, key groups that software-assigned study groups are interested in helping were not well-served by self-assigned study groups.

Specifically, in Fall 2020, women experienced significantly lower rates of positive interaction frequency (p=0.035), lower rates of comfort sharing ideas (p=0.008), and lower rates of comfort asking questions (p=0.001), in comparison to non-women students. This can be observed in Figure 26, in comparing the green bar associated with positive experience rates of women to the yellow bar of experience rates of men. No significant experience differences were found across student race in Fall 2020.

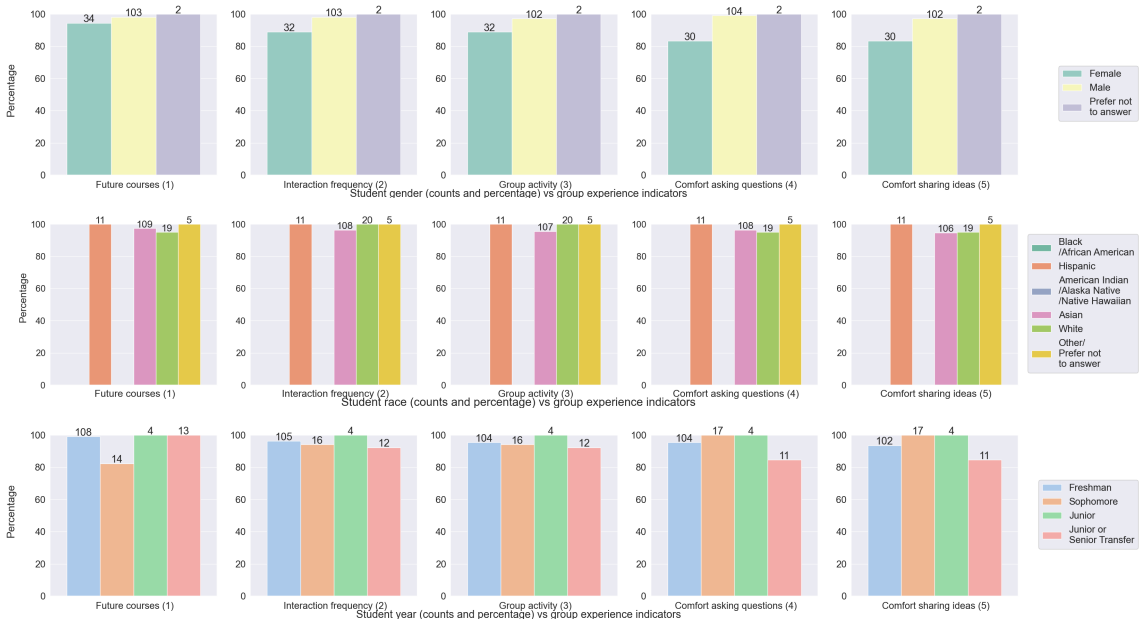



Figure 26: Drawn from self-assigned students in the EECS16A Fall 2020 dataset. Comparisons between students of differing demographics, in percentages ($y$-axis) answering positively to key study group quality questions, and counts (displayed on bars) answering positively to key study group quality questions. For student race, students of mixed race were counted in each of the racial categories they indicated they identified with.

In Fall 2021, it can be observed that there were significant differences in experience for Freshmen students. This possibly indicates that, for upperclassmen coming in with established peer networks, their experience has allowed them to choose students with whom they already work well. However, for Freshmen and Sophomores coming in to courses with established groups, their groups are not as likely to lead to successful outcomes, possibly due to not yet having met academically compatible classmates.

No differences in group experience were observed across gender racial subgroups within self-assigned groups, in Fall 2021 although it is worth mentioning that this is partially due to difficulty in drawing conclusions from the very small numbers of students of under-represented subgroups represented in the self-assigned dataset. For example, it can be observed in the gender comparisons in Figure 27 that women students seem to experience slightly lower quality rates of "Personal interation" and "Group participation", though the small sample size of $< 20$ women implies there is not a significantly low chance of this having occurred. The earlier discussed analysis with larger sample sizes from Fall 2020 supports that there are significant differences in experience for women.

In summary, the analysis reflects that self-formed groups are likely not sufficient to provide high-quality peer academic network experience, particularly not for students of under-represented demographics.

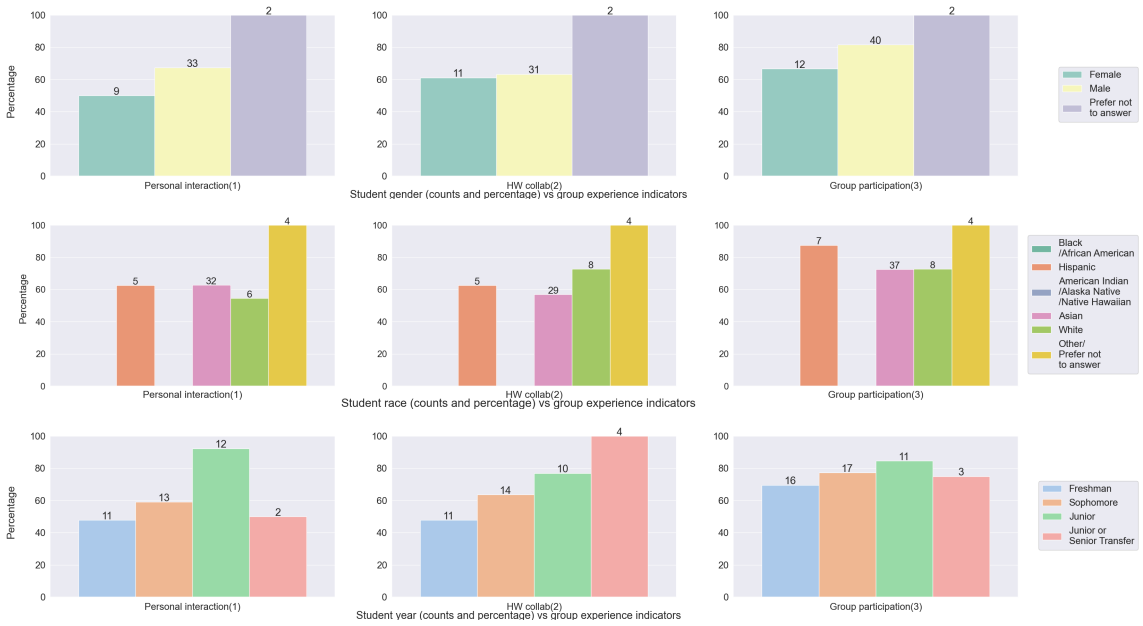



Figure 27: Drawn from self-assigned students in the Summary Indicators Fall 2021 Dataset. Comparisons between students of differing demographics, in percentages ($y$-axis) answering positively to key study group quality questions, and counts (displayed on bars) answering positively to key study group quality questions. For student race, students of mixed race were counted in each of the racial categories they indicated they identified with.

## 3.5 Anecdotal Analysis Students Not Opting for Software-Matched Study Groups

In wanting to make study group more accessible, as well as to analyze whether there were student needs that the group matching process has not yet been taking into account, open-ended feedback from students who did not participate in study group matching is summarized as follows:

- *"I believe that I have people I already know who I can work well with."*
- *"Working with others is waste of time."*

- *"I complete homework faster on my own"*
- *"I am usually able to get my questions answered in OH"*
- *"Everyone has different schedule, hard to connect"*
- *"not enough time"*
- *"Time conflicts"*

All other comments not included here express similar themes to the ones above. These themes include that students either feel they have a non-formalized peer network already, they feel it is specifically not worth making the time in their schedule to collaborate with other people, or they feel they cannot make time in their schedule even if they would like to.

For students in the first category, it would enable them to expand their networks or to engage with other students if it were a possibility to be matched with students to supplement their current group, rather than to provide an entirely new group. This feature has already been implemented and provided for students in the Spring 2022 run of software group matching, and analysis of its impact will follow in the future.

For students in the third category, it is likely not possible to enable study groups for them that would specifically fit their time schedule beyond making time-availability a priority in matching.

For students in the second category, the question lies in whether students who would not prioritize collaborative learning should be further encouraged to engage in it. It is already apparent that students experience low rates of response . We would encourage further future interviews and surveys to understand whether study groups could be better promoted to encourage such students to engage.

### 3.6 Conclusions and Next Steps

Our analysis has demonstrated that students, particularly students from under-represented demographics, draw benefit from software-matched study groups that meet consistently. However, for in-person contexts, there is evidence to suggest that students are less likely to meet with their groups, although in these semesters, students meeting with groups strongly correlates with positive experiences in other group quality indicators. Addressing lack of group meetup may be approached in multiple ways, both through providing matches that prioritize student availability, and through increasing the accessibility of interacting with groups once they are formed.

Key takeaways remain as follows:

- Given that group success is measured via both positive responses to quality indicators, as well as assessment grades, we find that success is correlated with a combination of measures of student comfort in the group, and frequency of group communication/interaction.

- Software-matched study groups, which include singleton elimination and matching on student group preferences, have equitable outcomes with no significant racial or gender demographic differences, across semesters. This supports the idea that matched groups, created by ensuring

against singletons and addressing student matching criteria, better alleviate equity-related issues that arise in self-formed study groups.

- Significant differences in student year indicate that singleton elimination should be performed on student year in a similar way to how it is performed in student race and gender.
- Feedback from students indicates a particular need for facilitating compatible meeting times and comfortable avenues of communication.
- There is a clear need for software-matched study groups in comparison to leaving students to self-form study groups, as self-formed study groups do not sufficiently or equitably serve demographics of students.

Independent of any grade or learning improvements, ensuring even a few students have better and happier college experiences through study groups, in a provably equitable way, is valuable. We hope that this can have a massive impact when scaled to many courses across institutions.

# Chapter 4: Offline Actor-Critic for Deep Clustering Policy Iteration

## 4.1 Introduction

This chapter approaches study group formation in large introductory EECS courses from a deep Reinforcement Learning (RL) perspective. In this model, the state space is a classroom of students with known study group preferences, arranged in a set of study group assignments, referred to as groupings. An action corresponds to any change in the classroom's groupings, with any policy defined as determining a new set of groupings. The reward function corresponds to the overall quality of assigned groups, measured as a function of responses to a group feedback survey. Iterating study groups through the semester allows instructors to form better groups according to student needs. This process can therefore be modeled as short RL agent trajectories, with trajectory length equal to the number of reassignments performed in a semester. An RL model would ideally be used in conjunction with re-deployments of the group feedback survey throughout a semester, and with student preferences being re-specified each time.

A clear limitation to applying RL in this data context is the short trajectory coupled with the long time scale, with steps in the trajectory occurring whenever groups are reassigned during the semester, at intervals of weeks or even months. Non-RL machine learning models could feasibly be used in this application, and will be explored in upcoming research not included in this project. However, in spite of these model drawbacks, the RL model remains promising in approximating the trajectory of regrouping students over the course of a semester, adapting to a student's changing needs, and providing feedback in training between a grouping actor and a group-quality estimator critic. Furthermore, after training a deep RL actor-critic model for creating student groups, the same model could be used in future semesters to form and improve groups.

This RL approach requires the design of a system that can best approximate both the reward function of a possibly unseen set of groupings, and the best estimate for the subsequent reassignment action to be taken given a current set of groupings. The task pairs high-dimensional input data from student group preference surveys, with a high-dimensional grouping action space, and allows rare opportunities for online training at a time scale of months. This task presents an interesting case application for designing a discrete RL offline actor for grouping, in combination with a group-quality evaluation critic. As such, this project will explore one variant for a deep RL algorithm, with specific interest in an offline actor-critic algorithm, towards the generation and evaluation of groupings.

### 4.1.1 Methods Summary

To develop an actor that generates classroom groupings, an Autoencoder model is chosen to embed student vectors in a latent feature space. The Autoencoder is trained on the task of reconstructing matching vectors from their latent space representations. After training, the encoded latent space

representations can be adapted to a training cycle, wherein high latent vector similarity will be trained to indicate high potential group compatibility.

Subsequently, a model performing unsupervised Deep Embedded Clustering (DEC) in the latent feature space was adapted as a policy to generate groupings given a class of students [52]. Many other non-clustering methods and algorithms are available for group formation, including partition-based algorithms, optimization algorithms over a group search space, and probabilistic models of student satisfaction with a group given assignment, to name a few. Clustering methods particularly pose issues with forming heterogeneous groups, forming many groups of set sizes, and ensuring against singleton students. However, given that DEC provided a viable grouping method that could be improved via gradient optimization in an RL context, it is implemented in this chapter with the intent of validating or rejecting its use.

For this reason, language in this chapter will refer to groups generated via DEC as clusters. A modified group selection method was introduced to the DEC algorithm to ensure group clusters of set sizes, where DEC would otherwise produce variable cluster sizes.

A deep neural network was employed as the value function critic model, regressing on group quality as estimated by each student given their group, and employing an offline value iteration update. These models were trained and hyperparameter-tuned separately. Re-initialized models were then deployed in an offline actor-critic training process adapted from AWAC [1], combining the optimized update functions with an added Bellman-based update method for the DEC actor.

The model is trained and evaluated on the 16A Fall 2020 dataset derived from the study groups application, though future work would ideally involve validating these methods on a larger dataset in this application.

### 4.1.2 Results Summary

Evaluation of the Autoencoder model resulted in high reconstruction accuracy (Fig. 28). This points to an Autoencoder being a well-suited model towards embedding student group matching vectors into a latent space.

The initial training of the DEC actor model resulted in plateauing loss, and no visible qualitative separability in the study group clusters formed in the latent space (Fig. 29). This indicates the DEC clustering method, and potentially other clustering methods, are not particularly well-suited to the group formation task on their own.

After a hyperparameter grid search and training, the best critic model NN architecture produced low RMSE in regressing on group quality given student matching interests and group assignments. These results indicate acceptable performance in estimating the quality of unseen groups, and motivate the exploration and interpretation of further machine learning models to approach this task.

Combined actor-critic training improved estimated performance of the output from the clustering actor, given optimized hyperparameters (Fig. 33).

The combination of an unsupervised clustering algorithm, with a critic trained to recognize

high quality states with high evaluation accuracy, is promising in demonstrating that clustering algorithms may not necessitate ground truth clustering indicators in order to be trained in a supervised way. Via appropriate modeling in a reinforcement learning context, the quality of a deep unsupervised clustering actor can be trained to respond to high dimensional states. Although online tuning should be performed for verification of these results and for further training, these results are exciting to consider in the context of any clustering application with high-dimensional data, off-policy action generation, and a need for mostly offline training with possibility to refine online[1].

## 4.2 Problem Formulation

In this study group matching context, the RL problem was formulated as follows:

- State space: a classroom of students who have been grouped. The state space includes information about each student's matching preferences, and the combination in which they are matched with other students. Each student is represented as a vector composed of demographic features, appended to matching features relating to the type of study group they would like to join. These features all correspond to questions asked in the group matching survey, in this case for the 16A Fall 2020 survey version (Appendix A.1). This input format allows us to incorporate student demographic and preference data into any evaluation or grouping model.

- Action space: a general re-grouping of all the students, given current groups and state space data. Given grouping is approached via clustering in this chapter, the intent is to form k clusters, where k = ($|s|$, total # of students) / ($|g|$, desired # of students per group). The dimensionality of the action space is very large, at $\prod_{i=0}^{k-1} \binom{|s|-i*|g|}{|g|}$ total possible group assignments for a restricted group size.

- Reward Function: the average group quality across all groups in the classroom of $|s|$ students. For a group, the quality is estimated as a weighted sum of the group's individual student responses to the group quality survey administered at the end of the semester. Selecting this weighted sum function is discussed in more detail in Section 4.6.1. A reward of zero is assigned to students who request reassignment, based on the assumption that students derive no benefit from a group that they choose to leave.

- Trajectory: A trajectory for a classroom is defined as the changes in groupings, up until the end of the semester. If a student chooses to remain with their group during a reassignment, until the end of the semester, the student's reward at each time step is assigned to the final group score they indicate at the end of the semester. If a student was reassigned to a group at any given time step, a reward of 0 is assigned as that student's group score.

[1]Chapter code can be found at: `https://github.com/ana-tudor/groups_rl/tree/master`

- Transition modeling: Given this specific application, the resulting state is deterministic given an action. Therefore, for the purpose of this project, there was no need to model a transition function.

## 4.3 Data

### 4.3.1 Data Sourcing

Several private datasets for this application have been assembled since Fall 2020, from the deployment of study group surveys and matching in EECS courses. However, due to the iterative improvement in survey version, as discussed in Chapters 2 and 3, different datasets are represented using different feature spaces. Therefore, the same reinforcement learning model cannot be applied indiscriminately to each semester's dataset. As such, in the initial trial training and validation, the largest consistent dataset was used. This dataset is sourced from students requesting study groups in the EECS 16A course administered in Fall 2020. Students answered detailed matching and demographics questions, were able to request reassignment up to three times over the course of the semester, and answered surveys describing categorical aspects of their study group experiences. For this dataset, the group feedback question wordings can be found in Appendix A, and the group matching question wordings can be found in Appendix C.

### 4.3.2 Input Format and Dataset Size

From the pilot dataset, all demographic and matching survey data was one-hot encoded. This resulted in an input dimension of 134 features for matching, with 513 total students who participated in the matching process, and provided consent for their information being used. Group numbers were appended as an additional feature per student vector, allowing the state space to be fully modeled as a 513x135 matrix. The one-hot encoding of the matching questions leaves room for the final model to assign different weightings to different response options in the same question multiple-choice or multiple-select question.

## 4.4 Student Vector Embeddings: Autoencoder

This project approaches the creation of groups based on student vectors embedded in a latent space, rather than creation of groups based on the full feature space as direct input. There is inherent value to generating a latent space of student vector embeddings: an embedding constitutes a distillation of the full feature space to key values which attempt to fully represent the original feature vectors. This may benefit in the potential interpretability of visualizing example cases of groups in a latent space, given we work in a feasible smaller set of dimensions. It may also benefit in the possibility of training a latent space to represent combinations of features in such a way that represents compatible students in proximity to each other in the latent space.

If used in a clustering context, the latent space would ideally be trained to model smaller distances as higher likelihood of positive group quality. The embedding of a classroom of students

into this latent space would intend to place student vectors closer together if they would be likely to function well in groups together.

A few options exist for generating and training such embeddings. Specifically, dimensionality reduction methods of many kinds exist, with PCA and deep network reduction methods being primary options. Options for generating embeddings in a way similar to Natural Language Processing (NLP) word embeddings, for example via methods such as skip-gram, word2vec, and BERT, were less appropriate, given those methods are often trained using syntactical sentence "context", but there is a lack of equivalent, frequently-recurring "context" for student matching vectors in group formation.

The embedding method used by DEC[52] was an Autoencoder, which uses a deep net encoder to embed feature vectors to low-dimensional latent space. In order to train the model, a decoder deep net with reversed architecture is run on embeddings, to attempt to recover the initial feature values, and both the encoder and decoder are gradient-optimized based on RMSE of decoder prediction per feature. Due to this model offering both of the primary benefits listed above, as well as being intended to be used with DEC (detailed as the choice of actor in Section 4.5), an Autoencoder was adapted as the embeddings model.

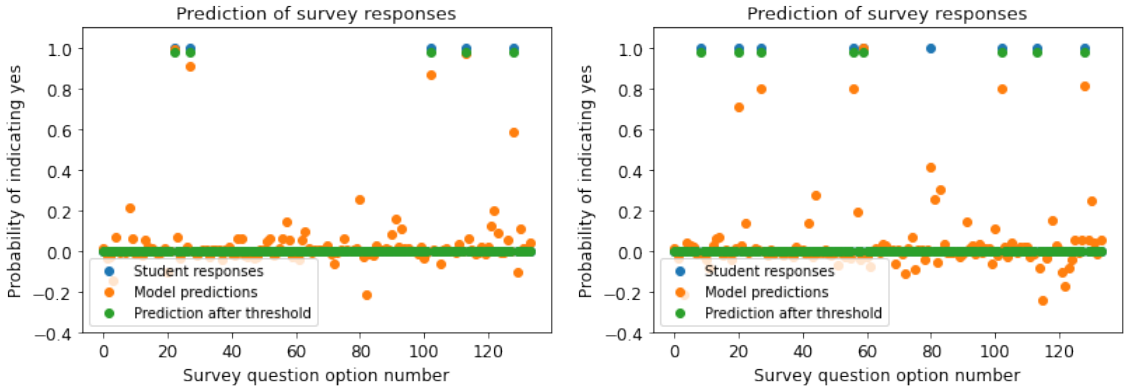



Figure 28: Sample Autoencoder predictions on evaluation set students. High accuracy of re-interpolating survey responses from embeddings is demonstrated in these cases, after thresholding probabilities of "True" responses to survey options above 0.5

### 4.4.1 Results from isolated Autoencoder training

In Autoencoder training, a grid search over architectures, comparing evaluation accuracies of reconstructing held-out student vector data, resulted in the following optimal architecture: hidden layer dimension = 24, number of hidden layers = 2, latent dimensions = 8, and layer activation as ReLU. A grid search over learning rate resulted in a learning rate of 1e-5 working best to avoid over-fitting.

This model performed quite well in recovering evaluation-set student matching vectors, at an evaluation loss of 0.024 MSE, and an accuracy of 97.4% when predictions are thresholded at 0.5. Figure 28 depicts two example reconstructions of evaluation-set student vectors, after encoding

the student matching responses into the eight-dimensional latent space. This presents a well-matched model towards the task of embedding student vectors in a latent space, and presents an argument that low-dimensional representations capture the variation in student matching vectors quite well. Such low-dimensional representations of students could potentially be used in many other formulations of group formation and evaluation, besides the RL perspective presented in this chapter.

## 4.5 Actor: Deep Embedding for Clustering

The task of selecting an appropriate model for the group formation actor was approached with the following priorities in mind:

1. An actor can form high quality groups according to some reasonable metric.
2. An actor can be trained via gradient optimization methods, in an RL context. This is primarily because any gradient-optimized method can be incorporated into an actor-critic RL training framework.
3. An actor can estimate the probability of a student being assigned to another group of students, in an RL context. This feature enables the actor to be trained based on expected reward given some probability of generating previously seen groups. This allows for the actor to be trained on only in-distribution groups with known rewards.
4. An actor can take into account the large-dimensional and rich feature space available, and distill key elements of survey responses to be incorporated during the matching process. A model offering weighting of different survey options can offer interpretability into the key features contributing to high quality study groups, and thus to the key features to be considered in group matching.
5. An actor can generate groups of set sizes, a necessary component for the standardization of study group sizing across a classroom, given size can impact the performance of a group[43].
6. An actor can take feature dissimilarity between students, or feature heterogeneity, into account as a positive factor when considering forming a group. For example, heterogeneity in learning styles has been shown to be beneficial for collaborative learning[53].
7. An actor can prevent against "singleton" students of any demographics, as singleton students have been shown to often have less favorable outcomes in study group experiences[29].

Clustering algorithms presented one feasible option for a grouping actor. If a clustering algorithm were to work within the context of a deep-network generated latent feature space, it would offer priorities 2 and 3 listed above, regarding gradient-optimization methods and weighting of the feature space respectively. Additionally, training a clustering actor via an RL framework would theoretically contribute towards satisfying priority 1.

However, a clustering-based actor is potentially not beneficial towards priorities 4, 5, or 6. Regarding priority 5, clustering algorithms are often based primarily on identifying groups of points based on similarity, generally unable to take heterogeneity into account as a certain "type" of combination of points. Clustering algorithms often do not worry about identifying clusters of set sizes, in opposition to priority 4 - these algorithms model clusters as "types" of points that may occur, with no regard to how many points may pertain to that "type". An argument can be made, for this application, towards certain "types" of groups being able to be identified via clustering, with different types of students falling near group centroids in a clustering space. However, any clustering method would require an augmentation to generate groups of set sizes based on students falling near "group" cluster centers. Finally, a clustering algorithm would certainly not be oriented towards guaranteeing priority 6.

With these considerations in mind, a clustering method should not be considered the ideal option for generating study groups. However, although suboptimal, testing a clustering method of group formation was still a potentially feasible option. It was therefore chosen as the approach for this chapter, with the intent to report on its effectiveness and viability.

The literature exploring RL methods for clustering or group formation is limited, but there exists extensive literature available describing deep methods for clustering [54, 55]. The gradient-enabled training of deep clustering methods, as well as the rich model complexity, both satisfy the priorities for our actor, and an unsupervised method would satisfy an additional priority. For these reasons, the model selected was the Deep Embedding for Clustering (DEC) [52] method.

### 4.5.1 Overview of DEC

The unsupervised clustering model outlined by DEC first fully trains an Autoencoder to learn latent-space embedding parameterizations of points $x_i$ to embeddings $z_i$, as described in Sec. 4.4. The model then iterates over clusters $\mu_j$ in the embedding space trying to maximize separability by minimizing in-cluster variation. This is done by minimizing the KL-divergence between the current Student-t distributions $q_{i,j}$ of latent space points to be assigned to each cluster, and a target tightened distribution $p_{i,j}$. The training algorithm is detailed in the "update_dec" function in Alg. 1.

Specifically, given student embeddings $z_i$ and initialized cluster centers $\mu_j$

$$q_{i,j} = \frac{(1 + \|z_i - \mu_j\|^2)^{-1}}{\sum_{j'}(1 + \left\|z_i - \mu_{j'}\right\|^2)^{-1}}, \quad f_j = \sum_i q_{i,j}, \quad p_{i,j} = \frac{q_{i,j}^2/f_j}{\sum_{j'} q_{i,j'}^2/f_{j'}}$$

The deep net architecture and soft distributions meant this algorithm could be adapted as the actor in an actor-critic algorithm, with group assignment action probabilities able to be estimated as a function of the target distribution probabilities that students are assigned to the same group. This is function is discussed further and defined in Section 4.7.1. Additionally, the tightening of the soft distribution q towards the target distribution p contributes to minimizing assignment probabilities for students farther from centroids. Finally, minimization of in-cluster variation through

unsupervised training, in combination with the prior outlined factors, made this method a viable candidate for attempting to generate groups, in a way which prioritizes student proximity to certain group types in the latent space.

### 4.5.2 Assignment Augmentation

This DEC model was designed for cases with relatively small numbers of clusters $k$. In early experiments with a very large $k$, as per the use case of study group generation, the model demonstrated repeated convergence on extremely similar centroids. Consequently, a large percentage of students would be attributed to one of $k$ very similar centroids, when grouped based only on the DEC algorithm.

Given study group generation requires a large number of distinct groups with sizes in a fixed range, a modified cluster centroid assignment mechanism was introduced for taking actions in the RL markov decision process. In this assignment process, a new grouping would be assigned deterministically as a function of the target distribution $p$ calculated above, up to a limited number of students assigned per group (Alg 1). Cluster centers would be re-calculated during every update.

Theoretically, the generation of clusters via this method would perform well for grouping similar students, or for grouping students who have been placed well in a latent space representing compatibility as proximity. However, given markedly different students are not likely to be represented very similarly in a embedding space generated via Autoencoder, this group generation scheme is not likely to create heterogeneous groups, even if such students would do well when grouped together.

**Alg 1: Modified DEC Update**

```
def update_dec( observation ):
        # Assign new centroids given encodings of new observations
        stud_assignments, mu = get_action_nearest_mu( encode( observation ) )

        #Encode observations, calculate soft and target distribution
        p, q = generate_p_and_q_distributions( observation )

        # Backpropagation update step, based on DEC KL-divergence loss
        loss = kl_loss(p, q)
        dec_optimizer.step( loss )

        return loss

def get_action_nearest_mu( observation ):

        p, q = generate_p_and_q_distributions( observation )
        encoded_observation = encode( observation )

        # Identify centroids with overall lowest probabilities of assignment
        total_probability_per_centroid = sum(p, dim=0)
        centroids_sorted = argsort(total_probability_per_centroid, descending=False)

        #Initialize assignment matrix to hold student vectors of each group
        groups = zeros( (number of groups, number of students per group, latent dimension size) )
```

```
student_assignments = zeros( (number of students, 1) )

for mu in centroids_sorted:
    # Select group_size number of students that are most likely to be
    #  assigned to this cluster
    students_of_centroid = argsort(p[:,mu], descending=True)[ :group_size]

    # Fill in assignment matrix with encoded student vectors for this centroid
    groups[mu] = encoded_observation[students_of_centroid]

    for student in students_of_centroid:
        # Assign this centroid to each student in the centroid
        stud_assignments[stud] = mu

        # Make sure the student cannot be reassigned to other centroids by
        #  over-writing their probability of assignment
        p[stud] = 0

# Reassign centroids as the mean of the students assigned to them
new_mus = groups.mean(dim=1)

return student_assignments, new_mus
```

### 4.5.3 Results from isolated DEC training

For validation of use of the DEC algorithm alone as a group-formation actor, the DEC actor was pre-trained via its specified update function (Alg. 1) and evaluated on a test set, before integrating it into an RL training context. A grid search over learning rate resulted in an optimal learning rate of 5e-5 for this dataset, with a best evaluation KL-divergence of 0.1686.

Training the DEC model alone on the dataset resulted in no clear visual clustering. Specifically, Fig. 29 includes an example of test set students, projected across two pairs of latent space dimensions, after the DEC actor had been trained with its own update function for four epochs. The figure demonstrates no clear visual clustering of what could be interpreted as either clusters of "group types" or "student types". Projection across all other pairs of latent space dimensions produced similar and even less separated results. This intuitive check indicates that this clustering procedure is likely not an appropriate fit for group formation, as it does not appear reasonable to identify types of high-quality groups via a clustering method.

However, the assignment procedure resulted in group assignments of student embeddings in some proximity to each other in the latent space. Fig. 29 illustrates students in formed groups often falling within $\frac{1}{2}$ of the maximal occurring distance from each other, across a given projection on two latent dimensions. However, as discussed, given that latent space similarity is not guaranteed to demonstrate group compatibility after DEC training, the value of this result cannot be interpreted.

Regardless, given the results here formed groups in a way that could be theoretically improved via gradient optimization methods, across potentially more useful objectives, the author continued forward with incorporating Group-Augmented DEC into the RL framework.

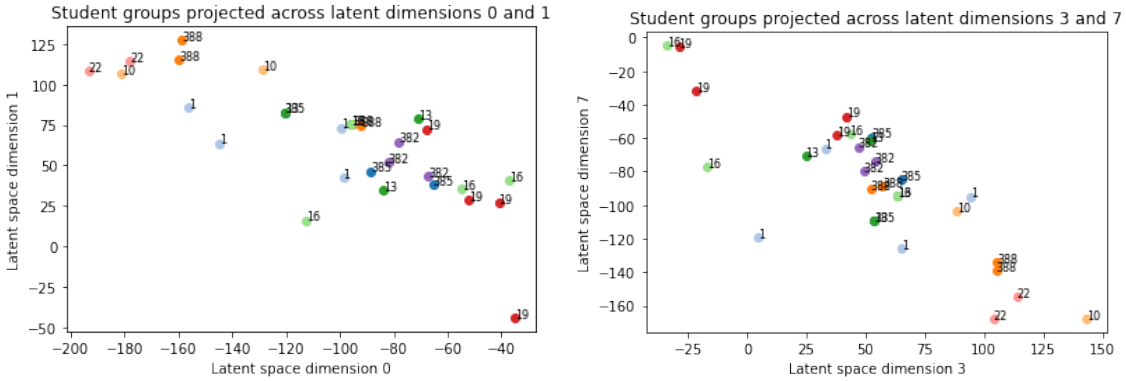



Figure 29: Student embeddings projected onto two example pairs of latent space dimensions, and colored according to group predictions of size 4, after DEC training and group assignment. Student group numbers are included as labels on the points.

## 4.6 Critic: Deep Network for Group Quality Estimation

Bringing a grouping actor into an RL context would be done with the following intuition: when training the results of DEC in an actor-critic context, embedding gradients may theoretically shift as a function of whether the assigned groups were of high quality. Specifically, if high quality clusters are chosen through the assignment process, embedding weights may be trained to place such well-matched students closer to each other. Conversely, generated low quality groups may result in training embeddings to separate students with lower compatibility.

However, in order to perform this process, a reward function would need to be defined for how high-quality a group should be considered to be, and a critic would need to be developed to evaluate the group quality of actor outputs. This section engages with the definition of a reward function, the formatting of data to present to a group quality evaluation critic, and the design of such a critic.

### 4.6.1 Reward function definition

We begin by defining reward function based on student responses to feedback surveys. As with the processing of group preferences and demographic data to create model input, student responses to final survey data were one-hot encoded to create indicator vectors of which group quality options they selected.

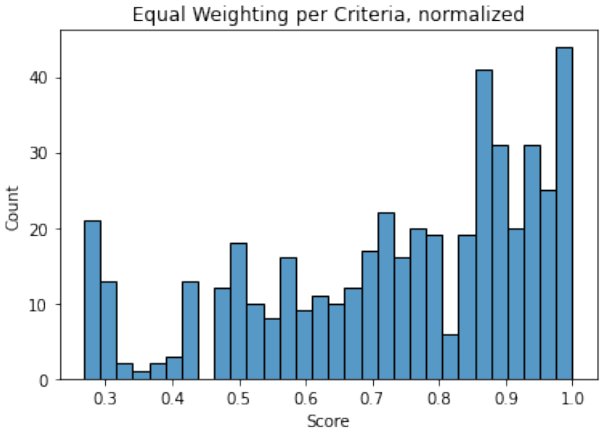
To obtain scalar scores per student, which summarize evaluations of an assigned group, a dot product was taken of the one-hot encoded feedback responses with a vector of linear weights. Such a vector of linear weights will be denoted as a "weighting". As an additional measure to ensure comparable gradient updates, scores were scaled from 0-1 across all student responses.

Several possible weightings were generated and investigated, to select one for use in model training. The final weighting was chosen evaluated based on how it interacted with the data. Specifically, we analyze differences in distributions of group scores per classroom of students, as

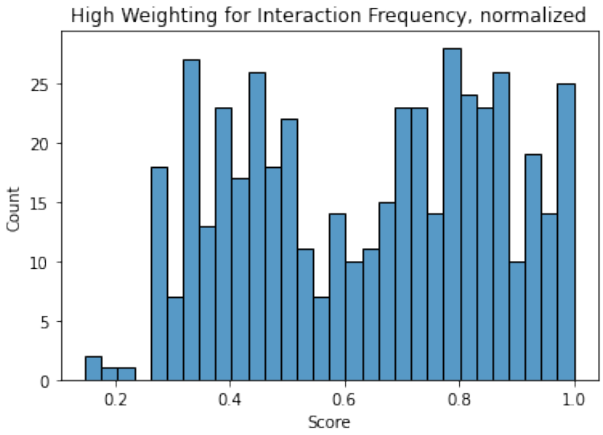
generated by each weighting.

Any metric to compare between distributions may be based on either:

- Intentionally prioritizing certain features, as informed by expert literature suggesting that certain group criteria should be weighted more highly in evaluating group quality score.
- A basis of comparing distributions of group scores in a classroom. Namely, variance in group scores from a resulting weighting may be higher or lower. An argument can be made that a higher variance distribution scores indicates larger differentiation in group scores per student. This differentiation may be beneficial for a model attempting to learn discriminating features between high and low quality groups.




(a) Weights of 3 were assigned to all question options surrounding demographic features. Weights of (0, 1, 2, 3) were assigned to options for each Likert-scale question present in the matching survey. Weights of (0, 1, 2, 3) were assigned to options which correspond to respectively frequency or number, in the questions "How often does your study group meet/interact/-text/chat?" and "Do you feel everyone in your group participates in the study group?". This weighting is denoted as "Equal Weighting". After the weighting is applied via dot product, scores are scaled from 0 to 1. The standard deviation of these scores over the classroom dataset is 0.208.




(b) Weights of 0.25 were assigned to all question options surrounding demographic features. Weights of (0, .5, 1, 2) were assigned to options for each Likert-scale question present in the matching survey. Weights of (0, 1, 2, 3) were assigned to options which correspond to respectively frequency or number, in the questions "How often does your study group meet/interact/text/chat?" and "Do you feel everyone in your group participates in the study group?". This weighting is denoted as "Group Interaction Frequency Weighting". After the weighting is applied via dot product, scores are scaled from 0 to 1. The standard deviation of these scores over the classroom dataset is 0.235.

Figure 30: Two example weightings are used to produce the visualized histograms of summarized group score per student, over the dataset. The high weighting of interaction frequency criteria results in a bimodal distribution, with a higher variance of scores in comparison to the "Equal Weighting" distribution.

In Chapter 2.1.1, it was discussed how experts in education have found varying contributing factors to study group quality. Primarily contributing factors include the overall social balance and composition[41], interaction frequency via in-person meetings and remote communications[42],

and group size[43]. Due to lack of strong consensus, taking any combination of these, or simply choosing one to weight more heavily, could be all argued to be beneficial. Whichever weighting is chosen, if it promotes one of these factors, it would theoretically serve in influencing the formation of groups towards that characteristic.

Regarding distributional comparisons, an example is visualized in Fig.30 - these histograms help visualize how it would impact the group score distribution to emphasize differing final features. We find that prioritizing frequency of group interactions and participation leads to distributions of group scores with highest variance. This focus ultimately resulted in the choosing of a weighting that emphasized group participation and interaction frequencies. This weighting is described as follows:

Weights of 0.25 were assigned to all question options surrounding demographic features. Weights of (0, .5, 1, 2) were assigned to options for each Likert-scale question present in the matching survey. Weights of (0, 1, 2, 3) were assigned to options which correspond to respectively frequency or number, in the questions "How often does your study group meet/interact/text/chat?" and "Do you feel everyone in your group participates in the study group?". This weighting is denoted as "Group Interaction Frequency Weighting".

This weight was arbitrarily chosen, as the one with highest variance among less than ten arbitrary possible weightings. No comprehensive search was performed, but due to both the relative higher variance and valid priority of group meeting frequency, this is deemed an acceptable weighting.

### 4.6.2 Bootstrapping via Data Permutation

The dataset is high-dimensional, but unfortunately relatively small in size, at a size of 525 consenting students who responded to both surveys. Additionally, a scheme needed to be devised to allow a critic model to take into account all students in a group in order to generate an estimate of the group quality.

In order to address both issues, group vectors were generated as follows: For each student, the feature vectors of group members are appended to the student's vector when predicting group quality per student via the critic. All possible permutations of group members are arranged to be appended, such that if there are $n$ members per group, an evaluation from one student about a group's quality generates $n - 1$ permuted data points. This permutation scheme results in an explosion of available training data, as different group permutations can be combined into the same classroom of students in $n \cdot (n - 1)$ ways.

The intuition behind this permutation lies in the idea that the value function model should take into account group assignments not as ground-truth categories, but as indicators towards which students they were assigned to. The combination of their features and other group member features should inform the group quality regression, and the order of the other group members should not matter. This would ideally lead to the value function model being able to optimally recognize features from all group members in a cohesive, non-overfitting way.

### 4.6.3 Critic Design Choice

The deep model architecture for the value function critic was designed with the goal of accurate regression on study group quality, in effect to accurately predict reward of out-of-distribution states. Although the results of such a critic on the output of the clustering actor are not typically used to actually train the actor, they are useful as indicators of how high quality the clustering actor's outputs are, as long as the input data is sufficiently diverse to represent the general distribution of students likely to be seen, and optimally bootstrapped to represent a wide variety of groups.

A few models were in consideration for the critic, including widely used offline Q-learning models that focus on best estimating the quality of an action in comparison to others. However, many of these models were not appropriate for the group generating application due simply to the extremely large action space that is not feasible to fully iterate over. For example, offline importance sampling [56] could not be used for this use case, as the length of rollouts in our study group dataset consisted of at most three total actions taken, none of which were necessarily expert actions. Thus, estimating the likelihood of out-of-distribution actions was not approachable as statistically, all actions tried by the actor were extremely likely to be out of distribution. Q-learning methods [57] specific to offline applications were also very difficult to adapt to this application, as they often assume the ability to optimize over the action space. However, in this use case, the action space is too immense to search over and perform an objective function maximization at every step.

The final architecture chosen was a 2-layer deep neural net model, regressing on each student's estimate of group quality. This critic is intended to estimate the value function $V(s)$, rather than the Q-value $Q(s, a)$. This regression on $V(s)$ is reasonable given direct data about $V(s)$, and given transitions are deterministic in this application, leading to an ability to estimate both Q-values and advantages directly based on $V(s)$.

As a final note, by regressing on student group evaluations, rather than average group evaluations, the critic would be theoretically allowed to provide an estimate of how a certain set of student preferences combined with demographics may lead to a particular experience for just one of the students. This promotes fitting to individual student needs, and could feasibly be used in other contexts of group formation that do not involve an RL training process, but do require the evaluation of potential group experience for a student.

### 4.6.4 Results from isolated Deep Network Critic training

The optimal reported hyperparameters for deep value critic network were at: learning rate = 0.01, number of iterations = 60000, with a reported best RMSE at 0.1242. Given group scores are predicted on a scale from 0 to 1, this RMSE represents that the average evaluation set error is within 13% of actual continuous group scores, an overall acceptable performance for predicting unseen group quality.

## 4.7 RL Actor-Critic Modeling and Results

Once an actor and critic had been designed, an appropriate actor-critic training process needed to be designed to take both into account.

The issues with adopting a Q-learning critic that optimizes over the action space, outlined in Section4.6.3, meant that any offline Q-learning model that required a maximization of Q-value over an action space was not feasible. Therefore, an advantage-based offline actor-critic method, namely the AWAC semi-offline Q-learning model, was adapted to this application [1].

The following definitions will be of use in this section, to contextualize the AWAC RL algorithm in the task of study group formation:

- State: a classroom of grouped students
- Reward: average group quality of a classroom of grouped students
- Critic: group quality estimator
- Actor: group formation algorithm
- Value function: the critic's estimate of the group quality for a class of grouped students
- Policy: the distribution over possible actions, as given by the actor
- Out-of-distribution state: a generated set of groupings that has not been seen in the training data

### 4.7.1 AWAC[1] Overview

AWAC uses the critic model to estimate advantage, $A^{\pi_k}(s, a)$, as a function of projected increase in value of a state, $s$, to the subsequent state $s'$, given some action $a$, and given an action probability distribution from which training data is sampled, $\pi_k$. The set of tuples of states and actions, which constitutes the training data, is denoted as $\beta$. The actor-generated action probability distribution is denoted as $\pi_\theta$.

For the group formation application, advantage is described using the group quality value function critic, as $A^{\pi_k}(s, a) = \gamma * V(s') - V(s)$, where $\gamma$ is the discount rate, and was considered a hyperparameter for this problem.

The AWAC algorithm incorporates the projected advantage into the Bellman equation to generate the following AWAC policy update [2]. Here,

$$\theta_{k+1} = \arg\max_\theta \quad \mathbf{E}_{s,a\sim\beta}\left[\log \pi_\theta(a|s) \exp\left(\frac{1}{\lambda} A^{\pi_k}(s, a)\right)\right]$$

The Lagrangian multiplier was set to $\lambda = 1$ for the group generation application.

AWAC is trained on the offline available permuted group data as if online training were occurring, and as if the step taken by the actor was the one available in the training data. Importantly, actions taken by the actor are not used in any gradient updates - instead, the target probability distribution of group assignments generated by the actor, $\pi_k$, is used to compute the probability

[2] Please refer to Appendix A of the AWAC manuscript[1] for a derivation of this update

that a training set action was taken. Implementation for this action probability computation, and its integration into an AWAC-based update for the actor, is available in Alg. 2 below:

**Alg 2: Actor AWAC Policy Update**

```
# Update the policy network utilizing AWAC update
def actor.update_policy_AWAC(observation, actions, advantages, eval=False):
        p_action_distribution, q, z = self.forward(observations)

        score = 0
        for actions in set( actions ):
            # Collect the distributions over cluster assignments
            #  for each student in the observed group
            action_stud_probs = p_action_distribution[actions==action]

            # First take the probability of these students being assigned to any given group,
            #  as the product over their probabilities of assignment to each group
            # Take the negative log of greatest probability of assignment to the same group
            group_probability = - log( max( product_over_centroids(action_stud_probs, dim=0) ) )

            score = score + group_probability * exp( advantages[actions==action] )

        # Take mean score over number of students
        loss = score / (observations.size()[0])

        # Take backpropagation update step, if in training mode
        if not eval: self.dec_optimizer.step( loss )
        return loss
```

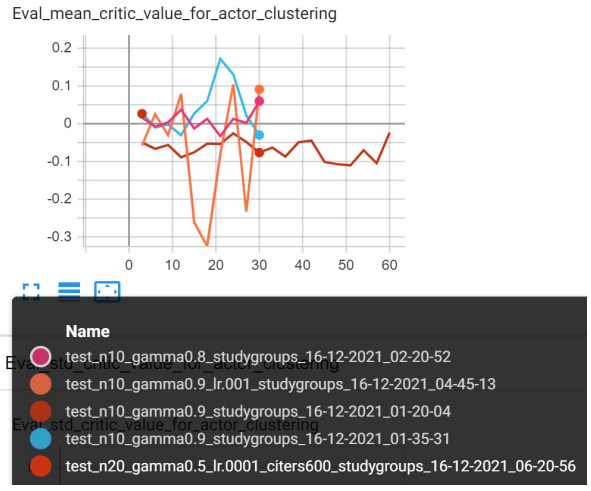
Each AWAC training cycle proceeds as follows: first, the critic is trained on the batch for a number of iterations at each policy step, to improve its estimation of the value function. Then, the actor update alternates between training the critic using the AWAC update, and using its KL-divergence distribution update. This was done to promote that embeddings maintain reasonable separation along the course of updating deep embeddings generation to vary towards more high quality clusters. The actor algorithm is described below (Alg. 3):

**Alg 3: Modified AWAC Training Cycle**

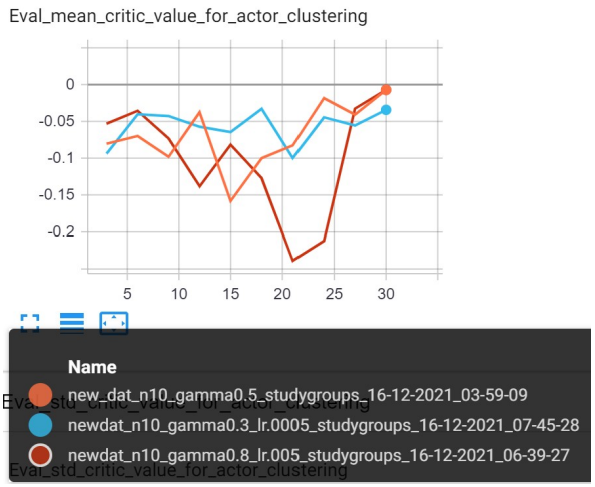
```
def train(observation, action, reward, next_observation):
        #Initiate actor-critic learning
        for i in critic_iterations:
            critic_loss = critic.update(observation, action, next_observation, reward)

        #Estimate the advantage of this action given a somewhat trained critic
        advantage = estimate_advantage(observation, next_observation, reward)

        for i in actor_iterations:
            if i%2==0:
                ac_loss1 = actor.update_dec(observation)
            else:
                ac_loss2 = actor.update_policy_AWAC(observation, action, advantage)
```




(a) Search trajectories with high intermediary rewards. Setting intermediary rewards for group reassignments to values corresponding to the final group outcome led to inconsistent improvement in training.




(b) Search trajectories with intermediary rewards set at zero. Training over 0 rewards for groups ending in reassignment resulted in lower overall scores, but more consistent group quality improvement over the course of model training.

Figure 31: Coarse grid search trajectories over combined actor-critic training. The y-axis corresponds to the mean value of the actor's generated groups, as evaluated by the actor.

### 4.7.2 Combined Actor-Critic Results

For the hyperparameter search over actor-critic training, the following hyperparameters were varied, with their abbreviations listed for later reference:

- lr: Bellman update learning rate
- ntu: Critic number of updates per general AWAC update step
- n: Actor number of updates per general AWAC update step
- $\gamma$, gamma: Discount factor

Initially, reward for groups that had been reassigned was set to the final group quality outcome at the end of the semester trajectory. However, training on these high intermediary rewards was leading to spuriously high results (Fig.31a), with unstable training. This led to a new data reward scheme (labeled "newdat" in figures): at some state, for a student who requested reassignment from their group, their group quality evaluation was set to zero. This decision was more consistent with intuition about study groups, since students requesting to leave their group should be considered to dislike working with their group.

After shifting this feature and analyzing the results on the overall lower intermediary rewards ( Fig. 31b), it was determined that very high gamma performed best, along with training the critic for a large number of steps, and training at a lower learning rate for more stable training. The high gamma value matched group formation problem intuition, as there is no uncertainty in moving on to subsequent state spaces, and propagating reward gradients from down the line is not negatively impactful in the case of very short trajectories. Although the actor should ideally

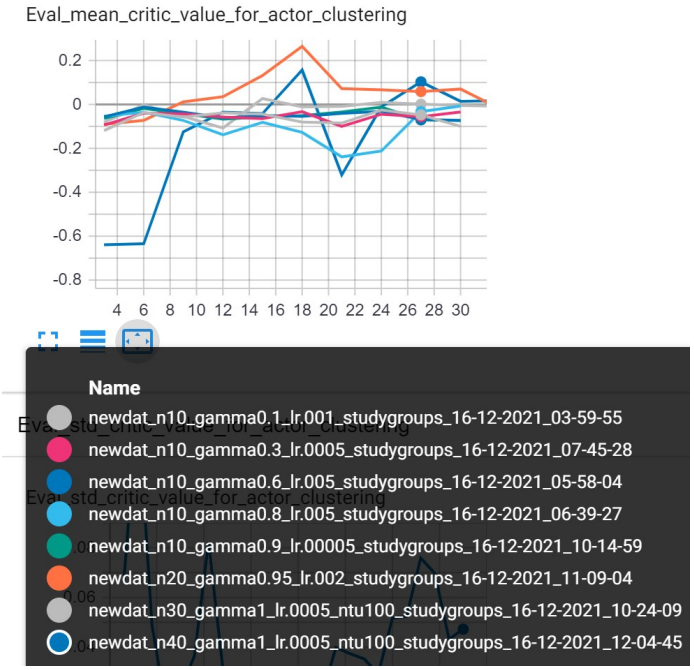



Figure 32: Final hyperparameter search. The model with $\gamma = 1$, learning rate = 5e-4, number of critic updates per AWAC update step = 100, number of actor updates per AWAC update step = 40, number of gradient steps per target update = 2 performed best.

prioritize immediate reward rather than long-term reward when forming quality study groups, for the sake of offline training, letting the actor discover rewards from a state later down the likely helps in moving quickly towards such a state.

The optimal reported hyperparameters for actor-critic training, after a fine-grain grid search (Fig.32), were at: $\gamma = 1$, learning rate = 5e-4, number of critic updates per AWAC update step = 100, number of actor updates per AWAC update step = 40, and number of gradient steps per target update = 2. This training trajectory provided unstable improvement, but overall long-term increase and positive terminal value performance after many training iterations.

Similar in performance was a slightly different model, with $\gamma = 0.95$, learning rate = 2e-3, number of critic updates per AWAC update step = 200, number of actor updates per AWAC update step = 40, and number of gradient steps per target update = 2. This model saw very consistent increase in critic-evaluated performance until a halfway point in training iterations, after which performance decreased steadily.

Below is reported a figure of averaged differently-seeded long runs given the first chosen set of optimal hyperparameters (Fig.33). Figure 33b shows a consistent decrease in AWAC loss on evaluation set data, indicating the gradual maximization in probability of higher-quality test set groups by the actor. In Figure 33a, the improvement in mean evaluation of the actor by the critic is considered to be another sign of improvement of the grouping actor through this RL-context. It corresponds to an overall decrease in critic loss over time, with the loss already starting at a

low point due to the critic having been pre-trained (Fig. 33c). The actor's loss function, the KL-divergence in clustering distributions, increases as training progresses, indicating that its loss to ensure tighter "clusters" is not being preserved over RL training. However, this increase in DEC loss is not of concern, since the other metrics are much more important indicators towards the improvement in quality of generated groups. Therefore, in combination, these figures point to an improvement in groups generated by the actor, primarily based on the critic's evaluation of the groups generated by the actor.

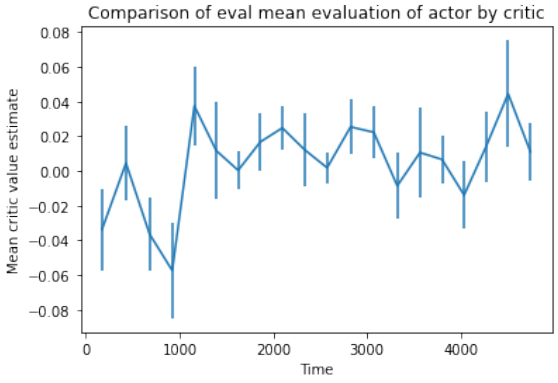
### 4.8 Conclusions and Next Steps

Although some improvement of clustering given states was demonstrated with optimal hyperparameters, learning was not very stable and had marginal improvement. Further work would require more hyperparameter optimization, as well as including a parameter to balance the actor update between AWAC loss and clustering separation loss. This would include theory work to analyze the current gradient behaviors of interchanging the update functions. Additionally, since the current critic intends to identify group quality given student features, it may be an option to share encoder architecture across the actor and critic, and have the critic add additional layers in a MTL-style architecture, with all actor-critic updates training both. Further developed methods should be tested with expanded datasets from future semesters.

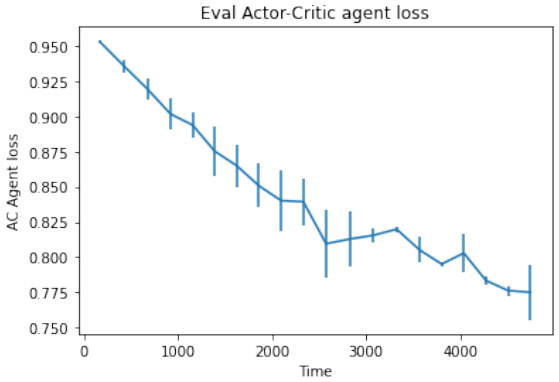
We leave a few notes of caution with using RL towards study group formation in the future:

- As noted in Section 4.5.3, the DEC actor by itself is not an appropriate clustering agent. There is some indicated improvement to its performance over the course of RL training, but any future implementation should explore alternate gradient-trainable group formation methods.

- The critic fits mainly to training data, so it is likely to be rating the actor more highly if the actor begins to assign students with similar characteristics to those previously seen together with higher probability. This overall would mean a high likelihood of this actor trying to generate groups similar to those which have been seen to work before. This is not necessarily a problem, but it does require a more rich and representative dataset, and ideally online training, to allow the model to learn to form out-of-distribution high-quality groups.

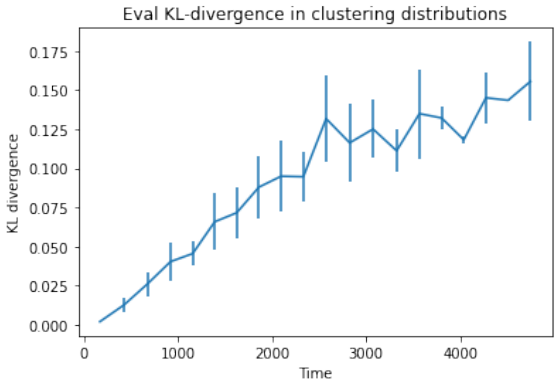



(a) Mean critic-predicted group quality of actor-generated groups, on the evaluation set. The y-axis represents mean critic value estimate per student given their assigned group, and the x-axis represents training iterations. A slight improvement over training is observed in critic-predicted group quality of actor-generated groups.




(b) AWAC agent loss value on the evaluation set, over the course of training. A steady decrease in AWAC evaluation loss occurs over the course of training.


(c) Deep net critic model loss value on the evaluation set, over the course of training. A slight decrease in critic evaluation loss occurs over the course of training, although the critic was pre-trained and therefore started at relatively low eval loss.




(d) DEC actor model loss value on the evaluation set, over the course of training. A dramatic increase in actor loss occurs over the course of training.

Figure 33: Average eval performance over combined actor-critic metrics, and individual model metrics, along the course of training.

# References


[1] A. Nair, M. Dalal, A. Gupta, and S. Levine, "Awac: Accelerating online reinforcement learning with offline datasets," *CoRR*, vol. abs/2006.09359, 2020.

[2] A. Mujkanovic and A. Bollin, "Improving learning outcomes through systematic group reformation-the role of skills and personality in software engineering education," in *2016 IEEE/ACM Cooperative and Human Aspects of Software Engineering (CHASE)*, (New York, NY), pp. 97–103, IEEE, IEEE, 2016.

[3] D. Dzvonyar, D. Henze, L. Alperowitz, and B. Bruegge, "Algorithmically supported team composition for software engineering project courses," in *2018 IEEE Global Engineering Education Conference (EDUCON)*, (New York, NY), pp. 1753–1760, IEEE, IEEE, 2018.

[4] H. H. Løvold, Y. Lindsjørn, and V. Stray, "Forming and assessing student teams in software engineering courses," in *International Conference on Agile Software Development*, (New York, NY), pp. 298–306, Springer, Springer, 2020.

[5] D. Dzvonyar, L. Alperowitz, D. Henze, and B. Bruegge, "Team composition in software engineering project courses," in *2018 IEEE/ACM International Workshop on Software Engineering Education for Millennials (SEEM)*, (New York, NY), pp. 16–23, IEEE, IEEE, 2018.

[6] C. Odo, J. Masthoff, and N. Beacham, "Group formation for collaborative learning," in *International Conference on Artificial Intelligence in Education*, (New York, NY), pp. 206–212, Springer, Springer, 2019.

[7] W. M. Cruz and S. Isotani, "Group formation algorithms in collaborative learning contexts: A systematic mapping of the literature," in *CYTED-RITOS International Workshop on Groupware*, (New York, NY), pp. 199–214, Springer, Springer, 2014.

[8] N. Maqtary, A. Mohsen, and K. Bechkoum, "Group formation techniques in computer-supported collaborative learning: A systematic literature review," *Technology, Knowledge and Learning*, vol. 24, no. 2, pp. 169–190, 2019.

[9] S. Borges, R. Mizoguchi, I. I. Bittencourt, and S. Isotani, "Group formation in CSCL: A review of the state of the art," in *Researcher Links Workshop: Higher Education for All*, (New York, NY), pp. 71–88, Springer, Springer, 2017.

[10] B. L. Putro, Y. Rosmansyah, *et al.*, "Group formation in smart learning environment: A literature review," in *2018 International Conference on Information Technology Systems and Innovation (ICITSI)*, (New York, NY), pp. 381–385, IEEE, IEEE, 2018.

[11] D. Lambić, B. Lazović, A. Djenić, and M. Marić, "A novel metaheuristic approach for collaborative learning group formation," *Journal of Computer Assisted Learning*, vol. 34, no. 6, pp. 907–916, 2018.

[12] T. Staubitz and C. Meinel, "Graded team assignments in MOOCs: Effects of team composition and further factors on team dropout rates and performance," in *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, (New York, NY, USA), pp. 1–10, Association for Computing Machinery, 2019.

[13] A. Ju, X. Fu, J. Zeitsoff, A. Hemani, Y. Dimitriadis, and A. Fox, "Scalable team-based software engineering education via automated systems," in *2018 Learning With MOOCS (LWMOOCS)*, (New York, NY), pp. 144–146, IEEE, IEEE, 2018.

[14] B. J. Zimmerman, "A social cognitive view of self-regulated academic learning.," *Journal of Educational Psychology*, vol. 81, no. 3, p. 329–339, 1989.

[15] J. Vassileva, "Toward social learning environments," *IEEE transactions on learning technologies*, vol. 1, no. 4, pp. 199–214, 2008.

[16] A. Calvó-Armengol, "Job contact networks," *Journal of economic Theory*, vol. 115, no. 1, pp. 191–206, 2004.

[17] A. Calvo-Armengol and M. O. Jackson, "The effects of social networks on employment and inequality," *American economic review*, vol. 94, no. 3, pp. 426–454, 2004.

[18] K. M. Turetsky, V. Purdie-Greenaway, J. E. Cook, J. P. Curley, and G. L. Cohen, "A psychological intervention strengthens students' peer social networks and promotes persistence in stem," *Science advances*, vol. 6, no. 45, p. eaba9221, 2020.

[19] D. T. Flynn, "Stem field persistence: The impact of engagement on postsecondary stem persistence for underrepresented minority students.," *Journal of Educational Issues*, vol. 2, no. 1, pp. 185–214, 2016.

[20] R. T. Palmer, D. C. Maramba, and T. E. Dancy, "A qualitative investigation of factors promoting the retention and persistence of students of color in stem," *The Journal of Negro Education*, vol. 80, no. 4, pp. 491–504, 2011.

[21] S. Mishra, "Social networks, social capital, social support and academic success in higher education: A systematic review with a special focus on 'underrepresented'students," *Educational Research Review*, vol. 29, p. 100307, 2020.

[22] Y. F. Niemann and J. F. Dovidio, "Relationship of solo status, academic rank, and perceived distinctiveness to job satisfaction of racial/ethnic minorities.," *Journal of Applied Psychology*, vol. 83, no. 1, p. 55, 1998.

[23] E. Spangler, M. A. Gordon, and R. M. Pipkin, "Token women: An empirical test of kanter's hypothesis," *American Journal of Sociology*, vol. 84, no. 1, pp. 160–170, 1978.

[24] C. Amelink, K. Davis, B. Ryder, and M. Ellis, "Exploring factors influencing the continued interest in a computer science major," in *2018 ASEE Annual Conference & Exposition*, (Salt Lake City, Utah), ASEE Conferences, 2018.

[25] I. J. Raabe, "Social exclusion and school achievement: Children of immigrants and children of natives in three european countries," *Child Indicators Research*, vol. 12, pp. 1003–1022, "Jun" 2019.

[26] J. Chen and D. Houser, "When are women willing to lead? the effect of team gender composition and gendered tasks," *The Leadership Quarterly*, vol. 30, no. 6, p. 101340, 2019.

[27] A. Born, E. Ranehill, and A. Sandberg, "Gender and willingness to lead: Does the gender composition of teams matter?," *Review of Economics and Statistics*, pp. 1–46, 2020.

[28] N. A. Buzzetto-More, O. Ukoha, and N. Rustagi, "Unlocking the barriers to women and minorities in computer science and information systems studies: Results from a multi-methodolical study conducted at two minority serving institutions," *Journal of Information Technology Education: Research*, vol. 9, no. 1, pp. 115–131, 2010.

[29] M. Thompson and D. Sekaquaptewa, "When being different is detrimental: Solo status and the performance of women and racial minorities," *Analyses of Social Issues and Public Policy*, vol. 2, no. 1, pp. 183–203, 2002.

[30] J. Maloy, M. B. Kwapisz, and B. E. Hughes, "Factors influencing retention of transgender and gender nonconforming students in undergraduate stem majors," *CBE—Life Sciences Education*, vol. 21, no. 1, p. ar13, 2022. PMID: 35044846.

[31] J. C. Garvey and S. R. Rankin, "Making the grade? classroom climate for lgbtq students across gender conformity," *Journal of Student Affairs Research and Practice*, vol. 52, no. 2, pp. 190–203, 2015.

[32] G. Tumushabe, "Inclusive and scalable study group formation," Master's thesis, EECS Department, University of California, Berkeley, May 2021.

[33] S. Kohli, N. Ramachandran, A. Tudor, G. Tumushabe, O. Hsu, and G. Ranade, "Inclusive study group formation at scale," 2022.

[34] S. Kohli, N. Ramachandran, A. Tudor, O. Hsu, G. Tumushabe, and G. Ranade, "16A Study Group Final Survey Evaluations Fall 2020." `https://drive.google.com/file/d/11volHEjTpMOvaxl_XYvvhqabsPg3BVyR/view?usp=sharing`, 2020.

[35] N. Diakopoulos, "Accountability in algorithmic decision making," *Commun. ACM*, vol. 59, p. 56–62, jan 2016.

[36] R. A. Layton, M. L. Loughry, M. W. Ohland, and G. D. Ricco, "Design and validation of a web-based system for assigning members to teams using instructor-specified criteria.," *Advances in Engineering Education*, vol. 2, no. 1, p. n1, 2010.

[37] J. L. Hertz, "Gruepr, a software tool for optimally partitioning students onto teams," 1 2022.

[38] G. Rasch, *Probabilistic models for some intelligence and attainment tests.* MESA Press, 1993.

[39] M. Wilson, *Constructing measures: An item response modeling approach.* Mahwah, NJ: Erlbaum, 2005.

[40] E. Hammar Chiriac, "Group work as an incentive for learning - students' experiences of group work," *Frontiers in Psychology*, vol. 5, 2014.

[41] E. Hammar Chiriac, "A scheme for understanding group processes in problem-based learning," *Higher Education*, vol. 55, no. 5, p. 505–518, 2007.

[42] J. A. Ansari and N. A. Khan, "Exploring the role of social media in collaborative learning the new domain of learning," *Smart Learning Environments*, vol. 7, no. 1, 2020.

[43] E. Baines, P. Blatchford, and A. Chowne, "Improving the effectiveness of collaborative group work in primary schools: Effects on science attainment," *British Educational Research Journal*, vol. 33, no. 5, p. 663–680, 2007.

[44] M. Reckase, "Item response theory: Parameter estimation techniques," *Applied Psychological Measurement*, vol. 22, pp. 89–91, 03 1998.

[45] D. O. Price, "Measurement and prediction. by samuel a. stouffer, louis guttman, edward a. suchman, paul f. lazarsfeld, shirley a. star, and john a. clausen studies in social psychology in world war ii, volume iv. princeton, n. j.: Princeton university press, 1950. 756 pp.," *Social Forces*, vol. 29, pp. 207–209, 12 1950.

[46] M. von Davier and J. Rost, *Polytomous Mixed Rasch Models*, pp. 371–379. New York, NY: Springer New York, 1995.

[47] L. L. Thurstone, *The measurement of Social Attitudes.* Bobbs-Merrill, 1963.

[48] D. T. Irribarra, M. Kendall, A. M. A. Reitze, and R. Freund, "Berkeley assessment system software (bass)," Mar 2021.

[49] B. D. Wright and G. N. Masters, *Rating scale analysis.* Mesa Press, 1982.

[50] W. Fisher, Jr, "Reliability statistics," *Rasch Measurement Transactions*, vol. 6, p. 238, 01 1992.

[51] R. J. Adams and S. T. Khoo, "Quest: The interactive test analysis system.," 1993.

[52] J. Xie, R. B. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," *CoRR*, vol. abs/1511.06335, 2015.

[53] E. Alfonseca, R. M. Carro, E. Martín, A. Ortigosa, and P. Paredes, "The impact of learning styles on student grouping for collaborative learning: A case study," *User Modeling and User-Adapted Interaction*, vol. 16, no. 3-4, p. 377–401, 2006.

[54] E. Bair, "Semi-supervised clustering methods," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 5, p. 349–361, Jul 2013.

[55] N. Grira, M. Crucianu, and N. Boujemaa, "Unsupervised and semi-supervised clustering: a brief survey," in *in 'A Review of Machine Learning Techniques for Processing Multimedia Content', Report of the MUSCLE European Network of Excellence (FP6*, 2004.

[56] R. Munos, T. Stepleton, A. Harutyunyan, and M. G. Bellemare, "Safe and efficient off-policy reinforcement learning," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, (Red Hook, NY, USA), p. 1054–1062, Curran Associates Inc., 2016.

[57] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative q-learning for offline reinforcement learning," 2020.

# Appendices

Appendix A: 16A Fall 2020 Feedback Form

---

# [EECS16A Fa20] Study Group Feedback

The respondent's email (**null**) was recorded on submission of this form.
* Required

1. Email *

We would like to use your feedback for improving the group-matching system for the future and the remainder of this class and we need your consent! Please check below if we may do this. Your participation will have no impact on your grade. The PI for the study is Prof. Ranade and the Protocol ID is 2020-08-13526."I consent to have anonymized information, feedback responses and scores used for research purposes so that the instructors may improve the efficacy of study groups in the future. I understand that no personally identifying information will be used. "

2. Anonymized feedback and class scores and information may be used to improve future study groups. We ask that you please consent so we can improve the experience for the next set of students! *

*Mark only one oval.*

- ◯ Yes, I consent
- ◯ No, I do not consent

3. Are you participating in a 16A study group? *

*Mark only one oval.*

- ◯ Yes
- ◯ No *Skip to question 15*

Study Group Feedback

4. Did you request your own study group members or were you assigned to a study group through the 16A matching mechanism? *

*Mark only one oval.*

- ◯ I requested my own group members
- ◯ I was assigned to a group through the 16A mechanism

5. How often does your study group meet/interact/text/chat? *

*Mark only one oval.*

- ◯ More than twice a week
- ◯ Twice a week
- ◯ Once a week
- ◯ Never

6. Do you feel everyone in your group participates in the study group? Participation can involve zoom meetings, exchanging chat/text/other messages, emails etc. *

*Mark only one oval.*

- ◯ Everyone regularly participates in the study group
- ◯ Most people participate in the study group
- ◯ Some people participate in the study group
- ◯ No one regularly participates in the study group

7. Do you think that most people in your group are comfortable sharing their ideas with the group? 1 means most people are uncomfortable, 4 means most people are comfortable. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| Many people are uncomfortable sharing their ideas | ◯ | ◯ | ◯ | ◯ | Most people are comfortable |

8. Are you comfortable sharing your ideas with the group? 1 means you are uncomfortable, 4 means you are comfortable. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| Many people are uncomfortable sharing their ideas | ◯ | ◯ | ◯ | ◯ | Most people are comfortable |

9. Do you think that most people in your group are comfortable asking questions in the group? 1 means most people are uncomfortable, 4 means most people are comfortable. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| Many people are uncomfortable sharing their ideas | ◯ | ◯ | ◯ | ◯ | Most people are comfortable |

10. Are you comfortable asking questions in the group? 1 means you are uncomfortable, 4 means you are are comfortable. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| Many people are uncomfortable sharing their ideas | ◯ | ◯ | ◯ | ◯ | Most people are comfortable |

11. Are you planning on taking future classes with people you met in your study group? *

*Mark only one oval.*

- ◯ Yes, definitely!
- ◯ I hope so, but we have not discussed it
- ◯ Not really

12. What are the best things that have resulted from your participation in the study group matching process? Any positive experiences we should know about? *

13. Was there anything you think you could have done differently this semester to have a better study group experience? *

14. Is there anything else you would like to let us know about your study group? Any negative experiences we should know about? *

No study group

15. Please let us know why you chose not to work with a study group this semester.

# Appendix B: 16B Fall 2021 Feedback Form

# Study Group Feedback

The respondent's email (**null**) was recorded on submission of this form.
* Required

1. Email *

2. First Name *

3. Last Name *

4. SID *

**We would like to use your feedback for improving the group-matching system for the future and the remainder of this class and we need your consent! Please check below if we may do this. Your participation will have no impact on your grade. The PI for the study is Prof. Ranade and the Protocol ID is 2020-08-13526.**
Anonymized feedback and class scores and information may be used to improve future study groups. We ask that you please consent so we can improve the experience for the next set of students!

5. I consent to have anonymized information, feedback responses and scores used for research purposes, regardless of any previous response, so that the instructors may improve the efficacy of study groups in the future. I understand that no personally identifying information will be used. *

*Mark only one oval.*

- ◯ Yes, I consent
- ◯ No, I do not consent

6. Did you work collaboratively with other students to complete assignments and/or study for this course? *

*Mark only one oval.*

- ◯ Yes
- ◯ No *Skip to question 14*

7. How did you form your study group? *

*Mark only one oval.*

- ◯ I was assigned to a group using the course system
- ◯ I formed my own study group and reported them to course staff in the group matching form
- ◯ I formed my own study group and did not report them in the group matching form
- ◯ I did not have or receive a fixed study group *Skip to question 14*

Study Group Feedback

Please answer the following questions about the most recent group you have been assigned to.

8. Group Interactions *

Interacting can involve meeting or talking online. Initiating study activities can involve proposing meeting to do coursework, or talking about coursework remotely.

*Mark only one oval per row.*

| | Never | A few times a month | Once a week | More than once a week |
|---|---|---|---|---|
| I interact with my study group | ◯ | ◯ | ◯ | ◯ |
| I initiate interactions with my group | ◯ | ◯ | ◯ | ◯ |
| Other group members initiate interactions | ◯ | ◯ | ◯ | ◯ |

9. Group Behavior *

*Mark only one oval per row.*

| | None | Few | Many | Most/All |
|---|---|---|---|---|
| I collaborate with my group on studying for _____ of the homeworks | | | | |
| I collaborate with my group on studying for _____ of the exams | | | | |
| _____ of the group members regularly participate in group interactions | | | | |
| Members generally respond to _____ of the meeting initiations or questions | | | | |

10. Group Experience *

*Mark only one oval per row.*

| | Strongly disagree | Disagree | Agree | Strongly agree |
|---|---|---|---|---|
| I wish I could have interacted with my group more frequently. | | | | |
| I perform better on assignments and/or exams when I collaborate with group members. | | | | |
| I feel comfortable asking questions in the group. | | | | |
| I feel comfortable with my role and contributions in this group. | | | | |
| I would like to work again with some or all of the people I met in my group, if possible. | | | | |

Any other comments?

Please feel free to add more details below about your experience this semester.

11. (Optional) What are the best things that have resulted from your participation in the study group matching process? Any positive experiences we should know about?

12. (Optional) Was there anything you think you could have done differently this semester to have a better study group experience?

13. (Optional) Is there anything else you would like to let us know about your study group? Any negative experiences we should know about?

Thank you for taking the time to answer this survey :)

No study group

14. (Optional) Please let us know why you were not able to work with a study group this semester.

# Appendix C: 16A Fall 2020 Matching Form

# Copy of 16A Group Matching Form

The respondent's email (**null**) was recorded on submission of this form.
* Required

1. Email *

2. First name *

3. Last name *

4. SID *

5. Would you like to be part of an EECS16A study group? (Answer yes even if you have an existing study group -- followup questions to come) *

   *Mark only one oval.*

   - ◯ Yes *Skip to question 6*
   - ◯ No

## Existing Study Group?

6. Do you have an existing study group of size 2-4 in mind? If you have group of 5 or more, we recommend you split into two groups of size 2 and 3 respectively. *

   *Mark only one oval.*

   - ◯ Yes *Skip to question 7*
   - ◯ No *Skip to question 10*

## Group Members

7. 2nd Group Member Berkeley Student Email (must be @berkeley.edu) *

8. 3rd Group Member Student Email (must be @berkeley.edu)

9. 4th Group Member Student Email (must be @berkeley.edu)

## Required Matching Questions

10. What is the UTC timezone offset closest to you? *

    *Mark only one oval.*

    - ◯ -7 (California/Vancouver)
    - ◯ -5 (Chicago)
    - ◯ -4 (New York/Toronto)
    - ◯ +1 (London/West Africa)
    - ◯ +8 (Hong Kong/China)
    - ◯ +9 (Tokyo)
    - ◯ +5.5 (India)

11. What year are you? *

*Mark only one oval.*

- ◯ Freshman
- ◯ Sophomore
- ◯ Junior
- ◯ Senior and above
- ◯ Transfer Student (Junior)
- ◯ Transfer Student (Senior)
- ◯ Graduate student

12. What courses have you completed (or passed out of) before this? *

*Check all that apply.*

- ☐ Math 1A
- ☐ Math 1B
- ☐ 61A
- ☐ Math 54 (Note Math 54 is not required at all for the L&S or EECS major and is not required for 16A)
- ☐ Linear Algebra course at a community college

13. What other classes are you currently taking?

*Check all that apply.*

- ☐ 61A
- ☐ 61B
- ☐ 61C
- ☐ 70
- ☐ Physics 7A
- ☐ Physics 7B
- ☐ Math 1B
- ☐ Math 54 (Note Math 54 is not required at all for the L&S or EECS major and is not required for 16A)

14. How much time are you hoping to put into 16A? *

*Mark only one oval.*

- ◯ Not very much
- ◯ A medium amount
- ◯ A significant amount

## Optional Matching Questions

If you would like to potentially improve the quality of your match, please fill out the following questions.

15. What times of the day do you prefer meeting for your study group?

*Mark only one oval.*

- ◯ Morning
- ◯ Afternoon
- ◯ Evening
- ◯ Night

16. How important is it to you that you are assigned to a group with one or more people that self-identify in terms of gender the same way as you? We will do our best to match you according to these preferences but may not always be able to.

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| Not important at all | ◯ | ◯ | ◯ | ◯ | Very important |

17. How important is it to you that your team is diverse and brings a variety of backgrounds to the group? We will do our best to match you according to these preferences but may not always be able to.

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| Not important at all | | | | | Very important |