Investigating the role of student preparation on cooperative grouping in an active learning classroom

Eric Burkholder
Department of Physics & Department of Chemical Engineering, Auburn University, 380 Duncan Drive, Auburn, AL 36849

R. Smith Strain
Department of Physics, Auburn University, 380 Duncan Drive, Auburn, AL 36849

Instructors new to active learning classrooms frequently ask how they should best structure groups in the classroom to ensure optimum learning. Groups within classrooms are complex social systems with many variables, so unfortunately there is no easy answer. Existing group-formation algorithms do not specify how groups should be structured; they only provide a way for instructors to specify their own algorithm based on factors like GPA or Gender. There are many dimensions of student thinking, motivation, and experience that may be relevant, but here we focus on one measurement that is relatively easy to measure: prior preparation. There have been some studies investigating the role of preparation in cooperative grouping, but each study seems to come to a different conclusion. Here we provide some evidence as to why that might be the case by investigating outcomes based on different measures of preparation and investigating the effects of cooperative grouping for different groups of students. We find that groups that are heterogeneous with respect to physics preparation tend to perform better. Additionally, we find that this effect is particularly pronounced for women and underrepresented students, but not for white men. This would seem to suggest that a reason for disagreements in the literature could be sensitive to how preparation is measured as well as the demographics of the study population.

2022 PERC Proceedings edited by Frank, Jones, and Ryan; Peer-reviewed, doi.org/10.1119/perc.2022.pr.Burkholder

I. INTRODUCTION

There is now substantial work showing that "active-learning," a broad term classifying teaching methods in which students are more actively engaged with material during class, is more effective than passive instruction [1]. There are many variations of active learning: some examples popular in physics include modeling instruction [2], peer instruction [3], and studio physics [4]. A common element among all these different modes of active learning is between-student interactions during class, whether that be in informal pairs (students sitting next to one another in lecture), or more formal (intentionally structured) groups. An outstanding controversy in the literature, and the subject of this paper, is how groups should be structured with respect to student preparation to ensure the best outcomes for all students – namely, should students of similar preparation levels be grouped together, or dispersed across different groups?

Group work in classrooms is often described through a socioconstructivist lens. Brookes et al. [5] describe how their interactive classroom treats knowledge as an ontological process and posits that students learn by engaging in that process [6]. This positions students as participants in a community, where students can be peripheral or central in the community, and thus they look at the kinds of participation to evaluate learning. Studies have pointed to the importance of community building rather than dividing up work [7].

The literature identifies structural and interpersonal factors which are likely to affect group performance [5]. It has been shown that task interdependence [8] [9] and reward interdependence are essential structural features for group performance. For the proposed research, this suggests that tasks in the classroom should be sufficiently difficult so as to require cooperation, and that there must be some reward (i.e., course credit) attached to group work.

A large interpersonal factor in determining group success is psychological safety [10] [11]. This is defined as"a shared belief that the team is safe for interpersonal risk taking," which "stems from mutual respect and trust among team members." Psychological safety has largely been measured by subject perceptions of group members [12], but Ref. [5] found that a sense of psychological safety and group functioning was correlated with group members not positioning themselves as experts, but rather by engaging in hedging statements and positioning themselves as intermediate experts or intermediate novices. This would seem to suggest that homogeneous groups with respect to preparation would function well, but Brookes et al. do not comment on how positioning is or is not correlated with measures of prior performance. For example, a highly-prepared student may be able to position themselves as less of an expert in some cases if they are aware of how their positioning affects others within the group.

Previous work investigating "ability-grouping" (hereafter preparation grouping) has largely taken place in K-12 classrooms. Slavin [13] conducted a survey of studies looking at the differences between homogeneous preparation grouping in math and whole-class instruction. They found that all students, regardless of preparation level, performed better in the grouped classroom. This is unsurprising as this is akin to comparing cooperative grouping to traditional instruction in the university classroom. It has been shown extensively that any inter-student interaction [14]; [3] is superior to whole-class lecture in university classrooms. This has been shown across fields and varying extents of active learning in the classroom [1].

Webb [15] conducted a study of small group learning in mathematics and reading in K-12 classrooms. She found that heterogeneous groupings with a narrow range of preparation seemed to function best (e.g. a low-preparation and medium preparation students working together). She found that high and low preparation students tended to form a teacher-student relationship when working together, which was correlated with increased performance for both low- and high-preparation students. However, the medium-preparation students were left out of this interaction, so they performed better in homogeneous groups. All studies they considered measured preparation using state standardized tests from various different locations in the United States, and there was a diverse range of students studied but the results were not disaggregated by demographics.

In contrast, a meta-analysis by Lou et el. [16] found a small effect size (Cohen's $d = 0.12$) favoring homogeneous grouping, but this homogeneous grouping favored medium- and high-preparation students over low-preparation students. However, the effect sizes across studies were not uniform, and in the studies examining math and science groups, they found that preparation-composition makes little to no difference on individual performance. Almost all studies they considered measured preparation using state standardized tests from various different locations in the United States, and there was a diverse range of students studied but the results were not disaggregated by demographics.

Heller and Hollbaugh [17] conducted a seminal study in Physics Education Research (PER) looking at cooperative grouping in the classroom – examining what types of problems were best suited for this type of learning, gender composition of groups, group size, and preparation-grouping. Through extensive qualitative analysis, they found that the optimum group size was 3 students, women function best in female-majority groups, and heterogeneous preparation groups performed better than low- and medium-preparation homogeneous groups, but equally well as high-preparation homogeneous groups. They again describe a student-teacher relationship that seems to benefit both low- and high-preparation students. They further identify common problems in groups (sometimes unrelated to preparation), such as dominance by one student or general conflict avoidance. Their solution was to have well defined roles (e.g., leader, scribe) that rotated throughout the term. They also found that there should be both individual and group-level deliverables to reward interdependence (the studies above largely considered individual metrics of performance only). In this study,

preparation was determined based on in-class exam scores (written by the instructors) and the population was predominantly white men.

More recently, Callan et al. [18] conducted a one of the largest studies in PER of cooperative grouping at the Colorado School of Mines. In this study, preparation was measured using standard physics concept inventories, and the population was a relatively high-achieving and predominantly consisted of white men. They found no difference between heterogeneous and homogeneous groups regardless of ability level or gender, though the methods they used to come to this conclusion were not clearly outlined in that paper.

These previous studies show the importance of studying the problem of preparation-grouping in greater detail because each study seems to come to a different conclusion. Across each of these studies, we felt that two salient details were missing that have been shown to have substantial impacts on student outcomes in the classroom: how preparation is measured and the demographics of the student population. The research questions we aim to answer here are:

1. Is the correlation between preparation-grouping and individual student performance sensitive to how preparation is measured?
2. Are different group compositions with respect to prior preparation more effective for different groups of students?

II. METHODS

We collected data from a single, highly interactive physics 1 course at a large public research university with a total enrollment of 57 students. The course was mostly in-person with some synchronous online sessions using breakout rooms in Zoom due to instructor illness. This course has 6 hours of contact time per week: 3 hours of "lecture" time during which students solve scaffolded, context-rich problems in groups of 6 with the guidance of an instructor and one Learning Assistant (there is assigned pre-reading to introduce basic content); one hour of recitation time which is dedicated to further problem-solving practice with a Teaching Assistant; and two hours of lab time (led by the Graduate Teaching Assistant). Note that in the in-class groups, all women were in female-majority groups in line with the recommendations of Dasgupta [19]. Each contact hour has an associated assignment which must be turned in for attendance credit, so attendance at all hours is high (roughly 90 %). During the first week of recitation, we administered a physics diagnostic survey to assess students' incoming level of physics preparation. This physics diagnostic is a validated exam that tests basic algebra and calculus concepts, as well as a range of conceptual and calculation questions covering Newton's laws, kinematics, statics, conservation of energy, and angular momentum [20]. This diagnostic was subject to both qualitative validation (think-aloud interviews with students) and quantitative validation (investigations of how well the exam and individual questions predicted course performance). There are weekly, graded homework assignments in this course, as well as three in-class exams, for which we allow quiz corrections to receive 50 % of lost points back.

The outcome variable of interest for this study is the individual quiz average score. This is a relatively high-stakes measure of individual performance in this course which should be more reflective of summative assessments in more traditionally taught courses. We use this outcome to provide the widest applicability of the results. The input variables we examined were students' individual physics diagnostic test scores, ACT scores, and first semester GPAs. We chose these variables because they are things that instructors are most likely to be able to gain access to in their teaching. In addition, we calculated the group-level variance in each of these measures and used it as an independent variable in the analysis (see below). Finally, we collected basic demographic information from students: a binary measure of gender (male/female), their first-generation status, and their race (which we aggregated into White/Asian and under-represented minority - URM). We note that these demographic variables may not reflect the rich variety of students' individual experiences, nor their individual challenges in physics. However, they are measures which are more accessible to instructors and provide a richer insight into the preparation-grouping data than had we not considered demographics at all. Furthermore, the small sample size did not allow us to investigate more detailed intersectionality because of the risk of identifying individual students.

For the first research question, we tested a linear model:

$$QuizScore = \beta_0 + \beta_1 Preparation + \beta_2 Prep.Variance, \quad (1)$$

where β_0 is the average quiz grade for a student with average preparation, β_1 is the correlation between prior preparation and quiz grade, and β_2 measures the correlation between within-group preparation variance and individual performance. We tested this model with three measures of preparation: diagnostic scores, ACT scores, and first-term GPA.

For the second research question, we extended this model to include demographic variables:

$$QuizScore = \beta_0 + \beta_1 Preparation + \beta_2 Prep.Variance + \beta_3 Demo. + \beta_4 Prep.Var \times Demo. \quad (2)$$

In Eqn. 2, β_3 measures the demographic gaps in performance while controlling for incoming preparation, and β_4 measures whether the within-group variance in preparation affects majority and minority students differently. Note that we scale all variables to be in units of standard deviations, so that the coefficients may be interpreted as effect sizes.

The demographics of this course was $\sim 20\%$ female (only binary measures of gender were provided to the instructor), $\sim 10\%$ URM and $\sim 10\%$ first generation (with relatively little overlap between minoritized populations). As mentioned

	Diagnostic	ACT	GPA
Intercept (β_0)	-0.0158 (0.147)	-0.000717 (0.138)	-0.00505 (0.132)
Prep. (β_1)	0.229 (0.149)	0.416 (0.137)	0.503 (0.139)
Prep. Var. (β_2)	0.139 (0.145)	-0.0327 (0.136)	-0.0172 (0.147)
R-squared	0.0767	0.184	0.254

TABLE I. Results of linear regression following Eqn. 1. Each column represents Eqn. 1 with the three different measures of incoming physics preparation. The standard error of each coefficient is in parentheses.

	Gender (F = 1)	URM	First Gen.
Intercept (β_0)	-0.217 (0.447)	0.0787 (0.140)	-0.0121 (0.160)
Diag. (β_1)	0.222 (0.150)	0.204 (0.137)	0.231 (0.170)
Diag. Var. (β_2)	0.108 (0.148)	-0.00991 (0.134)	0.129 (0.160)
Demo. (β_3)	-0.217 (0.474)	-1.22 (0.477)	-0.0442 (0.570)
Demo. $\times Var.(\beta_4)$	1.10 (0.873)	2.27 (0.941)	0.0858 (0.486
R-squared	0.114	0.271	0.254

TABLE II. Results of linear regression following Eqn. 2. Each column represents a different underrepresented group: women, URM students, and first-generation students.The standard error of each coefficient is in parentheses.

before, all women were in female-majority groups due to the robust prior research pointing to the benefits of such groupings. The rest of the groups were assigned randomly, which resulted in URM students and FG students all being the only person like them in that group.

III. RESULTS

The results provide the following answers to our research questions:

1. Yes, the way in which preparation is measured affects the correlation between preparation grouping and student performance.
2. Yes, heterogeneous groups are more effective for women and underrepresented students, while the grouping does not seem to matter for white and male students.

The results of the models described by Eqn. 1 are in Table 1. We do not provide p-values due to the small sample size and the recommendations of the American Statistical Association [21]. We see that all three measures of prior preparation are weakly correlated with quiz grades, though ACT score and GPA are more strongly correlated than the diagnostic score (this is by design, see Ref. [22]). However, we note that the only non-negligible effect of within-group preparation variance (effect size > 0.10) is for the diagnostic score. The results suggest that individual students may perform better when in more heterogeneous groups with respect to diagnostic scores (a proxy for prior physics preparation). It does not seem to be sensitive to more general measures of preparation (composite ACT score and GPA).

The results of the models described by Eqn. 2 are in Table 2. We only used diagnostic score as the measure of prior preparation as it was the only one for which within-group variation may be associated with the individual outcomes. Table 2 shows why those effects may have been small For men, as well as white and Asian students, the correlation between within-group preparation variation and individual performance is small (effect size < 0.15). However, for women and URM students, the correlation between within-group variation and individual performance is large (β_4), with a clear preference for groups that maximize the variation in incoming physics preparation. There does not seem to be a clear signal for first-generation students. Because most of the students in the course were white men, the overall correlation between preparation variance and performance was small.

Though these results are for a single, highly interactive classroom, some of the effect sizes are surprisingly large. Preparation grouping does not seem to matter for groups historically over-represented in STEM, but may make a substantial difference for historically marginalized groups. Women and students of color may benefit significantly from being in groups that are heterogeneous with respect to incoming physics preparation. These findings underscore the importance of this research question for equity in the classroom. Because historically marginalized students typically come in with lower levels of physics preparation [23] they stand to benefit the most from being in carefully designed groups.

IV. DISCUSSION

In this paper, we describe observational data related to within group variation in physics preparation in a physics classroom and how it is related to individual outcomes. We wish to emphasize that this is observational data from a small number of students and is not sufficient to draw policy recommendations from. It serves primarily to guide instructors and researchers to provide more robust data on the subject.

We found that the variability in preparation only mattered with respect to physics preparation, not more general measures of preparation. We expect that this is because the verbal and study skills measured by GPA and ACT composite scores may be less relevant to students when they are working on physics problems. Indeed, there is substantial evidence showing that different contexts activate different types of thinking in students [24]. There is also evidence that how students define "intelligence" is context dependent [25], so it is possible that students separate each others' "general intelligence" from their ability to do physics. Further interviews with students should confirm this. We hypothesize that students perceived one another to be on approximately equal footing intellectually due to the relatively homogeneous class composition, but may have acknolwedged different levels of physics preparation and been more open to hearing ideas and forming

student-teacher type relationships.

When we investigated the role of demographics in preparation-grouping, we found that women and URM students performed substantially better in heterogeneous groups with respect to physics preparation, while preparation variation did not matter for those historically over-represented in STEM. In general, the finding that heterogeneous groups perform better supports the "student-teacher" type of relationship described in previous literature [17]. We emphasize the importance that the only dimension of heterogeneity we investigated was preparation, not demographic composition. Indeed, in the groups were URM students were all the only one like them, we see that the average quiz scores are much lower for URM students in homogeneous groups. This may suggest that, beyond heterogeneous preparation grouping, URM students should not be isolated in individual groups.

In a larger sample, we would ideally see if preparation variation mattered for the least-prepared students and the most-prepared students differently. This could be done quantitatively with an additional interaction term in the model, but would require a sample size of at least 200 students to investigate with some reasonable statistical power.

In future investigations, this quantitative analysis should be paired with qualitative analysis, including videos of student interactions and interviews with various different students – particularly women and URM students – to understand what aspects of how their group functioned they felt seemed to best support their learning. For example, is the student-teacher relationship that is hypothesized always helpful? Or can it seem to be condescending under certain conditions. This is a rich area of future research that will certainly be of interest to instructors and physics education researchers.

V. CONCLUSIONS

The results of this study provide two data-driven suggestions to physics instructors as to how to ensure optimal group function within their classroom. First, take care to measure students' incoming physics preparation and place students in groups with wide variation in physics preparation. Second, take care to ensure that group norms are well established, and that students are engaging in productive and open dialogues in the classroom.

How applicable these findings will be to different groups of students and in different contexts remains to be seen. We have a study currently underway investigating preparation-grouping in different subjects and levels of active-learning STEM classrooms, which will provide more statistical power to analyze the results. We encourage more instructors who use cooperative group problem-solving in class to do these kinds of investigations, and to pay careful attention to how grouping affects different groups of students differently - particularly at institutions that are more diverse than Auburn. This will ultimately help to promote a more equitable and effective learning environment.

[1] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, "Active learning increases student performance in science, engineering, and mathematics," PNAS **111** (2014).

[2] D. Hesented, "Toward a modeling theory of physics instruction," American Journal of Physics **55**, 440–454 (1987).

[3] C. H. Crouch and E. Mazur, "Peer instruction: Ten years of experience and results," American Journal of Physics **69**, 970–77 (2001).

[4] K. Cummings, R. Thornton, and D. Kuhl, "Evaluating innovation in studio physics," American Journal of Physics **67** (1999).

[5] D. T. Brookes, Y Yang, and B. Nainabast, "Social positioning in small group interactions in an investigative science learning environment physics class," Physical Review Physics Education Research **17**, 010103 (2021).

[6] E. Etkina and A. Van Hevelen, in *Research-Based Reform of University Physics*, edited by E. F. Redish and P. J. Cooney (2007).

[7] D. Doucette and C. Singh, "Share it, don't split it: Can equitable group work improve student outcomes?" The Physics Teacher **60**, 166 (2022).

[8] R. Wageman, "Interdependence and group effectiveness," Administrative Science Quarterly **40**, 145 (1995).

[9] D. Mesch, D. W. Johnson, and R. Johnson, "Impact of positive interdependence and academic group contingencies on achievement," Journal of Social Psychology **128**, 345 (1988).

[10] A. Edmondson, "Psychological safety and learning behavior in work teams," Administrative Science Quarterly **44**, 350 (1999).

[11] J. Rozovsky, "The five keys to a successful google team," (2015).

[12] P. Van den Bossche, W. H. Gijselaers, M. Segers, and P. A. Kirschner, "Social and cognitive factors driving teamwork in collaborative learning environments: Team learning beliefs and behaviors," Small Group Research **37**, 490 (2006).

[13] R. E. Slavin, "Ability grouping and student achievement in elementary schools: A best-evidence synthesis," Review of Educational Research **57**, 293–336 (1987).

[14] L. Deslauriers, E. Schelew, and C. Wieman, "Improved learning in a large-enrollment physics class," Science **332**, 6031 (2011).

[15] N. M. Webb, "Task-related verbal interaction and mathematics learning in small groups," Journal for Research in Mathematics Education **22**, 336–89 (1991).

[16] Y. Lou, P. C. Abrami, Spence J. C., C. Poulsen, B. Chambers, and S d'Apollonia, "Within-class grouping: A meta-analysis," Review of Educational Research **66**, 423–58 (1996).

[17] P. Heller and M. Hollabaugh, "Teaching problem solving through cooperative grouping part 2: Designing problems and structuring groups," American Journal of Physics **60**, 637–644 (1992).

[18] K. E. Callan, B.R. Wilcox, and W. K. Adams, "Testing group composition within the studio learning environment," in

Physics Education Research Conference Proceedings (2017).

[19] N. Dasgupta, M. M. Scircle, and M. Hunsinger, "Female peers in small work groups enhance women's motivation, verbal participation, and career aspirations in engineering," PNAS **112** (2015).

[20] E. Burkholder, K. Wang, and C. Wieman, "validated diagnostic test for introductory physics course placement," Physical Review Physics Education Research **17**, 010127 (2021).

[21] *American Statistical Association releases statement of statistical significance and p-values: Provides principles to improve the conduct and interpretation of quantitative science* (ASA News, 2016).

[22] E. W. Burkholder, S. Salehi, S. Sackeyfio, N. Mohamed-Hinds, and C. Wieman, "An equitable and effective approach to introductory mechanics," (2021).

[23] S. Salehi, S. Cotner, and C. J. Ballen, "Variation in incoming academic preparation: Consequences for minority and first-generation students," Frontiers in Education **5**, 1–14 (2020).

[24] A. Elby and D. Hammer, "Epistemological resources and framing: A cognitive framework for helping teachers interpret and respond to their students' epistemologies," in *Personal epistemology in the classroom: Theory, research, and implications for practice*, edited by L. D. Bendixen and F. C. Feucht (Cambridge University Press, 2010).

[25] L. B. Limeri, J. Choe, H. G. Harper, H. R. Martin, A. Benton, and E. L. Dolan, "Knowledge or abilities? how undergraduates define intelligence," CBE Life Sciences Education **19** (2020).