Theories and patterns of productive group interaction in two-stage introductory physics exams

William D. Block[1] and Colin G. West[1]
[1]*Department of Physics, University of Colorado, Boulder, Colorado 80309, USA*

"Two-stage" exam formats, in which students take an exam and then repeat portions in a small-group setting, have been studied for years as a method to bring interactivity and peer-instruction into the classroom. Despite evidence of educational benefits, they have not yet seen widespread adoption at the scale of other active-learning interventions, perhaps in part because the mechanisms behind the potential benefits have been less thoroughly explored. We propose a theory of successful group exams in which students advocate for their chosen answers and deliberate thoughtfully among the options presented, without resorting to strategies like majority voting or deference to a single seemingly-strong student. By categorizing decision patterns across a sample of 59 group exam questions we provide statistical evidence that group exams can indeed support these kinds of pedagogically valuable interactions, and we examine the ways that different exam questions can influence productive or counterproductive interaction types.

2025 PERC Proceedings edited by Pawl, Zwolak, and Leak; Peer-reviewed, doi.org/10.1119/perc.2025.pr.Block

I. INTRODUCTION

Two-stage group exam formats, in which students first attempt an exam individually and then again in collaboration with a small group, are an increasingly-popular educational intervention designed to increase interactivity and peer-instruction, both of which are known to be valuable in STEM education. Current evidence is generally positive about the effects in affective domains [1, 2], and mixed-to-positive when it comes to effects on knowledge retention [3–5]. At a minimum, the technique shows promise for improving student learning and educational experience when done correctly, which raises the important question of how instructors can best deploy the technique to promote such benefits.

We hypothesize that much of the power of a two-stage exam format comes from cases where students make affirmative arguments for the answers they personally believe are correct (“advocacy"), or weigh the merits of competing positions presented by others (“deliberation"). Hence, in this work we explore two related questions: (RQ1) whether the two-stage format used here successfully created space for meaningful advocacy and deliberation and, (RQ2) whether certain features of exam questions like item difficulty and discrimination score enhance or inhibit such interactions.

II. BACKGROUND

Two-stage exams

Documented discussions of exam formats which include both individual and group components date back to at least the 1990s, when they were sometimes called “pyramid exams,” [6], and a modest body of literature exploring their impacts has emerged over the ensuing decades [1–5, 7–9]. In particular, two-stage examinations with group components were explicitly put forward in the field of physics education [10] as part of a broader package of reforms promoting active learning and peer interaction in the classroom [11], and may be particularly suitable in that context, since physics is a discipline where the benefits of peer-instructional activities have been powerfully demonstrated [12, 13].

Nevertheless, it is also known that variations in context and implementation can dramatically alter the students' experience of a given active-learning intervention [14]. Deciding how best to deploy such interventions in turn requires considering the possible mechanisms by which the benefits are intended to arise.

The role of advocacy and deliberation

One can imagine a two-stage exam which superficially raises test scores yet yields no additional learning gains. For example, groups might simply choose whichever answer a majority had selected during the individual stage. Alternatively, they might simply defer to the choice of some seemingly “strong” student in the group. We conjecture that for a group exam experience to unlock its potential, students must avoid these strategies, in favor of substantive discussion on the merits.

An anecdote from our own classroom may help demonstrate the plausibly high value of such interaction in this context. In a recent semester of an introductory mechanics course, students had confronted a classic exam problem that involved predicting the “hang time” of various projectiles. After giving their individual answers in the first stage of the exam, four students were deliberating during the group stage. Three of the students were advocating an incorrect answer, claiming to recall an in-class demonstration that would support their position (in fact, it is a documented problem in physics education that students will incorrectly remember the result of a class demonstration in order to harmonize it with their potentially-faulty intuitions! [15]). The fourth student objected that they recalled the experiment differently. At first, each student merely insisted on their recollection of events. Ultimately, however, the single student advocating for the correct answer provided a mathematical derivation which the others could not deny. After the exam, the students in this group were so eager to reconcile their conflicting recollections that they insisted on re-creating the experiment together in the lecture hall before going home, and were thrilled when the empirical result proved that they had made the right decision.

In this work, we investigate whether group exams can regularly create the conditions for this kind of earnest, meaningful discussion to emerge, including cases where minds are changed by correct scientific arguments. Then, we explore what sort of questions are most effective in promoting such discussions.

Course context

Our study took place in the second semester of a relatively large (N 190) introductory physics for life sciences course (“IPLS”) taught without calculus at a large R-1 university in the Western United States. Students in the course were evaluated on a combination of homework, lab work, and exams, with each of three midterms contributing 12 % of their final grade and a comprehensive final exam contributing 24 %). On a midterm exam, students were given 80 minutes to complete 20 multiple-choice problems, and then 40 minutes, working in groups of two-to-five students, to revisit the same 20 problems. These groups were not preconstructed but were left to the students to determine.

Students were told that, if their overall group score was better than their individual score, their final exam score would be entered as a weighted average of the two (85% individual and 15% group). In the event that a group's score was lower than an individual member's score, the individual score would stand. In practice, this policy gave license to roughly 10% of students who opted to skip the group stage.

III. METHODOLOGY

Our approach was inspired by the work of Martin [16] who extensively studied the nature of collaboration in group exams among college-level students in statistics and biology. Martin's work provides a statistically-powerful framework for asking exploratory questions. We ultimately deployed a considerably simplified version of the approach tailored to our research questions Building on Martin's finding that question-to-question variation can influence what happens in the group discussion [16], we set out to explore the scope and nature of that impact.

To begin, we collected data on every student's individual exam answers and every group's collective answer across 59 different questions (twenty from each midterm, less one question that was dropped because of a typo). For each instance of group interaction, we then classified the outcomes into one of four categories identified in Ref. [16]:

- **Productive Collaboration:** Although a majority of individuals got the question wrong, the group got the question correct. Generally, we assume this meant that very meaningful deliberation took place, since a minority viewpoint correctly won the day over an initially more popular perspective.
- **Consensus-Building:** A majority of individuals already had the question right, and the group got the question right. This outcome could be consistent with meaningful discussion but could also be indicative of blind majority voting.
- **Misunderstanding:** A majority of individuals had the question wrong, and the group proceeded to get the question wrong. Such cases may simply arise when no one in the group understood the correct answer. Alternatively, these could arise from simplistic majority voting.
- **Counterproductive Collaboration:** A majority of individuals got the question right, but ultimately the group got the question wrong. This (highly undesirable!) outcome could result from an insufficiently deliberative process (i.e. deferring to a perceived "strong" student who is in fact wrong)

Note that none of these categories alone is "best" in our advocacy-and-deliberation model of the group exam process. For example, while "productive collaboration" (correct minority wins over incorrect majority) can be a powerful indicator of collective intellectual development, it is also what we would expect in weak groups that repeatedly defer to a single strong student. Hence, one might hope to see a mixture incorporating both "productive collaboration" and "consensus building" (correct majority convinces incorrect minority). That combination suggests groups that sought correct answers regardless of whether that meant listening to one voice or many. A reasonably healthy group exam might result in consensus-building most of the time on easier problems, with some instances of productive collaboration on more challenging items, and few cases of misunderstanding or counterproductive collaboration.

For each question, we code the resulting interaction category for each of the groups. The result is that for each question we can show the percentage of each interaction type among the groups and qualitatively assess the "health" of the resulting mixture, in answer to RQ1.

Additionally, we can plot the percentage of each interaction type as a function of various properties of the exam questions, to examine the ways that variations in the questions affect the group exam. We choose two simple metrics to investigate for this: "difficulty" and "discrimination".

"**Difficulty**," is simply the fraction of students who got the question wrong during the individual exam. In theory this number could range from 0 to 100; on our exams, the range was roughly zero to 60. Naturally, a question which is so simple that nearly everyone gets it right does not create space for productive collaboration or consensus-building, because there is no one left to persuade. On the other hand, it would not be surprising if questions that are *too* difficult choke off all possibility of useful deliberation as well, if there are too few students in the groups who can advocate for the correct answer. "**Discrimination**" is, roughly, a measure of whether the exam item is assessing what the instructor thinks they are assessing. A high discrimination question is one which strong students are more likely to get right and weak students are more likely to get wrong (contrasted for example with an item that weak students could get right for the wrong reasons or which stronger students may overthink). There are a variety of statistical metrics available for this purpose including the discrimination index, the biserial correlation coefficient, and the point biserial correlation coefficient [17–20]. The differences between these can be subtly important in some contexts [21], but for a correlational study such as ours they are expected to be closely related [22]. We choose the point biserial correlation coefficient [20] in part because of its prevalence in commercial online learning systems.

This coefficient gives a correlation between a nominal, dichotomous variable (here "did the student get this item correct individually?") and a continuous variable in ratio scale (here "student's total score on the exam") and is calculated as

$$r_{pb} = \frac{\mu_1 - \mu_0}{\sigma_p}\sqrt{p(1-p)} \tag{1}$$

where μ_1 is the average total exam score of all students who answered the question correctly, μ_0 is the average total exam score of all students who answered the question incorrectly, σ_p is the population standard deviation of the total exam scores, and p is the percentage of students who answered the question correctly [20].

With difficulty and discrimination data for each question, we can then examine how they correlate with each of our four interaction modes, addressing RQ2.

IV. RESULTS AND DISCUSSION

First, we qualitatively assess the overall landscape of interaction categories by plotting the occurrence of each interaction type from question to question (see Fig. 1). In general, we see that the interaction modes are dominated by

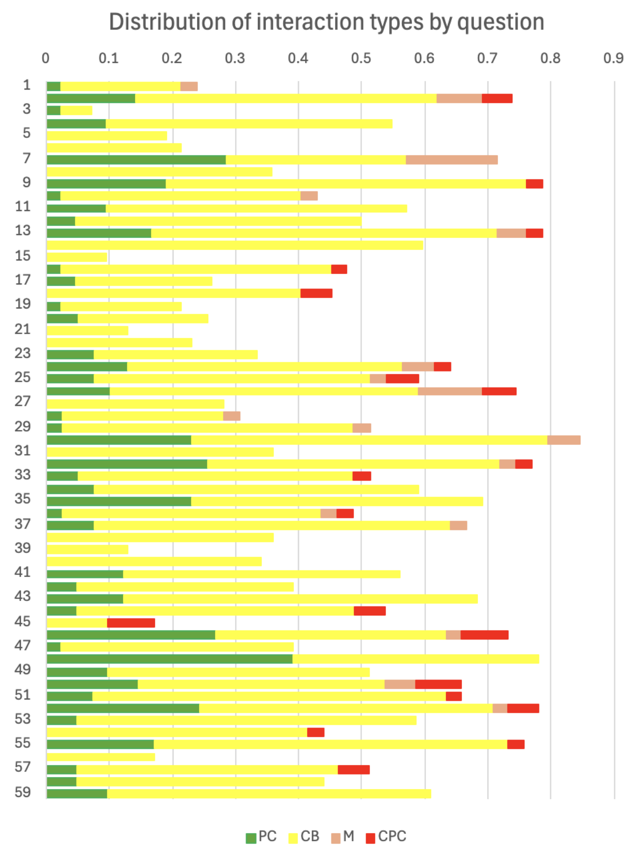

FIG. 1. (Color online). For each of 59 questions studied here (vertical axis), we plot the incidence of Productive Collaboration ("PC," green), Consensus-Building ("CB," yellow), Misunderstanding ("M," pink) and Counterproductive Collaboration ("CPC," red) in a stacked horizontal bar. Bars do not sum to 1 because of instances where no interaction necessarily took place (typically, because all students got the question right individually and also in the group exam).

consensus-building, which overall occurs 38% of the time. On its own, this could either be explained by healthy deliberation or by students defaulting to brute majority vote. However, the second most common form of interaction is productive collaboration, occurring 8% of the time, compared to only 1% each for misunderstanding and counterproductive collaboration. We take this to be a very positive sign for the health of the interaction landscape. Note: in the remaining 52% of cases, all students were initially correct and remained correct, so we have no "interaction" to classify.

A group-by-group analysis also showed a mixture of positive interaction types from question to question within an individual group, further supporting the conclusion that groups were not consistently defaulting to majority vote or other blanket strategies devoid of true discussion. Consider for example the categories of "consensus-building" and "misunderstanding. Since these two interaction types both involve majorities convincing the minority, if certain groups were frequently making their group exam decisions by a simple majority vote, then one would expect the rate at which a particular group engaged in consensus-building to correlated with the rate at which they engaged in misunderstanding. However, we found virtually no such correlation between the two ($R = 0.0008$). We similarly found no correlation between "productive collaboration" and "counterproductive collaboration", which might have signaled frequent deference to an individual, though these data were noisier as these particular outcomes were much more rare.

Of course, without direct observation or retrospective interviews, we do not know what happened in any *specific* group on any particular question. However, our evidence is intended to address RQ1 *in the aggregate* and suggests that the group exam format did create space favoring advocacy and deliberation, with no clear indications to the contrary.

Next, we examine trends between interaction types and the attributes of individual questions as measured by their difficulty (Fig. 2) and their discrimination (Fig. 3). We find first that productive collaboration displays a strong, linear correlation ($R = 0.86$) with item difficulty but that consensus-building, while also showing a tight relationship with difficulty, appears to peak around a difficulty of roughly 35 (that is, 35% of students getting the question incorrect on the individual exam). By contrast, the two negative interaction types are much more weakly correlated and with a much shallower slope. Of course, as questions get harder, it is reasonable that there are fewer groups able to form consensus around correct majorities. What is striking is that, as correct consensus-building declines, it is not replaced by *incorrect* consensus building ("misunderstanding") but rather by productive collaboration (compare the lower-left plot of Fig 2, in which misunderstanding is roughly flat as difficulty exceeds 35, and the upper-left plot of Fig. 2, in which productive collaboration continues to increase in that range).

Rates of productive collaboration and consensus-building are also correlated with item discrimination scores, though less strongly so ($R = 0.28$ and $R = 0.53$, respectively). In this case, both appear to increase across the range of discrimination scores in our sample, while negative interaction types are flat. We read this to suggest that when questions have clearer answers, discernible to strong students, those answers become more likely to win over the group. If the students' decision process was driven more by social dynamics or other spurious factors, we would not expect the outcomes to be influenced by the "quality" of the question at hand.

V. CONCLUSIONS AND OUTLOOK

If students in a group exam routinely deferred to a perceived group leader or to the unexamined majority consensus, we might expect a balanced mix of all four interactions types. One might also expect that counterproductive interactions would become increasingly common on difficult questions as the students' shallow decision methods began to fail,

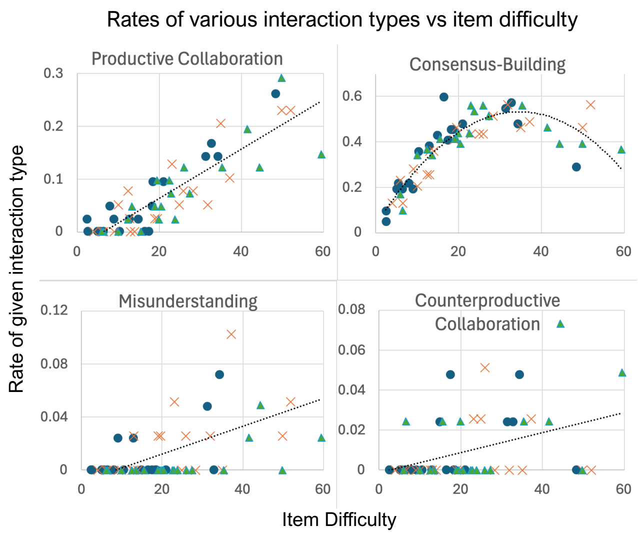

FIG. 2. (Color online). For each of the four group interaction types, the occurrence rate of that interaction type (vertical axes) is shown vs item difficulty (horizontal axes). Data from midterm 1 shown as circles, midterm 2 as crosses, and midterm 3 as triangles. Trend lines are linear regressions except for "consensus building," which is a quadratic fit.

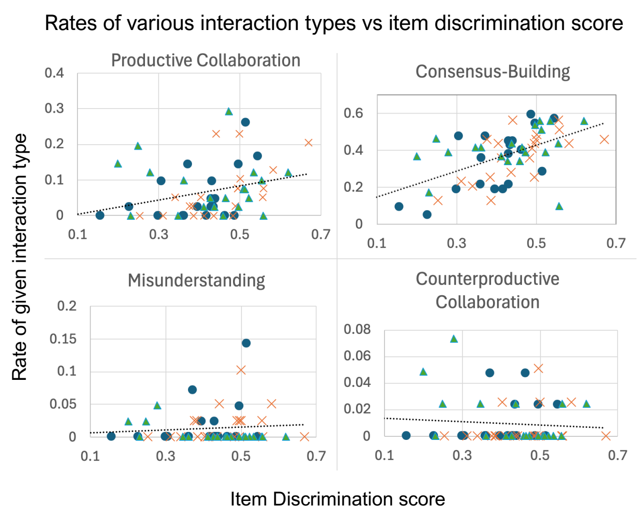

FIG. 3. (Color online). For each of the four group interaction types, the occurrence rate of that interaction type (vertical axes) is shown vs discrimination scores (horizontal axes). Data from midterm 1 shown as circles, midterm 2 as crosses, and midterm 3 as triangles. Trend lines are linear regressions.

and that group consensus would dominate regardless of the difficulty or discrimination of the question at hand. We do not observe any of this, instead finding evidence of students engaged in meaningful debate and tending often to settle on the correct answer even if it was not initially chosen by a majority of students. These observations in turn support the theory of "advocacy and deliberation" as a mechanism by which two-stage group exams can benefit students.

We speculate briefly about which features of the group exam design studied here might be important in producing these positive results. Certainly, it is essential that the format forces all the students in a group to come to a single answer, rather than letting them simply discuss as a group but then choose again individually. Presenting the students with the same (or very similar) questions on both the individual and group exam also seems likely to contribute to higher-level discourse and deliberation. More unclear is the effect of allowing students to choose their own groups; one could imagine self-selected groups encouraging deference to the group leader, but it is also plausible that the comfort of a self-selected group improves the quality of discussion.

These findings also have salience for instructors who wish to use group exam formats but without repeating the entire exam, since both item difficulty and item discrimination appear useful in identifying questions that will foster productive collaboration and consensus-building. For example, in the course studied here, we also used a two-stage structure in the final exam, but in that case, offered only ten of the thirty-two final exam questions on the group stage. These ten questions were chosen primarily on the basis of projected difficulty and discrimination score (using a combination of data from past exams and instructor intuition as a guide). Qualitatively, we also chose questions which hinged on conceptual points rather than complex calculations, or which presented clash between intuition and formal results. We observed even higher rates of productive collaboration and consensus building on the final compared to the midterms.

There are several clear limitations to this study. Chiefly, we are drawing conclusions based only on high-level statistical inference; a much stronger picture of the dynamics could be drawn by coupling this kind of analysis with direct observations and targeted interviews, which is the subject of an ongoing work. Deeper statistical insights might also be gained about the factors that support a positive group exam experience from a regression analysis of the sort performed in Ref [16], weighing factors like group construction or question content. Of course we note also that factors beyond the nature of the exam itself (student demographics, classroom culture, instructor effects, etc) may play a significant role in creating the results we see here. Nevertheless, these positive findings should encourage others to consider adopting group exam formats in their classrooms and to perform additional research into the nature and mechanisms of their effects.

ACKNOWLEDGMENTS

The authors are grateful to Noah Finkelstein and Charles Rogers for introducing them to the group exam concept and to Andrew Martin for fruitful discussion of his prior work.

[1] B. P. Rempel, M. B. Dirks, and E. G. McGinitie, Two-stage testing reduces student-perceived exam anxiety in introductory chemistry, Journal of Chemical Education **98**, 2527 (2021).

[2] K. A. Fournier, J. Couret, J. B. Ramsay, and J. L. Caulkins, Using collaborative two-stage examinations to address test anxiety in a large enrollment gateway course, Anatomical sciences education **10**, 409 (2017).

[3] J. Ives, Measuring the learning from two-stage collaborative group exams, arXiv preprint arXiv:1407.6442 (2014).

[4] G. Kinnear, Two-stage collaborative exams have little impact on subsequent exam performance in undergraduate mathematics, International Journal of Research in Undergraduate Mathematics Education **7**, 33 (2021).

[5] R. N. Cortright, H. L. Collins, D. W. Rodenbaugh, and S. E. DiCarlo, Student retention of course content is improved by collaborative-group testing, Advances in Physiology Education **27**, 102 (2003).

[6] D. Cohen and J. Henle, The pyramid exam, in *FASEB JOURNAL*, Vol. 12 (FEDERATION AMER SOC EXP BIOL 9650 ROCKVILLE PIKE, BETHESDA, MD 20814-3998 USA, 1998) pp. A1321–A1321.

[7] P. G. Zimbardo, L. D. Butler, and V. A. Wolfe, Cooperative college examinations: More gain, less pain when students share information and grades, The Journal of Experimental Education **71**, 101 (2003).

[8] T. R. C. Lee, M. Pye, O. Lilje, H. D. Nguyen, S. Hockey, M. De Bruyn, and F. T. Van den Berg, Two-stage examinations in stem: A narrative literature review, International Journal of Innovation in Science and Mathematics Education **30** (2022).

[9] T. R. Lee, M. Pye, O. Lilje, H. D. Nguyen, S. Hockey, M. de Bruyn, and F. T. van den Berg, Two-staging a comeback: A review of two-stage exams from 1996 to 2022, in *Proceedings of The Australian Conference on Science and Mathematics Education* (2023) pp. 47–47.

[10] C. E. Wieman, G. W. Rieger, and C. E. Heiner, Physics exams that promote collaborative learning, The physics teacher **52**, 51 (2014).

[11] C. Wieman, *Improving how universities teach science: Lessons from the science education initiative* (Harvard University Press, 2017).

[12] C. H. Crouch and E. Mazur, Peer instruction: Ten years of experience and results, American journal of physics **69**, 970 (2001).

[13] P. Zhang, L. Ding, and E. Mazur, Peer instruction in introductory physics: A method to bring about positive changes in students' attitudes and beliefs, Physical Review Physics Education Research **13**, 010104 (2017).

[14] C. Turpen and N. D. Finkelstein, Not all interactive engagement is the same: variations in physics professors' implementation of peer instruction, Physical Review Special Topics—Physics Education Research **5**, 020101 (2009).

[15] K. Miller, N. Lasry, K. Chu, and E. Mazur, Role of physics lecture demonstrations in conceptual learning, Physical Review Special Topics—Physics Education Research **9**, 020113 (2013).

[16] A. P. Martin, A quantitative framework for the analysis of two-stage exams., International Journal of Higher Education **7**, 33 (2018).

[17] R. M. Thorndike, G. K. Cunningham, R. L. Thorndike, and E. P. Hagen, *Measurement and evaluation in psychology and education* (Macmillan Publishing Co, Inc, 1991).

[18] S. Henrysson, Gathering, analyzing, and using data on test items, Educational measurement **2**, 130 (1971).

[19] S. Matlock-Hetzel, Basic concepts in item and test analysis., (1997).

[20] V. LeBlanc and M. Cox, Interpretation of the point-biserial correlation coefficient in the context of a school examination, The Quantitative Methods for Psychology **13**, 46 (2017).

[21] Y. Attali and T. Fraenkel, The point-biserial as a discrimination index for distractors in multiple-choice items: Deficiencies in usage and an alternative, Journal of Educational Measurement **37**, 77 (2000).

[22] L. M. Aleamoni and R. E. Spencer, A comparison of biserial discrimination, point biserial discrimination, and difficulty indices in item analysis data, Educational and Psychological Measurement **29**, 353 (1969).