A Quantitative Framework for the Analysis of Two-Stage Exams

Andrew P. Martin[1]

[1] Department of Ecology and Evolutionary Biology, University of Colorado, Boulder CO, USA

Correspondence: Andrew Martin, Department of Ecology and Evolutionary Biology, University of Colorado, Boulder CO, 80309 USA

Received: July 7, 2018 Accepted: July 19, 2018 Online Published: August 2, 2018

doi:10.5430/ijhe.v7n4p33 URL: https://doi.org/10.5430/ijhe.v7n4p33

Abstract

Two-stage exams have gained traction in education as a means of creating collaborative active-learning experiences in the classroom in a manner that advances learning, positively increases student engagement, and reduces test anxiety. Published analyses have focused almost exclusively on the increase in student scores from the first individual stage to the second collaboration stage and have shown clear positive effects on gains in student scores. Missing from these analyses is a comprehensive evaluation of the effects of individual preparation, the characteristics of questions, and small group composition on the outcomes two-stage exams. I developed a simple quantitative framework that provides a flexible approach for estimating and evaluating the effects of individuals, questions, and groups on student performance. Additionally, the framework yields statistics appropriate for making inferences about productive collaboration, consensus-building, and counter-productive interaction that happens within small groups. Analyses of 12 exams across two courses and 2 years using the quantitative framework revealed considerable variation for all three of these effects within and among exams. Overall, the results highlight the value of quantitative estimation of two-stage exams for gaining perspective on the effects of individuals, questions, and groups on student performance, and facilitates data-driven revision of assessments, curricula, and teaching strategies towards achieving gains in students' collaborative skills.

Keywords: two-stage exams, collaboration, active learning, quantitative framework

1. Introduction

Collaborative learning in small groups is a common strategy associated with active learning in education (see Cohen 1994; Prince 2004). There are positive and negative outcomes of students working in small groups that depend, in large part, on the behavioral dynamics composition of small groups (Dohlmans and Schmidt 2006; Eddy et al. 2015; Grunspan et al. 2016; Buchenroth-Martin et al. 2017), preparation for engaging in productive interaction (Dohlmans et al. 2001), and the cognitive challenge of prompts designed to elicit student engagement in learning (Crowe et al. 2017), among other factors. It is clear from the literature collaborative work within small groups enables a rich and sometimes productive learning environment that can achieve greater learning than possible in the absence of student interactions (Springer et al. 1999).

One strategy that has gained traction for leverage the power of collaborative learning is two-stage exams. Two-stage exams—also called collaborative exams—involve students completing an exam individually (stage 1) followed by small groups of students working together to complete the same—or a very similar—exam a second time (stage 2) (Stearns 1996; Zipp 2007; Wieman et al. 2014). Two-stage exams have been effective for fostering productive engagement, presumably because the exam format incentivizes students to listen to their peers and discover competing rationales for differences in peer answers across diverse questions (Bruno et al. 2017; Sandahl 2010; Wieman et al. 2014). Moreover, collaborative exams have been reported to improve student performance on subsequent individual exams and enhance student learning across low, middle, and high achiever categories (Gilley and Clarkston 2014). However, there is also evidence two-stage exams do not lead to retention on content knowledge (Leight et al. 2012).

Typically, two-stage exam data are analyzed by comparing the proportions of students with a correct answer for the individual and group portions of the exam as a means of estimating student gains (Cortright et al. 2003; Zimbardo et al. 2003; Gilley and Clarkston 2014; Fengler and Ostafichuk 2015; Bruno et al. 2017; Jang et al. 2017). Additionally, several studies revealed two-stage exams had positive effects on student engagement and perceptions of assessments (Cortright et al. 2003; Zimbardo et al. 2003; Reiger and Heiner 2014; Weiman et al. 2014; Bruno et al. 2017). Yet,

two-stage exams are explicitly designed to foster collaborative learning; consequently, there are multiple factors that influence the performance of students that remain hidden from simple summaries of the proportion of correct answers and surveys of student perceptions. Missing from the published studies is a more formal quantitative framework for assessing the variation in estimates of collaboration across questions, groups, and exams.

Here I use the results from two-stage exams for two courses across two years to introduce a quantitative framework for data analysis. The prevailing perspective from published studies of two-stage exams is that most, or all, of the estimated effects of two-stage stem from the effect of individual performance (stage 1 results). The purpose of my investigation was to estimate the magnitude of the variation in the performance of students for three important aspects of two stage exams: the preparation of individuals within a group (the effect of individuals during stage 1), the particular challenge posed by questions (the effect of questions), and the collaboration of individuals within small groups (the effect of groups). Although I was not motivated to test a particular hypothesis, my prediction, based on previous studies, is that I would see consistent and large effects of individuals on performance for two stage exams with neglible effects of question and groups. I was interested in describing variation for these important aspects of active learning that influence student learning with the goal of estimating and evaluating aspects of student interactions that happens during the engaging active learning opportunities afforded by two-stage exams. Moreover, the results should prove valuable for data-driven revision of curriculum, assessments, and teaching strategies aimed at leveraging the power of peer instruction.

2. Methods

2.1 Participant Characteristics

The institutional review board indicated the work was exempt from human subjects' research. Participants in the study included all students in two courses across two years (2016-2017) and included two midterms and a final exam in each of the courses for each year for a total of 12 exams. One course was an introductory statistics course (abbreviated Stats) and the other an upper-division course on evolutionary biology (abbreviated Evol). Both courses implemented student-centered curricula, a claim supported, in part, by COPUS (Smith et al. 2013) and MIST (Durham et al. 2018) observation data (data available on request). Importantly, students in both classes were frequently asked to engage in small group interactions centered on constructing and interpreting graphs, making claims from evidence, solving problems, and other core competencies as part of the regular curriculum. There were, however, important differences between the two courses. Stats was taught in a room with individual tables and the tables could be moved into small groups with each student facing other students. In contrast, Evol was taught in a larger room with large, round tables, and the number of students at each table varied depending on class size. For large classes, there were as many as nine at most tables, consisting of three groups of three. Furthermore, students in small groups did not face each other as directly as in the Stats class because students in a group were typically seated adjacent to each other on one side of a large round table. Finally, because of the geometric configuration of students, the opportunity for inter-group interaction was much greater in the Evol class than in the Stats class and the probability of inter-group interaction was likely higher with larger numbers of groups (as a consequence of larger class sizes). An additional distinction between the two classes was that Stats was mostly 2nd-year students whereas Evol was mostly 3rd year students. The ratio of self-identified females and males were similar between the two courses and was about 1.4 to 1 (females to males). Importantly, the focus of this work was towards development of a quantitative framework and was not implemented to investigate the behavior or performance of individual students.

2.2 Implementing Two-stage Exams

Two-stage exams in both courses included only selected-response questions (i.e. multiple choice). Clickers were used for stage 1 and IF-AT ® cards (Epstein Educational Enterprises) were used for recording group answers during stage 2. For each exam, the same questions were used for both stages but the order of answers for each question was shuffled between stage 1 and stage 2. Students were warned the order of answers differed between the individual and group portions of the exam. Typically, students were allowed 90 to 120 seconds to answer a question using clickers during stage 1. The answers to each question were not revealed to students after stage 1. Prior to the group portion of the exam, students were urged to listen to and paraphrase each other's claims, justify their choice of answers, challenge each other's claims, and establish consensus before answering. Each question was projected onto a screen one at a time and students typically answered each question after about two-three minutes (± one minute) of discussion. Before advancing to the next question, I asked whether students needed more time and typically allocated more time depending on the number of groups still working on an answer. For instance, if there was more than one group still working, I would let them go and ask again after about a minute. If a single group was still engaged, I would typically warn them they had 30 seconds remaining.

The small collaborative groups were assembled by student choice. All two-stage exams were administered to students that had worked within the same group for at least 4 weeks. Group membership varied between exams, although to different degree for different groups within a course. While individual membership in groups was recorded, the dynamics of individuals within groups were not explored in this study. (For a network perspective on the dynamics of group membership in the Evol course see Buchenroth-Martin et al. 2016.)

For each question and individual, I recorded the year, course, individual identity, exam number (midterm 1, midterm 2, or final), question number, group number, each individual's answer, each individual's answers, and whether individuals answered each question correctly (coded as 0 and 1 for incorrect and correct, respectively). For the group scores from the IF-AT cards, I recorded the number of answers scratched off. For some types of analyses, when the data were analyzed, group scores greater than 1 were coded as 0. These data enabled calculation of the proportion of each group that answered a question correctly during stage 1, referred to as the individual score, and whether a group answered correctly during stage 2, referred to as the group score, for each question.

2.3 Quantitative Framework

I developed a framework for the analysis of two-stage exam data. The framework for categorizing student outcomes is illustrated in figure 1. The example of the illustrated framework was designed based on a group size of three, the most common group size in this study. Stage 1 results were summarized as the proportion of individuals in a group with the correct answer: these data are used as a predictive variable (Figure 1). For a group size of three, there were four possible outcomes from stage 1: 1) all students answered a question incorrectly and therefore the proportion of individuals answering correctly was zero; 2) one out of three students answered correctly and the proportion of individuals answering correctly was 0.33; 3) two out of three students answered correctly and the proportion of individuals answering correctly was 0.67; and 4) all three students answered correctly and the proportion of individuals answering correctly was 1. After completing the exam individually, students in each group worked together and completed the exam again (stage 2). There were two possible outcomes for each question from the group portion: either the question was answered incorrectly ($y = 0$) or correctly ($y = 1$)(see figure 1). Depending on the specific outcomes from the two stages, student performance was classified into four categories, and each category represents an operational definition of interaction type. First, productive collaboration was evident if less than 50% of individuals within a group answered a question incorrectly during the first individual stage but the group answered the question correctly (the upper left quadrant in figure 1). Second, consensus building or mutual understanding was evident if greater than 50% of individuals answered correctly and the group answered correctly (the upper right quadrant in Figure 1). Third, a lack of student understanding (misunderstanding) was evident if less than 50% of individuals answered incorrectly during the first stage and the group answered incorrectly (the lower left quadrant in figure 1). Finally, counter-productive interaction was evident if a majority of students answered correctly during the first stage but the group answered incorrectly (the lower right quadrant in figure 1). Importantly, productive collaboration can involve building consensus; the difference between the ways these two have been defined here is that consensus emerges from groups when a majority of students were correct during stage 1 whereas productive collaboration is defined for cases in which less than half of the students were correct during stage 1, a distinction that recognizes different interactions may be occurring as a consequence of agreement among group members following stage 1.

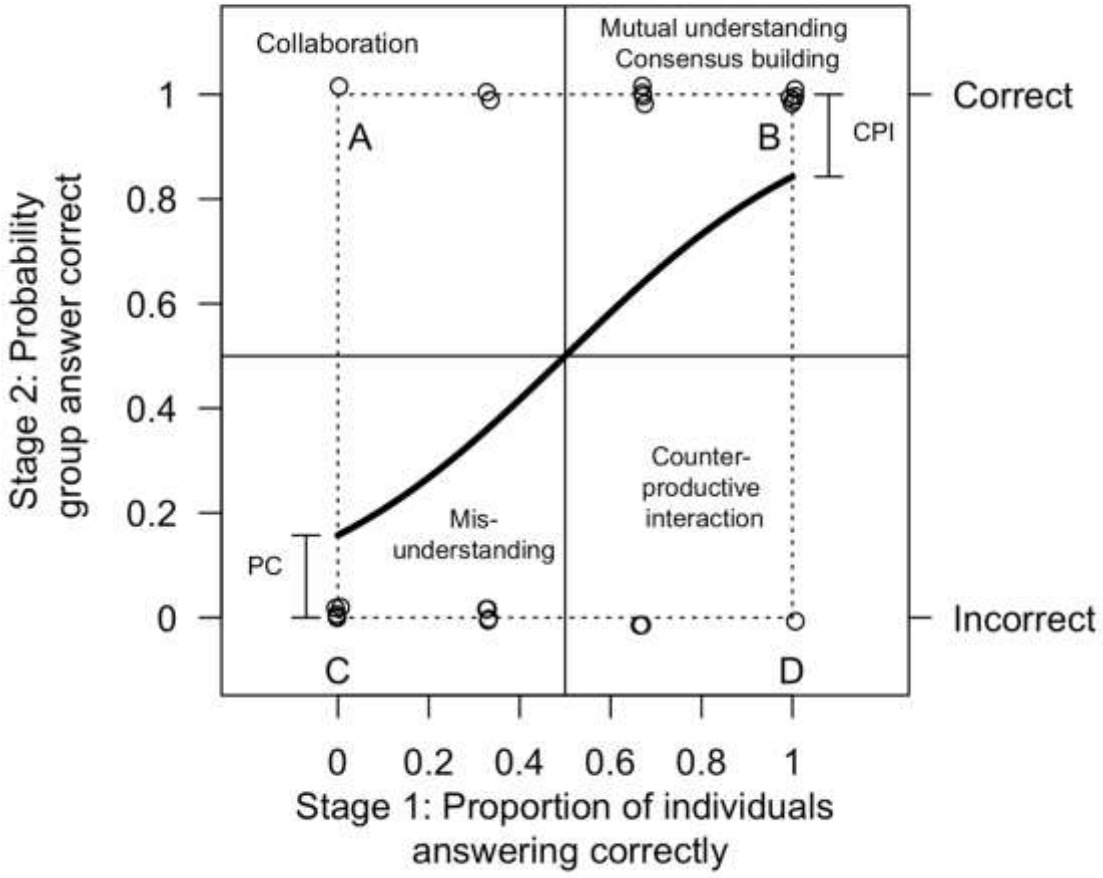


Figure 1. Illustration of the quantitative framework for the analysis of two-stage exams

The x-axis is the proportion of individuals within a group that answered the question correctly. The y-axis is the probability the group answer is correct. The data for the group score are binomial: 0 = incorrect and 1 = correct (right axis). Each open circle is the outcome from a two-stage exam for one group on a single question. For instance, the single point at x = 0 and y = 1 (identified with the letter A) represents a single group in which all individuals answered incorrectly during the first stage of the exam but the group answered the question correctly during the second stage of the exam. The multiple points at x = 1 and y = 1 (indicated by the letter B) identify questions in which all members of a group answered correctly during stage 1 and the group answered correctly. The multiple points at x = 0 and y = 0 (indicated by the letter C) identify questions in which all members of a group answered incorrectly during stage 1 and the group answered incorrectly. The single point at x = 1 and y = 0 (identified with the letter D) represents a single group in which all individuals answered correctly during the first stage of the exam but the group answered the question incorrectly during the second stage of the exam. PC indicates the value calculated using equation 1 (see methods). Finally, CPI indicates the value calculated using equation 3 (see methods). The solid black line shows the predicted values for the 24 data points shown.

Performance on two-stage exams was quantified using binomial regression because the group score could be either 0 or 1. The y-intercept from regression estimated the probability of a correct group answer (G = 1) (from stage 2) when all individuals answered incorrectly from stage 1 (I = 0). The equation for the y-intercept, namely

$$\mathrm{PC}\ (\mathrm{G} = 1|\mathrm{I} = 0) = \exp(B_0)\ /(1 + \exp(B_0)) \quad (1)$$

where B_0 is the y-intercept from the linear model analysis, provided the basis for estimating productive collaboration (PC), assuming individuals in a group changed their answers from incorrect to correct as a consequence of

construction interactions (e.g. collaboration).

Another relevant metric from regression was the amount the estimated probability of a correct answer deviated from one when all individuals in a group answered correctly during stage 1. The probability of a correct group answer from stage 2 (G = 1) when all individuals answered correctly during stage 1 (I = 1) is

$$\mathrm{P(G = 0|I = 1)} = \exp(B_0 + B_1)/(1 + \exp(B_0 + B_1)) \quad (2)$$

where B_0 and B_1 are the first (y-intercept) and the second (slope) coefficients from logistic regression, respectively. This probability deviated from one when all individuals in a group answered correctly during stage 1 but the group answer from stage 2 was incorrect. Thus, the deviation from one provides evidence of counter-productive interactions (CPI) and was estimated as

$$\mathrm{CPI} = 1 - \exp(B_0 + B_1)/(1 + \exp(B_0 + B_1)). \quad (3)$$

Note that in figure 1, the inference of CPI scales with the value for equation 3.

Finally, I calculated the 50% response probability (R50) as

$$\mathrm{R50} = \mathrm{P(G = 1|I = 0.5)} = \exp(B_0 + B_1 * 0.5)/(1 + \exp(B_0 + B_1 * 0.5)). \quad (4)$$

R50 represents the estimated group performance at the transition between the delineation of collaboration (upper left quadrant in figure 1) and consensus-building or mutual understanding (upper right quagrant in figure 1). Here, I use 50% response rate as an estimate of the combined effects of collaboration and consensus-building from stage 1 to stage 2 of the exam. These three metrics—PC, CPI, and R50—provide informative metrics about the performance of individuals and groups during two-stage exams.

I constructed several visualizations of the data that differed from typical graphs of two-stage exam data. First, I plotted the mean, min, and max values for the proportion of individuals correct during stage 1 for each group across all 12 exams to show the variation in individual performance across groups for each exam. Second, I subjected the data to a simple logistic regression with group and individual scores for each group and plotted the predicted values for all 12 exams to show the variation in the dependence of group scores on individuals among groups for each exam. Finally, I constructed a histogram of R50 calculated for all group for each of the exams to show the variation in the combined effects of collaboration and consensus-building among groups for each exam.

2.4 Analytical Models

The purpose of this study was to evaluate effects of different variables on student performance during the group (stage 2) exam. There were four important variables with likely effects on group performance: the proportion of individuals that answered correctly during stage 1 (individual score), group identity (group id), question (question id), and group size. All of these were considered as random effects because each was assumed to have some unknown variance greater than zero.

For each of the 12 exams, I evaluated eight different logistic regression models using the Akaike Information Criterion (AIC)(Burnham and Anderson 2002). All models included individual score; models differed with respect to the inclusion of the three other random effects (question id, group id, and group size). The models were evaluated in R using the glmer function in the lme4 library (Bates et al. 2015). AIC scores were calculated for each of the 8 models for all exams. I calculated the difference in AIC score from the best model (ΔAIC) for each set of eight AIC scores, resulting in 12 ΔAIC scores for each model (one ΔAIC score for each exam). Models were compared based on the mean of the 12 ΔAIC scores and the number of exams in which a particular model had the smallest (best) AIC score. Finally, for each exam, I summarized the variance estimates of each effect (using the standard deviation), summed the variances across all variables, and calculated the percent of the total variance for each included variable. Importantly, I was not interested in testing a particular hypothesis and therefore did not focus on interpretation of p-values or other traditional methods of statistical significance, in part, because of non-independence of data renders p-values mostly uninterpretable.

2.5 Summarizing the Data and Evaluating Associations between Variables

For each exam I estimated PC, CPI, and R50 using two different models: one with individual score + question id and one with individual score + group id. I constructed plots of these data as a means of showing the variance in PC and CPI in relation to question and group for each of the 12 exams. I also examined the data for a single exam in more detail to reveal variation in performance among groups, among questions, and among individuals. I focused on a single exam to illustrate the power of the analysis for revealing the effects of student preparation (individual score), question, and group on collaborative exam performance. For evaluating the performance among groups, I calculated

the mean for the individual and group scores for each group across all questions and plotted the data relative to the case in which individual and group performance were identical. From these data, I used logistic regression to focus attention on the variation in performance among select groups. I also calculated the mean for the individual and group scores for each question across all groups and plotted the data relative to the case in which individual and group performance were identical. From these data, I used logistic regression to focus attention on the variation in performance among selected questions. Finally, I used a non-parametric Spearman rank correlation to assess whether there was evidence of association between the mean individual and group scores for both the analysis of groups and questions.

Finally, I evaluated whether there was an associated between the diversity of student responses and the ease of each question from stage 1 and measures of collaboration using a non-parametric Spearman rank correlation. Diversity of responses was estimated as $1 - \sum p_i^2$, where p_i^2 is the frequency of each response. Ease was estimated as the proportion of individuals with a correct answer during stage 1. Additionally, I extracted cases in which diversity of student response during stage 1 was zero (all individuals answers the same answer) and compiled the number of cases in which the stage 1 answers were wrong (all individuals in the group answered the same wrong answer) and the number of cases in which the stage 1 answers was correct (all individuals in the group answered the same correct answer) and compared these values to the group scores as a means of further highlighting evidence of collaboration and counter-productive interactions.

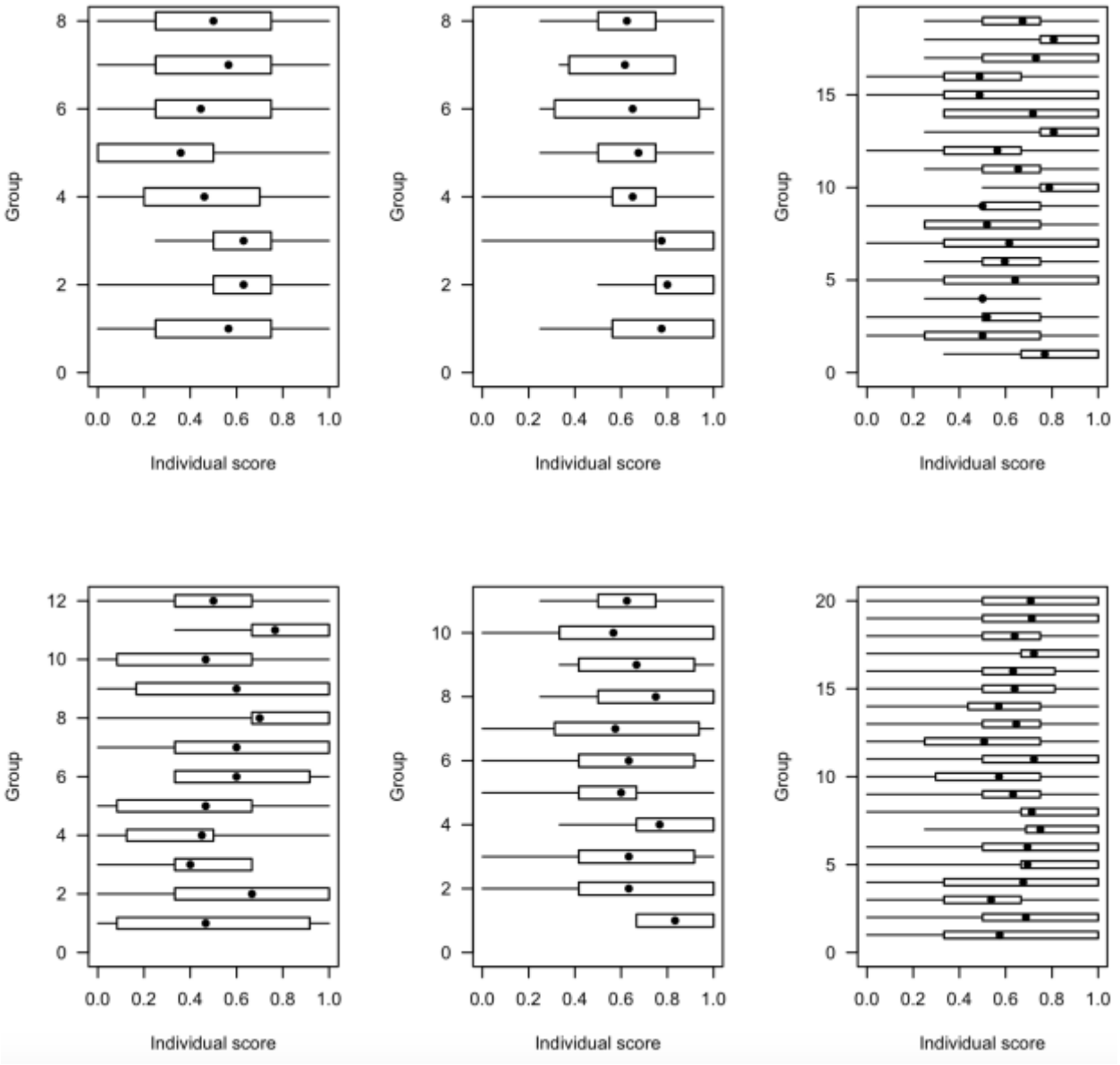
3. Results

3.1 Overall Summary across All Exams

Across all 12 exams, there were 11634 combined individual and group exam scores generated from 753 students that answered 182 questions in 236 groups. Although group sizes of three or four were emphasized, for the 236 different groups summed across 12 exams, there were 7 groups of size two, 179 of size three, 48 of size four and one each of sizes five and six. For the 11634 recorded scores, there were 1284 (11.1%) cases in which an individual was incorrect and his or her group was incorrect, 3350 (28.8%) cases in which an individual was incorrect but his or her group was correct, 380 (3.3%) cases in which an individual was correct but his or her group was incorrect, and 6620 (56.9%) cases in which an individual and their group were both correct. Overall, 56.9 + 3.3 = 62.2% of the questions were answered correctly during stage 1 and 28.8 + 56.9 = 85.7% of questions were answered correctly during stage 2 across all 753 students. Thus, there was an overall normalized gain of 62%. At a group level, comparison of average individual and group exam scores revealed three different outcomes: 1) the group score was greater than the maximum average individual score, supporting an inference of collaboration; 2) the group score was equal to the maximum average individual score; and 3) the group score was less than the maximum average individual score, supporting an inference of counter-productive interaction. Across all 236 groups, there were 162 (68.8%) instances of outcome 1, 54 (22.7%) instances of outcome 2, and 20 (8.6%) instances of outcome 3. These data revealed that, on average, students answered more questions correctly on the group (stage 2) exams than on the individual exams.

3.2 Variation among Exams

Visualization of the results from stage 1 for each group across all 12 exams revealed individual scores across all questions within each group typically varied from 0 to 1 and that the mean scores varied among groups (Figure 2). There were some notable differences in individual scores from stage 1 across exams. For example, individual scores varied from 0 to 1 across all questions for 7 out of 8 groups on exam 1 but varied from 0 to 1 for only 2 out of 8 groups on exam 2; similarly, the maximum individual scores for 8 out of 33 groups were all less than 1 on exam 11 but were less than 1 for only 1 group on exam 12 (Figure 2). Overall, the data revealed marked heterogeneity among groups for the stage 1 performance of individuals within groups (Figure 2).


Group
Individual score
0.0 0.2 0.4 0.6 0.8 1.0
0 2 4 6 8
0 5 10 15
0 2 4 6 8 10 12
0 2 4 6 8 10
0 5 10 15 20

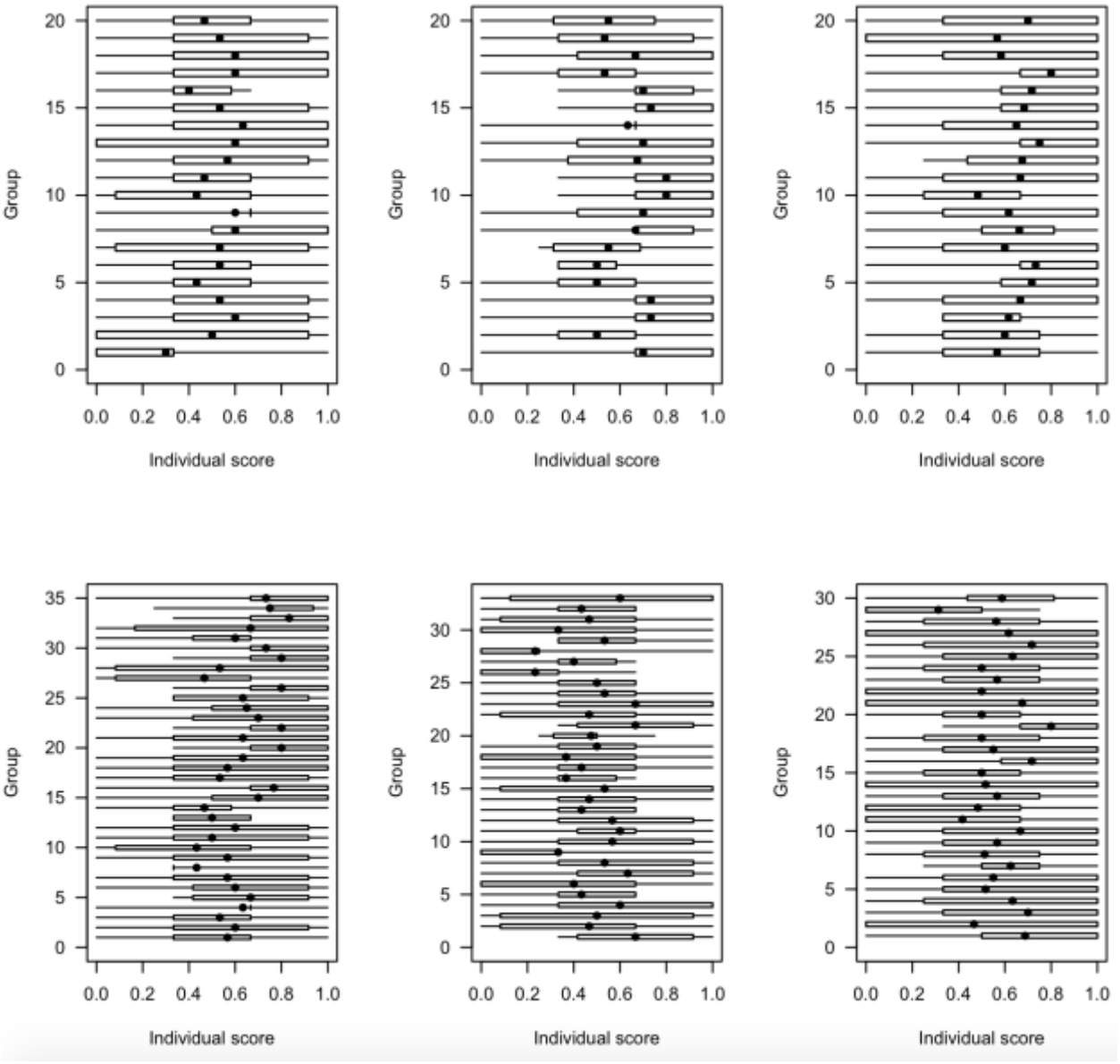


Figure 2. The distribution of scores from stage 1 for each of the groups on all 12 exams

The visualizations show the range (thin horizontal lines), 25% and 75% quartiles (rectangles) and means (filled circle) of the individuals scores across all question for each group for each of the 12 exams.

The heterogeneity among groups for individual scores within groups was also evident in the graphs of the predicted values from logistic regression for each group on each exam (Figure 3). There were a few particularly noteworthy results. First, there was at least one group with a negative slope on three of the 12 exams (exams 2, 4 and 11) (Figure 3). Second, there was at least 1 group with a perfect group score across all questions on each of 11 out of 12 exams: the only exception was exam 1. Third, there was also at least 1 group for each of 11 out of 12 exams in which there was a marked switch from 0 to 1 in the predicted group score value: the only exception was exam 1.

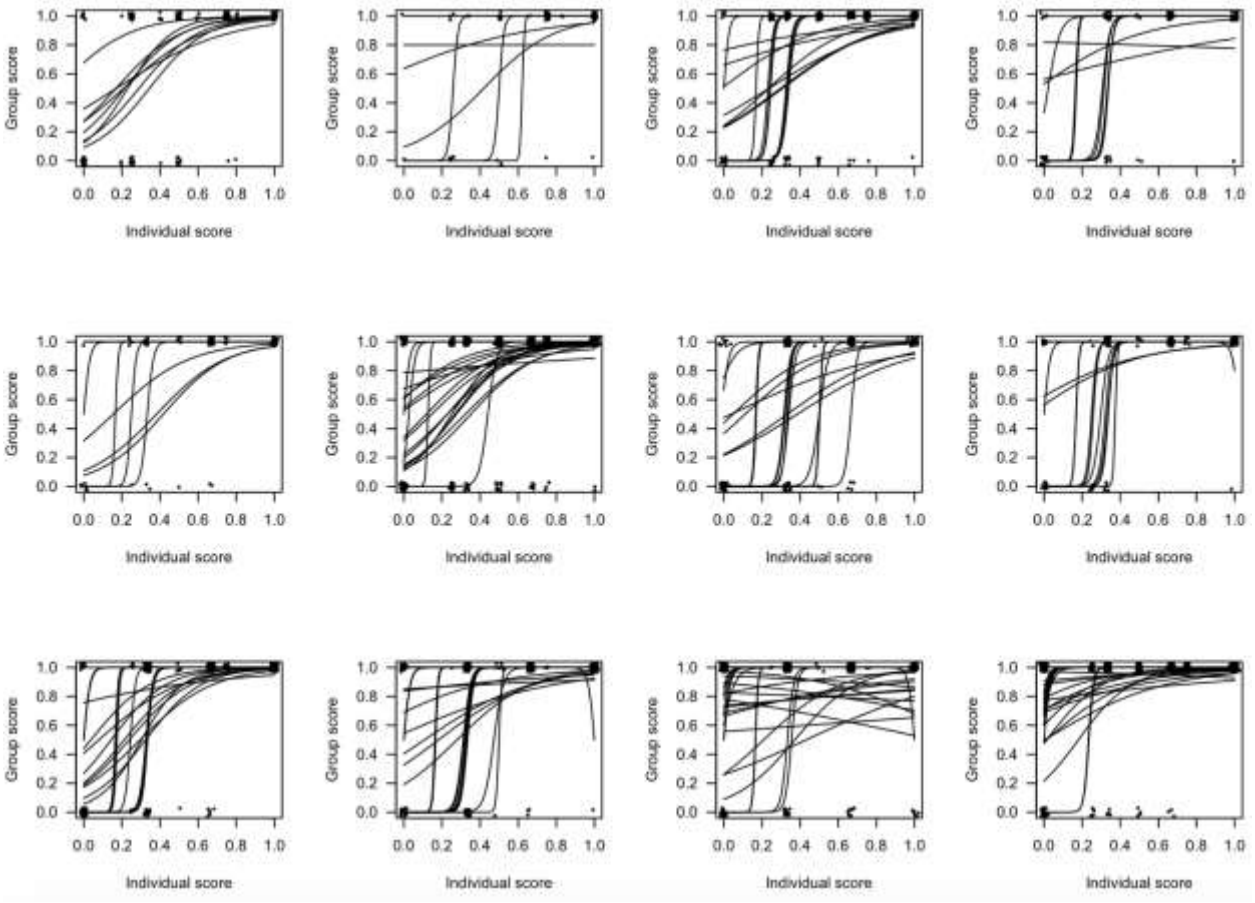


Figure 3. Predicted values for group scores for each group on each of the 12 exams

Plots of the individual and group scores (small points) for all questions and groups for each of 12 exams with the predicted values for each group shown as a logistic curve showing the variation in performance among groups within and among exams.

The distribution of R50 values for each group on each exam clearly revealed differences in the outcomes of two-stage exams across the 12 exams (Figure 4). For two exams, R50 was one or nearly one for all for groups (e.g. exams 8 and 12). For other exams, R50 was widely different across groups (e.g. exams 2, 7, 10 and 11). For all exams, except exam 1, the mode for R50 was 1. Overall, the compiled graphs showing the predicted values from logistic regression and the corresponding histograms for R50 for each group on each exam revealed heterogeneity in performance among groups within and among exams.

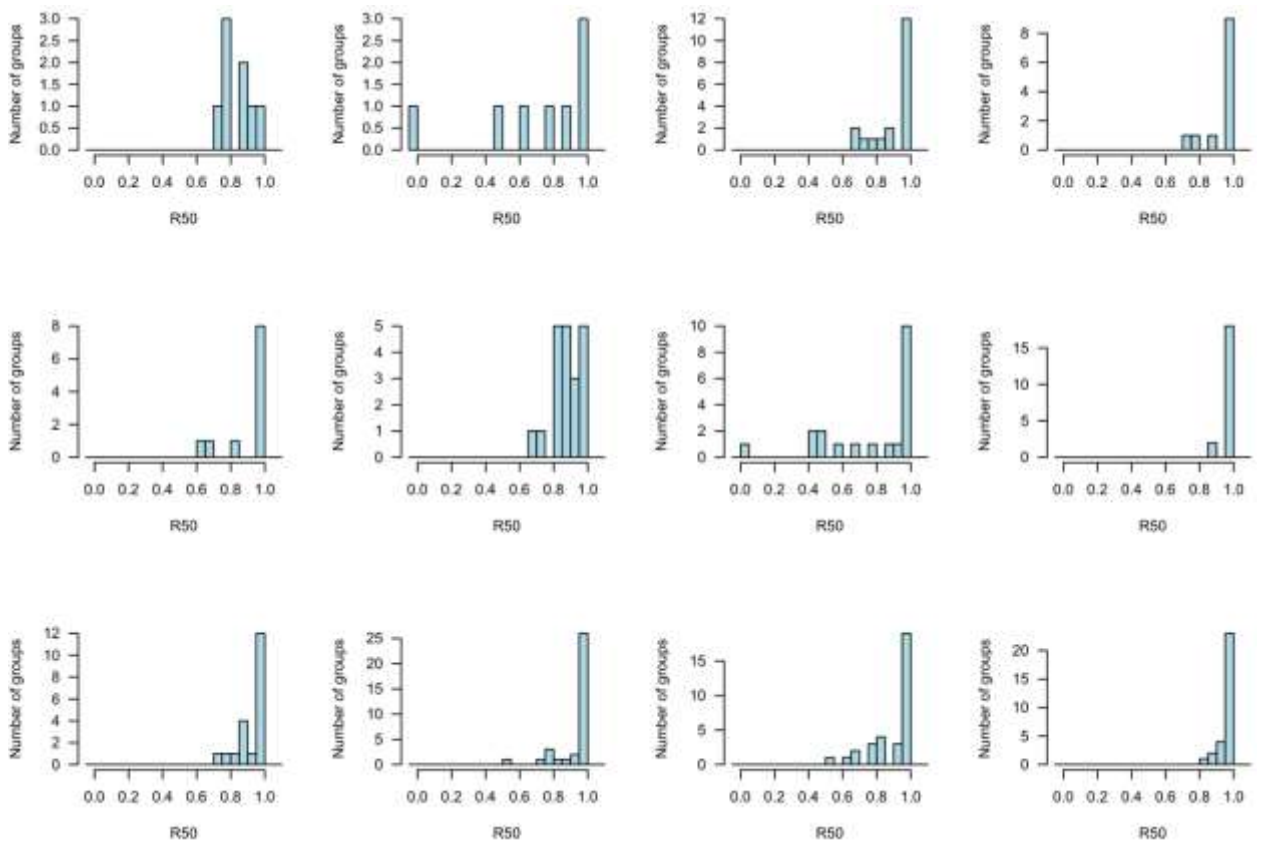


Figure 4. Histogram of R50 scores for different groups from each of the 12 exams

3.3 Evaluation of Alternative Models

To assess the relative contribution of individual performance, question, and group on stage 2 outcomes, I compared 8 different random-effects logistic regression models fitted to the two-stage data and evaluated using the Akaike Information Criterion (AIC)(Table 1). On average, the best model based on mean ΔAIC scores included the proportion of individuals in a group correct (PIC) and question identity. This model was the best model (defined as having the lowest AIC score) for four of the 12 exams (Table 1). The second-best model, based on the mean ΔAIC score, included PIC, question identity and group identity as random effects: this model was the best model for two of the 12 exams (Table 2). The simplest model (only PIC as the included random effect) was the best model for five of the 12 exams and was ranked fifth overall based on the mean ΔAIC score (Table 1). None of the models with group size included as a predictive variable was the best model for any one of the exams. Overall, model comparison using AIC suggested individual scores, question id and group id had demonstrable effects on group performance.

Table 1. Comparison of different logistic models using AIC

Model	Model description	K	Difference from minimum mean AIC	Relative weight	Number AIC scores = minimum
1	Only PIC	2	3.54	0.170	5
2	PIC + Group size	3	5.51	0.064	0
3	PIC + Group id	3	4.81	0.090	1
4	PIC + Question id	3	0.00	1	4
5	PIC + Group id + group size	4	6.77	0.034	0
6	PIC + Group size + question id	4	1.96	0.375	0
7	PIC + Group id + question id	4	1.03	0.600	2
8	PIC + Group id + question id + group size	5	2.99	0.224	0

Model description shows the predictive variables included as random effects in the model. PIC (the proportion of individuals correct from stage 1) was included in all models. K is the number of parameters. For each model, the mean AIC score was calculated for all 12 exams. The difference from minimum mean AIC is the difference in mean AIC score from the model with the minimum mean AIC value. Relative weight was calculated as the negative exponent of the difference in mean AIC scores from the model with the minimum mean * 0.5. Number AIC score = minimum is the number of exams in which the model had the lowest AIC score.

I evaluated each exam using a model with individual scores, questions, and groups as random effects, recorded the estimated variance for each of the effects, and calculate the percent of each estimated variance relative to the sum of the variance across all effects (Table 2). While there was a general trend in the rank order of variance, with individual scores > questions > groups, the relative variance for each variable differed considerably among exams, and in some cases, the rank order of variances changed (Table 2). For instance, the variance for questions was greater than individual scores on exam 2 and 11, and the variance for groups was greater than questions on exams 5 and 7 (Table 2). Finally, for some exams, the estimated variances for groups was zero (exams 1, 4, 8, and 12) and was zero for questions on one of the exams (exam 5).

Table 2. Estimates of the random effects variances

Exam	PIC	Q	G	Sum	% PIC	% Q	% G
1	1.726	0.187	0	1.913	90.2	9.8	0
2	1.248	1.808	0.851	3.907	31.9	46.3	21.8
3	1.560	0.307	0.110	1.977	78.9	15.5	5.6
4	2.287	0.94	0	3.227	70.9	29.1	0
5	2.016	0	0.518	2.534	79.6	0	20.4
6	1.625	1.011	0.532	3.168	51.3	31.9	16.8
7	2.654	0.263	1.000	3.917	67.8	6.7	25.5
8	2.675	0.623	0	3.298	81.1	18.9	0
9	2.556	0.991	0.234	3.781	67.6	26.2	6.2
10	1.679	1.189	0.355	3.223	52.1	36.9	11.0
11	0.501	0.945	0.807	2.253	22.2	41.9	35.8
12	1.629	0.9	0	2.529	64.4	35.6	0
Mean	1.846	0.763	0.367	2.977	63.2	24.9	11.9

Numbers are the estimated standard deviations of effect for the proportion of individuals correct from stage 1 (PIC), question (Q), and group (G). Sum is the sum of the standard deviations and %PIC, %Q, and %G are the corresponding percent of the summed standard deviation of each effect.

3.4 Effects of Questions and Groups among Exams

To investigate the effects of questions and groups, I used random effects models with individual score and question id for one set of analyses and individual score and group id for another set of analyses and extracted PC, CPI, and R50 from the resulting model predictions. Figure 5 shows the data for PC and CPI. PC exhibited marked variation among questions depending on exam. For example, there were small effects of question on PC for exams 1 and 3 relative exams 2, 6, and 10 (Figure 5A). The pattern for the variation in the effects of group on PC was similar to the estimated effects of questions, although there were many more cases in which there was no effect of group: the variance of group effect was zero or near zero for exams 1, 2, 3, 4, 8, 9, 10 and 12 (Figure 5B; Table 2). Variation in CPI among questions was evident for exams 2, 4, 6, 10 and 11 (Figure 6C) and among groups for exam 11 (Figure 6D). These data, and the values for R50, revealed considerable heterogeneity in quantitative estimates of collaboration among questions and among groups for each of the 12 exams.

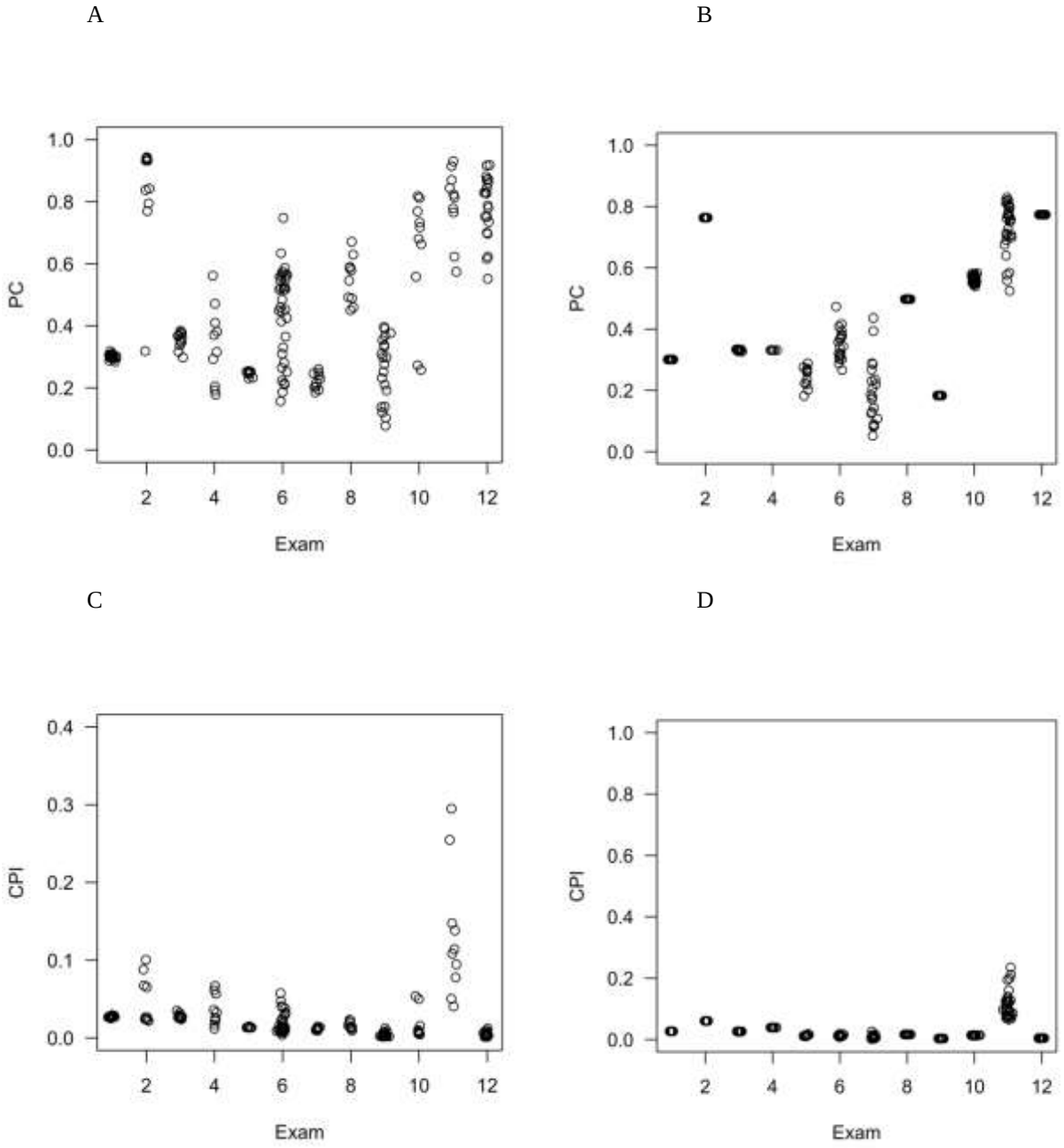


Figure 5. Plots of PC and CPI

The two top graphs (A and B) show the PC values for each question (A) and each groups (B). The two bottom graphs (C and D) show CPI values for each question (C) and group (D). Values were jittered slightly to show all the data.

3.5 Variation within a Single Exam

I highlighted variation in performance among questions and groups for exam 11 in part because the variance estimates for individual score, question, and group were similar (Table 2). (Importantly, though, there was variation in the performance on the two-stage exams among groups and among questions for all exams.) I plotted the mean individual and group scores for each question across all groups and each group across all questions for exam 11. For all questions, the mean group score was greater than the mean individual score. There was not, however, an association between the mean stage 1 scores and mean stage 2 scores (Spearman $r = 0.474$, $p \approx 0.167$); moreover, the magnitude of the difference from the expectation of no difference in performance between stage 1 and stage 2 varied across questions (Figure 6). To get a better idea of the performance for stage 1 and 2 for different questions, I constructed a simple logistic regression separately for questions 3, 4, and 10 (Figure 7). For question 3, all 35 groups answered correctly during the group (stage 2) portion of the exam even though, on average, the proportion of individuals that answered correctly during the individual (stage 1) portion of the exam was 0.5 (Figure 6). By contrast, the logistic model for question 10 had a negative slope that was due, mostly, to three groups in which all individuals answered the correctly during stage 1 but answered incorrectly during stage 2 (Figure 7). The predicted values for question 4 were intermediate between the responses for question 3 and 10 (Figure 7). All three questions (see supplementary information for all 10 questions that comprised exam 11) were designed using the 3D-LAP protocol—they focused on a big idea, emphasized a core disciplinary practice, and involved a crossing-cutting concept (Laverty et al. 2016)—and each one proved to elicit different dynamics within groups during stage 2.

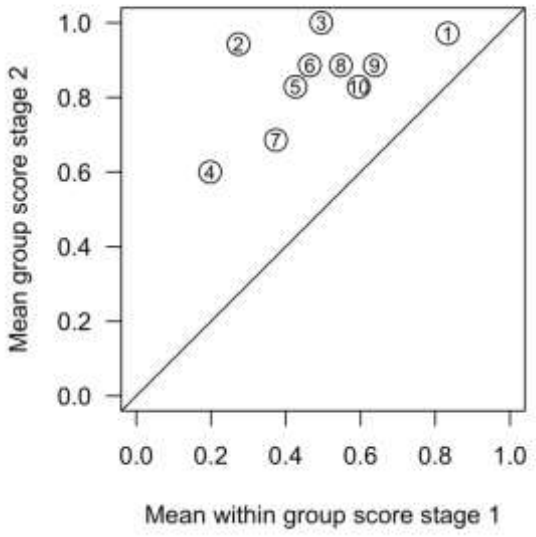


Figure 6. Graph of the mean within group score from stage 1 and the mean group score from stage 2 for each question on exam 11

Numbers inside of the open circle indicate question number and the diagonal line shows the expectation for the case in which there was no change in scores between stage 1 and stage 2 of the exam.

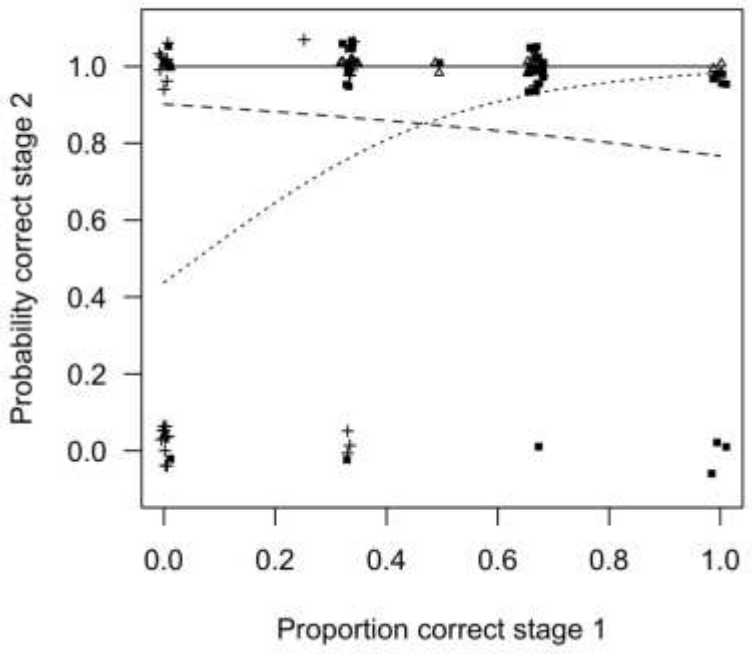


Figure 7. Visualization of the data and predicted values for three questions from from exam 11

Predicted values (lines on the graph) were generated using logistic regression with individual scores and question as random effects. Only the data and predicted values for three questions are shown: question 3 (solid line and triangles), 4 (short dashed line and plusses) and 10 (long dashed line and filled squares).

The mean group score was greater than the mean individual scores across all groups (Figure 8). There was not, however, a detectable positive association between the mean individual score from stage 1 and the mean group score for stage 2 among groups (Spearman $r = 0.203$, $p \approx 0.242$). I constructed simple logistic regression models for three different groups (groups 29, 30 and 35) to reveal some of the variation in dynamics among groups (Figure 9). Group 30 was perfect for all questions during stage 2 even though the mean proportion of correct answers during stage 1 was less than 0.5. The predicted values for group 29 suggest that if at least 1 person chose the correct answer during stage 1, the group converged on the correct answer during stage 2. (Note that in group 29 at least one individual answered a question correctly during stage 1 because none of the x-axis scores for the filled squares are less than 0.33.) Finally, group 35 revealed an intermediate pattern due, in more or less equal parts, to one question in which the majority of individuals answered correctly during stage 1 but the group answered incorrectly, and to one case in which all individual answered incorrectly during stage 1 and the group answer correctly.

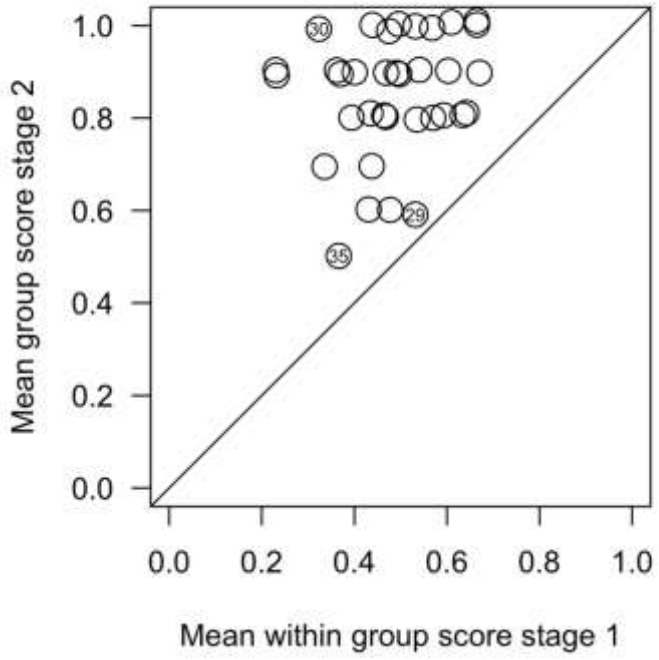


Figure 8. Graph of the mean individual scores from stage 1 and mean group scores from stage 2 on exam 11 for each of the 35 groups

Circles were jittered slightly to better show the data. Numbers within circles identify three groups highlighted in figure 9 and the diagonal line shows the expectation for the case in which there was no change in scores between stage 1 and stage 2 of the exam.

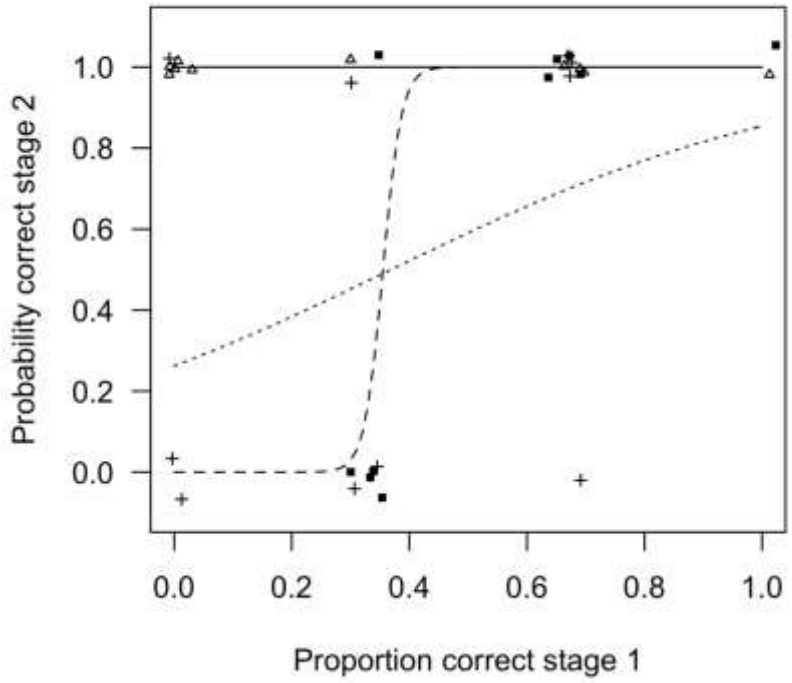
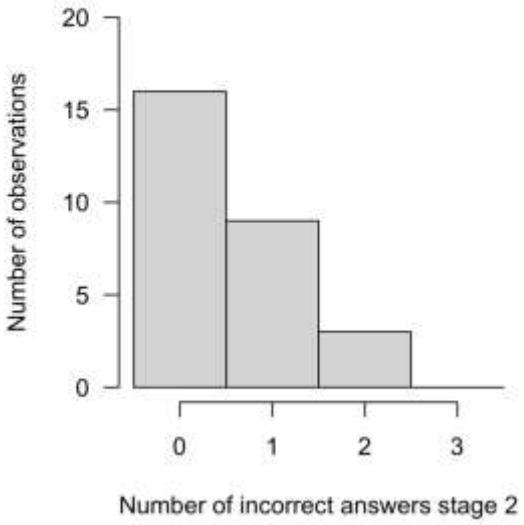


Figure 9. Visualization of the data and predicted values for three groups from from exam 11

Predicted values (lines on the graph) were generated using logistic regression with individual scores and question as random effects. Only the data and predicted values for three groups are shown: group 30 (solid line and triangles), 29 (short dashed and plusses), and 35 (long dashed line and filled squares) from exam 11.

Using Spearman's non-parametric test of association, I did not detect an association between the diversity of response during stage 1 and collaboration statistics (R50, PC, and CPI) for either questions or groups across all exams (data not shown). I also did not detect an association between the ease of questions and collaboration statistics (data not shown).

Finally, I extracted cases in which all individuals in a group chose the same incorrect answer during stage 1 but answered correctly during stage 2 (Figure 10, top) and cases in which all individuals in a group chose the correct answer during stage 1 but answered incorrectly during stage 2 (Figure 10, bottom). These data provide another means of estimating collaboration and counter-production inreractions, respectively. Particularly noteworthy was the discovery that out of 28 cases in which all individuals in a group answered the same incorrect answer during stage 1, there were 16 (57%) cases in which the question was answered correctly during stage 2. These data suggest students productively engaged with each other towards a common goal.



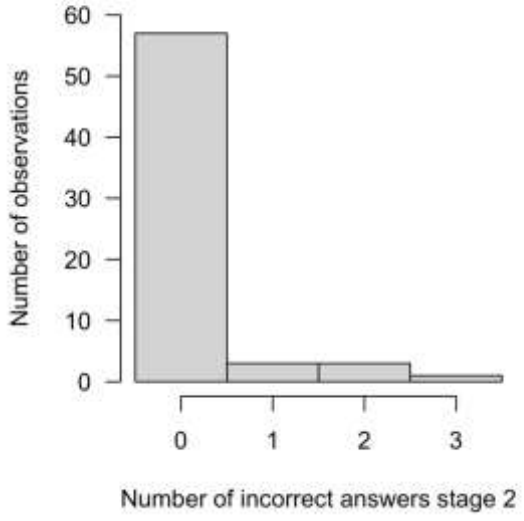


Figure 10. Visualization of the distribution of number of incorrect answers for stage 2 when all individuals in a group answered the same answer during stage 1

Above. Number of cases in which all individuals in a group answered the same *incorrect* answer during stage 1. The number of cases in which there were zero incorrect answers during stage 2 provides evidence of collaboration. Below. Number of cases in which all individuals in a group answered the same *correct* answere during stage 1. The number of cases in which there were 1 or more incorrect answers provides evidence of counter-productive interactions.

4. Discussion

Active learning typically involves having students construct their understanding often as a consequence of working in small groups of peers on authentic problems aligned with relevant learning goals. During active learning there are three factors that can influence the dynamics and outcomes of student work in small groups. First, each student comes to the interaction with different levels of understanding (or mastery) of the content and relevant disciplinary

practices; thus, there is an effect of each individual's knowledge on the outcome of working in s small group. Second, the cognitive challenge of a question can vary (Laverty et al. 2016) in ways that challenge the students in a group differently and collectively; thus, there is an effect of the prompt that stimulates active learning on the outcome of small group work. Finally, the effectiveness of how students work together within small groups can influence the outcome. This triad of effects can generate considerable complexity that was manifested as heterogeneity for these three factors within and among exams.

There were several outcomes of my investigation of two stage exams. First, binomial regression models yielded quantitative estimates of productive collaboration, consensus-building, and counter-productive interactions. These three measures expand the scope of inferences that can be gained from two-stage exams relative to simple comparisons of the proportion of correct answers between stages 1 and 2 typically reported. Second, there were marked differences in the effects of individuals, questions, and groups on two-stage exam scores within and among exams, an observation suggesting the active learning associated with two-stage exams may be relatively complex due to the triad of individual, question, and group effects. Third, variation in effects among questions and groups can provide valuable information for data-driven revision of teaching, assessments, and curricula.

4.1 Quantitative Estimates of Collaboration, Consensus-Building, and Counter-Productive Interaction

Two-stage exams can provide information about student interactions in small groups working towards common goals. I proposed three statistics calculated from a binomial regression model that may provide relevant information for making inferences about student performance from two-stage exam data. The probability of a correct group answer when all individuals within the group answered incorrectly estimates productive collaboration (the PC statistic). Productive collaboration is especially apparent when all individuals in the group answered the same wrong answer during stage 1. I operationally defined productive collaboration in this way because students had to arrive at a correct answer from unanimous incorrect answers based solely on what they knew, suggesting that students built sufficient knowledge collaboratively from each individual's incomplete understanding of a problem's solution to arrive at the correct solution. The probability of a correct answer when 50% of the individuals answered the question correctly during stage 1 (the R50 statistic) supports an inference of consensus building. I operationally defined R50 as a measure of consensus because in the majority of cases not all students in a group answered correctly during stage 1; thus values of R50 > 0.5 provides an indication that students correctly settle on a correct answer after sharing knowledge and perspectives. Finaly, the probability of an incorrect answer when 100% of the individuals answered correctly during stage 1 supports an inference of counter-productive interaction (CPI). Although CPI was rare, it did happen and may signal that more attention on practicing collaborative behaviors is required for productive two-stage exams. Overall, these three statistics can form the basis for inferences about the effects of individuals, questions, and groups on learning.

4.2 Variation among Exams

The most conspicuous result from this study was heterogeneity among exams. I discovered differences in the estimated variance for random effect predictive variables in the quantitative model, variation in quantitative estimates of collaboration among exams, and variation in quantitative estimates of collaboration among groups and questions within exams. Contrast between exams was most evident for comparisons of successive exams in a single course during one semester. For instance, the magnitude of the variance in random effects differed markedly between exams 1 and 2 (midterms 1 and 2 in the lower-level "Stats" course). Additionally, the rank order of the magnitude of variance among the random effects changed depending on exam. For exam 1, the individual scores accounted for 90% of the sum of the variance across effects whereas for exam 2 individual scores accounted for 32% of the sum of the variances; furthermore, for exam 2, the variance for the random effects of question was greater than the individual scores during stage 1. Similar patterns were evident for the other three classes from which two-stage exam data were obtained.

These results suggest there are interactions between individual knowledge, the cognitive challenge of questions, and the group dynamics, a hypothesis that aligns with the conviction teaching and learning are complex human endeavors influenced by a variety of context-dependent factors. Previous claims two-stage exams elevate student engagement and improve student scores markedly underestimate the effect of two stage exams. There is clearly much more going on during two-stage exams. Thus, like other methods of "active learning", two-stage exams create complexity in the classroom that has largely gone unrecognized and unquantified from previous studies. Moreover, the outcome of the complex interactions varies across exams. I don't have any clear answers about why there is so much variation. With the aid of the analytical framework, it should be possible to design some controlled studies with the aim of better understanding the intra- and interpersonal dynanmics underlying the heterogeneity of estimated effects.

4.3 Variation among Questions

Two-stage exams can reveal differences in the effects of questions on student performance. This information may be useful for understanding multi-dimensional properties of assessment and direct revision towards developing questions that elicit collaboration. Laverty et al. (2016) documented the dimensionality of questions can vary from zero to three depending on the rubric underlying the creation of engaging questions. When we add student interaction, an additional dimension of assessment emerges, including negotiation and consensus-building: two complex human attributes that depend on a host of learned behaviors. Thus, it is not surprising I discovered different questions can result in widely different quantitative estimates of collaboration. One value of comparing quantitative estimates of collaboration (e.g. PC, CPI and R50) for different questions from a model developed for a single exam is that I can focus my attention on pursuing qualitative analysis of student response to pairs of questions that differ markedly in the estimated effects. Review of several questions that generated very different estimates of collaboration failed to reveal any overt features that might trigger differences in behaviors and performance within small groups. All questions were quantitative in nature, involved model-thinking of some sort, and were answered correctly by at least one out of five and mostly by the majority of students during the individual (stage 1) portion of the exam. Given the results for these questions, it would be worthwhile to record student conversations and interview students about their prior knowledge and interactions during stage 2. Thus, I can imagine combining quantitative analysis with qualitative studies to better understand how particular questions influence collaboration and test specific hypotheses about the effect of questions on student learning.

4.4 Variation among Groups

While the modal value of R50 was 1 for all but one exam, there was variation among groups, and the heterogeneity among groups varied across exams. Variation among groups in the change in scores from the individual to group portions of two-stage exams has been observed before. However, the quantitative framework portrayed more detail in the performance of groups than has been reported in the literature. For instance, I examined two groups with similar gains from the individual to group portions of the exam, yet logistic regression revealed very different performance profiles between the two groups. Ideally, quantitative models for different groups can be compared in the context of qualitative data from student transcripts and interviews to more fully understand how collaboration happens and why it differs among groups. Of particular relevance would be comparison between one group in which all individuals within a group answered the same incorrect answer but nonetheless answered the question correctly as a group (PC = 1) , and one group in which all individuals within a group answered the same correct answer but nonetheless answered the question incorrectly as a group (CPI = 1).

4.5 Limitations of the Approach

Quantitative studies can show whether there are effects of question or group on estimates of collaboration but typically do not provide any information about the underlying mechanism or process producing estimated effects. Thus, quantitative approaches have inherent limitations. Ideally, the quantitative framework I described should be combined with qualitative approaches that enable inference of the underlying cause of the estimated effects. Qualitative study could involve analysis of videos, student interviews, and/or transcripts of student interactions transcripts. For instance, Summer and Volet (2010) analyzed student interactions from videos and interviews and revealed high variance among small peer groups for two key dimensions of collaborative learning: the level and regulation of content processing. The level of content processing refers to whether discussion about a topic remains largely descriptive and focused on facts and definitions (low-level) or includes evidence of inference, explanation, model-building, information integration and synthesis (high level). This axis of interaction is similar to Bloom's levels of cognitive challenge (Crowe et al. 2008). The regulation of content processing refers to how students engage with each other and corresponds in many ways with Chi et al.'s (2014) ICAP framework. Low level regulation of content includes acceptance of claims without challenge or asking for clarification (ICAP's "passive" learning). High level regulation of content involves co-regulation and co-development of understanding as a consequence of interactions, elaboration, challenge, and exchange (ICAP's "interactive" learning).

4.6 Evaluating the Value of Quantitative Data

The value of two-stage exams and analysis of the data depends on purpose. If the purpose of implementing two-stage exams is to elevate student interaction and reduce the anxiety and stress associated with assessment, then simply implementing two-stage exams is sufficient and it is not necessary to collect and analyze the data because, as several studies have demonstrated, student response to the exams is generally positive. However, if the purpose of two-stage exams is to estimate aspects of student collaboration in small groups, an important component of many active learning strategies, then it strikes me as important to gain as much information as possible. Doing so requires going

beyond simple comparisons of the proportion of correct answers on the two stages. I developed a quantitative framework for extracting relevant information from two stages exams about the effects of 1) student knowledge estimated by the proportion of individuals correct during stage one, 2) specific questions, and 3) group composition. Additionally, the quantitative results revealed lots of variation in student performance both within and among exams, suggesting that two-stage exams, like most activities that require student interaction, elicit complex interpersonal dynamics.

At first pass, the various visualization generated from the data may trigger productive reflection and revision of teaching and learning practices for the educator and students, respectively. For instance, the picture of the distribution of individual scores across groups (e.g. figure 2) might be useful for the educator as an estimate of the variability of the cognitive challenge posed by the set of questions; these data also can be shared with students as a means of emphasizing the value of preparing for an exam. Similarly, visualizations of predicted values for different groups (e.g. figure 3) can provide educators with comparative data for identifying high- and low-collaborative groups that may prompt interventions for both highlighting the behavioral characteristics of high-performing groups and creating the basis for a reflective meeting with students in low-performing groups. Finally, compilations of R50, PC, or CPI can provide information about the collaborative value of a two-stage exams. These data might be particularly relevant for analysis of the same set of questions over time in the context of revisions of teaching strategies or curriculum revision, or for evaluating different exams within or among classes. For example, I highlighted three questions from exam 11 with variable effects as a first step towards data-driven and productive revision of the question, curricula, or teaching strategies that might improve both the individual and group scores. After whatever revision happens, the quantitative profiles generated using the quantitative framework for the pre- and post-revision can be compared to assess the effects of data-driven revision.

4.7 Evidence of the Complexity of Students Working in Small Groups

This study was motivated by recognition of the value of collaboration as a core undergraduate learning goal. There is a general push to adopt active learning strategies in classrooms which often involves asking students to work in small groups and develop consensus answers that make sense and demonstrate understanding. Student interactions in small groups can create complex dynamics in ways that can both enhance and inhibit learning. Studies of behavioral consistency indicate individual interactions vary as a consequence of changes of context (i.e. the topic and social milieu), interaction partners (e.g. the ratio of males to females, racial identities), and the behavioral characteristics of focal individuals (Funder 2006). This triad of interaction variables can generate significant variation in the behavior of individuals working in small groups (Funder 2009); heterogeneity that may underlie the variable effects of questions and groups within and among exams documented in this study. The discovery of large but inconsistent effects of questions and groups coupled with published work on social identity effects on student interactions in small groups (Eddy et al. 2014, 2015; Eddy and Hogan 2014; Buchenroth-Martin et al. 2016; Grunspan et al. 2016) underscores instructors need to pay particular attention to what students are being asked to do and encourage behaviors that foster productive interactions in a socially engaging and rewarding environment. The instructor can, for instance, emphasize evidence-based warrants in debates about claims (Knight *et al.* 2013) and provide cues and models of productive interaction emphasizing mutual respect, listening with intention, valuing peers, and being prepared to contribute towards the products of group work. Quantitative estimates of collaboration with question and group included as random effects may be useful for data-driven revision of curriculum, identifying pairs of questions that result in high collaboration scores and counter-productive interactions, and identifying groups that may benefit from interventions with the purpose of improving student performance. This latter issue is particularly important because the existence of counter-productive interaction may signal conflict among students within groups. In these cases, when the data suggest counter-productive interactions happened, it may be important to observe particular groups and record interaction dynamics, talk with students individually and ask whether they feel comfortable and valued, emphasize the characteristics of effective group work using behavioral modeling, and perhaps change up group membership in a way that yields more productive groups.

4.8 Conclusions

Applying a quantitative framework for the analysis of individual and group performance enables estimating gains in collaboration useful for assessment of students, emphasizing metacognition, and data-driven revision of curricula and teaching strategies. The value of quantitative estimates of collaboration using two-stage exams depends, of course, on what an instructor wants to learn about their teaching or student performance. Because one of my main goals in my classes was to create opportunities for practicing collaboration, I used two stage exams as means for both promoting and assessing this core competency. My discovery the data can be used for quantitative estimation of

collaboration opens the door for future studies aimed at testing the effects of altering curricula and teaching strategies to achieve high level interactions underlying the emergence of co-created innovations (Chi et al. 2016) from peers working in small groups towards relevant goals. The real challenge now is creating relevant and informative assessments.

Acknowledgments

The work was supported, in part, by support from the Arts and Science Support for Education Through Technology (ASSETT). Several anonymous reviewers provided comments and suggested reading that improved (IMO) the manuscript, Anne-Marie Hoskinson first introduced me to two-stage exams, and the Science Education Initiative and several of my colleagues (Nancy Emery, Lisa Corwin, Brett Melbourne, Kendi Davies, Nichole Barger, Jenny Knight, Sarah Wise, and Amanda McAndrew) helped me become a better educator. To all I am grateful.

References

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal Statistical Software, 67*, 1-48. https://doi.org/10.18637/jss.v067.i01

Bruno, B. C., Engels, J., Ito, G., Gillis-Davis, J., Dulai, H., Carter, G., Flectcher, C. & Bottjer-Wilson, D. (2017). Two-stage exams: A powerful tool for reducing the achievement gap in undergraduate oceanography and geology classes. *Oceanography, 30*. https://doi.org/10.5670/oceanog.2017.241

Buchenroth-Martin, C., DiMartino, R, & Martin, A. P. (2017). Measuring student interactions using networks: Insights into the learning community of a large active learning course. *Journal of College Science Teaching, 46*, 90-99. https://doi.org/10.2505/4/jcst17_046_03_90

Burnham, K. P., Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd edn. Springer, New York.

Chi, M. T. H., Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219-243. https://doi.org/10.1080/00461520.2014.965823

Chi, M. T. H., Kang, S., & Yaghmourian, D. L. (2016). Why students learn more from dialogue than monologue videos: Analyses of peer interactions. *Journal of the Learning Sciences, 46*(1), 10-50. https://doi.org/10.1080/10508406.2016.1204546

Cohen, E. (1994). Restructuring the classroom: conditions for productive small groups. *Review of Educational Research, 64*, 1-35. https://doi.org/10.3102/00346543064001001

Cortright, R. N., Collins, H. L., Rodenbaugh, D. W. & DiCarlo, S. E. (2003). Student retention of course content is improved by collaborative-group testing. *Advances in Physiological Education, 27*(4), 102-108. https://doi.org/10.1152/advan.00041.2002

Crowe, A., Dirks, C. & Wenderoth, M. P. (2017). Biology in Bloom: Implementing Bloom's taxonomy to enhance student learning in biology. *CBE-Life Sciences Education, 7*, 368-381. https://doi.org/10.1187/cbe.08-05-0024

Dolmans, D. J. H. M., Wolfhagen, I. H. A. P., van der Vleuten, C. P. M. & Wijnen, W. H. F. (2001) Solving problems with group work in problem-based learning: hold on to the philosophy. *Medical Education* 35: 884–889. https://doi.org/10.1046/j.1365-2923.2001.00915.x

Dolmans, D. H. J. M. & Schmidt, H. K. (2006). What do we know about cognitive and motivational effects of small group tutorials in problem-based learning. *Advances in Health Sciences Education, 11*, 321-336. https://doi.org/10.1007/s10459-006-9012-8

Durham, M. F., J. K. Knight & Couch, B. A. (2018). Instrument for scientific teaching (MIST): A tool to measure the frequencies of research-based teaching practices in undergraduate science courses. *CBE-Life Sciences Education, 16*, 1-14. https://doi.org/10.1187/cbe.17-02-0033

Eddy, S. L., Brownell, S. E. & Wenderoth, M. P. (2014). Gender gaps in achievement and participation in multiple introductory biology classrooms. *CBE-Life Sciences Education, 13*(3), 478-492. https://doi.org/10.1187/cbe/13-10-0204

Eddy, S. L., Brownell, S. E., Thummaphan, P., Lan, M. C. & Wenderoth, M. P. (2015). Caution, student experience may vary: social identities impact a student's experience in peer discussions. *CBE-Life Sciences Education, 14*, ar45. https://doi.org/10.1187/cbe.15-05-0108

Eddy, S. L. & K. A. Hogan. (2014). Getting under the hood: How and for whom does increasing course structure work? *CBE-Life Sciences Education, 13*(3), 453-468. https://doi.org/10.1187/cbe.14-03-0050

Fengler, M., & Ostafichuk, P. M. (2015). Successes with two-stage exams in mechanical engineering. *Proceedings 2015 Canadian Engineering Education Association Conference, 15*, 1-5. https://doi.org/10.24908/pceea.v0i0.5744

Funder, D. C. (2006). Towards a resolution of the personality triad: Persons, situations and behaviors. *Journal of Research in Personality, 40*, 21-34. https://doi.org/10.1016/j.jrp.2005.08.003

Funder, D. (2009) Persons, behavior and situations: An agenda for personality psychology in the postwar era. *Journal of Research in Personality, 43*, 120-126. https://doi.org/10.1016/j.jrp.2008.12.041

Gilley, B. H. & Clarkston, B. (2014). Collaborative testing: Evidence of learning in a controlled in-class study of undergraduate students. *Journal of College Science Teaching* 43(3): 83-91. https://doi.org/10.2505/4/jcst14_043_03_83

Grunspan, D. Z., Eddy, S. L., Brownell, S. E., Wiggins, B. L., Crowe, A. J. & Goodman, S. M. (2016). Males under-estimate academic performance of their female peers in undergraduate biology classrooms. *PlosOne, 11*(2), e0148405. https://doi.org/10.1371/journal.pone.0148405

Jang, H., Lasry, N., Miller, K., & Mazur, E. (2017). Collaborative exams: Cheating? Or learning? *American Journal of Physics, 85*, 223-227. https://doi.org/10.1119/1.4974744

Knierim, K., Turner, H., & Davis, R K. (2015). Two-stage exams improve student learning in an introductory geology course: logistics, attendance, and grades. *Journal of Geoscience Education, 63*, 157-164. https://doi.org/10.5408/14-051.1

Knight, J. L, Wise, S. B., Southard, K. M. (2013). Understanding clicker discussions: Student reasoning and the impact of instructional cues. *CBE Life Sciences Education, 14*(1), 1-12. https://doi.org/10.1187/cbe.13-05-0090

Laverty, J. T., Underwood, S. M., Matz, R. L, Posey, L. A. et al. (2016). Characterizing college science assessments: The three-dimensional learning assessment protocol. *PLoS ONE, 11*(9), e0162333. https://doi.org/10.1371/journal. pone.0162333

Leight, H., Saunders, C., Caikins, R. & Withers, M. (2012). Collaborative testing improves performance but not content retention in a large-enrollment introductory biology class. *CBE-Life Sciences Education, 11*(4), 392-401. https://doi.org/10.1187/cbe.12-04-0048

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education, 93*, 223-231. https://doi.org/10.1002/j.2168-9830.2004.tb00809.x

Reiger, G., & Heiner, C. W. (2014). Examinations that support collaborative learning: The students' perspective. *J. College Science Teaching, 43*, 41-47. https://doi.org/10.2505/4/jcst14_043_04_41

Sandahl, S. S. (2010). Collaborative testing as a learning strategy in nursing education. *Nursing Education Perspective, 31*(3), 142-147.

Smith, M. K., Jones, F. H. M., S. L. Gilbert, & Weiman, C. E. (2013). The classroom observation protocol for undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *CBE-Life Sciences Education, 12*, 618-627. https://doi.org/10.1187/cbe.13-08-0154

Springer, L., Stanne, M. E. & Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of Educational Research, 69*, 21-51. https://doi.org/10.3102/00346543069001021

Stearns, S. (1996) Collaborative exams as learning tools. *College Teaching, 44*(3), 111-112. https://doi.org/10.1080/87567555.1996.9925564

Summer, M., & Volet, S. (2010). Group work does not necessarily equal collaborative learning: evidence from observations and self-reports. *European Journal of Psychology Education, 25*, 473-492. https://doi.org/10.1007/s10212-010-0026-5

Weiman, C. E., Rieger, G. W. & Heiner, C. E. (2014). Physics exams that promote collaborative learning. *The Physics Teacher, 52*. https://doi.org/10.1119/1.4849159

Zimbardo, P. G., Butler, L. D., & Wolfe, V. A. (2003). Cooperative college examinations; More gain, less pain when students share information and grades. *Journal of Experimental Education, 71*, 101-125. https://doi.org/10.1080/00220970309602059

Zipp, J. F. (2007). Learning by exams: The impact of two-stage cooperative tests. *Teaching Sociology, 35*(1), 62-76. https://doi.org/10.1177/0092055X0703500105