# Two-Stage Collaborative Exams have Little Impact on Subsequent Exam Performance in Undergraduate Mathematics

**Citation for published version:**
Kinnear, G 2020, 'Two-Stage Collaborative Exams have Little Impact on Subsequent Exam Performance in Undergraduate Mathematics', *International Journal of Research in Undergraduate Mathematics Education*. https://doi.org/10.1007/s40753-020-00121-w

**Digital Object Identifier (DOI):**
10.1007/s40753-020-00121-w

**Link:**
Link to publication record in Edinburgh Research Explorer

**Document Version:**
Peer reviewed version

**Published In:**
International Journal of Research in Undergraduate Mathematics Education

International Journal of Research in Undergraduate Mathematics Education
https://doi.org/10.1007/s40753-020-00121-w

# Two-Stage Collaborative Exams have Little Impact on Subsequent Exam Performance in Undergraduate Mathematics

George Kinnear[1]

Accepted: 7 August 2020Published online: 17 September 2020

## Abstract

In this paper, I investigate whether two-stage exams aid learning in undergraduate mathematics, as measured by students' performance on subsequent exams. In a two-stage exam, students complete the exam individually then form into groups to solve it again, with grades based on a combination of the two stages. Previous research in other disciplines has found mixed results about their effect on subsequent performance, and little is known about their use in undergraduate mathematics. Here, I report on three studies which investigate the use of two-stage exams in different undergraduate mathematics contexts. The first two studies replicate observational methods from previous research, and find indications of a positive impact from group collaboration. The third study is experimental and finds that, in a delayed post-test, there is no difference in performance between students who answered related questions in a two-stage exam format and a control group which had no collaborative second stage. The findings suggest that two-stage exams may have little impact on longer-term learning of mathematics, but instructors may still wish to use them to emphasise a collaborative classroom pedagogy.

**Keywords** Collaborative exams · Two-stage exam · Assessment · Evaluation

## Introduction

Exams are widely used to assess mathematics students in higher education. In a survey of modules offered at UK institutions, "nearly 70% ... use closed book examinations for at least three quarters of the final mark" (Iannone and Simpson 2012, p4). At the same time, active learning is increasingly seen as an important part of STEM education (Freeman et al. 2014), for instance through the use of flipped classrooms and peer instruction (Mazur 1997; Maciejewski 2015). The idea of a two-stage exam

✉ George Kinnear
G.Kinnear@ed.ac.uk

[1] School of Mathematics, University of Edinburgh, Edinburgh, UK

is to build a collaborative element into the assessment process, by having students take an exam individually, then solve the exam again in groups, with their grade determined by a combination of the scores at both stages. For a class which uses peer collaboration as part of the pedagogical approach, having a collaborative element as part of the assessment helps to emphasise its importance, and contributes to the constructive alignment of the course (Biggs and Tang 2011).

There are at least three reasons why adding a collaborative second stage to a traditional exam might be expected to benefit students' learning, and therefore their performance on a subsequent exam. First, the group discussion could function in a similar way to peer instruction where students benefit both from hearing explanations from their peers and from communicating their own understanding (Smith et al. 2009). Second, students may benefit from immediate feedback from their peers (Epstein et al. 2002). Third, students may have lower anxiety in two-stage exams due to having the opportunity to correct mistakes and benefit from the support of their peers, which may help them to perform better (Kapitanoff 2009).

On the other hand, including a collaborative second stage is likely to mean that there is less time available for the individual assessment, or that time is taken away from other learning activities. There may also be concern about students benefiting unfairly from the efforts of their peers (Jang et al. 2017). So, it is important to understand whether, on balance, the two-stage exam format has a positive effect.

Two-stage exams have been studied in a wide range of disciplines. I review these results in more detail in Section "Measuring Learning from Collaborative Exams". In short, the studies describe many positive aspects of two-stage exams, but findings about their impact on students' learning have been mixed. In fact, one study concluded "instructors who wish to use this method in their classes should determine, rather than assume, that it improves their students' learning" (Leight et al. 2012, p. 400). To date, there has been little investigation of the use of two-stage exams in mathematics, and the only study published so far "did not see better retention by the treatment groups" (Garaschuk and Cytrynbaum 2019, p1076).

Here, I present three studies in a range of contexts in undergraduate mathematics. These replicate the design of studies from other disciplines, to investigate the effects of two-stage exams in mathematics.

In the next section, I review the existing literature on two-stage exams and present the specific research questions that will be addressed in this paper.

## Investigating Collaborative Exams

### Assessment as Learning

The distinction between "assessment *of* learning" and "assessment *for* learning" is often made, and resonates with the distinction between the summative and formative purposes of assessment (Wiliam and Black 1996). Additionally, we may consider "assessment *as* learning": this term has at least three different interpretations. In the education literature, it is most commonly used to refer to the process of students developing metacognitive, self-assessment skills (Earl 2004). For instance, Sadler (1989) suggests teachers should aim for students to "eventually become independent of the teacher and intelligently engage in and monitor their own development" (p141). A second interpretation casts "assessment as learning" in a pejorative sense, "where assessment procedures and practices come completely to dominate the learning experience, and 'criteria compliance' comes to replace 'learning' " (Torrance 2007, p282).

A third, more positive, interpretation of "assessment as learning" is inspired by findings from cognitive science, which show that assessment activity can be used to enhance learning. For instance, the well-established *testing effect* shows that "testing not only measures knowledge, but also changes it, often greatly improving retention of the tested knowledge" (Roediger and Karpicke 2006, p181). These ideas are starting to influence course design in higher education, under the name "test-enhanced learning" (Brame and Biel 2015), and the use of two-stage assessments is an example of this.

## Measuring Learning from Collaborative Exams

In seeking to measure learning, we should be mindful that "Performance is what we can observe and measure during instruction or training. Learning—that is, the more or less permanent change in knowledge or understanding that is the target of instruction—is something we must try to infer" (Bjork and Bjork 2014, p57). Furthermore, in mathematics education, distinctions can be made between instrumental and relational understanding (Skemp 1976), or between conceptual and procedural knowledge (Rittle-Johnson et al. 2015). Our ability to measure learning is therefore partial, but we can use performance on tests and exams as a proxy. Indeed, this is the approach taken in the body of existing research on collaborative exams, which I shall now review.

The simplest approach to measuring the effect of two-stage exams on students' learning is to compare the assessment scores from the first and second stages. Perhaps unsurprisingly, there tends to be a significant improvement in performance from stage one (individuals) to stage two (groups). This is a consistent finding in various studies across disciplines (e.g. Cortright and et al. (2003), Giuliodori et al. (2008), Zipp (2007)), but presents weak evidence in favour of the approach having an effect on individuals' learning.

A more sophisticated approach to measuring learning is to introduce a third assessment stage, taken individually, in order to determine whether the improved performance in the second stage is sustained over time and without the help of a group. This is the approach taken in the studies reported later in this paper. In particular, I replicate aspects of the design used by Zipp (2007), where students took four exams during the semester followed by a cumulative final exam (made up of questions from the previous exams), all given in two-stage form. Students' individual performance in the cumulative final was significantly better if they had been in a group which answered the same question correctly in the second stage rather than incorrectly, and independent of how the individual answered in the first stage. Zipp posits that the group discussion can help to reinforce correct answers, and that "participating in a group that answered a question correctly helped students who initially got a question wrong to learn the material well enough to transfer it to the final exam" (Zipp 2007, p73).

An enhanced approach can be used when there are multiple assessment events during the course – this enables a crossover design, where the assessments can be given in traditional or two-stage form, and students swap between these two conditions for the different assessment events. Such a crossover design was used in an introductory biology class by Leight et al. (2012), who found that students in the two-stage condition did not perform significantly better than those in the traditional condition, on a subsequent test of the same material. Gilley and Clarkston (2014) used a similar design, with all students taking a retest at stage two in either 'group' or 'individual' conditions, in order to control for time on task and the testing effect. They found that only the 'group' condition had a significant boost in performance at stage two, and that this level of performance was maintained in the third stage. Cao and Porter (2017a) replicated this design and similarly found that the 'group' condition performed significantly better in a quiz two weeks later. However, the difference had disappeared by the final exam (three or seven weeks later).

A crossover design can also be implemented during a single assessment event by using different versions of the same assessment. Ives (2014) created three conditions by producing three different group exam papers for use in stage two, each containing a different subset of the original questions. Additionally, the questions in the third stage were different questions on the same concepts ("near transfer") in order to "remove the confounding factor of question recall" (p2). This was repeated in two midterms, with the finding that, when answering questions in the third stage test, students performed significantly better if they had discussed the question in the second midterm (1-2 weeks before), but not if they had discussed it in the first midterm (6-7 weeks before). Together with the result of Cao and Porter (2017a), this suggests the learning gain from two-stage exams is short-lived.

## Group Dynamics

If it were simply the case that weaker students were copying the strongest student in their group, then it would be expected that a substantial number of these stronger students would not improve their scores. However, Leight et al. (2012) found that "the vast majority (93.9%) of students scored higher on their group exams than on their individual exams" (p395). Moreover, Garaschuk and Cytrynbaum (2019) identified that, out of 36 groups studied, 15 performed better than the top individual.

Looking in more depth at the dynamics within groups, Jang et al. (2017) compared the performance of groups based on how many members gave a correct response at stage one. They found that, among groups where no students had the correct answer, 25% produced a correct answer at stage two, while groups with two or more students who had the correct answer were successful 96% of the time. Levy et al. (2018) took this further, defining a measure of 'collaborative efficiency' which shows how much the group gained at stage two as a proportion of the maximum possible gain based on combining the best of their stage one responses. The study found that groups attained an average of 67% of the maximum possible gain, suggesting that the group collaboration was "substantially (albeit partially) effective in improving students' examination performances and taking advantage of the aggregate knowledge and skills of the group's individual members" (Levy et al. 2018, p8).

### Affective Factors

Much previous work on two-stage exams has discussed their positive effect on student attitudes. For instance, Kapitanoff (2009) found a reduction in test anxiety compared with a traditional exam, and Levy et al. (2018) found that around two thirds of students reported two-stage exams were less stressful than a normal exam. Other studies have also found students responding positively to various aspects of the two-stage exam format (Cortright and et al. 2003; Leight et al. 2012; Rieger and Heiner 2014).

However, Iannone and Simpson (2015) found that students' views about assessment practices are not uniform across disciplines, with mathematics students tending to "prefer traditional assessment methods over more innovative ones" (p1062). Thus, it remains to be seen whether mathematics students also have positive views of the two-stage format.

### Investigating Collaborative Exams in Mathematics

So far, only one study has investigated the impact of two-stage exams in mathematics (Garaschuk and Cytrynbaum 2019). That study used a crossover design, similar to Leight et al. (2012), and found the two-stage exam format "does not have a strong influence on long-term retention" (p1067) as measured by performance on an exam at the end of the course.

Here, I report on work that furthers this investigation. I address the following research questions:

1. Do two-stage collaborative exams have an effect on students' learning of mathematics, as measured by subsequent exam performance?
2. Does the pattern of answers in the individual and group stages provide any evidence of collaboration within the groups?
3. Do mathematics students have positive views about two-stage exams?

In the next section, I give an overview of the methodology that I will use across three different studies. These studies are then presented in Sections "Study 1: Widening-Access Summer School", "Study 2: Final Year Geometry Module", "Study 3: Second Year Probability Module".

## Methodology

To investigate the research questions, I undertook three separate studies in different contexts. The first was a widening-access summer school course that I taught. The two-stage format was introduced in 2015 and personal communication with those involved suggested it had been a success. The desire to make this judgment more rigorous led to the research project in subsequent years. The study reported in

Section "Study 1: Widening-Access Summer School" was conducted in 2017 and 2018. Further opportunities to study the effect of the exam format arose when a colleague subsequently decided to use the two-stage format in two courses which were being created/re-developed. One is a final-year pure mathematics course, where two stage exams were used as a regular part of an active classroom pedagogy (see Section "Study 2: Final Year Geometry Module"). The other is an introductory course in probability, where a mid-term exam was given in two-stage format, and the large class size made it feasible to implement a controlled study (see Section "Study 3: Second Year Probability Module".

The data analysis is conducted using Bayesian methods. One motivating factor is that, in Study 1, the initial sample size (with data from 2017 only) was too small to give conclusive results. There are many possible interpretations of this lack of evidence (Dienes 2014), so it is not possible to distinguish 'no evidence of an effect' from 'evidence of no effect'. Using null-hypothesis significance tests, it is not possible to "top up" with additional data and re-test at the same significance level, since the tests are sensitive to the stopping rule used to gather the data (Dienes 2011, p278). With a Bayesian approach, more data can be gathered to inform conclusions. This enabled me to gather further data in 2018 to reach a conclusive result. Furthermore, the Bayesian approach gives "a richly informative posterior distribution over the joint parameter space. Every parameter in the model has an exact credible interval" (Liddell and Kruschke 2018, p339); this enables us to look at the strength of evidence for differences between different groups of students, without needing to worry about corrections for multiple comparisons.

In particular, I take a parameter estimation approach. For instance, a parameter of interest will be the difference in the proportion of correct answers between control and intervention groups. In estimating these parameters, there is uncertainty:

> "In a Bayesian framework, the uncertainty in parameter values is represented as a probability distribution over the space of parameter values. Parameter values that are more consistent with the data have higher probability than parameter values that are less consistent with the data." (Kruschke 2018, p271)

The Bayesian approach takes an initial 'prior distribution' of probabilities and updates this in light of the data to produce a 'posterior distribution' of probabilities. Full details of the posterior distributions I obtain are given in the Appendix. To summarise these distributions in the text, I report the modal (i.e. most probable) value and the 95% highest density interval (HDI). The HDI is the range of values around the modal value which contain 95% of the mass of the posterior distribution, and has the property that no value outside the HDI is more credible than a value inside the HDI (Kruschke 2018). So, for instance, if the difference between experimental and control groups is estimated as +.21 with 95% HDI [.15, .26], this would be viewed as evidence in favour of the intervention since the HDI takes only positive values.

## Study 1: Widening-Access Summer School

The first study took place during a summer school for students in the local area who are set to take up the offer of a place at university. Students were invited to take part based on characteristics such as being the first in their family to attend university. The summer school ran full time for seven weeks, and students studied three courses in parallel. Results from each course were sent to the students' universities, so most students were committed to doing well on the courses. The mathematics course revised topics from high school mathematics (like algebra) and included topics which were new to most of the class (e.g. combinatorics and probability). The class met for 3 hours at a time, on Monday mornings and Thursday afternoons.

### Method

Students took a mid-course test during the class time on Thursday of week 5, which was given in two-stage form. This test did not count toward the result of the course, but students were motivated to take it seriously as practice for the final exam during week 7, which was worth 40% of the grade (the remainder of the grade was based on various items of coursework). In the mid-course test, the first stage was a 1-hour individually written exam consisting of 11 questions, and this was followed by the second stage where students formed into groups of three and solved four of the questions again. Students knew that, if their group score was better than their own, it would replace their individual score for that question. The exams were marked by a team of graduate students and were ready to be returned the following week. During the class on Monday of week 6 (before the exams were returned, and before any feedback was given), the students were asked to complete two of the questions again, individually and under exam conditions. This provides data at 'stage three' from which we can assess students' individual learning gains (RQ1). In addition, some questions on the final exam were matched to earlier questions (see Table 1 for details), giving data at 'stage four'.

The data also enables us to investigate the effectiveness of group collaboration (RQ2). A model for this comes from Levy et al. (2018) which defines a measure of the 'collaborative efficiency' of the group, and also from Jang et al. (2017) which looks for evidence of collaboration in the pattern of individual and group responses.

To assess students' views of the two-stage exam format (RQ3), a questionnaire was constructed and administered during the final class session (at which point the students had completed both the practice and final exams in two-stage form). This consisted of 10 statements with a 5-point agree/disagree scale. The statements were based on those used by Leight et al. (2012) (which were in turn a re-expression of those used by Cortright et al. 2003)) but the wording was simplified in many cases. The full set of questions is given in Table 2.

The same method was used in 2017 (30 students) and 2018 (22 students), and the data were combined for analysis. The combined data from 2017 and 2018 has 47 students who completed stages one and two, 38 at stage three and 44 at stage four, all of whom gave consent for their data to be included in the study.

**Table 1** Questions used in Study 1

| Question | Stages 1-3 | Stage 4 |
|---|---|---|
| 1a | Write in the form $ax^n$, where $a$ and $n$ are numbers: $\frac{(3x)^5}{9x^2}$ | - |
| 1b | Write in the form $ax^n$, where $a$ and $n$ are numbers: $2x^{-3}\left(4x^{\frac{1}{2}}\right)^2$ | Write $3x^{-2}\left(2x^{\frac{1}{3}}\right)^2$ in the form $ax^n$ where $a$ and $n$ are numbers. |
| 2a | 200 raffle tickets are entered in to a draw. In how many different ways could the first, second and third prizes be distributed? | A tourist plans to visit Scotland's 7 cities in turn over the course of their holiday, but they have yet to decide on the order in which to visit them. How many options can they choose from? Explain your reasoning. |
| 2b | In how many ways can a jury of twelve people be chosen from a group of eighteen potential jurors? | A game is played using Scrabble tiles which spell out the word LUCKYWORD. The tiles are placed in a bag, and the player draws out two of the tiles at random. The player wins £1 if their pair of tiles includes the letter K, and wins £5 if their pair is LW; otherwise they win nothing. How many possible pairs of tiles are there? |
| 2c | You have eight tiles, each showing one letter as follows: $\boxed{D}\ \boxed{O}\ \boxed{U}\ \boxed{G}\ \boxed{H}\ \boxed{N}\ \boxed{U}\ \boxed{T}$ In how many distinct ways can you arrange (permute) the tiles? | - |
| 11 | Given that $\sum_{k=1}^{n}(2k+1) = 675$, what is the value of $n$? | Given that $\sum_{k=1}^{n}(4k+3) = 592$, what is the value of $n$? |

**Table 2** Questionnaire statements, which were presented with a 5-point agree/disagree scale

| | |
|---|---|
| The collaborative exam process... | |
| 1 | was easy to understand |
| 2 | was more stressful than traditional exams |
| 3 | was fair |
| 4 | helped me to learn |
| 5 | made me less nervous about taking the exam |
| 6 | should be used in more courses |
| During the group discussions... | |
| 7 | I was an active participant |
| 8 | I learned from others in the group |
| 9 | every group member contributed |
| 10 | I felt more confident about the answers |

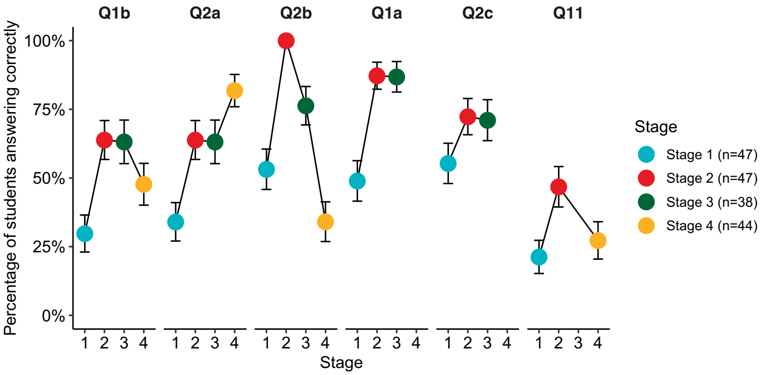


**Fig. 1** Percentage of students answering correctly at each stage (error bars show standard errors). Only stage 2 was taken in groups. Questions at stages 1-3 were identical; these were matched to similar stage 4 questions on the same topic

## Results

The percentage of correct responses for each question and stage is shown in Fig. 1. The mean scores at each stage show a similar pattern to that found in previous studies (e.g. Gilley and Clarkston 2014, Figure 3). In particular, group performance at stage two tends to be higher than individual performance, and this is mostly maintained by individuals answering the same questions at stage three. The individual performance at stage four, which used related questions on the same skills, shows mixed results: on Q2a, the performance is higher than at stage two, while on Q1b, Q2b and Q11 it has decreased from stage two (with performance on Q2b below the level at stage one).

While there is no control group, I can begin to address RQ1 by replicating the analysis of Zipp (2007). To do this, the student responses are split into four groups based on whether they are correct at stages one and two. I then consider the proportion of correct responses at stage three from each group, as shown in the third column of Table 3. I then model the data assuming the value in each cell is drawn from a binomial distribution with probability of success $\theta_i$, $i = 1, 2, 3, 4$, and use a Bayesian approach[1] to estimate the values of the $\theta_i$.

## Retention

Following Zipp (2007), I compare entries with the same stage one correctness, as this gives some measure of the possible impact of the group discussion. The expectation is that students will perform better at stage three if they are in a group which was correct at stage two. This is indeed the case: for students who were incorrect at stage one, the proportion of correct answers is higher at stage three when the stage two

[1] This was based on the `bayes.prop.test` function from the *BayesianFirstAid* R package (http://sumsar.net/blog/2014/06/bayesian-first-aid-prop-test/), but the model was modified to use a Beta(2,2) prior (rather than the default uniform prior) based on the understanding from previous uses of the test that students' mean scores tended to be around 50%, so the $\theta_i$ are likely to be close to 0.5.

**Table 3** Proportion of correct responses by individuals at Stages 3 and 4 (as a percentage and as a raw proportion) based on correctness at Stages 1 and 2

| Stage 1 | Stage 2 | Stage 3 | Stage 4 |
|---|---|---|---|
| Incorrect | Incorrect | 29.4 (10/34) | 31.5 (17/54) |
| | Correct | 73.4 (47/64) | 49.2 (30/61) |
| Correct | Incorrect | 40.0 (2/5) | 75.0 (3/4) |
| | Correct | 89.7 (78/87) | 59.6 (34/57) |

response was correct (the most probable value of $\theta_2 - \theta_1$ is +.40, with 95% HDI [.23, .58]); similarly for students correct at stage one (+.46, 95% HDI [.13, .74]). Looking at the overall contrast based on correctness at stage two (i.e. combining the two 'incorrect' and two 'correct' groups, regardless of their stage one performance), the 'correct' group has a higher proportion of correct responses at stage three (+.43, 95% HDI [.23,.59]). Considering only stage one correctness, there is no evidence of a difference in proportions. The contrast gives the difference between 'incorrect' and 'correct' as +.17 but the 95% HDI is [-.04, .32] so a zero or negative difference would be equally credible.

A similar approach was used for the data from stage four, shown in the final column of Table 3. Here, the Bayesian proportion test gives the difference between $\theta_1$ and $\theta_2$ as +.18 (95% HDI [.01,.33]), while $\theta_3$ to $\theta_4$ is -.08 (95% HDI [-.36,.30]). The first result is positive, but since one end of the HDI is very close to 0, we cannot rule out a zero or small negative effect with great confidence. The second result is negative, but the HDI is very wide so the only firm conclusion we can draw is that more data is needed.

## Group Dynamics

There are 90 instances of a group responding to a question at stage two, and, following Jang et al. (2017), I analyse these with respect to the number of students in the group who were correct at stage one (see Fig. 2). Of the 90 group responses, 20 (22%) were from groups where no student was correct at stage one, and these groups gave the correct answer 20% of the time. This shows that there are occasions when the group collaboration leads to correct answers being generated. In a further 35 groups (39%), there was only one student who had the correct answer at stage one and, among these groups, 80% gave the correct answer at stage two. This suggests that groups were generally able to discuss and identify that this response was the correct one.

Another way to analyse this data is suggested by Levy et al. (2018), who define the 'collaborative efficiency' as the proportion of the possible gain actually attained by the group. The possible gain (referred to as the "super surplus" by Levy et al. (2018, p5)) is computed as the difference between the mean individual score of the students in the group and the score of the 'super student', which is the sum of the maximum score on each question in the group. The 'super student' represents the score that

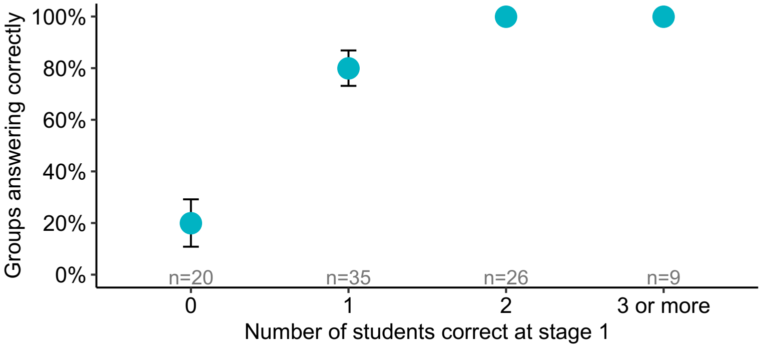


**Fig. 2** Percentage of groups answering correctly at stage 2, based on the number of students in the group who were correct at stage 1 (error bars show standard errors). The number of groups in each case is shown on the axis

would be possible at stage two if the students were able to share their correct answers perfectly in collaboration (though higher scores are possible, e.g. if all students got one question wrong at stage one and they collaborate to find the correct answer at stage two). The mean collaborative efficiency of the groups in this study was 0.95, suggesting that the groups were able to achieve almost all of the possible gain. Two groups had collaborative efficiency higher than 1, providing further evidence that the group discussions can generate answers which are better than those achieved by the individuals at stage one.

## Student Views

The survey was completed by 48 students, and the results are shown in Fig. 3. This shows an overwhelmingly positive response to the exam format (bearing in mind that Q2 is phrased negatively).

## Discussion

The results at stage three are consistent with the findings of Zipp (2007), with students more likely to give a correct answer if they were in a group which was correct at stage two. This suggests that the group discussions are informing students' learning, at least as far as the stage three test which took place four days after the group discussion. Retention was measured over a longer time interval, with stage four consisting of matched questions on the exam 11 days after stages one and two. However, the Bayesian analysis gave an inconclusive result suggesting that more data is needed. A possible reason is that there is some evidence of a question effect. Unlike other questions, Q2b saw a drop in performance at stage three compared with stage two. Performance on this question also fell at stage four, where the questions used were not identical but were on the same topics. The stage four variant may have higher difficulty due to being part of a more contextual question (see Table 1).

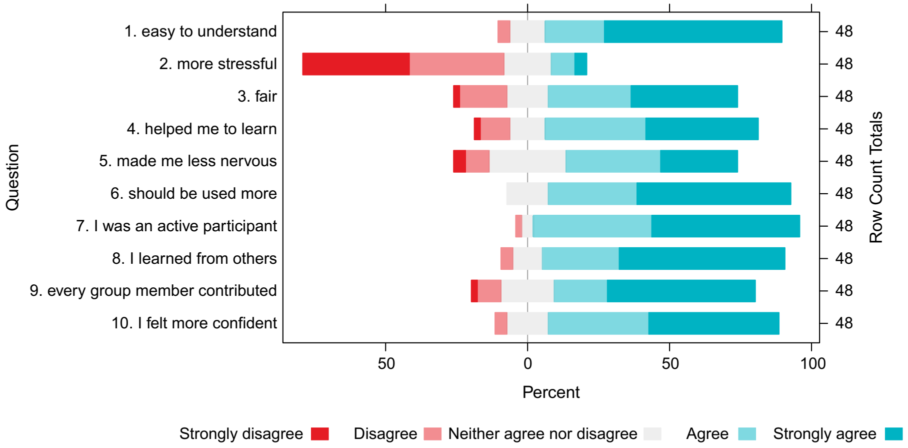


**Fig. 3** Percentage of respondents giving each response

Student views were measured through a survey. The results are consistent with findings in other disciplines in suggesting that most students found the exam format less stressful than a traditional exam.

## Study 2: Final Year Geometry Module

Further study became possible when a colleague decided to use two-stage exams in a new module on geometry, offered as an option to final-year students. The course uses a flipped classroom pedagogy, and two-stage exams were used each week as a quiz on topics from the reading. Such frequent use of two-stage exams is an interesting pedagogical choice and may show how they can be used in course design to foster student collaboration. A revision quiz in the final week was composed of questions from previous weeks, providing a natural ‘stage three’ measure which enables a similar analysis to study 1.

### Method

Each week[2] students individually completed a quiz consisting of five multiple-choice questions, before forming into groups and solving the questions again. The groups used special scratchcards[3] to record their answers, so that groups could try again if their initial answer was incorrect. Students received a mark out of 20 for each quiz and the best 9/11 scores contributed 10% to the course grade. The score out of 20

[2] except week 4, when there was a problem with the printed test papers

[3] “Immediate Feedback Assessment Technique” or “IF-AT” scratchcards (Epstein et al. 2002). These are multiple-choice answer forms where the student indicates their answer by scratching off a particular position on the scratchcard. The scratchcards are pre-printed with stars in certain positions to indicate correct answers.

came from the sum of the individual scores (5 questions, 2 marks each) and the group scores (5 questions, 2 marks for correct answer on first attempt, 1 mark for correct answer on second attempt). The final quiz in week 11 consisted of 10 questions which had been asked in previous quizzes. This quiz was also a two-stage exam. So, in addition to individual results at stage three, there are group results at stage four.

Groups normally consisted of four students who were assigned to work together throughout the semester. However, there was some fluidity to account for absences so groups sometimes had three or five members (and on a couple of occasions, two members). By the end of the course, 21 students were enrolled, although data was collected from a total of 24 students over the course of the semester.

## Results

### Retention

The results, shown in Fig. 4, are much more variable than in Study 1. The group performances (at stages two and four) do appear to be typically better than the corresponding individual performances. This is consistent with previous findings, but it is not uniformly true (Q2 stage 1-2, Q6/Q9 stage 3-4). The retention from stage two to stage three also appears to be highly variable, with some seeing large gains (Q2) and others seeing large falls (Q3, Q5).

There are 130 complete cases of attempts at stages 1-3 and they are summarised in Table 4. While this shows the same pattern of differences in the raw proportions as in Table 3 from the previous study, there is insufficient evidence to support the claim that the proportions in each cell are different. The Bayesian analysis gives 95% HDIs for the differences between each proportion (see Appendix), and all of these HDIs contain 0, except for the difference between the first and last (+.31, 95% HDI [.11, .49]). Looking at the overall contrast based on correctness at stage two

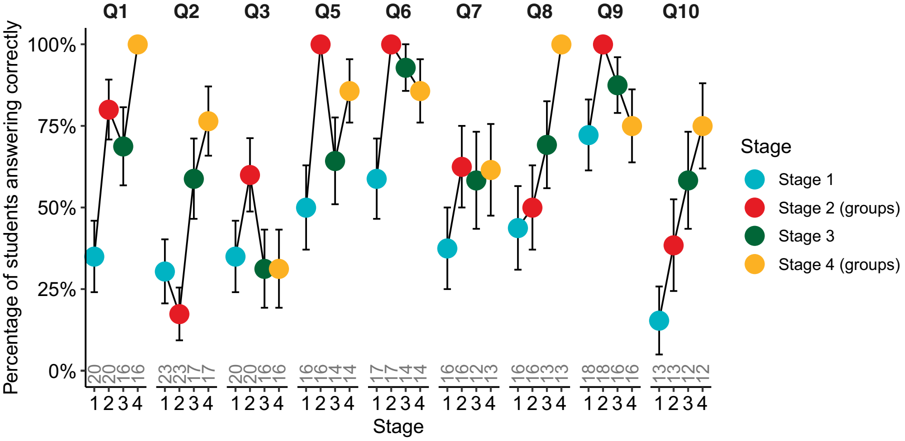


**Fig. 4** Percentage of students answering correctly at each stage (error bars show standard errors). The number of students is shown below each point. Stages 2 and 4 were taken in groups

**Table 4** Proportion of correct responses by individuals at Stage 3 (as a percentage and as a raw proportion) based on correctness at Stages 1 and 2

| Stage 1 | Stage 2 | Stage 3 |
|---|---|---|
| Incorrect | Incorrect | 47.2 (17/36) |
| | Correct | 63.2 (24/38) |
| Correct | Incorrect | 66.7 (4/6) |
| | Correct | 80.0 (40/50) |

(i.e. combining the two 'incorrect' and two 'correct' groups, regardless of their stage one performance), the 'correct' group appears to have a higher proportion of correct responses at stage three (+.16). While there is considerable uncertainty about this (the 95% HDI is [-.02, .35] which spans 0), over 95% of the credible mass is towards a higher proportion so there is tentative evidence in this direction.

## Group Dynamics

There are 250 instances of a group responding to a question at stage two. Of these, 29 (12%) were from groups where no student was correct at stage one, and these groups gave the correct answer 17% of the time, with a further 24% giving the correct answer on the second attempt (as shown in Fig. 5). This shows that there are occasions when the group collaboration leads to correct answers being generated. This is also reflected in the collaborative efficiency (Levy et al. 2018) of the groups: 7 of the 46 groups had collaborative efficiency larger than 1 and the mean across all groups was 0.86.

Out of the 250 group responses, there were 96 instances of groups having no single modal response among the students' stage one responses, and in these cases the group response was correct 78% of the time. When there was a modal response, the group did not always use it. There were 110 instances of the group giving the modal response, and this was correct 84% of the time. There were 44 instances of the group deciding not to use the modal response – in 35 of these cases, the modal response was

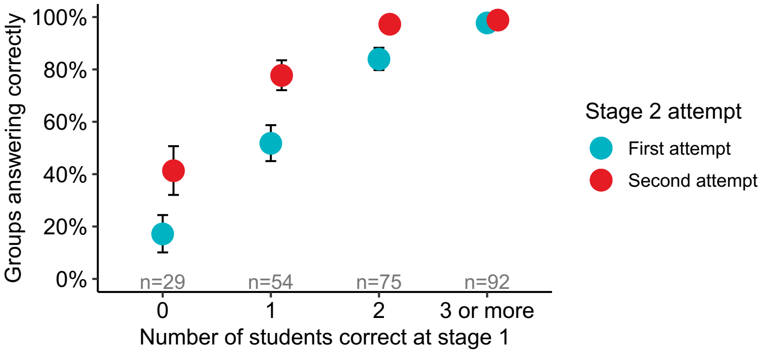


**Fig. 5** Percentage of groups answering correctly at stage 2 on their first and second attempts, based on the number of students in the group who were correct at stage 1 (error bars show standard errors). The number of groups in each case is shown on the axis

in fact incorrect, and the alternative group response was correct 54% of the time. This is further evidence that the groups genuinely engage in discussion and collaborate to produce the group response, rather than simply repeating answers from stage one.

### Discussion

The analysis of success at stage three as a function of stages one and two showed a similar pattern to the previous study but the results were inconclusive here, most likely because of the small sample size. This is reflected in the Bayesian analysis by the wide HDIs on the posterior distributions.

The variable retention from stage two to stage three could be due to the topics covered by each question, however a post-hoc analysis of the questions (see Appendix) does not suggest any clear explanation. Future exploratory work could seek to investigate possible explanations based on the topics or types of questions. Other factors, such as the relative emphasis placed on the different topics in the course between stages two and three, are likely to be important but were not controlled for.

## Study 3: Second Year Probability Module

A two-stage exam was also added to an introductory course in probability, for students in their second year of study. The class is large (301 students), and taught as a flipped classroom with interactive engagement in lectures. The two-stage exam was used for a midterm in week 6 (of 11), and took place in the normal lecture theatre during a 50 minute period. This is similar to the way that two-stage exams are used with large classes at other institutions (Wieman and Gilbert 2014).

This presented an opportunity to replicate the crossover design used by Ives (2014), by giving different subsets of questions to groups at stage two. Groups were thus in the experimental condition for questions which they discussed at stage two, and in the control condition for questions which were omitted from their stage two exam paper.

### Method

The midterm consisted of six questions: half of these were multi-stage questions which required the students to answer either two or four sub-parts. Examples of the questions are given in the Appendix. The questions were permuted into 16 different random orders, and these different versions of the test were distributed randomly to students to mitigate against the potential for copying in the lecture theatre venue. Students were given 25 minutes to complete these questions and write their responses onto the question sheet. The question sheets were collected and were subsequently hand-marked by the students' regular tutors, with each question scored out of four.

Students formed into self-selected groups to complete stage two, resulting in 65 groups of 4 and 7 groups of 3. The stage two paper consisted of four of the six questions (two of the three single-response questions, and two of the three multi-part questions but with only the final answer asked for) which were presented as multiple-choice questions. There were nine different variants of the stage two paper, and these were distributed randomly to groups. Students had 15 minutes to complete stage two, with answers given on IF-AT scratchcards. The answer options for Q6 were in the wrong order so the IF-AT cards gave misleading feedback to students; for this reason, Q6 was removed from the analysis. Each question at stage two was scored out of two marks, with one mark available for responses which were correct on the second try. The stage two score counted for 30% of the midterm, unless the student's stage one score was better on its own.

Stage three was provided by the final exam, taken just under seven weeks after the midterm. The lecturer identified exam questions which tested the same concepts. In some cases, these were sub-parts of exam questions so all scripts were checked for the appropriate working.

The midterm was completed by 281 students, of whom 254 gave consent for their data to be included in this study (though 3 of these students did not go on to take the final exam).

## Results

The results are shown in Fig. 6. These show the typical pattern that group performance at stage two is higher than individual performance at stage one. They also show gains from stage two to stage three on Q3/Q4/Q5 but not on Q1/Q2. It appears that there is not much difference between the experimental and control groups, and to check this rigorously I used an extended version of the Bayesian proportion test from the preceding sections. This models the proportion of correct answers expected at stage three for the six groups shown in Table 5. The contrast for the key comparison, between experimental and control groups, shows the proportion for the experimental group is -0.02 (95% HDI [-0.104,0.025]) relative to the control group – this suggests

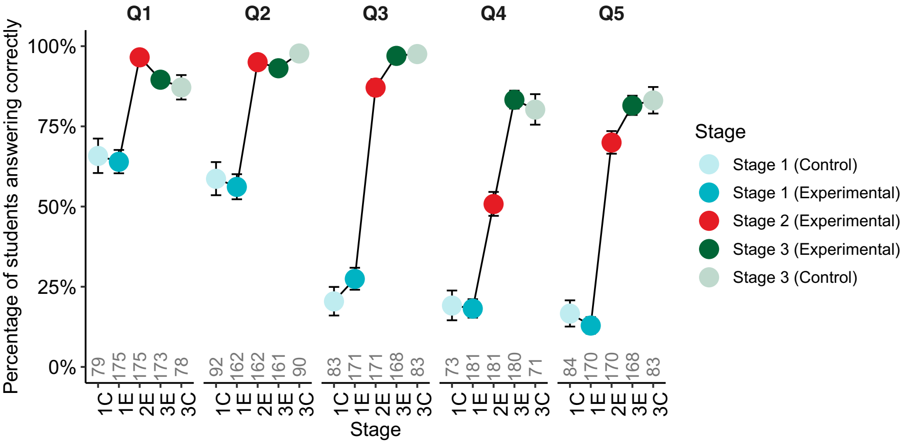


**Fig. 6** Percentage of students answering correctly at each stage (error bars show standard errors). The number of students is shown below each point. Stages 1 and 3 were individual, and stage 2 was taken in groups (by the experimental group only)

**Table 5** Proportion of correct responses by individuals at Stage 3 (as a percentage and as a raw proportion) based on correctness at Stages 1 and 2

| Group | Stage 1 | Stage 2 | Stage 3 |
|---|---|---|---|
| Experimental | Incorrect | Incorrect | 83.9 (141/168) |
| | | Correct | 87.1 (330/379) |
| | Correct | Incorrect | 100.0 (7/7) |
| | | Correct | 93.6 (277/296) |
| Control | Incorrect | - | 87.1 (222/255) |
| | Correct | - | 94.0 (141/150) |

the experimental group did slightly less well at stage three, although the 95% HDI spans 0 so a small positive effect is also credible. Overall, it appears the stage two discussion had a negligible effect on performance at stage three.

## Group Dynamics

There are 229 instances of a group responding to a question at stage two. Of these, 69 (30%) were from groups where no student was correct at stage one, and these groups gave the correct answer 39% of the time, with a further 23% giving the correct answer on the second attempt (as shown in Fig. 7). This shows that there are occasions in which the group collaboration leads to correct answers being generated.

The collaborative efficiency (Levy et al. 2018) of the groups was very high, with the mean being 1.46 and the value being greater than 1 for 60% of groups. This likely reflects the relatively low stage 1 scores on Q3-5 which saw substantial improvement at stage two.

## Student Views

The same questionnaire from Study 1 was used in an online survey which students were invited to complete at the end of the course, 5 weeks after the midterm. The

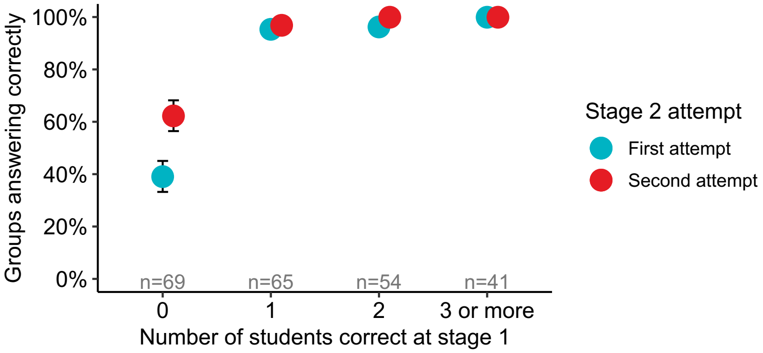


**Fig. 7** Percentage of groups answering correctly at stage 2 on their first and second attempts, based on the number of students in the group who were correct at stage 1 (error bars show standard errors). The number of groups in each case is shown on the axis

survey was completed by 32 students (out of 254 students who participated in the study). This low response rate means the results cannot be regarded as representative of the whole class. Nevertheless, there is a clear picture from the results that the students' views are less positive than in Study 1 (Fig. 8). The survey also provided a space for free-text comments[4] and the comments left there shed light on the reasons behind this: almost all of the comments mentioned a lack of time to complete stage one, and many students also commented that they found it hard to work as a group in the lecture theatre.

## Discussion

Overall, there was no difference between the experimental and control groups in their subsequent exam performance. This is consistent with previous findings (Ives 2014; Cao and Porter 2017a). One possible reason for this lack of effect is that students who were incorrect at stage one may have sought out more help or revised this topic, regardless of whether it was discussed at stage two. Additionally, any effect from the two-stage exams may have been eclipsed by the effect of general exam revision (and students may well have used the midterm as indicative of important topics to revise for the exam). This issue has been identified in previous research on the testing effect, with Wooldridge et al. (2014) noting the hope that reduced confidence after a test "propels students to devote additional efforts toward learning the information" (p220). A further possible reason for the lack of difference between control and experimental groups is the ceiling effect observed in some of the stage three questions; with the control group achieving near perfect scores, there was little room for the experimental group to do better (Fig. 6).

The implementation of two-stage exams in this case was different to the previous studies, with them taking place as a one-off activity in a lecture theatre rather than as part of other group work taking place in the course. As in other experimental work, the available evidence does not rule out the possibility that "a longer and differently designed intervention" (Jones et al. 2019, p12) would produce an effect. Moreover, feedback from students suggests that the available time for this exam was too short, and this will have impacted on the nature of discussion students were able to have at stage two – several students reported not having enough time for stage one, so they will have had little to bring to the group discussion of some questions. Some indication of this can be seen in the relatively high 'collaborative efficiency' scores, as noted in Section "Group Dynamics", which show that groups were able to go well beyond what the students had collectively achieved in stage one.

It is interesting to note that Fig. 7 is very similar to Figure 3 of Jang et al. (2017), which is interpreted as "all students (on average) gain from collaborating" (p226). Yet, the present study found that the two-stage exam had no effect on learning (as measured by a delayed post-test). This underlines the importance of including experimental methods in the range of methods used to evaluate interventions (Alcock et al. 2013).

[4] "If you have any further comments about the two-stage test format, please share them here."

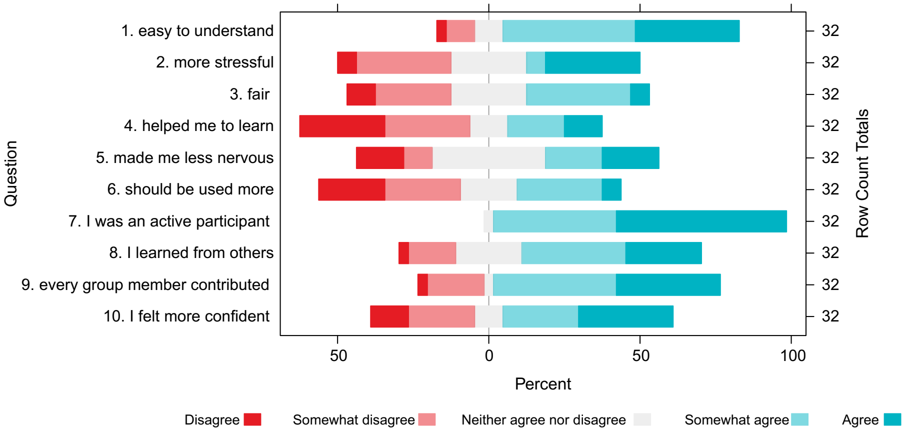


**Fig. 8** Percentage of respondents giving each response

## General Discussion

Across three studies, I investigated the effectiveness of two-stage exams in mathematics. While in each case there was some evidence of improved group performance, and some indications of collaboration within groups, the most compelling result comes from the controlled experiment in Study 3 which found that there was a negligible impact on students' longer-term learning, as measured by their performance on a subsequent exam.

In studies 1 and 2, the small class sizes limited the strength of evidence available, although the Bayesian approach made it possible to gather more data in study 1 to clarify the results. The data on student views could also have been more comprehensive: while there was a high response rate (and overall positive reaction) in Study 1, this was not the case in Study 3. So, while there is some indication that mathematics students view two-stage exams in a similar way to students in other disciplines, further work should be done to confirm this.

The results were also affected by the reliance on a small number of questions, which in practice had varying levels of difficulty. In particular, some of the questions in Study 3 had very high scores and this ceiling effect may have obscured any difference between the experimental and control groups. Future work should seek to use questions of a suitable difficulty level; indeed, there are some suggestions in the literature that the best questions for group discussion in two-stage exams are the more difficult ones (Jang et al. 2017), although this has not been systematically tested.

Different patterns of results across stages 1-3 for different questions suggest that there may be a 'question effect' on the effectiveness of the two-stage format, with certain questions more amenable to group discussion and to prompting individual student learning. For mathematics in particular, it could be the case that more procedural questions offer less opportunity for fruitful discussion than conceptual questions – for instance, there may be a different effect for questions in two-stage exams based on their categorisation in the MATH taxonomy (Smith et al. 1996), or depending on whether the questions are multiple-choice or constructed-response (Sangwin and

Jones 2017). This was not something I set out to explore in the studies reported here, and a post-hoc analysis of the questions based on whether they saw gains following group discussion does not reveal any obvious pattern. It may also be the case that the two-stage format is effective for some topics and not others. These issues could be systematically explored in further exploratory work.

There are three additional aspects which could be investigated in more depth in further work. First, the suggestion that any performance gains decay over time (Ives 2014), which could be investigated using a similar design to that used in Study 2 but with a larger class and multiple questions each week. Second, the effect of group formation on results could be explored, following the surprising suggestion that "student groups in collaborative exams should be formed to be homogeneous" with respect to students' prior performance (Cao and Porter 2017b, p157); in particular, a qualitative study based on close observation of the group discussions could investigate the influence of interpersonal factors on learning. Third, the effectiveness of variations on the method could be investigated, such as adding a phase of instructor explanations either after or instead of the second stage (Smith et al. 2011).

These results add to a growing set of findings that two-stage exams have a negligible effect on students' long-term learning (Leight et al. 2012; Ives 2014; Cao and Porter 2017a; Garaschuk and Cytrynbaum 2019). This is despite some indication that the group discussion does involve genuine collaboration and generation of novel answers, which has been viewed in other studies as an indicator of the success of the method (Jang et al. 2017; Levy et al. 2018). Of course, the negligible effect can also be viewed as not causing harm, so lecturers may still wish to use two-stage exams (which by many accounts are popular with students) in order to foster a collaborative pedagogy.

**Additional Information** Raw data and analysis scripts associated with this paper can be found at https://doi.org/10.17605/OSF.IO/SXR9U.

## Appendix

### Test materials for Study 2

1. An infinite chess board (with the squares alternating between black and white in the usual way) gives a plane pattern with which signature?
   (A) ∗222 (B) ∗2222 (C) ∗442 (D) 2∗22 (E) 4∗2

2. The point group of the wallpaper group 22∗ is:

   (A) $D_1$ (just the identity and a reflection)

(B) $D_2$
(C) A group containing just the identity and rotation by $\pi$
(D) A 1-element group (i.e. just the identity)
(E) A group containing just the identity, a reflection and a rotation by $\pi$.

3. The spherical symmetry group $2{*}3$ has how many elements?
(A) 12 (B) 9 (C) 6 (D) 4 (E) 3

5. Take a Klein bottle, add a handle and punch three holes in it. The resulting surface has Euler characteristic
(A) -8 (B) -7 (C) -6 (D) -5 (E)-4

6. Consider a proper circle (i.e. a circle which is not a line) in $\mathbf{C}_\infty$ with a line passing through its centre. What is the best description of the possible configurations after applying a Möbius transformation?

(A) Two circles with one passing through the centre of the other
(B) A proper circle and a line passing through its centre
(C) A proper circle and a line not necessarily passing through its centre
(D) Two circles meeting orthogonally
(E) Two circles meeting but not necessarily orthogonally

7. The orbifold Euler characteristic of the orbifold for $*552$ is
(A) $-1/20$ (B) $-1/10$ (C) $1/10$ (D) 1 (E) 2

8. The curve in H corresponding to the segment of the x-axis in D is

(A) the x-axis
(B) the imaginary axis within H
(C) the line $y = 1$ in H
(D) the part of the circle $|z - i| = 1$ within H
(E) the part of the circle $|z| = 1$ within H

9. The generators and relations $\alpha^2 = \beta^2 = \gamma^2 = \delta^2 = \alpha\beta\gamma\delta = 1$ define which symmetry group:
(A) $*2222$ (B) $2{*}22$ (C) $22{*}$ (D) 2222 (E) $**$

10. The group $2{*}$ has how many 2-colourings? (A) 4 (B) 3 (C) 2 (D) 1 (E) 0

### Test materials for Study 3

Question 1

Stage 1. A gate on a footpath starts each day closed. After that, each person using the gate behaves as follows. If the gate was closed when they arrived at it, they close it behind them with probability 2/3. If however the gate was open when they arrived, they close it behind them with probability 1/3. What is the probability that the third person to use the gate on a given day leaves it open? (Answer exactly.)

Stage 2. (A) $\frac{1}{2}$, (B) $\frac{9}{27}$, (C) $\frac{13}{27}$, (D) $\frac{15}{27}$, (E) None of the other answers is correct

Stage 3. Each day independently I decide to travel to work either by car or by bus. I choose the car with probability 4/5 and the bus with probability 1/5. If I travel by car, I arrive on time 90% of the time and if I take the bus I arrive on time 50% of the time.

One day, I arrive at work late. What is the probability that I took the bus?

*Here, responses were checked for the intermediate calculation of* $\mathbb{P}(late) = \frac{9}{50}$

Question 2

Stage 1. A random variable X on the interval [0, 1] has density function

$$\mathrm{f}_{\mathrm{X}}(x) = kx^2(1-x)$$

where k is a constant you should determine. Find the expected value and variance of X. *Here, the mark for the variance was used to represent correctness at stage 1, since this corresponds to stage 2.*

Stage 2. What is the variance of X? (A) None of the other answers are correct, (B) $\frac{1}{25}$, (C) $\frac{2}{5}$, (D) $\frac{1}{5}$, (E) $\frac{1}{300}$

Stage 3. A random variable X on [0, 1] has pdf $kx^2$, where $k$ is a constant you should determine. Use an approximation by a normal distribution to approximate the probability that the mean of 10 independent samples from this distribution is between 0.8357 and 0.8602 (A table of values for $\Phi$ can be found below.)

*Here, responses were checked for the intermediate calculation of* $\mathrm{Var(X)} = \frac{3}{80}$

Question 3

Stage 1. You are talking with your friend on your mobile phone as you both arrive at bus stops in separate cities where (independently) buses arrive according to a Poisson distribution at a rate of one every 4 minutes. What is the probability that after 5 minutes, at least one bus has arrived at both your bus stops. Answer accurate to at least three decimal places.

Stage 2. The answer is in the interval: (A) $(0.5, 0.52]$ (B) $(0.52, 0.54]$, (C) $(0.54, 0.56]$, (D) $(0.46, 0.48]$, (E) $(0.48, 0.50]$

Stage 3. During a thunderstorm, lightening flashes are occurring randomly at an average rate of one per minute. What is the probability of there being exactly three flashes in two minutes?

Question 4

Stage 1. In the usual sense of a limit,

(a) what is the probability that a random positive integer is divisible by at least one of 3, 4 and 6;

(b) what is the probability that a random positive integer is divisible by two of the numbers 3, 4 and 6 but not by all three?

*Here, the mark for (b) was used to represent correctness at stage 1, since this corresponds to stage 2.*

Stage 2. In the usual sense of a limit, what is the probability that a random positive integer is divisible by exactly two of the numbers 3, 4 and 6 but not by all three?

(A) None of the other answers is correct, (B) $\frac{1}{3}$, (C) $\frac{1}{4}$, (D) $\frac{1}{6}$, (E) $\frac{1}{12}$

Stage 3. In the usual sense of a limit, what is the probability of a random natural number being divisible by at least one of 6 and 10?

Question 5

Stage 1. I roll two D3. Let X be the minimum of the two rolls. What is the expected value of X and what is its variance?

Suppose now that I roll a pair of D3 five times and let Y be the mean (i.e. the average value) of the resulting five "2-dice minimums". What is the expected value of Y and what is its variance?

*Here, the mark for (b) was used to represent correctness at stage 1, since this corresponds to stage 2.*

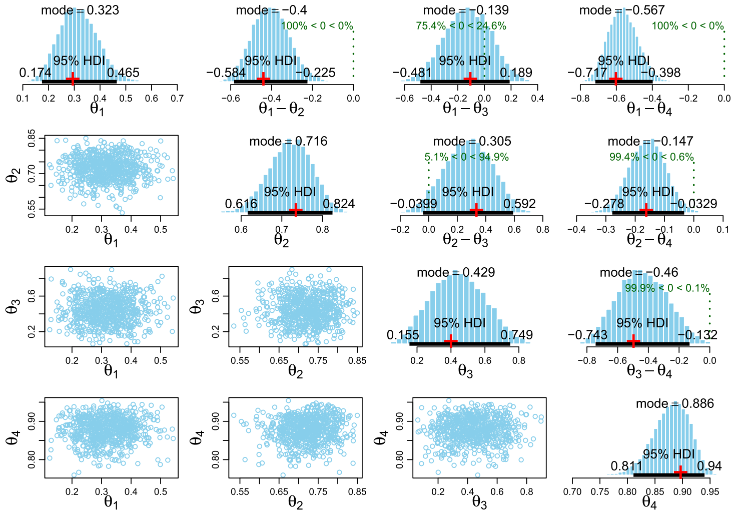
Stage 2. The variance of Y is: (A) $\frac{4}{3}$, (B) $\frac{26}{9}$, (C) $\frac{7}{12}$, (D) $\frac{38}{405}$, (E) None of the other answers is correct

Stage 3. I have three cards numbered 1, 2, 3. I choose a card uniformly randomly and call the number on it X. I then randomly choose one of the two remaining cards and call the number on it Y. Compute the covariance of X and Y and also their correlation.

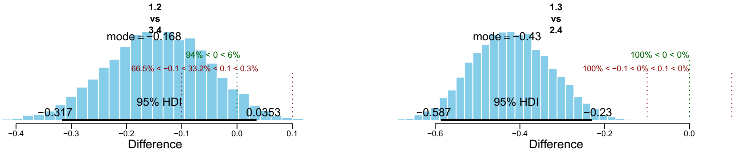
*Here, responses were checked for the intermediate calculation of* $\mathbf{E}(XY) = \frac{11}{3}$

## Bayesian analysis: Study 1

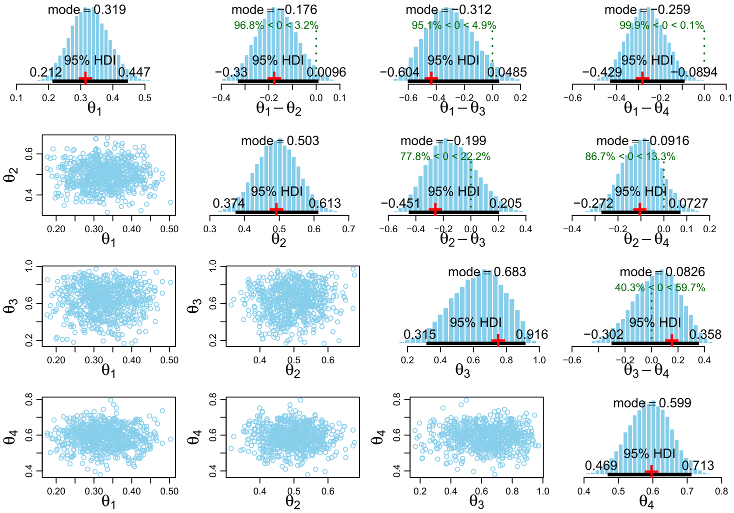
Posterior distributions for the parameters $\theta_i$ (giving the probability of correctness at stage three for group $i$ in Table 3) and differences between these. Observed values are marked with a red + on the axis.



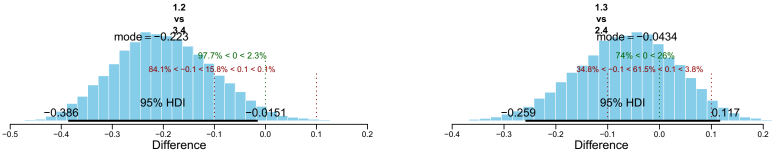
Contrasts between combinations of cells:



Posterior distributions for the parameters $\theta_i$ (giving the probability of correctness at stage four for group $i$) and differences between these.

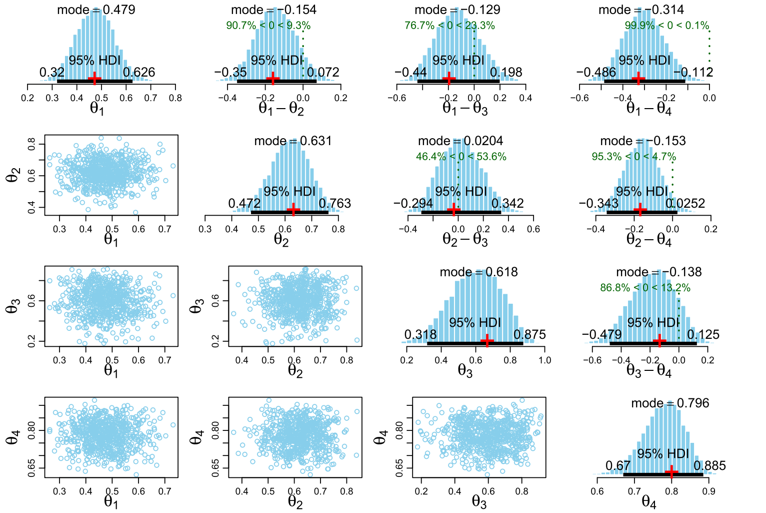


Contrasts between combinations of cells:

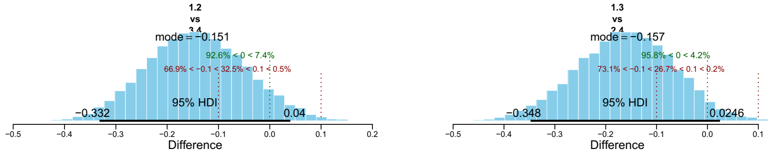


## Bayesian Analysis: Study 2

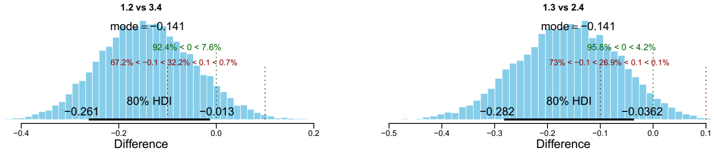
Posterior distributions for the parameters $\theta_i$ (giving the probability of correctness at stage three for group $i$ in Table 4) and differences between these. Observed values are marked with a red + on the axis.



Contrasts between combinations of cells:

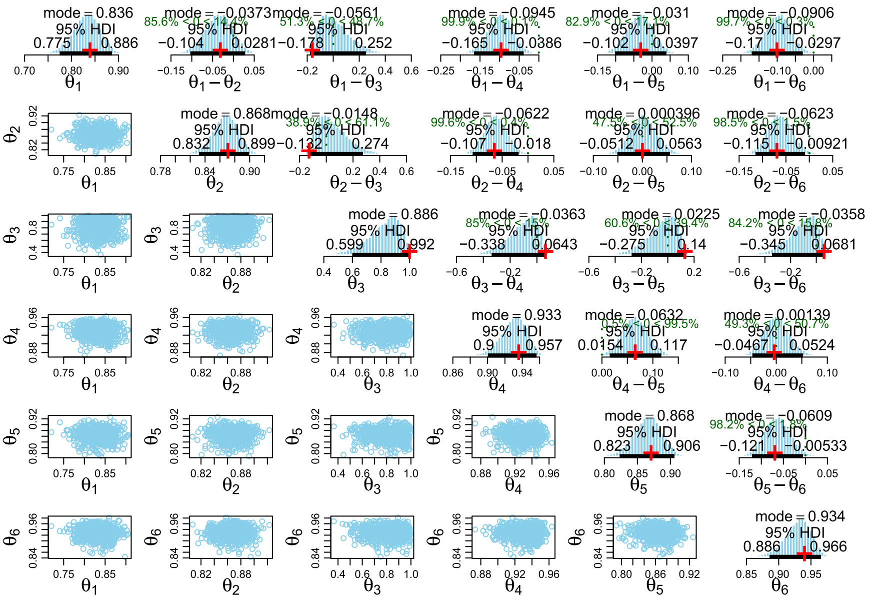


Contrasts with 80% HDIs:

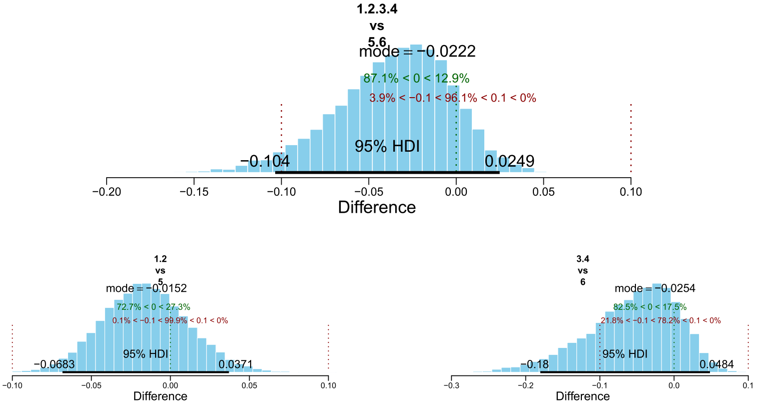


## Bayesian Analysis: Study 3

Posterior distributions for the parameters $\theta_i$ (giving the probability of correctness at stage three for group $i$ in Table 5, with the Control group split by correctness at stage one into groups $i = 5$ for incorrect and $i = 6$ for correct) and differences between these. Observed values are marked with a red + on the axis.



Contrasts in Table 5:



## References

Alcock, L., Gilmore, C., Inglis, M. (2013). Experimental methods in mathematics education research. *Research in Mathematics Education*, *15*(2), 97–99.

Biggs, J., & Tang, C. (2011). Teaching for quality learning at university: what the student does. McGraw-Hill/Society for Research into Higher Education/Open University Press.

Bjork, E.L., & Bjork, R.A. (2014). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In Gernsbacher, M.A., & Pomerantz, J. (Eds.) *Psychology and the real world*. 2nd edn. (pp. 59–68). New York. Chap. 5.

Brame, C.J., & Biel, R. (2015). Test-enhanced learning: the potential for testing to promote greater learning in undergraduate science courses. *CBE Life Sciences Education*, *14*(2), 14:es4.

Cao, Y., & Porter, L. (2017a). Evaluating student learning from collaborative group tests in introductory computing. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education - SIGCSE '17* (pp. 99–104). New York: ACM Press.

Cao, Y., & Porter, L. (2017b). Impact of performance level and group composition on student learning during collaborative exams. In *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education - ITiCSE '17* (pp. 152–157). New York: ACM Press.

Cortright, R.N., et al. (2003). Student retention of course content is improved by collaborative-group testing. *AJP: Advances in Physiology Education*, *27*(3), 102–108.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781.

Earl, L.M. (2004). *Assessment as learning : using classroom assessment to maximize student learning*, (p. 132). Thousand Oaks: Corwin Press.

Epstein, M. et al. (2002). Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. *The Psychological Record, 52*(2).

Freeman, S. et al. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8410–5.

Garaschuk, K.M., & Cytrynbaum, E.N. (2019). Feasibility and effectiveness of group exams in mathematics courses. *PRIMUS*, *29*(10), 1061–1079.

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.

Giuliodori, M.J., Lujan, H.L., DiCarlo, S.E. (2008). Collaborative group testing benefits high- and low-performing students. *AJP: Advances in Physiology Education*, *32*(4), 274–278.

Iannone, P., & Simpson, A. (2015). Students' preferences in undergraduate mathematics assessment. *Studies in Higher Education*, *40*(6), 1046–1067.

Iannone, P., & Simpson, A. (2012). A survey of current assessment practices. In Iannone, P., & Simpson, A. (Eds.) *Mapping university mathematics assessment practices, Chap. 1* (pp. 3–15).

Ives, J. (2014). Measuring the learning from Two-Stage collaborative group exams. In *Physics education research conference 2014. PER Conference. Minneapolis, MN* (pp. 123–126).

Jang, H. et al. (2017). Collaborative exams: Cheating? Or learning? *American Journal Of Physics*, *85*(3), 223–227.

Jones, I. et al. (2019). Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? *British Educational Research Journal*, *45*(3), 662–680.

Kapitanoff, S.H. (2009). Collaborative testing. *Active Learning in Higher Education*, *10*(1), 56–70.

Kruschke, J.K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280.

Leight, H. et al. (2012). Collaborative testing improves performance but not content retention in a large-enrollment introductory biology class. *Cell Biology Education*, *11*(4), 392–401.

Levy, D., Svoronos, T., Klinger, M. (2018). Two-stage examinations: Can examinations be more formative experiences? In: Active Learning in Higher Education, p. 146978741880166.

Liddell, T.M., & Kruschke, J.K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348.

Maciejewski, W. (2015). Flipping the calculus classroom: an evaluative study. *Teaching Mathematics and its Applications*, *19*(4), hrv019.

Mazur, E. (1997). *Peer instruction: A user's manual*, (p. 253). Upper Saddle River: Prentice Hall.

Rieger, G.W., & Heiner, C.E. (2014). Examinations that support collaborative learning: The students' perspective. *Journal of College Science Teaching*, *43*(4), 41–47.

Rittle-Johnson, B., Schneider, M., Star, J.R. (2015). Not a one-way street: Bidirectional relations between procedural and conceptual knowledge of mathematics. *Educational Psychology Review*, *27*(4), 587–597.

Roediger, H.L., & Karpicke, J.D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210.

Sadler, D. (1989). Royce formative assessment and the design of instructional systems. *Instructional Science Instructional Science*, *18*(2), 119–144.

Sangwin, C.J., & Jones, I. (2017). Asymmetry in student achievement on multiple-choice and constructed-response items in reversible mathematics processes. *Educational Studies in Mathematics*, *94*(2), 205–222.

Skemp, R.R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching*.

Smith, G. et al. (1996). Constructing mathematical examinations to assess a range of knowledge and skills. *International Journal of Mathematical Education in Science and Technology*, *27*(1), 65–77.

Smith, M.K. et al. (2011). Combining peer discussion with instructor explanation increases student learning from In-Class concept questions. *CBE - Life sciences education*, *10*, 55–63.

Smith, M.K. et al. (2009). Why peer discussion improves student performance on in-class concept questions. *Science*, *323*, 5910.

Torrance, H. (2007). Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education: Principles, Policy & Practice*, *14*(3), 281–294.

Wieman, C., & Gilbert, S. (2014). The teaching practices inventory: a new tool for characterizing college and university teaching in mathematics and science. *CBE Life Sciences Education*, *13*(3), 552–69.

Wiliam, D., & Black, P. (1996). Meanings and Consequences: a basis for distinguishing formative and summative functions of assessment?. *British Educational Research Journal*, *22*(5), 537–548.

Wooldridge, C.L. et al. (2014). The testing effect with authentic educational materials: a cautionary note. *Journal of Applied Research in Memory and Cognition*, *3*, 214–221.

Zipp, J.F. (2007). Learning by exams: The impact of two-stage cooperative tests. *Teaching Sociology*, *35*(1), 62–76.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.