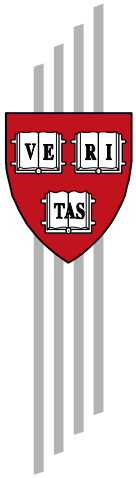
Two-Stage Examinations: Can Examinations Be More Formative Experiences?

Dan Levy, Mae Klinger and
Theodore Svoronos

CID Faculty Working Paper No. 363
September 2018



Working Papers

Center for International Development
at Harvard University

Article

active learning in higher education

Active Learning in Higher Education
1–16

DOI: 10.1177/1469787418801668
journals.sagepub.com/home/alh
SAGE

Two-stage examinations: Can examinations be more formative experiences?

Dan Levy
Harvard University, USA

Theodore Svoronos
Harvard University, USA

Mae Klinger
Harvard University, USA

Abstract

Two-stage examinations consist of a first stage in which students work individually as they typically do in examinations (stage 1), followed by a second stage in which they work in groups to complete another examination (stage 2), which typically consists of a subset of the questions from the first examination. Data from two-stage midterm and final examinations are used to assess the extent to which individuals improve their performance when collaborating with other students. On average, the group (stage 2) score was about one standard deviation above the individual (stage 1) score. While this difference cannot be interpreted as the causal effect of two-stage examinations on learning, it suggests that individuals experienced substantial performance gains when working in groups in an examination. This average performance gain was comparable with the average difference between the top performer of the group in stage 1 and the group's stage 1 average, and was equivalent to about two-thirds of the difference between the "super student" score (i.e. the sum of the maximum score for each question in stage 1) and the group's stage 1 average. This last result suggests that group collaboration takes substantial (albeit partial) advantage of the aggregate knowledge and skills of the group's individual members. Student feedback about their experience with two-stage examinations reveal that that these types of examinations are generally perceived to be more helpful for learning and are less stressful than traditional examinations. Finally, using data on group gender compositions, we investigate the potential role of gender dynamics on group efficiency.

Keywords

two-stage exam, collaborative learning, collaborative efficiency

Corresponding author:
Dan Levy, John F. Kennedy School of Government, Harvard University, 79 John F. Kennedy Street, Cambridge, MA 02138, USA.
Email: dan_levy@hks.harvard.edu

The promise of two-stage examinations

A two-stage examination is a type of learning assessment that can go beyond the evaluative purpose of traditional examinations and offers the potential to foster both learning and collaboration. It consists of an individual part, in which students complete their examinations as in traditional testing, followed by a group or collaborative part, in which students discuss a subset of the questions in the first stage and, after reaching consensus, submit their second stage responses as a group. The central idea is that, by fostering debate and the exchange of ideas in small groups within an examination setting, students are able to learn from their peers and ultimately improve their learning. The potential of two-stage examinations to improve learning is supported by foundational research on the importance of cooperative learning (Herrmann, 2013; Johnson and Johnson, 1999; Slavin, 1996; Springer et al., 1999) and timely feedback (Bransford et al., 2004; Gibbs and Simpson, 2005).

The literature on two-stage examinations and collaborative testing largely focuses on three different issues: performance, retention (i.e. knowledge recalled), and student anxiety. The premise in these studies is that two-stage examinations, because of their incorporation of a second component that includes collaboration, allows students to learn from their peers, filling each other's gaps in knowledge, while at the same time putting less pressure on the individual student. These studies also seek to answer the question of whether students are willing to collaborate and whether this collaboration benefits both high- and low-performing students.

The early literature on two-stage examinations and performance (Balch, 1992; Billington, 1994; Webb, 1993) identified positive gains on performance from two-stage examinations with the gains being largely concentrated among low performers in the first study. Later studies have confirmed the positive effects of two-stage examinations using experimental or quasi-experimental designs across different fields and student performance levels (Bloom, 2009; Gilley and Clarkston, 2014; Giuliodori et al., 2008; Leight et al., 2012; Meseke et al., 2008; Yuretich et al., 2001; Zipp, 2007). Wieman et al. (2014) found that this type of examination fostered collaboration and increased learning without hindering the assessment of individual performance that is the goal of traditional examinations. In other words, while students learned from each other, they did not become "free-riders." Instead, students learned from their mistakes and this in turn facilitated their retention. These positive effects are supported by a related literature of the effects of collaborative learning, an approach underpinning two-stage examinations. For example, Wieman and Perkins (2005) found that under traditional methods (standard lecturing techniques) students learned, on average, 30% less. Smith et al. (2009) found that the number of students that answered a question correctly improved after peer discussion. Deslauriers et al. (2011) showed that using instruction based on research in cognitive psychology and physics education increased both engagement and learning.

Giuliodori et al. (2009) found that students with incorrect responses switched their answers more often than students with correct responses, implying that group feedback helped students learn, particularly in the case of those with inadequate performance. This is in line with evidence of strong peer effects, where students benefit more from being grouped with stronger peers (De Paola and Scoppa, 2010). Meseke et al. (2009) controlled for differences in study habits and also found positive effects for collaborative testing. Cranney et al. (2009) analyzed whether a "testing effect," that is, repeated testing as a way of increasing retention of the material, was confounded with collaboration in a two-stage examination setting. By introducing both new and old questions in the second stage, they attempted to identify the effect of collaboration apart from repetition and found a positive effect.

However, there are also risks associated with collaborative testing. For example, Moore (2010) documented instances of both free-riding and intragroup conflict. Soetanto and MacDonald (2017)

documented the types of obstacles that students experience during group work activities, and concluded that these obstacles tend to increase over time and that different interventions prompted different patterns of obstacle development. Hall and Buzwell (2012) found that free-riding was the most common concern expressed by students regarding group work. Furthermore, the methods by which students form groups may result in suboptimal collaboration. For example, Freeman et al. (2017) demonstrated that self-selected groups tended to be homogeneous in terms of gender, ethnicity, and performance, while Takeda and Homberg (2014) showed that groups with more homogeneous gender representation exhibited lower levels of collaboration during group work, though this was not consistently associated with lower overall group performance. Weldon and Bellinger (1997) found that while group scores were higher than individual scores for each member, they appeared to be below the pooled results of the students (note that "pooled results" match to the concept of the super student we use in this article), which was taken as evidence of suboptimal collaboration. Such "collaborative inhibition" (Masanobu and Saito, 2004) can be explained by the above-mentioned free-riding, but the authors also suggested cognitive factors similar to "part-set" cueing, where cues from one group member disrupt other members' cognitive retrieval strategies. Such suboptimal collaboration was also observed by Takahashi and Saito (2004), but they also found that introducing a 1-week delay between the individual and group parts of the examination reduced the inhibition to collaborate. This is consistent with the results in Yu et al. (2010), which examined positive spillover effects from two-stage examination in midterm tests to final examinations, and Centrella-Nigro (2012), which assigned students to small groups to retake the same test.

Looking beyond performance, the effect of two-stage examinations on retention is mixed. Cortright et al. (2003) reported on a study of a group of students that, after being tested individually, completed the same tests in pairs. Four weeks later, students were rendered a new examination on the same material, and the students that worked in groups achieved an average score of 81.3% compared with 63.5% for the set of students that only worked individually. Similar outcomes were later replicated for groups of four to five students by Glass et al. (2013) and Rivaz et al. (2015). In the second case, however, group-induced retention was lower than the pooled results for the entire cohort, which was interpreted as a sign of some degree of inhibition; students, especially high-performing ones, did not collaborate as much as they could. On the other hand, Leight et al. (2012) found no statistically significant effect on retention for a similar design in which students were tested in groups immediately after they completed their individual examinations.

Research conducted on two-stage examinations has also explored the consequences that they have on student engagement and stress. Yuretich et al. (2001) found that interest in the class moved from an average of 3.3 to 3.5 on a 0–5 scale. Yu et al. (2010) found that three-quarters of students in a class reported positive attitudes toward the implementation of two-stage examinations. This appears to be a trait of collaborative testing more broadly. Martin et al. (2014) identified positive consequences from collaboration in one-stage testing, reporting that 83% of students stated that collaborative examinations increased their confidence in their own knowledge. Studies, such as Breedlove et al. (2004), however, found no significant difference in test anxiety between students who collaborate on their examination and students who work alone.

Finally, there remain several gaps in the existing literature on two-stage examinations. First, not much research has been done about students' perceptions of the usefulness and fairness of two-stage examinations. Second, while research has documented some of the learning gains of two-stage examinations, less work has been done around developing measures of effectiveness of student collaboration in the group stage of the examination. Finally, further research is needed to learn about the role of gender dynamics on group collaboration in two-stage examination settings.

Methodology

Two-stage examinations were administered 11 times to five successive cohorts of students between 2013 and 2017. The examinations were a part of the grading assessment for three courses at the Harvard Kennedy School of Government at Harvard University in the United States: the first-semester required statistics course (API-209) in the Master in Public Administration in International Development (MPA/ID); the first-semester required statistics course (API-201) in the Master in Public Policy (MPP); and the second-semester required econometrics course (API-210) in the MPA/ID. In 2013, students took a two-stage final examination while in 2014–2017, students participated in two-stage examinations for both the midterm and the final examination. The final examination in all the six cases was cumulative and included material already covered in the midterm examination.

The first stage involved an individual closed-book examination. Students were permitted to have two-page reference sheets consisting of statistical formulas and definitions of concepts prepared by them on their own. The first stage lasted 80 minutes for the midterm examinations and 120 minutes for the final examinations. Immediately after they turned in their examinations, students were asked to work in preassigned groups of four or five in order to complete a collaborative part (stage 2) comprising a subset of identical questions from the individual examination. Students were not allowed to check or review the individual examinations they had just turned in.

The second stage lasted for 35 minutes for the midterm examination and 55 minutes for the final examination. Hence, the ratio of time for the individual stage relative to the group stage was roughly 2–1, and so was the ratio of the number of questions between the stages. Each group submitted one examination and cross-group collaboration was prohibited. The students knew before the examination took place that there would be a group stage, but they did not learn about which group they were assigned to until the individual stage was over. At the end of stage 2, students completed an anonymous survey about their experience with the two-stage, collaborative examination format. The responses to this survey were reviewed and coded by two independent raters to identify key themes, and the average results and interrater reliability were calculated across the two raters.

The students were assigned randomly to the groups for the stage 2 examination, but were stratified to make sure that at least one student in the group was in the top 40% of performance prior to the examination. This was done to avoid some groups consisting of only low-performing students (by luck of the draw). Measurement of prior performance was based on the results from problem sets completed prior to the examination, in the case of the midterm, and performance in the midterm examination, in the case of the final.

Student performance data

The data on examination scores were organized in the following way for producing the key results: scores for questions that appeared in stage 1 of the examination but not in stage 2 were discarded, so that the scores from the two stages were directly comparable. The stage 1 score in the examination refers to the sum of the scores on the questions that were common across stages only. Then for each group in each examination, we calculated four numbers, expressed in a normalized score out of 100:

- *Individual average score*: average of the stage 1 scores for individuals in that group (i.e. the sum of group members' stage 1 scores divided by the number of group members).
- *Group score*: score that the group obtained in stage 2 of the examination.
- *Top student score*: highest stage 1 score in that group (i.e. we calculated stage 1 scores for all individuals in the group and picked the highest score).

- *Super student score*: sum of the highest stage 1 score for each question in the examination (i.e. for every question, pick the highest stage 1 score in the group and then add up all the highest scores for that group).

In addition to top and super scores, we also defined top surplus and super surplus (we called it a surplus and not a gain since the top and super scores were determined in stage 1, not stage 2, and they can be considered the surplus knowledge that the top and (hypothetical) super student brought to the group's stage 1 average) as the difference between the top or super score and the individual average score, expressed in exam-specific standard deviations (SDs).

Finally, we also introduced two outcome indicators, both at the group level:

- *Gain*: as an absolute measure of improvement, this is the difference between the second stage group score and the individual average score, measured in exam-specific SDs.
- *Collaborative efficiency*: coined in loose opposition to "collaborative inhibition" as a relative measure of improvement, this is how much a group "caught up" with the above-defined super student. It is the gain divided by the super surplus (i.e. a ratio with the difference between the second stage score and the individual average score in the numerator and the difference between the super student score and the individual average score in the denominator).

The measures above are summarized in Box 1.

Box 1. Key measures related to test scores in two-stage examinations.

Raw scores

Score obtained in question k by student i from group j in stage $s = Q_{kijs}$

Score for student i from group j in stage $s = X_{ijs} = \sum_{k=1}^{m} Q_{kijs}$

Processed scores

Individual average score$_j$ = $IAS_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij1}$

Group score$_j$ = $GS_j = X_{ij2}$

Top student score$_j$ = $TSS_j = MAX_{i=1,\ldots n_j} X_{ij1}$

Super student score$_j$ = $SSS_j = \sum_{k=1}^{m} MAX_{i=1,\ldots n_j} \left(Q_{kij1}\right)$

Measures of differences between stages 1 and 2

$Gain_j = GS_j - IAS_j$

$Top\_Surplus_j = TSS_j - IAS_j$

$Super\_Surplus_j = SSS_j - IAS_j$

Measures of collaboration

Collaborative efficiency = $\frac{Gain_j}{Super\_Surplus_j} = \frac{GS_j - IAS_j}{SSS_j - IAS_j}$

where n_j is the size of group j (generally 4 or 5) and m is the number of questions in the stage 2 examination (also generally around 4 or 5)

Student experience data

In addition to the analysis of student performance described above, qualitative student feedback data collected after all but the 2017 API-210 final examination were analyzed. This survey was filled out by 802 out of 830 students. While five questions were multiple-choice, three were open-ended feedback questions. This gave students the opportunity to provide open-ended qualitative comments, an opportunity which 355 of the 802 students (44%) took.

Two independent raters coded the students' responses, identifying key themes, and classified them as positive, neutral, or negative overall. To assess the level of agreement, two measures of interrater reliability were calculated:

- *Joint probability of agreement*, which is simply the percentage of the time that the two raters agree;
- *Cohen's Kappa*, a more robust measure that takes into account the level of agreement that would be expected to occur by chance.

Results

Table 1 provides the results of an analysis of student performance data, with the average of the normalized individual stage 1 and stage 2 group scores shown in columns (A) and (B), respectively. Individual and group performances are compared with the scores of the top student within each group, shown in column (C), and the super student score, shown in column (D).

Table 1. Results by class, year, and examination.

					Average score of				Differences in average scores* (in SDs)			
		Sample size			Stage 1	Stage 2	Stage 1	Stage 1				
Class	Exam	n	N	SD	Indiv [A]	Group [B]	Top [C]	Super [D]	Gain [B − A]	Top surplus [C − A]	Super surplus [D − A]	Collaborative efficiency
API 209	2013 Fin	73	18	13.8	69.5	84.2	83.0	90.5	1.1	1.0	1.5	0.70
	2014 Mid	70	18	19.2	67.3	83.9	82.2	93.9	0.9	0.8	1.4	0.62
	2014 Fin	70	18	16.1	66.1	80.7	83.2	90.4	0.9	1.1	1.5	0.60
	2015 Mid	60	15	18.6	74.6	93.0	89.3	96.3	1.0	0.8	1.2	0.85
	2015 Fin	60	15	13.0	75.0	87.5	86.5	92.2	1.0	0.9	1.3	0.73
	2016 Mid	73	18	21.8	57.9	80.9	81.8	89.4	1.1	1.1	1.4	0.73
	2016 Fin	74	18	14.6	71.3	83.2	86.0	93.3	0.8	1.0	1.5	0.54
API 201	2016 Mid	139	34	21.9	69.2	85.5	89.4	97.1	0.7	0.9	1.3	0.58
	2016 Fin	144	36	14.6	59.9	80.7	74.6	87.0	1.4	1.0	1.9	0.77
API 210	2017 Mid	67	17	17.9	65.2	82.1	81.8	90.0	0.9	0.9	1.4	0.68
	2017 Fin	69	18	16.0	63.6	80.0	80.5	89.7	1.0	1.1	1.6	0.63
All examinations		899	225	18.2	66.6	83.5	83.1	91.8	1.0	1.0	1.5	0.67

n = number of students; N = Number of groups.
*All differences are statistically significant, with t-statistics of 5 and higher.

Table 1 shows that, across all 11 examinations, groups outperformed individual scores by about 1 SD (column (B−A)). The magnitude of the difference represented around 18 points on a scale of 100, and is relatively large when compared with the performance gains associated with most education interventions. Moreover, the difference was similar to the average difference between the top student and the individual average (column (C−A)). While the group slightly outperformed individual members by a little more than the top student does, this difference of 0.041 SDs was not statistically significant. This result suggests that on average, each member of the group caught up completely with the top student, but did not advance further.

These results hold when looking at each examination separately. In fact, with the exception of the 2016 API-201 midterm and final examinations and the 2016 API-209 final examination, there was little variation in the difference between group performance and individual performance. The difference was between 0.9 and 1.1 SDs for all other examinations, even though individual standardized scores varied significantly from test to test depending on the difficulty of each examination. On the other hand, there was more variation in the degree to which group scores closed the gap between individual scores and the super student score, with the difference between columns (D−A) and (B−A) ranging between 0.2 and 0.7 SDs, corresponding to, respectively, 85% and 54% of this gap being closed. The former (85%) corresponded to the 2015 API-209 midterm examination, where group scores significantly outperformed individual scores and groups were able to closely match the super student score, meaning that collaboration was significantly more effective in that case compared with the other examinations. In contrast, the latter case (54%) was the 2016 final examination for API-209, where group scores only closed roughly half the gap with the super student score, suggesting that collaboration was less effective.

About four-fifths of students improved their grades, and the gain for the average student was significant, in both a statistical and a practical sense. But not all students gained equally (the distribution of gains at the individual level is shown in Figure 1): 22% of individuals had gains in scores of over 2 SDs or 36 points, 3% of individuals even had gains over 3 SDs or 54 points, 7% of students had the exact same score, and 14% scored lower than their individual stage 1 grade (note that individuals with group scores at or below their individual scores did not have their final examination grades affected in any negative way). Only 1.4% had a group score that was more than 1 SD worse than their individual score.

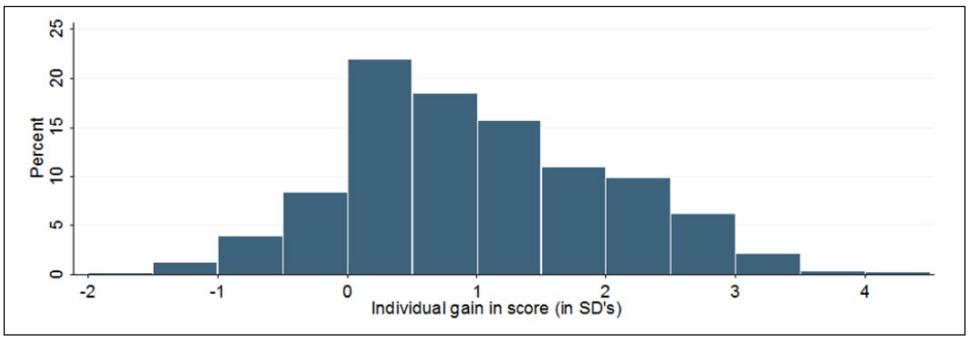


Figure 1. Individual improvements in score from individual stage 1 to group stage 2.

At the group level, Figure 2 reveals that only 13 out of 225 groups had a second stage group score that was lower than their first stage average group score, while 9 groups outperformed their first stage average by more than 2 SDs.

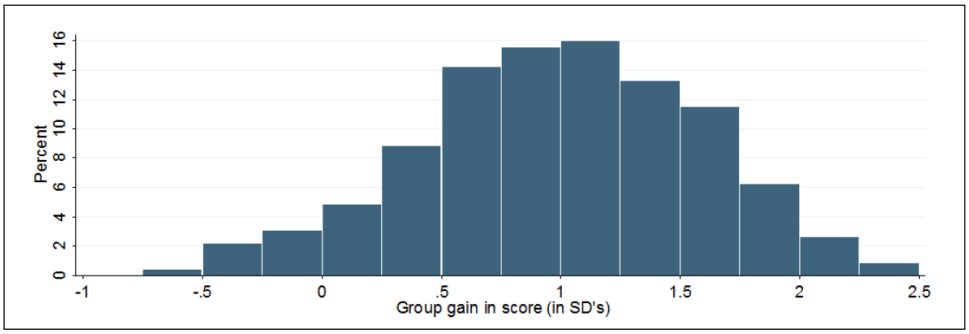


Figure 2. Group improvements in score from stage 1 to stage 2.

Still at the group level, Figure 3 gives the distributions of scores for (1) group average of stage 1 individual scores, (2) the stage 2 group scores, (3) top student scores, and (4) super student scores. This figure reveals a longer bottom tail in the distributions for the group score and the top score relative to the distributions for the individual average score and the super student score.

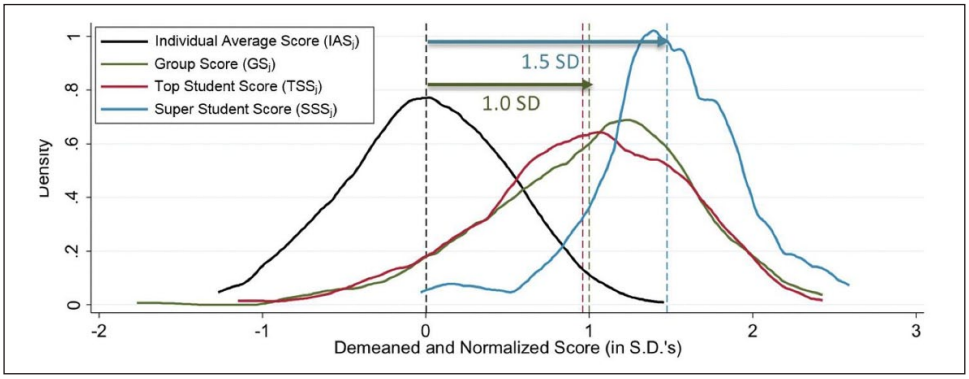


Figure 3. Distributions of stage 1 and stage 2 group scores.

The fact that on average the group score was below the super student suggests that there was still some room for improvement in how students collaborated, since the groups did not manage to replicate the best responses among their members for every question. As indicated earlier, we used a measure we call collaborative efficiency to assess how much the group "caught up" with the super student, that is, how much the group closed the gap between individual scores and the super student score. Collaborative efficiency is calculated as the gain divided by the super surplus (i.e. the ratio of (B–A) over (D–A) in Table 1), which was equal to 67% (i.e. 1 SD divided over 1.5 SDs). This result suggested that group collaboration in two-stage examinations were substantially (albeit partially) effective in improving students' examination performances and taking advantage of the aggregate knowledge and skills of the group's individual members.

While one might expect collaborative efficiency to have been (1) positive, as groups improved their performance and (2) below 1, as the most efficient groups successfully extracted all knowledge from the group members, the histogram in Figure 4 shows that some groups had a collaborative efficiency below 0 or others above 1. Instead of asking how much the group "caught up" with the super student (i.e. collaborative efficiency: gain divided by super surplus), we can also ask how

much the group "caught up" with the top student, a measure given by dividing the gain by the top surplus. This distribution, similar to Figure 4, is given in Figure 5. It tells us that, with the median at 1, half the groups did better than their top student and the other half did worse. Similar to our earlier findings, 13 groups out of 225 did not outperform their stage 1 group average. For 25 groups (i.e. 1 out of 9 groups), group work led to a score *higher than* the super student score, implying that collaboration resulted in new insights and knowledge creation: these students did *better* than just taking the best stage 1 answer for each question.

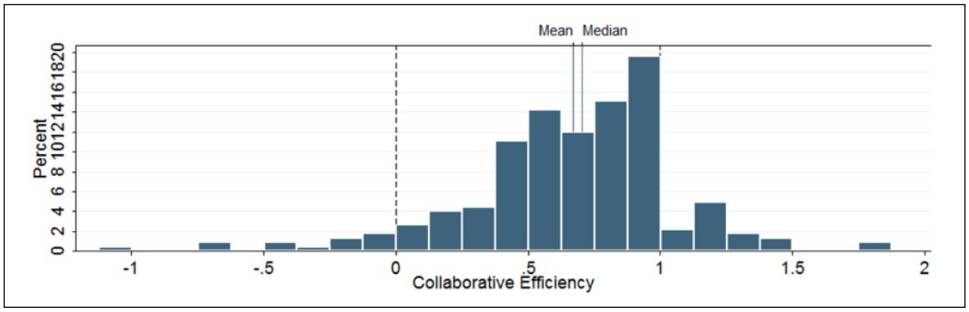


Figure 4. Collaborative efficiency: stage 2 score relative to super student score.

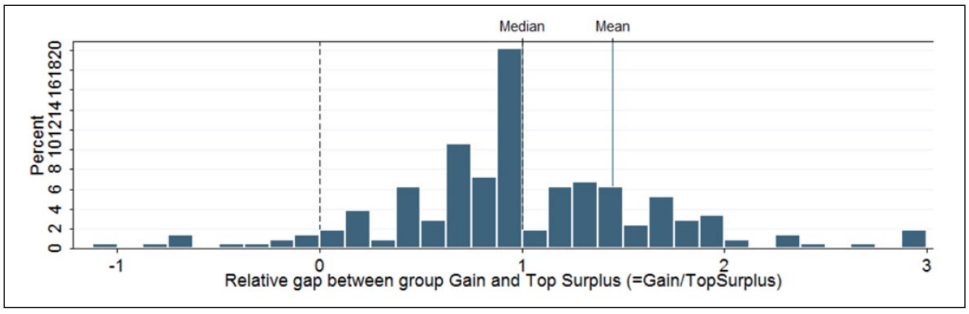


Figure 5. Closing the gap with top student: stage 2 score relative to top student score.

Finally, we explored the extent to which some factors might be associated with higher gains (between stage 2 and stage 1) and collaborative efficiency, and did not detect any noteworthy patterns (see Tables 2 and 3 for details).

Table 2. Predictors of score gains dependent variable is score gain (= stage 2 − stage 1).

Variables	(1)	(2)	(3)	(4)	(5)
Stage1 group average	−0.347***	−0.087	−0.084	−0.078	−0.089
	(0.090)	(0.095)	(0.093)	(0.092)	(0.092)
Top surplus	0.258***		−0.021	−0.014	−0.007
	(0.086)		(0.093)	(0.096)	(0.098)
Super surplus		0.624***	0.636***	0.647***	0.620***
		(0.088)	(0.102)	(0.107)	(0.115)

(Continued)

Table 2. (Continued)

Variables	(1)	(2)	(3)	(4)	(5)
Done 2SE before					0.053
					(0.077)
Cohort FE	No	No	No	Yes	Yes
Constant	0.750***	0.076	0.077	0.096	0.103
	(0.097)	(0.134)	(0.135)	(0.158)	(0.158)
Observations	225	225	225	225	225
R-squared	0.143	0.274	0.274	0.290	0.291

"Done 2SE before" is a dummy variable indicating whether the student did a two-stage examination in the past. "Cohort FE" refers to cohort fixed effects, which are dummy variables that allow us to control for the time-invariant, unobserved characteristics of each cohort.
Robust standard errors in parentheses.
***$p < 0.01$.
**$p < 0.05$.
*$p < 0.1$.

Table 3. Predictors of collaborative efficiency dependent variable is collaborative efficiency.

Variables	(1)	(2)	(3)	(4)	(5)	(6)
Stage 1 group average	−0.085	−0.083	−0.083	−0.089	−0.080	−0.082
	(0.075)	(0.085)	(0.084)	(0.087)	(0.085)	(0.085)
Top surplus	0.004		0.001	0.005	−0.013	−0.002
	(0.064)		(0.061)	(0.063)	(0.062)	(0.061)
Super surplus		0.006	0.006	−0.008	0.012	0.013
		(0.092)	(0.094)	(0.101)	(0.094)	(0.093)
Done 2SE before				0.027		
				(0.063)		
Top is female					−0.084	
					(0.059)	
fem_presence 0 to 4					0.011	
					(0.032)	
1.fem_presence						0.078
						(0.110)
2.fem_presence						0.051
						(0.099)
3.fem_presence						0.095
						(0.106)
4.fem_presence						−0.237
						(0.169)
Constant	0.674***	0.668***	0.668***	0.672***	0.693***	0.599***
	(0.079)	(0.157)	(0.158)	(0.159)	(0.176)	(0.187)
Observations	221	221	221	221	221	221
R-squared	0.042	0.042	0.042	0.043	0.051	0.066

All regressions are robust OLS using cohort fixed effects. "Done 2SE before" is a dummy variable indicating whether the student did a two-stage examination in the past. "Top_Female" is a dummy for whether top stage 1 scorer in the group is a female. For a brief explanation of cohort fixed effects, refer to the notes in Table 2.
Robust standard errors in parentheses.
***$p < 0.01$.
**$p < 0.05$.
*$p < 0.1$.

At the end of each of the stage 2 examinations (with the exception of the 2017 API-210 final examination), students were asked to complete a short survey about their perceptions of the two-stage examination. 802 out of 830 students taking these examinations responded to the post-examination survey (97%), and the response rate exceeded 99% for all but one of the surveys.

After 3 of the 11 examinations (API-209 final examinations for 2013, 2014, and 2015), students were asked whether the two-stage examination should be offered in the future. Of the 203 students taking these examinations, 201 responded (99%). Across all 3 years, 80% of the students who took the two-stage examination and responded to the survey recommended that it should be implemented again in the future. Support for two-stage examinations improved over time (from 72% in 2013% to 83% in 2014% and 87% in 2015), suggesting that refinements in implementation may have contributed to higher student support.

As described in Figure 6, most students reported that the two-stage examination was more helpful for learning than a normal examination (84%), less stressful than a normal examination (67%), that their groups worked together in a mostly equal and fair way (84%), and that the process of coming to a consensus was mostly smooth (67%). Only 2% of students reported very asymmetrical group dynamics or major disagreements in coming to a consensus. However, there was a significant minority (11%) that reported that the two-stage approach was more stressful than a normal examination.

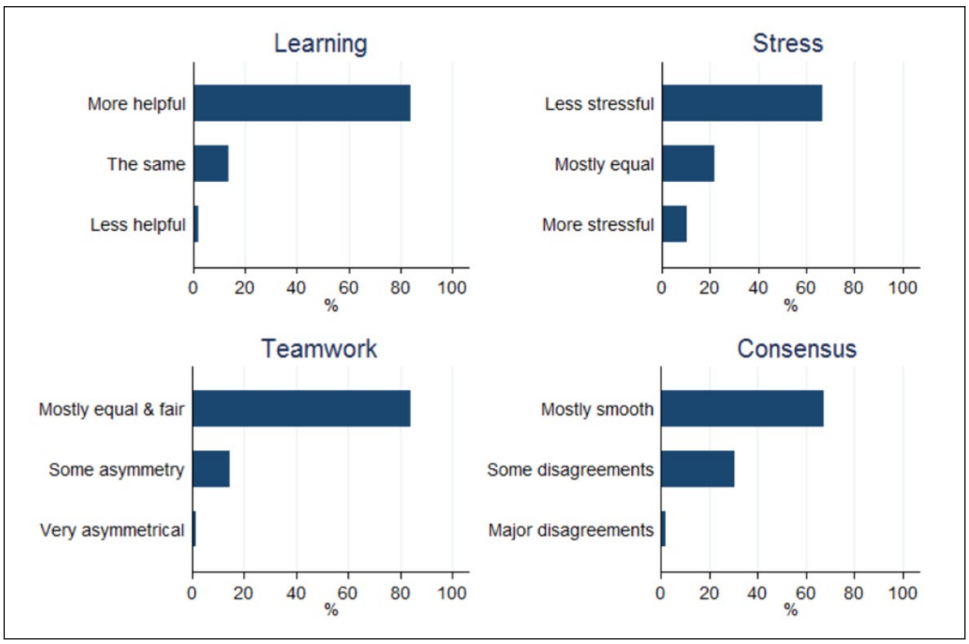


Figure 6. Student feedback for two-stage examinations.

Students from the three API-209 final examinations from 2013 to 2015 were also asked open-ended questions about the effectiveness of two-stage. Out of 201 students, 174 students (87%) provided a substantive response related to what was effective, and 98 students (49%) provided a substantive response related to what was ineffective. These responses were coded according to their content and several key themes emerged (numbers reported are averages across the two independent raters): 82 students (41%) reported that they benefited from rich discussion and exposure

to different problem-solving approaches; 58 students (29%) reported that peer discussion helped to solidify their understanding of the subject matter; and 48 students (24%) reported that they benefited from receiving immediate feedback and discovering their mistakes in stage 1. On the other hand, 24 students (12%) reported that there was not time for adequate discussion to come to a consensus; 22 students (11%) reported a negative feeling from receiving immediate feedback; and 15 students (7%) reported poor group dynamics, such as strong dissent among members, inequality in members' levels of effort, or clashes in personality. Other common themes included remaining confusion even after the group stage, fatigue from repeating the same questions in the first and second stages, and challenges from multiple groups working in the same space.

To assess the level of agreement between the two raters, two measures of interrater reliability were calculated. Across the six major themes identified above, the joint probability of agreement as proportion of the time that the two raters agree was 0.84, and Cohen's Kappa, a more robust measure, was 0.62.

Gender

Before exploring whether gender composition affected performance gains in two-stage examinations, note that there did not seem to be a difference in performance between male and female students. For the whole sample, there were 504 male students (56.4%) and 390 female students (43.6%). The gender difference in stage 1 scores was not statistically significantly different from 0 at the 10% level. In addition, stage 2 group scores were very similar for groups in which the top scoring students were, respectively, male versus female. A significant difference could have suggested that either males or females are intrinsically less listened to in a group than the other sex.

Looking at how the gender composition within a group related to second stage scores and gain, we first classified the groups into five categories: all male, male majority, balanced, female majority, and all female (while some groups had five students, most groups had four, which means that in most cases the balanced groups had two men and two women, and the majority groups had a 1–3 or 3–1 composition). Given the randomization of groups, the baseline measures were similar across categories. The stage 2 performance was also similar across the five categories with one exception: groups that consisted only of female students scored on average significantly lower on the second stage compared with the other groups. This finding was also confirmed in regressions controlling for a host of factors (see Table 4). However, the statistically significant results for all-female groups were only based on a limited sample of eight groups, which makes us hesitate to put too much weight on this finding.

Table 4. Relationship between gender and performance in two-stage examinations dependent variable is gain (= stage 2 − stage 1).

Variables	(1)	(2)	(3)	(4)
Stage 1 group average	−0.078	−0.080	−0.077	−0.080
	(0.092)	(0.092)	(0.094)	(0.094)
Top surplus	−0.014	−0.027	−0.010	−0.017
	(0.096)	(0.096)	(0.096)	(0.094)
Super surplus	0.647***	0.638***	0.632***	0.645***
	(0.107)	(0.108)	(0.109)	(0.107)
Top is female		−0.093		
		(0.069)		

Table 4. (Continued)

Variables	(1)	(2)	(3)	(4)
fem_presence 0–4			−0.034	
			(0.037)	
1.fem_presence				0.077
				(0.132)
2.fem_presence				0.001
				(0.120)
3.fem_presence				0.077
				(0.123)
4.fem_presence				−0.473**
				(0.234)
Cohort FE	Yes	Yes	Yes	Yes
Constant	0.096	0.153	0.160	0.051
	(0.158)	(0.162)	(0.179)	(0.190)
Observations	225	221	221	221
R-squared	0.290	0.289	0.287	0.313

Regression (1) is the same as Table 2 Regression (4). "Fem_presence0_4" is a dummy variable for whether there are any females in the group. "Top_Female" is a dummy for whether top stage 1 scorer in the group is a female. "1.fem_presence" is a female dummy variable for whether there is exactly one female student in the group. Similar definition for "2.fem_presence" "3.fem_presence," and "4.fem_presence". For a brief explanation of cohort fixed effects, refer to the notes in Table 2.
Robust standard errors in parentheses.
***$p < 0.01$.
**$p < 0.05$.
*$p < 0.1$.

Conclusion and discussion

This article contributes to the existing literature on two-stage examinations in the following ways: first, we introduced a new metric, the super student score, composed of the best answers among all members of the group for each individual question. This allowed us to create an indicator to benchmark the effectiveness of student collaboration: collaborative efficiency. We think this measure can become a useful metric for others interested in assessing gains from collaboration in two-stage examinations and other collaborative learning approaches. Second, while two-stage examinations have been studied before, this study is, to the best of our knowledge, the first to do so with a relatively large sample of students. Third, we combined quantitative analyses of examination scores and other measures with qualitative analyses based on student structured and unstructured feedback. Finally, to the extent possible given the examination setup and available data, we analyzed group dynamics, including gender dynamics, to explore what factors may affect group efficiency.

The results suggested that groups substantially improve their performance when going from the individual stage of a two-stage examination (stage 1) to the group stage of the examination (stage 2) by 1 SD, or 18 points, on average. On average, a group was able to close about two-thirds of the gap between the group average stage 1 score and the super student score. This result and further analysis suggested that the group takes advantage of much of the aggregate knowledge and skills of its individual members. Furthermore, student feedback on two-stage examinations was predominantly positive, with most students reporting that these examinations were more helpful for learning, less stressful, and should be continued in the future. Students identified several key

themes that highlight what makes a two-stage examination effective, including exposure to different problem-solving approaches, peer discussion that helped to solidify understanding of the subject matter, and immediate feedback. On the other hand, students also identified features of the two-stage examination that are ineffective, including inadequate discussion time to come to a consensus, adverse emotional impact from immediate feedback, and poor group dynamics.

The key limitations of the study are the following. First, the study was conducted in one professional school in one university in one country/cultural context and so the extent to which the findings would apply to other types of schools or universities is unknown. In particular, the implementation of two-stage examinations in this context involved awarding up to 10% of the examination grade to individuals for work that was done in groups. There are some contexts in which awarding any grades/marks which count toward the GPA or award of an individual student for work that was done in groups is simply not feasible or legal. Second, the study identified a very large performance gap between the second and first stages of the examination, but we could not assess what fraction of this gap represented the causal effect of two-stage examinations on learning, given that this would require a research design in which some students were assigned to take a two-stage examination and others were not, and this was not feasible in this context. Third, while some of the lessons learned helped us improve implementation over time and seemed generalizable to other settings, it is hard to know *ex ante* which implementation features would be most critical in different contexts. Fourth, the pedagogy employed in the courses encourages collaborative learning throughout the course, so the two-stage examinations came as a natural extension, and it is unclear whether the findings from this study generalize to courses with pedagogies that do not employ collaborative learning on a regular basis. Finally, the gender results were based on a very small sample.

Further research could help shed some light on the generalizability of these findings. In particular, research done in institutional contexts very different from ours could be particularly helpful in informing the pedagogic value of two-stage examinations and the key factors needed for their successful implementation. Furthermore, more research on the dynamics that make some groups particularly effective would be helpful, particularly in dynamics involving gender.

Finally, our main messages for faculty members and instructors considering implementing two-stage examinations are as follows. First, these kinds of examinations can extend your efforts to promote active learning and reflection in your courses. Second, the gains in performance between the first and second stages can be very large. Third, using a measure like the one we used to assess collaborative efficiency can help you assess the extent to which your students are taking advantage of the skills and knowledge of their classmates. Finally, collecting feedback from students can help you assess the value that students see in two-stage examinations, and improve the future implementation in a way that could result in greater learning benefits for your students.

Acknowledgements

We are very grateful to the Vice-Provost for Advances in Learning (VPAL) Research Committee at Harvard University for generous financial support, to Evan Abramsky, Hector Cordero, Alfonso De La Torre, Jonathan Lehe, and Vincent Vanderputten for excellent research assistance, and to hundreds of students at the Harvard Kennedy School for their participation and feedback. We are also thankful to Josh Goodman, Andrew Ho, Dick Light, Richard Zeckhauser, and participants at several seminars at Harvard University for helpful comments, and to Eric Mazur and Carl Wieman, who provided us with the initial encouragement to use two-stage examinations and with insights that helped us design and implement these.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: The author(s) received financial support from the Vice-Provost for Advances in Learning (VPAL) Research Committee at Harvard University for the research conducted in this article.

References

Balch W (1992) Effect of class standing on students' predictions of their final exam scores. *Teaching of Psychology* 19(3): 136–41.

Billington R (1994) Effects of collaborative test taking on retention in eight third-grade mathematics classes. *The Elementary School Journal* 95(1): 23–31.

Bloom D (2009) Collaborative test taking: Benefits for learning and retention. *College Teaching* 57(4): 216–20.

Bransford J, Brown A and Cocking R (2004) *How People Learn: Brain, Mind, Experience, and School*. Washington, DC: National Academies Press.

Breedlove W, Burkett T and Winfield I (2004) Collaborative testing and test anxiety. *Journal of Scholarship of Teaching and Learning* 4(2): 33–42.

Centrella-Nigro A (2012) Collaborative testing as posttest review. *Nursing Education Perspectives* 33(5): 340–1.

Cortright R, Collins H, Rodenbaugh D, et al. (2003) Student retention of course content is improved by collaborative-group testing. *Advances in Physiology Education* 27(1–4): 102–8.

Cranney J, Ahn M, McKinnon R, et al. (2009) The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology* 21(6): 919–40.

De Paola M and Scoppa V (2010) Peer group effects on the academic performance of Italian students. *Applied Economics* 42(17): 2203–15.

Deslauriers L, Schelew E and Wieman C (2011) Improved learning in a large-enrollment physics class. *Science* 332(6031): 862–4.

Freeman S, Theobald R, Crowe AJ, et al. (2017) Likes attract: Students self-sort in a classroom by gender, demography, and academic characteristics. *Active Learning in Higher Education* 18(2): 115–26.

Gibbs G and Simpson C (2005) Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education* 1(1): 3–31.

Gilley B and 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, Lujan H and DiCarlo S (2008) Collaborative group testing benefits high- and low-performing students. *Advances in Physiology Education* 32(4): 274–8.

Giuliodori M, Lujan H and DiCarlo S (2009) Student interaction characteristics during collaborative group testing. *Advances in Physiology Education* 33(1): 24–9.

Glass A, Ingate M and Sinha N (2013) The effect of a final exam on long-term retention. *The Journal of General Psychology* 140(3): 224–41.

Hall D and Buzwell S (2012) The problem of free-riding in group projects: Looking beyond social loafing as reason for non-contribution. *Active Learning in Higher Education* 14(1): 37–49.

Herrmann K (2013) The impact of cooperative learning on student engagement: Results from an intervention. *Active Learning in Higher Education* 14(3): 175–87.

Johnson D and Johnson R (1999) Making cooperative learning work. *Theory into Practice* 38(2): 67–73.

Leight H, Saunders C, Calkins R, 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.

Martin D, Friesen E and De Pau A (2014) Three heads are better than one: A mixed methods study examining collaborative versus traditional test-taking with nursing students. *Nurse Education Today* 34(6): 971–7.

Masanobu T and Saito S (2004) Does test delay eliminate collaborative inhibition? *Memory* 12(6): 722–31.

Meseke C, Bovée M and Gran D (2009) Impact of collaborative testing on student performance and satisfaction in a chiropractic science course. *Journal of Manipulative and Physiological Therapeutics* 32(4): 309–14.

Meseke J, Nafziger R and Meseke C (2008) Facilitating the learning process: A pilot study of collaborative testing vs individualistic testing in the chiropractic college setting. *Journal of Manipulative and Physiological Therapeutics* 31(4): 308–12.

Moore L (2010) Students' attitudes and perceptions about the use of cooperative exams in an introductory leadership class. *Journal of Leadership Education* 9(2): 72–85.

Rivaz M, Momennasab M and Shokrollahi P (2015) Effect of collaborative testing on learning and retention of course content in nursing students. *Journal of Advances in Medical Education & Professionalism* 3(4): 178–82.

Slavin R (1996) Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary Educational Psychology* 21(1): 43–69.

Smith M, Wood W, Adams W, et al. (2009) Why peer discussion improves student performance on in-class concept questions. *Science* 323: 122–4.

Soetanto D and MacDonald M (2017) Group work and the change of obstacles over time: The influence of learning style and group composition. *Active Learning in Higher Education* 18(2): 99–113.

Springer L, Stanne M and Donovan S (1999) Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of Educational Research* 69(1): 21–51.

Takahashi M and Saito S (2004) Does test delay eliminate collaborative inhibition? *Memory* 12(6): 722–31.

Takeda S and Homberg F (2014) The effects of gender on group work process and achievement: An analysis through self- and peer-assessment. *British Educational Research Journal* 40(2): 373–96.

Webb N (1993) Collaborative group versus individual assessment in mathematics: Processes and outcomes. *Educational Assessment* 1(2): 131–52.

Weldon M and Bellinger K (1997) Collective memory: Collaborative and individual processes in remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23(5): 1160–75.

Wieman C and Perkins K (2005) Transforming physics education. *Physics Today* 58(11): 36–41.

Wieman C, Rieger G and Heiner C (2014) Physics exams that promote collaborative learning. *The Physics Teacher* 52(1): 51–3.

Yu B, Tsiknis G and Allen M (2010) Turning exams into a learning experience. In: *Proceedings of the 41st ACM technical symposium on computer science education*, Milwaukee, WI, 10–13 March, pp. 336–40. New York: ACM Press.

Yuretich R, Khan S, Leckie R, et al. (2001) Active-learning methods to improve student performance and scientific interest in a large introductory oceanography course. *Journal of Geoscience Education* 49(2): 111–9.

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

Author biographies

Dan Levy is a senior lecturer in Public Policy and Faculty Chair of the Harvard Kennedy School of Government's SLATE (Strengthening Learning and Teaching Excellence) Initiative, and teaches courses in quantitative methods, policy analysis, and program evaluation. Address: John F. Kennedy School of Government, Harvard University, 79 John F. Kennedy Street, Mailbox 67, Cambridge, MA 02138, USA. [Email: dan_levy@hks.harvard.edu]

Theodore Svoronos is a lecturer at the Kennedy School of Government, and teaches courses in statistics and econometrics, and develops new digital and online teaching materials for graduate students. Address: John F. Kennedy School of Government, Harvard University, 79 John F. Kennedy Street, Mailbox 67, Cambridge, MA 02138, USA. [Email: theodore_svoronos@hks.harvard.edu]

Mae Klinger is a digital learning designer at the SLATE Initiative, designs and develops digital learning materials and supports the use of learning data in graduate courses. Address: John F. Kennedy School of Government, Harvard University, 79 John F. Kennedy Street, Mailbox 67, Cambridge, MA 02138, USA. [Email: mae_klinger@hks.harvard.edu]