

Female Peers under Weakest-Link Incentives: Evidence from Randomized Classroom Groups*

José María Cabrera Alejandro Cid

Universidad de Montevideo

September 2026

Abstract

We study how peer composition affects individual effort under weakest-link incentives, where each group member's reward is lost if anyone falls short. In two small randomized classroom field experiments with undergraduate students in Uruguay, students assigned to joint-liability incentives were randomly formed into trios. Homework was submitted and graded individually. Students assigned at least one female peer earned homework grades 0.62 standard deviations higher than those assigned only male peers, before any bonus was added. The estimate is positive in both experiments and for male and female students, with no robust difference by own gender. Homework submission also rises, though less robustly; midterm scores, final course grades, and cumulative GPA show no clear improvement. These findings show a short-run response in individual homework performance, but do not establish persistent learning or a particular peer-interaction mechanism. Under weakest-link incentives, group composition can change the performance elicited by a fixed reward rule.

Keywords: peer effects; gender composition; joint liability; academic incentives; field experiment; higher education.

JEL codes: C93; I21; I23; J16.

1 Introduction

Many contracts make each participant's reward depend on whether every member of a group performs. Examples include production teams, group lending, and classroom incentives.

*Cabrera: jmcabrera@um.edu.uy | Cid: acid@um.edu.uy.

These arrangements create a basic organizational problem: changing who belongs to the group may alter individual effort even when tasks and rewards remain fixed. Peer composition may affect beliefs about whether the group will qualify, as well as opportunities for assistance, monitoring, or social pressure. It is therefore potentially part of the incentive technology, not merely a background characteristic of the team.

Economics describes as *weakest link* environments in which collective success is pinned down by the lowest individual contribution (Hirshleifer, 1983; Van Huyck, Battalio, and Beil, 1990). The experiments studied here implement a discrete threshold version of this logic. The reward is not a continuous function of minimum effort, but any member’s failure to meet the requirements eliminates it for everyone. Although team incentives can change individual effort (Babcock et al., 2015; Blimpo, 2014), we know less about whether randomly changing team composition changes behavior under a fixed weakest-link rule.

We study this question using two randomized classroom field experiments at Universidad de Montevideo in Uruguay. Students were randomly assigned to a joint-liability grade incentive, an individual grade incentive, or a pure control. Within the joint-liability arm, they were also randomly partitioned into groups of three. Homework remained an individual task, but a student earned a grade bonus only if all three group members submitted the work, attended class, and passed the relevant assessment. Random group formation generated variation in whether a student had zero, one, or two female peers while holding the incentive rule fixed.

The analysis covers 81 students in 27 joint-liability trios. Students assigned at least one female peer earned homework grades 0.62 standard deviations higher than students assigned none, before the experimental bonus was added. A complementary linear comparison from zero to two female peers yields an increase of 0.91 standard deviations. The homework-grade estimate is positive in both experiments and remains present in the broader comparison with at least one female peer.

The improvement is concentrated in outcomes closest to the intervention. Submission rates also rise, although the evidence is weaker, while midterm performance, final course grades, and cumulative GPA show no clear improvement. Homework grades improve for both male and female students, with no robust difference by own gender. Taken together, the evidence points to short-run effort or homework quality rather than durable learning.

The paper connects research on peer composition with research on group incentives. Studies of gender peer effects typically vary exposure across grades, classrooms, tutorial sections, or seating neighborhoods (Lavy and Schlosser, 2011; Lu and Anderson, 2015; Oosterbeek and van Ewijk, 2014). Recent evidence from university teams and tutorials remains context dependent (Contreras et al., 2025; Zhang, 2025; Shan, 2026). In a closely related

field experiment, [Hoogendoorn, Oosterbeek, and van Praag \(2013\)](#) randomly varied the gender composition of student business teams and found that teams with an equal gender mix outperformed male-dominated teams in joint production. [Lamiraud and Vranceanu \(2018\)](#) randomly formed student teams of different gender compositions in a business simulation and studied their collective decisions and performance. Our setting differs because homework is produced individually, while the reward is shared and lost when any group member fails to meet the requirements.

The incentive literature shows that responses depend on both the contract and the population ([Angrist, Lang, and Oreopoulos, 2009](#); [Bigoni et al., 2015](#)). All-or-none rewards can improve coordination while they are in place ([Hamman, Rick, and Weber, 2007](#)), and real-effort experiments compare group and individual incentives under weak-link technologies ([Bortolotti, Devetag, and Ortmann, 2016](#)). In two graduate-school field experiments, [List and Shah \(2022\)](#) vary weakest-link and best-shot rewards across study groups. We study a different margin: who is placed in a group when each person’s reward depends on everyone. The randomized field setting identifies the effect of composition under that contract, although it cannot separately identify assistance, monitoring, beliefs, or social pressure.

Section 2 describes the experiments and the weakest-link rule. Section 3 presents the data, estimands, and randomization inference. Section 4 reports the main composition effects. Section 5 examines heterogeneity and the pattern across outcomes. Section 6 concludes.

2 Institutional Setting and Experimental Designs

The experiments were implemented in undergraduate courses at Universidad de Montevideo in Uruguay. Ethical approval for fieldwork was obtained from the Board of Universidad de Montevideo (n. FCEE 01/11). They were part of a sequence of classroom interventions that studied homework and academic incentives. The present paper combines the 2011 and 2013 cohorts, both of which contain joint-liability, individual-incentive, and pure-control arms. [Cid and Cabrera \(2025\)](#) previously analyze the 2011 experiment, focusing on the average effects of joint liability relative to individual incentives and control. The 24 students assigned to joint liability in 2011 therefore enter both studies, but the randomized margin and estimand differ: the earlier paper compares incentive arms, whereas we use peer gender composition within the joint-liability arm. The present paper also adds a separate 2013 experiment.

2.1 The 2011 experiment

The 2011 experiment covered two courses. Students were assigned to one of three grading regimes. In the joint-liability arm, groups contained three students. Each student received a 20 percent increase in the homework component of the course grade if every member of the trio submitted the required work, attended class, and passed. In the individual-incentive arm, the same 20 percent increase depended only on the student’s own compliance. The control arm followed the ordinary grading rule. Students completed eight homework assignments, and the course grade combined the midterm assessment (70 percent) and homework (30 percent).

Assignment used rerandomization based on 18 baseline variables. The experiment randomized 49 students: 24 to joint liability, 13 to individual incentives, and 12 to pure control. The joint-liability arm formed eight trios. Two additional students entered after randomization and are excluded from the randomized analyses. The appendix reports the acceptance rule and the pre-assignment covariate values used to implement it.

2.2 The 2013 experiment

The 2013 experiment covered four courses and retained the same three-arm structure. The joint-liability rule again offered a 20 percent homework bonus and formed groups of three. The individual arm offered a 10 percent bonus based on the student’s own performance; the pure control received no experimental bonus.

Assignment again used rerandomization, based on 13 baseline variables and four pre-defined comparisons across arms. The experiment randomized 111 students: 57 to joint liability, 27 to individual incentives, and 27 to pure control. All 111 are retained in the randomized sample, and the joint-liability arm contributes nineteen trios to the composition analysis. The appendix provides the complete acceptance rule and implementation details.

Students enrolled in more than one participating course were assigned to a single experiment, so no individual appears twice in the analysis.

2.3 Weakest-link incentives and random trio formation

Conditional on assignment to the joint-liability arm, students were ordered at random within course and partitioned into consecutive groups of three. This procedure generates random variation in peers’ predetermined characteristics. For student i , let F_{ig} denote the number of female students among the other two members of group g . The primary composition exposure is $S_{ig} = F_{ig}/2$, which takes values zero, one-half, and one. Across both cohorts,

35 students have no female peers and 46 have at least one. Among female students, the corresponding counts at zero, one, and two female peers are 8, 18, and 3; among male students they are 27, 16, and 9.

Homework was graded before any bonus was applied by teaching assistants and instructors from other courses who were not involved in the experiment and were blind to treatment assignment; any earned bonus was added separately. Assignments were submitted individually without identifying the experimental trio. Ordinary collaboration was encouraged under the same rules in every arm, and copying was sanctioned identically across arms. Students were not required to complete homework with their assigned peers, but the design left room for the informal peer monitoring and encouragement that the group-contingent incentive was intended to elicit, such as checking whether peers were keeping up with the assignments.

The design is naturally described as a threshold weakest-link incentive. A student's homework remained individual, but the reward was lost if any group member failed to meet the conditions. Unlike the canonical minimum-effort game, the realized course output is not mechanically equal to the lowest member's effort. The analogy concerns the incentive's conjunctive structure: every member must cross the relevant thresholds, so any member can block the common reward. Female-peer composition may therefore affect effort through several channels, including mutual assistance, monitoring, social pressure, or beliefs about peers' reliability. The experiment identifies their combined behavioral effect but does not randomize or measure those mechanisms separately.

3 Data and Empirical Strategy

3.1 Samples and outcomes

The harmonized dataset contains 160 originally randomized students: 49 from 2011 and 111 from 2013. The composition analysis is restricted by design to the 81 students assigned to joint liability. Our primary outcome is the student's average homework grade before the experimental bonus. It is available for 77 of the 81 joint-liability students. The secondary family contains the homework submission rate ($N = 77$), standardized midterm performance ($N = 77$), final course grade ($N = 76$), and cumulative post-treatment GPA ($N = 81$). Baseline GPA and cumulative post-treatment GPA are complete. We do not drop students because they failed to answer a follow-up survey when their administrative outcomes remain observed.

Proximal outcomes have a small amount of missing data, while the final-exam measure is less complete and is reserved for the appendix. Missingness does not vary detectably with

treatment, own gender, or their interaction. We report the outcome-specific sample size in every table and do not impute outcomes; the appendix provides the full accounting.

We report grades in their original units and in standard-deviation units. Standardization uses the mean and standard deviation among all randomized arms within each cohort-course stratum, before restricting the data to joint liability. This transformation is fixed with respect to the random partition of joint-arm students.

3.2 Estimands and empirical specifications

For outcome Y_{ig} , the pooled composition specification is

$$Y_{ig} = \alpha + \beta S_{ig} + \delta Female_i + X_i' \gamma + \lambda_{c(i)} + \varepsilon_{ig}, \quad (1)$$

where S_{ig} is female-peer share, $Female_i$ is the student's own recorded sex, X_i contains baseline GPA with a cohort-specific slope, and $\lambda_{c(i)}$ denotes cohort-course indicators. The coefficient β measures the average change in Y_{ig} associated with increasing S_{ig} from 0 to 1, which corresponds to moving from two male peers to two female peers under the linear specification. Equivalently, the model assigns an average change of $\beta/2$ to replacing one male peer with one female peer. We also estimate the model separately by cohort, using baseline GPA and course indicators.

We complement the linear specification with an indicator for at least one female peer. This contrast uses substantially more support at each exposure level and does not impose equal incremental effects of the first and second female peer. It supplies the headline magnitude because it is based on the broader comparison between students assigned no female peers and those assigned at least one. The linear specification remains a complementary summary of the ordered exposure. Because the treatment is a random partition rather than an independently assigned student-level indicator, all inference re-creates the full grouping procedure.

To examine heterogeneity by own sex, we augment equation (1) with $Female_i \times S_{ig}$. We report the peer-composition slope for men, the slope for women, and their difference. This heterogeneity is secondary. We apply the same outcome domains separately to the direct difference and to the sex-specific slopes.

3.3 Inference and multiple testing

Conventional standard errors are clustered by randomized trio. They summarize sampling uncertainty under a regression approximation, but our primary p-values come from condi-

tional randomization inference.

Under the sharp null that group composition has no effect, observed outcomes and baseline characteristics remain fixed when students are repartitioned into trios. We independently permute students within each cohort-course joint arm, divide the randomized ordering into groups of three, reconstruct each student’s female-peer share, and re-estimate the statistic.

This procedure conditions on the number and sex of students in each cohort-course and respects the dependence created by sharing peers. It is also used for the binary any-female-peer contrast and is repeated separately by cohort. The appendix reports simulation counts, seeds, and the plus-one correction.

We organize the outcomes into three domains. Homework performance contains the grade before the bonus and submission. Performance in the specific course contains the midterm and final course grade; the latter includes homework and the experimental bonus and is not an independent measure of learning. Cumulative GPA is a separate, broader outcome. We report unadjusted two-sided randomization p-values and Romano–Wolf stepdown adjustments based on common random partitions within each domain. For the primary overall estimates, the adjustment also accounts for reporting both the linear and binary exposure summaries. Homework grade is the primary outcome because it is the closest measure of sustained individual effort that cannot be generated mechanically by the bonus. The overall composition effect is primary and own-gender heterogeneity is secondary. For the interaction statistic, randomization inference is exact under the sharp null of no composition effect, but it is not an exact test of equal gender-specific effects under an unknown nonzero common effect.

Gender and group composition were included among the planned analyses in the 2011 protocol, but the peer-composition hypothesis was not specified in advance for the 2013 experiment. Random assignment of students to groups identifies the composition effects in both cohorts; however, we do not interpret the pooled result as a test of a hypothesis specified in advance for both. This distinction affects the interpretation of the pooled evidence, not its causal identification within either cohort.

The initial analysis plan emphasized the female-specific composition slope. The estimates, however, are positive for both male and female students. We therefore treat the average composition effect as primary and own-gender heterogeneity as secondary. This hierarchy does not alter the sample, specifications, or outcomes.

4 Main Results

We present the main result in two complementary ways. First, we use the full randomized variation in whether a student has zero, one, or two female peers and summarize the implied linear change from two male peers to two female peers. Figure 1 shows the support and adjusted pattern behind that comparison, while Panel A of Table 1 reports estimates across the three outcome domains. Second, because the endpoint with two female peers is sparse, we compare students assigned at least one female peer with those assigned none. Panel B reports this more broadly supported contrast for all students and, as secondary evidence, for women. Both approaches focus first on homework grades before the bonus is added.

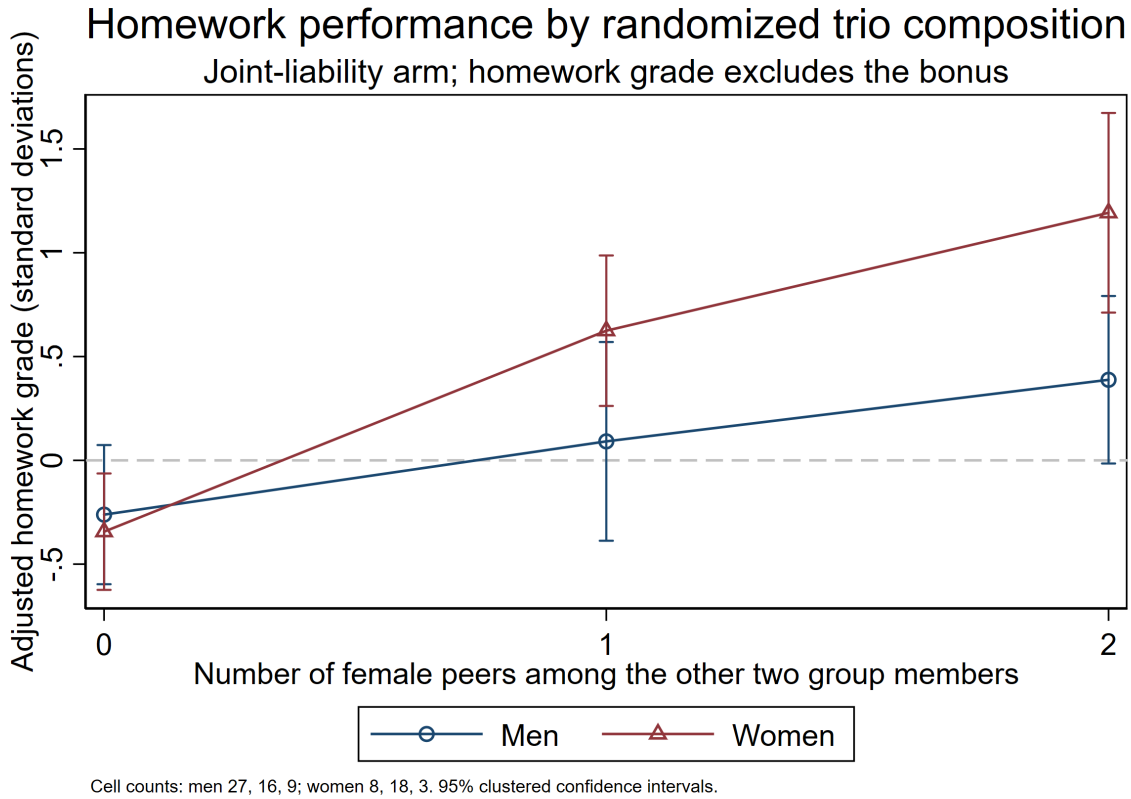


Figure 1: Homework performance by randomized trio composition

4.1 Linear female-peer exposure

Figure 1 first displays the support of the composition treatment. Raw homework means for students with zero, one, and two female peers are 7.54, 7.62, and 8.07 among men and 6.69, 8.61, and 8.77 among women. The figure reports adjusted categorical predictions rather than

Table 1: Gender composition and academic outcomes within joint-liability trios

Outcome or contrast	Effect	Effect (SD)	RI p	RW p	N
<i>Panel A. Linear female-peer exposure</i>					
Homework grade (no bonus)	2.245** (0.624)	0.913** (0.256)	0.006	0.015	77
Homework submission	0.158** (0.046)	0.739** (0.229)	0.020	0.028	77
Midterm performance	0.229 (0.232)	0.222 (0.230)	0.460	0.602	77
Final course grade	0.989 (0.485)	0.478 (0.226)	0.108	0.182	76
Cumulative GPA	0.172 (0.152)	0.121 (0.112)	0.190	0.190	81
<i>Panel B. At least one female peer versus none: homework grade</i>					
All students	1.561** (0.478)	0.624** (0.186)	0.006	0.015	77
Women only (secondary)	1.922* (0.512)	0.827* (0.255)	0.049	0.090	29

Notes: Panel A effects correspond to moving from zero to two female peers. Panel B compares students assigned at least one female peer with students assigned none. Models control for own sex when applicable, include cohort-course indicators and cohort-specific baseline-GPA slopes, and cluster conventional standard errors by randomized trio; standard errors are in parentheses. Standardized outcomes use all randomized arms within each cohort-course. RI p -values, Romano–Wolf (RW) adjustments, and stars refer to the original-unit specifications; standardized estimates summarize magnitude. RI conditions on random trio formation within cohort-course. For overall effects, RW uses common partitions and adjusts jointly within the relevant outcome domain and across the linear and binary exposure summaries. The secondary women-only row is adjusted within the homework domain. Stars: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

imposing a linear relationship. It also shows the wider confidence interval at two female peers for women, where the cell contains only three observations.

Panel A of Table 1 reports the pooled linear estimate. Moving from zero to two female peers increases homework grades before the bonus by 2.245 points on the 0–12 scale. The standardized estimate is 0.913 standard deviations. Conditional randomization inference gives $p = 0.0064$, and the Romano–Wolf adjusted value is 0.015.

The homework-grade estimate is positive in each experiment. Moving from zero to two female peers raises homework grades by 5.666 points in 2011 ($p_{RI} = 0.0118$) and by 1.838 points in 2013 ($p_{RI} = 0.0316$). Because the 2011 experiment contains only eight trios and the two estimates differ substantially in magnitude, we interpret this pattern as replication of the sign rather than evidence of a common effect size. Appendix Table A1 reports cohort-specific estimates for all outcomes and the leave-one-trio-out diagnostics.

4.2 Any female peer versus none

The categorical support motivates a less parametric contrast. Thirty-five students have two male peers and 46 have at least one female peer. Adjusting for own sex, baseline GPA, cohort, and course, the latter group scores 1.561 points, or 0.624 standard deviations, higher on homework before the bonus. Panel B of Table 1 reports this contrast. Repartitioning students into trios gives $p_{RI} = 0.0064$; the Romano–Wolf adjusted value is 0.015. This result is central because it rests on broad support across the two exposure categories rather than on the sparsely populated women-only endpoint with two female peers.

Among women alone, having at least one female peer raises the homework grade by 1.922 points, or 0.827 standard deviations. Its design-based p-value is 0.0491 and becomes 0.090 after adjustment within the homework domain. The women-only contrast is therefore secondary and does not establish robust heterogeneity by own gender.

Appendix Table A3 reports the same binary contrast for all five outcomes, both for all students and for women.

5 Heterogeneity and Outcome Patterns

The main results show that peer gender composition affects homework grades under the joint-liability incentive. We next ask two questions about that response. First, does the effect differ between male and female students? We examine this question by interacting female-peer composition with the student’s own gender. Second, is the response confined to homework, or does it extend to submission, exam performance, final course grades, and cumulative

GPA? Comparing effects across these outcomes helps distinguish a change in short-run effort or homework quality from broader course performance or cumulative achievement. Figure 2 summarizes the composition effects across outcomes and cohorts. These exercises can characterize who responds and where the response appears, but they do not separately identify monitoring, assistance, social pressure, or other mechanisms.

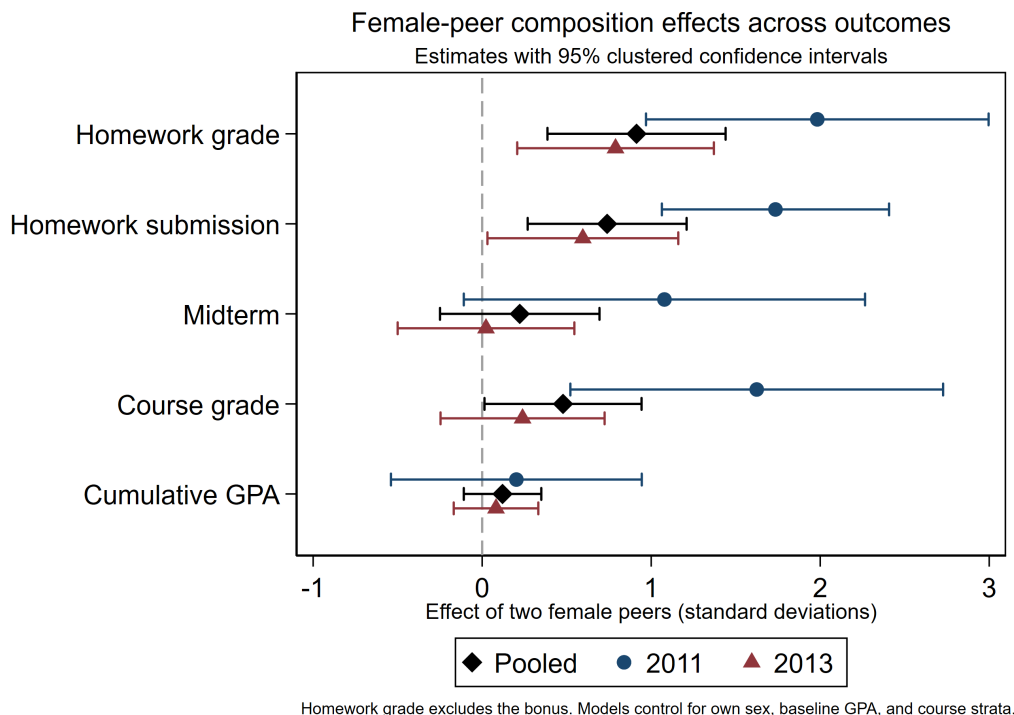


Figure 2: Outcomes and randomized female-peer composition

5.1 Heterogeneity by own gender

Appendix Table A2 reports the sex-specific slopes for all five outcomes and direct tests of the differences between them. The interacted model yields a 3.759-point homework slope for women and a 1.732-point slope for men. Both point estimates are positive. The randomization p-value for their difference is 0.177 and its Romano–Wolf adjusted value is 0.197. Homework submission follows the same qualitative pattern: the estimated slopes are 0.193 for women and 0.146 for men, with no evidence of a difference. The course-grade difference is larger in levels and has an adjusted value of 0.043. Because the interaction test is exact only under the global sharp null, we treat this secondary result as suggestive rather than as a decisive test of unequal effects. Taken together, the results do not support a general claim that female peers benefit only female students.

5.2 Homework submission, performance, and later outcomes

Panel A of Table 1 reports the overall estimates for the outcomes discussed in this subsection. Female-peer share increases the homework submission rate by 0.158, equivalent to 0.739 standard deviations (randomization $p = 0.020$). The Romano–Wolf adjusted value is 0.028, so submission provides corroborating evidence within the homework domain. Because homework grades exclude the bonus, neither result can be an arithmetic consequence of adding the experimental reward to the grade.

The estimated effect on the standardized midterm is 0.229 standard deviations ($p_{RI} = 0.460$). The final course-grade effect is 0.989 points ($p_{RI} = 0.108$), and the cumulative-GPA effect is 0.172 points ($p_{RI} = 0.190$). All three confidence intervals include zero, and their Romano–Wolf adjusted values range from 0.182 to 0.602. The outcome pattern is therefore most consistent with altered homework effort or quality, not a demonstrated improvement in exam performance or persistent achievement.

5.3 Interpreting the homework response

Higher homework grades before the bonus and higher submission rates show a response in students’ own work under the shared threshold. Each student’s reward depended on both partners’ compliance even though homework was completed individually. Random composition identifies the total behavioral response to female peers in this environment. Several channels could produce it, including assistance, expectations, monitoring, and social pressure, but interactions within trios were not measured consistently across cohorts. Because composition was randomized only within joint-liability groups, the experiments also do not establish whether the same effect would arise under other incentive arrangements.

6 Conclusion

Two randomized classroom field experiments let us ask whether the people who share a group-contingent reward affect performance on an individual task. Students in joint-liability trios earned a bonus only if all three met common requirements, yet homework remained an individual assignment. Those assigned at least one female peer earned homework grades 0.62 standard deviations higher than those assigned none, before the bonus was added. The gradient across zero, one, and two female peers is positive in both experiments.

Homework submission also rose, placing the response in students’ work rather than in the mechanical addition of a bonus. The pattern is concentrated in homework: midterm performance, final course grades, and cumulative GPA show no clear corresponding gains.

Together, these outcomes point to short-run effort or homework quality under joint accountability, even without a requirement to work on assignments with the assigned trio.

The finding matters for educators and organizations that make each person’s reward depend on every group member’s performance. Such arrangements make partners consequential even when tasks are assigned separately. The experiments show that changing who shares the reward can change the behavior elicited by a fixed incentive rule. Under weakest-link incentives, group formation is part of the incentive design.

References

- Angrist, Joshua, Daniel Lang, and Philip Oreopoulos. 2009. [Incentives and Services for College Achievement: Evidence from a Randomized Trial](#). *American Economic Journal: Applied Economics* 1(1): 136–163.
- Babcock, Philip, Kelly Bedard, Gary Charness, John Hartman, and Heather Royer. 2015. [Letting Down the Team? Social Effects of Team Incentives](#). *Journal of the European Economic Association* 13(5): 841–870.
- Bigoni, Maria, Margherita Fort, Mattia Nardotto, and Tommaso G. Reggiani. 2015. [Cooperation or Competition? A Field Experiment on Non-Monetary Learning Incentives](#). *The B.E. Journal of Economic Analysis & Policy* 15(4): 1753–1792.
- Blimpo, Moussa P. 2014. [Team Incentives for Education in Developing Countries: A Randomized Field Experiment in Benin](#). *American Economic Journal: Applied Economics* 6(4): 90–109.
- Bortolotti, Stefania, Giovanna Devetag, and Andreas Ortmann. 2016. [Group Incentives or Individual Incentives? A Real-Effort Weak-Link Experiment](#). *Journal of Economic Psychology* 56: 60–73.
- Cid, Alejandro, and José María Cabrera. 2025. [The Use of Innovative Incentives in the Classroom to Explore the Impact of Peer Monitoring on Academic Achievements](#). *Journal of Behavioral Education* 34(1): 163–183.
- Contreras, Valentina, Chiara Orsini, Berkay Özcan, and Johann Koehler. 2025. [Effects of Team Diversity on Individual Performance and Voice: A Field Experiment of Group Composition by Gender and Language](#). *Labour Economics* 95: 102763.
- Hamman, John, Scott Rick, and Roberto A. Weber. 2007. [Solving Coordination Failure with “All-or-None” Group-Level Incentives](#). *Experimental Economics* 10(3): 285–303.
- Hirshleifer, Jack. 1983. [From Weakest-Link to Best-Shot: The Voluntary Provision of Public Goods](#). *Public Choice* 41(3): 371–386.
- Hoogendoorn, Sander, Hessel Oosterbeek, and Mirjam van Praag. 2013. [The Impact of Gender Diversity on the Performance of Business Teams: Evidence from a Field Experiment](#). *Management Science* 59(7): 1514–1528.
- Lamiraud, Karine, and Radu Vranceanu. 2018. [Group Gender Composition and Economic Decision-Making: Evidence from the Kallystée Business Game](#). *Journal of Economic Behavior & Organization* 145: 294–305.
- Lavy, Victor, and Analia Schlosser. 2011. [Mechanisms and Impacts of Gender Peer Effects at School](#). *American Economic Journal: Applied Economics* 3(2): 1–33.
- List, John A., and Rohen Shah. 2022. [The Impact of Team Incentives on Performance in Graduate School: Evidence from Two Pilot RCTs](#). *Economics Letters* 221: 110894.
- Lu, Fangwen, and Michael L. Anderson. 2015. [Peer Effects in Microenvironments: The Benefits of Homogeneous Classroom Groups](#). *Journal of Labor Economics* 33(1): 91–122.
- Oosterbeek, Hessel, and Reyn van Ewijk. 2014. [Gender Peer Effects in University: Evidence from a Randomized Experiment](#). *Economics of Education Review* 38: 51–63.

- Shan, Xiaoyue. 2026. [Gender Diversity Improves Performance but Reinforces Gendered Roles](#). CESifo Working Paper 12565.
- Van Huyck, John B., Raymond C. Battalio, and Richard O. Beil. 1990. [Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure](#). *American Economic Review* 80(1): 234–248.
- Zhang, Yuxuan. 2025. [Gender Peer Effects in University Courses: Evidence from Tutorial Groups](#). *Labour Economics* 97: 102804.

Appendix

A Additional Results and Sensitivity Analyses

The peer-composition hypothesis was among the planned analyses in the 2011 protocol but was not specified in advance for the 2013 experiment. The analysis uses three outcome domains: homework grade and submission; midterm and final course grade; and cumulative GPA as a singleton. Homework grade is the primary outcome. We report unadjusted randomization p -values and Romano–Wolf stepdown adjustments based on common partitions. The overall estimates also account for the two reported exposure summaries.

The estimand hierarchy was amended after the initial results were examined. The initial plan emphasized the female-specific composition slope, whereas the estimates showed positive homework slopes for both sexes, no multiplicity-robust difference between them, and stronger support for the overall comparison of any female peer versus none. We therefore treat the overall composition effect as primary and own-gender heterogeneity as secondary. This change in emphasis did not alter the sample, primary outcome, covariates, or specifications.

A.1 Cohort replication and influence diagnostics

The cohort comparison reinforces the distinction between a replicated sign and a stable magnitude. Overall homework-grade and submission effects are positive in both years. In 2011, moving from zero to two female peers raises the submission rate by 0.489 ($p_{RI} = 0.016$); the corresponding 2013 estimate is 0.110 ($p_{RI} = 0.090$). Final course-grade effects are also positive in both cohorts but are detectable only in the smaller 2011 experiment. Cumulative-GPA estimates are small and imprecise in both years.

The leave-one-trio-out diagnostics cover both primary homework estimands and the larger female-specific slope. Omitting each trio in turn leaves the overall linear effect between 1.857 and 2.649 points, compared with 2.245 in the full sample. The binary effect remains between 1.200 and 1.768 points, compared with 1.561 in the full sample. The corresponding range for the female-specific slope is 3.415 to 4.700 points, compared with 3.759. These checks do not overcome the small number of randomized groups, but no single trio generates any of the reported homework patterns.

Table A1: Replication across cohorts and influence diagnostics

Outcome	2011		2013	
	Effect	RI p	Effect	RI p
<i>Panel A. Linear female-peer exposure</i>				
Homework grade (no bonus)	5.666	0.012	1.838	0.032
Homework submission	0.489	0.016	0.110	0.090
Midterm performance	1.116	0.093	0.023	0.951
Final course grade	2.756	0.014	0.602	0.419
Cumulative GPA	0.210	0.564	0.136	0.334
<i>Panel B. Leave-one-trio-out sensitivity</i>				
Diagnostic	Full sample	Minimum	Maximum	Omissions
Overall linear homework effect	2.245	1.857	2.649	27
At least one female peer	1.561	1.200	1.768	27
Female-specific homework slope	3.759	3.415	4.700	27

Notes: Panel A effects correspond to moving from zero to two female peers, control for own sex and baseline GPA, and include course indicators. Randomization inference recreates trio formation separately within each cohort-course. Panel B reports the full-sample estimates and the ranges obtained when each randomized trio is omitted in turn. The first two rows are the primary homework estimands; the female-specific slope is secondary. Homework grades exclude the experimental bonus.

A.2 Heterogeneity by own gender

Table A2 complements the discussion in Section 5.1 by reporting the estimated composition effect separately for male and female students and testing the difference directly for each outcome.

Table A2: Female-peer effects by own gender

Outcome	Male effect	Female effect	Difference	RI p	RW p	N
Homework grade (no bonus)	1.732 (0.704)	3.759 (0.708)	2.027 (0.812)	0.177	0.197	77
Homework submission	0.146 (0.055)	0.193 (0.041)	0.047 (0.058)	0.598	0.598	77
Midterm performance	0.035 (0.247)	0.788 (0.337)	0.753 (0.400)	0.090	0.090	77
Final course grade	0.483 (0.493)	2.438 (0.832)	1.955 (0.842)	0.026	0.043	76
Cumulative GPA	0.158 (0.176)	0.215 (0.144)	0.057 (0.163)	0.772	0.772	81

Notes: Effects correspond to moving from zero to two female peers. The difference is the female effect minus the male effect. Models include cohort-course indicators and cohort-specific baseline-GPA slopes. Conventional standard errors, clustered by randomized trio, are in parentheses. RI p -values condition on random trio formation within cohort-course and are exact under the sharp null of no composition effect. They are not exact tests of equal gender-specific effects under an unknown nonzero common effect. Romano–Wolf (RW) adjusts the difference statistic within each outcome domain using common partitions. Homework and final course grades are measured in points, homework submission is a rate, midterm performance is standardized, and cumulative GPA is measured in points.

A.3 Binary contrast across outcomes

Table A3 reports the binary any-female-peer comparison across all outcomes. Homework submission also survives adjustment for the homework domain. The course outcomes and cumulative GPA do not provide stable evidence across the two exposure summaries, and no women-only estimate is conclusive after adjustment.

Table A3: Binary female-peer contrast across outcomes

Outcome	Effect	RI p	RW p	N
<i>Panel A. All students</i>				
Homework grade (no bonus)	1.561** (0.478)	0.006	0.015	77
Homework submission	0.099** (0.033)	0.033	0.033	77
Midterm performance	0.148 (0.176)	0.477	0.602	77
Final course grade	0.725 (0.394)	0.077	0.153	76
Cumulative GPA	0.160 (0.098)	0.082	0.117	81
<i>Panel B. Women only (secondary)</i>				
Homework grade (no bonus)	1.922* (0.512)	0.049	0.090	29
Homework submission	0.093* (0.042)	0.099	0.099	29
Midterm performance	0.447 (0.149)	0.227	0.227	29
Final course grade	1.422 (0.406)	0.097	0.134	29
Cumulative GPA	0.111 (0.105)	0.573	0.573	29

Notes: Effects compare students assigned at least one female peer with students assigned none. Models control for own sex when applicable, include cohort-course indicators and cohort-specific baseline-GPA slopes, and cluster conventional standard errors by randomized trio; standard errors are in parentheses. RI p -values condition on random trio formation within cohort-course. For all students, Romano–Wolf (RW) uses common partitions and adjusts jointly within the relevant outcome domain and across the linear and binary exposure summaries. For women, RW adjusts the binary estimates within each outcome domain. Homework and final course grades are measured in points, homework submission is a rate, midterm performance is standardized, and cumulative GPA is measured in points. Stars refer to RW-adjusted p -values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B Experimental Assignment and Data Construction

The analysis sample contains the 160 students originally randomized in the two experiments. Table B1 reports their allocation by cohort and arm. In 2011, two students added after randomization, one to the individual-incentive arm and one to control, are excluded throughout. The 2013 sample contains all 111 randomized students. Students enrolled in more than one participating course were assigned to only one experiment and therefore enter the analysis once.

Table B1: Randomized sample by cohort and arm

Cohort	Joint liability	Individual incentive	Pure control	Total
2011	24	13	12	49
2013	57	27	27	111
Total	81	40	39	160

Notes: The joint-liability arm contains eight randomized trios in 2011 and nineteen in 2013. The two post-randomization additions in 2011 are excluded from the randomized sample.

Balance in 2011 is evaluated using the covariate values available at assignment, including five peer-background values that were subsequently updated in the administrative records.

C Balance, Missingness, and Attrition

The 2011 assignment program repeatedly drew candidate three-arm assignments until 18 baseline variables satisfied a $p > 0.10$ balance rule for each of three arm-versus-rest comparisons, for 54 restrictions in total. In 2013, candidate assignments had to satisfy the same threshold for 13 covariates and four contrasts—joint versus control, individual versus control, joint versus individual, and any incentive versus control—for 52 restrictions. The program retained 300 acceptable assignments and selected number 53 at random. The realized assignment satisfies the stated rule in each cohort.

Baseline GPA and cumulative post-treatment GPA are observed for every student in the randomized sample. Proximal outcomes are incomplete for a small number of students: the composition analysis contains 77 observations for homework grade, homework submission, and the midterm; 76 for final course grade; and all 81 for cumulative GPA. In the 2013 full sample, five homework outcomes, four midterms, and sixteen final exams are missing. In the pooled randomized sample, joint tests of treatment-arm indicators and their interactions with sex, conditional on sex and cohort-course, do not reject equality across arms. The analysis neither imputes these outcomes nor drops students with observed administrative

outcomes because they failed to complete a follow-up survey. Survey nonresponse is not treated as attrition from the administrative GPA outcome. Within the composition sample, the homework grade is missing for two of 35 students assigned no female peers and two of 46 assigned at least one.

As a falsification check, female-peer share does not significantly predict baseline GPA under the same conditional repartitioning procedure ($p_{RI} = 0.380$ for the pooled slope and $p_{RI} = 0.139$ for the female-specific slope). The latter point estimate is nevertheless nontrivial in this small sample. For that reason, all post-treatment composition estimates include baseline GPA, with cohort-specific slopes in pooled models, and the unadjusted specifications are treated only as sensitivity checks.

D Conditional Randomization Procedure

For composition effects, outcomes, baseline variables, course membership, and own sex remain fixed. Within each cohort-course joint-liability arm, students are randomly reordered, divided into consecutive groups of three, and assigned the implied female-peer share. The regression statistic is then re-estimated. Pooled tests use 10,000 random partitions; cohort-specific tests use 5,000 per cohort. All tests are two-sided and use the plus-one correction. For pooled results, the same partitions generate the reference distributions for both female-peer share and the indicator of at least one female peer.

Romano–Wolf adjustments use the same partitions jointly. Coefficients are scaled by their randomization standard deviations, and the stepdown procedure compares each statistic with the largest remaining statistic in its family. For the overall estimates, the procedure also treats the linear and binary exposure summaries jointly.

The simulations use fixed seeds recorded in the replication documentation. Conventional standard errors cluster at the randomized trio.