Incentives, Interaction, and Peer Effects*

Matias Busso Veronica Frisancho Shirley Redroban

September 15, 2026

Abstract

This paper studies how peer interaction and incentives jointly shape performance in team settings. We conducted two framed field experiments with seventh-grade students in public middle schools in Mexico City, randomly assigning students to one of three peer-interaction regimes during a practice activity: self-selected pairs (endogenous pairing), randomly assigned pairs (exogenous pairing), and randomly assigned pairs whose final grade was the average of both partners' individual scores (shared grades). Exogenous pairing increases performance during practice by about 0.15σ but has no effect on final exam outcomes. Endogenous pairing reduces final achievement by roughly 0.07σ. In contrast, shared grades generate large practice gains (about 0.33σ) and increase final exam performance by about 0.11σ. Effects are heterogeneous by pair composition: endogenous pairing is particularly harmful for mixed-ability pairs and for pairs without pre-existing social ties. At the same time, shared grades improve outcomes for low-ability students paired with high-ability peers. We develop a conceptual framework in which peer interaction affects outcomes through regime-specific effort costs, social responses during interaction, and the persistence of peer-generated inputs. The results reject a benchmark in which students respond only to direct grading incentives and highlight how institutional design determines whether peer interaction enhances or undermines performance.

JEL classifications: C93, I21, I28, O15.
Keywords: peer-to-peer learning, peer effects, class assignment, Mexico.

*Matias Busso: Research Department, Inter-American Development Bank (mbusso@iadb.org). Veronica Frisancho: Chief Economist, Development Bank of Latin America and the Caribbean, (vfrisancho@caf.com). Shirley Redroban: Research Department, Inter-American Development Bank and University of Maryland (shirleyre@iadb.org). We are very grateful to the Autoridad Educativa Federal de la Ciudad de Mexico of the Secretaria de Educacion Publica de Mexico. In particular, we thank the support throughout the project of Sofia Leticia Morales Garza, Maria del Carmen Ayllon, Arturo Velatti, Norma Alegria, and Homar Ortiz. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the the Inter-American Development Bank, its Board of Directors, or the countries they represent.

1 Introduction

Peer effects shape behavior in a wide range of economic environments. Workers collaborate in teams, firms link compensation to group performance, students learn together, and individuals make decisions within social networks. In each of these settings, individual effort and performance depend not only on one's own ability, but also on the characteristics of peers and the institutional rules governing interactions or incentives. A central question, therefore, arises:

Understanding this question is challenging because peer effects combine multiple mechanisms. Interaction may alter effort costs, generate social pressure or accountability, facilitate knowledge transmission, or introduce coordination frictions such as free-riding. Moreover, outcomes may depend on how peers are matched and how rewards are structured. Distinguishing among these channels requires settings in which both peer composition and incentive schemes can be experimentally manipulated.

Education offers a natural environment for examining these channels, as classrooms are spaces characterized by repeated interactions, teamwork, and performance comparisons. Yet evidence on peer-to-peer learning remains mixed. While some studies document positive spillovers from higher-ability peers (Hoxby, 2000; Berlinski et al., 2023), others find limited (Feld and Zölitz, 2017) or heterogeneous effects (Gu, 2023). The literature also highlights that peer effects are context-dependent (Sacerdote, 2011), yet there is limited evidence of how incentive alignment and match formation interact to determine whether peer collaboration enhances learning.

In this paper, we examine how peer-to-peer learning happens. We implement two framed field experiments with seventh-grade students in public middle schools in Mexico City. The intervention varies both (i) how peers are matched during a practice activity and (ii) how final exam grades are determined. Students participate in a one-day session that includes a standardized lesson, a practice test completed either individually or in pairs, and a final exam taken individually. We experimentally assign students to one of three peer interaction

structures: self-selected pairs (endogenous pairing), randomly assigned pairs (exogenous pairing), and randomly assigned pairs whose final grade is the average of both partners' individual scores (shared grades). A control group works individually throughout. This design allows us to isolate three distinct dimensions of peer interaction: match formation, peer composition, and incentive alignment.

We document three main findings. First, peer interaction alone is not sufficient to improve learning. Random pairing increases performance during the practice stage by approximately 0.15σ, but these gains do not translate into gains on the final exam. This pattern indicates that interaction can affect contemporaneous behavior without generating durable learning.

Second, allowing students to choose their own partners does not enhance performance and, in fact, reduces final exam scores by about 0.07σ. The negative effects are concentrated among pairs lacking pre-existing social ties and among mixed-ability matches. Although endogenous pairing generates socially proximate matches on average, it also produces a subset of low-quality pairings characterized by coordination problems and disruptions. These relational frictions undermine learning persistence.

Third, and most strikingly, linking students' grades within randomly assigned pairs substantially improves both practice and final performance. Shared grades increase practice scores by roughly 0.33σ and final exam scores by about 0.11σ relative to the control group. Because shared grades mechanically weaken individual incentives under a standard model of self-interested effort, this result rejects a benchmark in which students respond only to direct grading incentives. Instead, the evidence suggests that aligning incentives within the pair affects social responses and enhances the benefits of peer interaction.

We develop a conceptual framework to discipline the interpretation of these findings. In the model, peer interaction can affect outcomes through three channels: variation in effort costs during practice, social responses that depend on partner characteristics and relational proximity, and the productivity and persistence of peer-generated inputs. The shared grades regime provides a sharp test of a pure incentives benchmark: if students respond only to the

marginal return to their own grade, performance should decline when grades are averaged. The fact that performance instead improves implies that social accountability and incentive alignment play a central role.

Heterogeneity patterns reinforce this interpretation. Low-ability students paired with high-ability peers experience large gains under shared grades, while endogenous pairing is particularly harmful for mixed-ability pairs without prior social ties. Familiarity amplifies the positive effects of shared grades but does not independently generate gains absent incentive alignment. These results indicate that peer effects depend jointly on partner ability, relational proximity, and institutional structure.

This paper contributes to several strands of the literature on peer effects. It builds on the large body of work identifying peer effects in educational settings (Sacerdote, 2011, 2014), by showing that peer exposure alone is not sufficient to generate persistent learning gains. Consistent with recent evidence documenting heterogeneous and context-dependent effects (Feld and Zölitz, 2017; Booij et al., 2017), we show that the incentive environment governing interaction is central to whether peer effects emerge and persist.

The paper also speaks directly to the literature on peer group formation and endogenous sorting. Carrell et al. (2013) demonstrate that peer effects are influenced by group formation strategies and that efforts to design peer composition can fail if social interactions adapt endogenously. Our findings complement this insight by holding peer composition constant while varying the incentive structure and the peer-matching regime. We show that even with identical ability exposure, outcomes differ sharply depending on whether students self-select into pairs or face shared incentives. In this sense, we provide evidence that peer effects are not solely a function of composition but also of the institutional rules governing peer interaction.

Our findings further contribute to the literature investigating the mechanisms underlying peer effects. While several studies attribute peer spillovers to peer-to-peer interaction or social networks (Wu et al., 2023; Coveney and Oosterveen, 2021), relatively less attention

has been paid to how incentive structures shape those interactions. By holding peer composition constant and varying the incentive regime, we isolate the role of shared performance incentives in governing effort provision within pairs. Our findings show that incentives do not merely alter individual effort mechanically; they shape the nature of collaboration itself. Finally, consistent with recent reviews emphasizing the importance of understanding heterogeneity and mechanisms in peer effects (Barrios-Fernandez, 2023), our results clarify how peer effect estimates should be interpreted. We document a sharp distinction between contemporaneous collaboration effects and persistent learning gains. Peer interaction without aligned incentives improves performance during joint work but does not translate into higher individual performance once collaboration ends. In contrast, shared incentives generate gains that persist beyond the collaborative stage. This suggests that variation in institutional design may underlie the substantial heterogeneity in peer effect estimates documented in the literature.

More broadly, our findings suggest that peer-to-peer learning is not inherently beneficial. Its effectiveness depends on how interaction is structured and how incentives govern collaborative effort. Even modest changes to grading rules can fundamentally alter how peer interaction translates into individual performance. Peer effects are not purely compositional; they are mediated by institutional factors. Developing effective collaborative environments requires understanding how incentives, peer matching, and social responses influence each other.

The remainder of the paper is structured as follows: Section 2 describes the experimental design. Section 3 presents the conceptual framework. Section 4 describes the empirical strategy. Section 5 presents the results. Section 6 discusses the role of peer ability. Section 7 shows possible mechanisms at play, and Section 8 concludes.

2 Experimental Design

2.1 Setting and Intervention

The Mexican public school system comprises four educational levels: preschool (beginning at age three), elementary school (grades one through six), middle school (grades seven through nine), and high school (grades ten through twelve). Each year, approximately 160,000 students enter middle school in Mexico City, the country's largest school district, which serves over two million students. About 85% of these students attend public schools.

Our intervention targeted seventh-grade (first-year middle school) students enrolled in public middle schools in the 2015-2016 school year. The main objective was to examine how peer interactions shape performance through a short, structured mathematics activity. All participating students took part in a one-day session that included a math lesson, a practice test, and an individual final exam. The experimental design manipulated both the structure of peer interaction during the practice test and the incentive environment surrounding the final exam, enabling us to isolate the mechanisms through which peer effects operate.

Specifically, we implemented a control group and three treatment arms designed to experimentally vary peer interactions during the practice test.

(1) Control group: Students worked individually in the practice test.

(2) Endogenous pairing: Students selected their own partner and worked together in the practice test. [1]

(3) Exogenous pairing: Students were randomly paired to work together in the practice test. [2]

(4) Exogenous pairing plus shared grades: Students were randomly paired as in treatment

[1] Students were instructed to select a partner of their choice before beginning the practice exercise. No additional guidance was provided.

[2] In treatment arms involving exogenous pairing, students were instructed to work with the peer seated next to them; seating was arranged alphabetically.

arm (3) and were informed that their final grade would be the average of both partners' individual scores.

Under all treatment arms, students completed the final test individually. Panel A of Table 1 summarizes the main features of each treatment arm, including whether the practice test was completed individually or in pairs, how pairs were formed (endogenously or exogenously assigned), and the grading structure. The intervention was designed to either foster different forms of peer interaction or remove them entirely, as in the control group.

Table 1: Research Design

	Control	Exogenous pairing	Endogenous pairing	Exogenous pairing + shared grades
Panel A: Description				
Practice test:	Individual	Pairs	Pairs	Pairs
Pairs formed:	-	Randomly	By students	Randomly
Final grade:	Individual	Individual	Individual	Pair's average
Panel B: Probability of being assigned to each treatment arm:				
Experiment 1	1/3	1/3	1/3	
Experiment 2	1/3	1/3		1/3
Panel C: Sample size				
# schools	48	48	24	24
# homerooms	144	144	72	72
# students	1686	1562	792	713

Notes. The table presents the research design for the two experiments and the corresponding treatment arms: Control, Exogenous pairing, Endogenous pairing, and Exogenous pairing with shared Grades. Panel A describes the administration of the practice test, the procedures used to form pairs, and the method for computing final grades after the assessment. Panel B reports the probability of assignment to each experiment. Panel C summarizes the number of participating schools, homerooms, and students.

To deliver the intervention, we implemented two experiments. Under the first experiment, each student faced a one-third probability of being assigned to the control, endogenous pairing, or exogenous pairing– that is, treatment arms (1), (2), or (3). Under experiment 2, students faced a one-third probability of being assigned to the control, exogenous pairing, or exogenous pairing with shared grades– in other words, treatment arms (1), (2), or (4). The implementation followed a standardized protocol across all participating schools. Upon arrival, students proceeded to their homerooms as usual. Soon after, students from three different homerooms were brought together into a larger shared space, where they

were randomly assigned to one of the corresponding experimental groups, depending on the experiment they were participating in. Students were then taken to a separate classroom corresponding to their assigned experimental group. Figure 1 shows a schematic overview of the intervention rollout.

Each group received a 15-minute mathematics lesson focused on fractions and decimals, delivered by a trained instructor. The lesson covered fundamental concepts, including definitions, basic operations (addition, subtraction, multiplication), and conversions between fractions and decimals. The choice of this content was made in coordination with local school authorities, who highlighted the pedagogical importance of these topics for students in seventh grade. The instructor was a qualified mathematics teacher with prior experience teaching middle school students.

Following the lesson, students completed a 15-minute practice test with 13 questions on fractions and decimals. Depending on their assigned condition, the practice test was completed either individually or in pairs. Afterward, all students took a 15-minute individual final exam consisting of 10 questions on the same topic. After completing the final exam, students went back to their usual homeroom and continued their regular school routines. Note that students were told that their final exam scores would be shared with their math teacher.

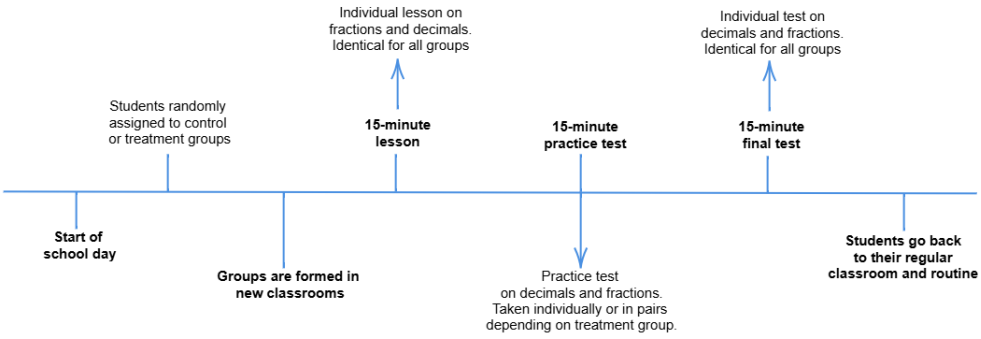
Following the lesson, students completed a 15 minute practice test consisting of 13 questions on fractions and decimals. Depending on their assigned condition, students completed the practice test either individually or with a partner. During the practice test, instructor involvement was intentionally limited and followed the same protocol across experimental conditions. Students with questions were required to raise their hand and state the question aloud, and the instructor responded aloud so that all students in the classroom could hear both the question and the response. The instructor did not confirm whether individual answers were correct, grade students' work, or provide worked solutions. Thus, instructor assistance operated as a common source of clarification across conditions rather than as

student or pair specific feedback.

Immediately after the practice test, all students completed a 15 minute individual final exam consisting of 10 questions on the same material. Unlike during the practice stage, students were not permitted to ask questions or receive assistance from their partner during the final exam. No answers or solutions to the practice test were revealed before the final exam. Students were informed that their final exam scores would be shared with their mathematics teacher. After completing the exam, students returned to their usual homerooms and resumed their regular school activities.

Student survey responses indicate that the mathematics lesson was implemented consistently and was well received by students. Appendix Figure A.1 presents students' assessments of lesson quality. Most students rated the lesson's quality as high. Additionally, when questions arose during the lesson, more than 92% of students rated the substitute teacher's responses as good or very good (see Appendix Table A.2).

Figure 1: Intervention description



Notes: The Figure presents the structure of the intervention. During the 15-minute practice test, students in the control group worked individually. In the endogenous pairing group, students worked in self-selected pairs. In the exogenous pairing treatment, students worked in randomly assigned pairs. In the exogenous plus shared grades condition, students also worked in randomly assigned pairs but were informed that their final exam grade would be the average of their own and their partner's score.

2.2 Randomization

The experimental sample consists of seventh-grade students in 48 public middle schools selected randomly at the strata level in Mexico City during the 2015–2016 academic year. These schools were originally part of a broader pilot program involving 171 public schools that tested alternative grouping strategies based on prior student performance. As part of that pilot, schools were stratified into 57 groups, each comprising three schools with similar baseline characteristics, including school size, socioeconomic composition, and academic achievement (see Busso and Frisancho (2025) for more details).

From the 57 original strata, we randomly selected 16 strata (48 schools in total) to participate in our intervention. Once selected, the 16 strata were randomly assigned to one of two field experiments: 8 strata (24 schools) to Experiment 1 and the remaining 8 strata (24 schools) to Experiment 2. Experiment 1 consisted of three peer interaction conditions: control, endogenous pairing, and exogenous pairing, while Experiment 2 included control, exogenous pairing, and exogenous pairing with shared grades. In both experiments, three homerooms were chosen in each school to participate in the intervention.

The design of the intervention was guided by both logistical constraints and statistical power considerations. Our initial objective was to evaluate four distinct peer interaction models within each school. However, implementing all four treatment arms simultaneously was not feasible given classroom capacity limitations and the need to maintain adequate sample sizes per condition. To address these constraints, we organized the study into two separate experiments, each comprising the three treatment arms mentioned above. The exogenous pairing condition was included in both experiments to provide a common benchmark, enabling direct comparisons between endogenous pairing and shared grades under a consistent reference point.

Within each school, students were assigned to treatment arms through a quasi-random procedure. Students were first ordered alphabetically by last name and then sequentially assigned

to one of the three treatment conditions.[3] This approach ensured balance in group sizes and was feasible to implement on site.

Panel B of Table 1 reports the assignment probabilities to each treatment arm by experiment. Panel C presents the distribution of schools, homerooms, and students across treatments. Because the control and exogenous pairing conditions were implemented under both experiments, twice as many schools and homerooms participate in these arms relative to endogenous pairing and shared grades conditions.

2.3 Pairs information

Table 2 presents descriptive statistics on the characteristics of student pairs formed under each treatment arm. The differences across treatments confirm that the experimental design successfully generated meaningful variation in peer composition.

Under the endogenous pairing condition, where students selected their own partners, we observe some evidence of assortative matching on social and observed characteristics. A large majority of pairs were formed by students of the same gender (87%), from the same homeroom (84%), and with whom they had a prior relationship (90%). These patterns suggest that when given the opportunity, students strongly prefer to work with peers they are familiar with, reflecting a desire for comfort, trust, or ease of communication.

Interestingly, however, the extent of sorting by academic ability in the endogenous pairing condition is relatively small. Although high-low ability pairings were slightly less common than in the exogenous pairing conditions (40% vs. 48% in exogenous pairing), the overall distribution of ability types remains broadly similar. This suggests that while middle school students prioritize social familiarity when choosing a peer, they place less weight on their partner's academic ability, at least in this short-term, low-stakes setting. This may reflect limited observability of academic skill, or that students simply value social comfort over potential academic gains when working in pairs.

[3] In cases where the number of students was not divisible by three, the remaining students were assigned

Table 2: Pairs' Characteristics

	Exogenous pairing (1)	Endogenous pairing (2)	Shared Grades (3)	p-value (1)=(2) (4)	p-value (1)=(3) (5)	p-value (2)=(3) (6)
Ability						
High-High	0.33	0.36	0.32	0.16	0.76	0.15
High-Low	0.48	0.40	0.46	0.00	0.51	0.01
Low-Low	0.20	0.25	0.22	0.00	0.24	0.19
Gender						
Female-Female	0.22	0.45	0.29	0.00	0.00	0.00
Male-Female	0.46	0.13	0.46	0.00	0.90	0.00
Male-Male	0.32	0.42	0.25	0.00	0.00	0.00
Homeroom						
Same homeroom	0.23	0.84	0.18	0.00	0.00	0.00
Different homeroom	0.77	0.16	0.82	0.00	0.00	0.00
Relationship						
Known	0.43	0.90	0.39	0.00	0.07	0.00
Unknown	0.57	0.10	0.61	0.00	0.07	0.00
Observations	1538.00	786.00	702.00			

Notes. The table reports the characteristics of student pairs formed under each treatment condition. Columns (1), (2), and (3) present averages of ability, gender composition, homeroom assignment, and pre-existing relationship indicators for pairs in the Exogenous Pairing, Endogenous Pairing, and Exogenous Grading + Shared Grades treatments, respectively. Column (4) reports p-values for tests of equality in means between Columns (1) and (2), Column (5) between Columns (1) and (3), and Column (6) between Columns (2) and (3).

Table 2 also shows that both exogenous pairing and exogenous pairing plus shared grades created more heterogeneous pairs. These groups had a significantly higher proportion of mixed-gender, mixed-homeroom, and unknown-peer matches. While both treatments used the same assignment method, the shared grades condition produced slightly fewer male-male or known-peer pairs, possibly reflecting differences in the initial composition of the origin homerooms.

Taken together, these patterns confirm that the experimental design generated substantial variation in the types of peer interactions students experienced. This variation allows us to test not only whether working with a peer matters, but also whether the characteristics of the peer and the incentive structure governing that interaction influence short-run academic outcomes.

to the control group.

2.4 Data

We use a combination of administrative records, surveys, and student performance assessments collected before and during the implementation of the framed experiments.

Administrative records: We use student-level application records from the centralized admission system in place to regulate admission to public middle schools in Mexico City. These records include students' scores on the standardized admission exam (IDANIS), cumulative elementary school GPA, and basic demographic information such as gender, date of birth, and whether the student has any special learning or physical need.[4] The registration form also captures household characteristics including the education level and age of each parent, household size, and whether the student lives with both parents. Additional information gathered in the form includes anthropometric measures (height and weight) and the number of siblings already enrolled in the student's preferred schools.

We also rely on administrative records at the school level to collect additional information such as school size, location, and past average academic performance in the admission exam.

Students tests: To construct our main outcome measures, we administered two academic assessments during the intervention: a 15-minute practice test and a 15-minute final test. Both assessments were developed based on Mexico's national standardized testing framework, particularly the IDANIS exam. Since students are familiar with this format, the likelihood of test misinterpretation or confusion is minimal.

The practice test consisted of 13 questions, and the final test included 10 questions, both covering the same material introduced during the instructional session. All students received identical instructions and completed the assessments under the same conditions. The primary outcome is the standardized score on the final test, defined as the number of correct answers normalized by the control group's mean and standard deviation. We also analyze the standardized score on the practice test to document performance during the interaction

[4]The IDANIS is a standardized, high-stakes exam administered at the end of Grade 6 to allocate students to middle schools.

stage.

Students' survey: After completing the final test and prior to students returning to their homerooms, all students filled out a short questionnaire. This survey collected feedback on the perceived quality and difficulty of the lesson, as well as the effectiveness of the instructor. Students assigned to the endogenous and exogenous treatment arms (i.e., not in the control group) were also asked about their experience working with a partner. This included whether they knew their partner beforehand, how they collaborated during the practice test, and their expectations regarding their partner's performance. These responses allow us to examine the potential mechanisms behind treatment effects.

3 Conceptual Framework

This section presents a simple conceptual framework to discipline the interpretation of the experimental results. Its goal is not to provide a complete model of learning or peer interaction, but rather to establish a clear behavioral benchmark and to clarify how departures from this benchmark can be interpreted using the experimental design.

Environment and Timing In our setting, students participate in two stages. In the first stage, they complete a practice activity under a peer interaction regime. In the second stage, they take a final exam individually, after peer interaction has ended. Only the final exam grade is reported to the mathematics teacher.

Let X_{0i} denote student i's outcome in the practice stage and X_{1i} denote the final exam score. Peer interaction affects behavior only in the practice stage; the final exam is taken in isolation.

Students are randomly assigned to one of four regimes during the practice stage: a control group that works individually, endogenous pairing (EP), exogenous pairing (ExP), and exogenous pairing with shared grades (SG).

Production and Effort Each student i has baseline ability a_i. During the practice stage, the student chooses effort $e_{0i} \geq 0$. Practice-stage output is given by

$$X_{0i} = a_i + \gamma_0 e_{0i} + \psi_r(a_j, k_{ij}) + \varepsilon_{0i},$$

where $\gamma_0 > 0$ denotes the productivity of effort during practice, and $\psi_r(a_j, k_{ij})$ captures peer-related inputs that arise under regime r. These inputs may depend on the peer's ability a_j and on whether the peer is known to the student (k_{ij}), reflecting factors such as explanations, coordination, or the quality of interaction. The term ε_{0i} captures idiosyncratic shocks to practice performance.

In the final stage, students take the exam individually and choose effort $e_{1i} \geq 0$. Final exam performance is given by

$$X_{1i} = a_i + \gamma_1 e_{1i} + \lambda_r \, \psi_r(a_j, k_{ij}) + \varepsilon_{1i},$$

where $\gamma_1 > 0$ denotes the productivity of effort on the final exam. The parameter $\lambda_r \in [0, 1]$ is a *pass-through term*, which governs the extent to which peer-related inputs generated during the practice stage persist into final exam performance. A higher λ implies that learning, strategies, or other inputs acquired during peer interaction translate more strongly into individual performance on the exam. When $\lambda = 0$, peer interaction affects only practice-stage outcomes; when $\lambda > 0$, practice-stage peer inputs carry over to the final stage. The term ε_{1i} represents idiosyncratic shocks to final performance.

Grading Rules and Student Maximization Problem Let G_i denote the grade reported to the teacher. Under individual grading (control, EP, ExP),

$$G_i = X_{1i}.$$

Under shared grades (SG),

$$G_i = \frac{1}{2}X_{1i} + \frac{1}{2}X_{1j}.$$

Equivalently,

$$G_i = (1 - \sigma_r)X_{1i} + \sigma_r X_{1j},$$

where $\sigma_r = 0$ under individual grading and $\sigma_r = \frac{1}{2}$ under shared grades.

Students choose effort to maximize utility from the reported grade net of effort costs and social responses arising during the practice stage:

$$\max_{e_{0i}, e_{1i} \geq 0} \; G_i - \frac{c_r}{2}e_{0i}^2 - \frac{c_1}{2}e_{1i}^2 + \theta_r \, \mathcal{S}_i(a_j, k_{ij}),$$

where $c_r > 0$ denotes the cost of exerting effort during the practice stage (which may vary across regimes), and $c_1 > 0$ denotes the cost of exerting effort on the final exam. The term $\mathcal{S}_i(a_j, k_{ij})$ captures social responses arising from peer interaction during practice. These may reflect social pressure, reputational concerns, encouragement, or other social forces operating within the pair. Social responses may depend on peer characteristics, such as ability a_j, and on the nature of the relationship between partners, captured by k_{ij}. The parameter θ_r is regime-specific, and its sign is a priori ambiguous.

Social responses operate only during the practice stage, when interaction occurs, and are absent during the final exam, which is taken individually.

Benchmark: Incentives Only We begin with a benchmark in which students are purely self-interested ($\theta_r = 0$) and peer interaction affects behavior only through the incentives embedded in the grading rule. In particular, peer interaction does not affect effort costs ($c_r = c$ for all r), the pass-through term ($\lambda_r = \lambda$ for all r), or the production of practice-stage inputs beyond what is mechanically captured by grading incentives.

Under this benchmark, the shared-grades regime mechanically reduces incentives. The

marginal benefit of individual performance is

$$\frac{\partial G_i}{\partial X_{1i}} = 1 - \sigma_r,$$

which equals $1/2$ under SG, compared to 1 under individual grading. Holding effort costs and the persistence of practice-stage inputs fixed, this incentive reduction predicts lower optimal final effort and lower final exam performance under shared grades relative to individual grading.

The purpose of this benchmark is to generate a sharp null hypothesis: if students respond only to grading incentives and peer interaction does not affect costs, persistence, or social preferences, shared grades should reduce performance.

Mechanisms Beyond the Benchmark The benchmark corresponds to a set of restrictions on the maximization problem described above: $\theta_r = 0$, $c_r = c$, and $\lambda_r = \lambda$ for all r, with peer interaction affecting outcomes only through the grading rule. Relaxing these restrictions allows peer interaction to affect behavior through regime-specific variation in effort costs (c_r), social responses (θ_r), and the productivity or persistence of peer inputs (ψ_r, λ_r).

First, peer interaction may alter the effective cost of exerting effort during practice. Working with a peer may reduce cognitive or coordination costs, increase engagement, or otherwise make effort less costly. In the model, this corresponds to variation in c_r across regimes.

Second, peer interaction may generate social responses. When effort is observable within the pair, students may anticipate reputational concerns, peer pressure, encouragement, or other social forces linked to their partner's characteristics or the nature of the relationship. In the model, this operates through $\theta_r \mathcal{S}_i(a_j, k_{ij})$. Depending on the sign of θ_r, social responses may affect incentives to exert effort during practice. Because interaction occurs only in the practice stage, these forces are absent in the final exam.

Third, peer interaction may affect the quality of practice-stage inputs $\psi_r(a_j, k_{ij})$ and the extent to which those inputs persist. In the model, this operates through both $\psi_r(\cdot)$ and the

regime-specific pass-through term λ_r. Even if grading incentives are weaker, peer interaction may improve outcomes by generating higher-quality practice-stage inputs or by increasing the extent to which those inputs carry forward into final performance.

Empirical Implications Let $\beta_r^{(0)}$ denote the estimated effect of regime r on practice outcomes X_{0i} and $\beta_r^{(1)}$ the corresponding effect on final exam outcomes X_{1i}, each relative to the control group.

Because both SG and ExP involve random assignment of peers, comparing these regimes isolates the role of the grading rule (σ_r) while holding peer composition fixed. Under the benchmark model, where $\theta_r = 0$, $c_r = c$, and $\lambda_r = \lambda$, shared grades should reduce performance relative to exogenous pairing due to weaker individual incentives. If shared grades perform similarly or better, the benchmark is rejected, implying that at least one of the additional channels in the maximization problem, variation in effort costs (c_r), social responses (θ_r), or the productivity and persistence of peer inputs (ψ_r, λ_r), must be operative.

Comparing effects on practice and final outcomes helps distinguish among these channels. Large effects on X_{0i} combined with smaller or zero effects on X_{1i} suggest mechanisms operating contemporaneously during peer interaction without persistence, consistent with changes in c_r or θ_r that affect practice effort but do not translate into durable inputs. In contrast, positive effects on X_{1i} indicate that practice-stage inputs persist through the regime-specific pass-through term λ_r.

Heterogeneity by peer characteristics further disciplines interpretation. Stronger effects when peers are socially close or familiar are consistent with social responses operating through $\mathcal{S}_i(a_j, k_{ij})$. Stronger effects when paired with higher-ability peers, particularly on final outcomes, point to productive collaboration operating through $\psi_r(a_j, k_{ij})$ and persistence through λ_r.

Because effort is not directly observed, treatment effects on final exam performance may reflect both persistence and changes in final-stage effort. The empirical analysis therefore

examines observable proxies for effort to assess whether treatment effects primarily operate through practice-stage behavior or through persistent inputs generated during interaction.

4 Empirical Strategy

We estimate the impact of different peer interaction structures on student academic performance by pooling data from both experiments and estimating average treatment effects using the following OLS specification:

$$Y_{ihs} = \alpha + \beta_{EP} D_{ihs}^{EP} + \beta_{ExP} D_{ihs}^{ExP} + \beta_{SG} D_{ihs}^{SG} + \delta X_{ihs} + \mu_h + \varepsilon_{ihs} \tag{1}$$

where Y_{ihs} is the outcome of interest for student i in homeroom h in school s. The coefficients β_{EP}, β_{ExP}, and β_{SG} capture the effects of assignment to the endogenous pairing, exogenous pairing, and exogenous pairing plus shared grades treatment conditions, respectively. The reference group consists of students in the control condition, who completed all tasks individually and our main outcome of interest is the final test score standardized by the mean and standard deviation of the control group.

We also include additional individual and homeroom characteristics as controls, denoted by vector X_{ihs}. This vector comprises the student's baseline ability using the aggregate admission standardized test score. We also include homeroom fixed effects, μ_h, to account for random assignment at the homeroom level and we calculate robust standard errors.

In sum, we estimate our treatment effects by measuring differences in the final test score among students of the same homeroom assigned to different working arrangements.[5]

4.1 Experimental Validity

To evaluate the validity of the random assignment, Table 3 reports the baseline characteristics of students across treatment groups. Column 1 shows the mean for each variable in the

[5] Appendix Tables A.3 and A.4 report balance separately for each experiment.

control group, while Columns 2 to 4 present the estimated differences between each treatment condition and the control group. All specifications control for homeroom fixed effects and use robust standard errors.

Table 3: Balance

Baseline variable	Mean control	Exogenous pairing	Endogenous pairing	Shared grades	Observations
	(1)	(2)	(3)	(4)	(5)
Baseline ability	31.019	0.544	–0.256	0.470	4606
	[10.746]	[0.288]*	[0.373]	[0.402]	
Years of educ., mother	10.043	0.151	0.237	–0.003	4606
	[3.225]	[0.110]	[0.148]	[0.142]	
Living with both parents	0.687	0.022	–0.011	0.014	4606
	[0.464]	[0.016]	[0.022]	[0.023]	
Household size	4.766	–0.041	0.002	–0.008	4606
	[1.596]	[0.057]	[0.079]	[0.075]	
Male	0.491	0.058	–0.021	–0.002	4606
	[0.500]	[0.018]***	[0.024]	[0.025]	
Older than 12	0.032	0.001	0.003	–0.001	4606
	[0.177]	[0.006]	[0.008]	[0.009]	
Primary school GPA	8.636	0.028	0.014	0.025	4606
	[0.621]	[0.020]	[0.027]	[0.027]	
Special needs	0.023	–0.008	–0.002	–0.011	4606
	[0.149]	[0.005]*	[0.006]	[0.007]	

Note: The table reports the mean of each covariate in the control group and the estimated treatment effects for the Exogenous Pairing, Endogenous Pairing, and Exogenous Pairing + Shared Grades groups. The raw initial exam score ranges from 0 to 60. All specifications include classroom fixed effects and the full set of baseline controls. Robust standard errors are used in all specifications. * significant at 10%; ** significant at 5%; *** significant at 1%.

We find no systematic differences between students who were randomized to each treatment arm. For most characteristics, we cannot reject the null hypothesis that the average baseline student-level mean is equal across each treatment arm and the control group. Students in all treatment arms have similar baseline academic achievement as measured both by the standardized admission test score and their primary school GPA.[6] We also rule out differences related to family background: the average control student has parents with 10 years of education, lives in a household with nearly 5 members, and has a 68% chance of living with both parents.

The only statistically significant differences at the 5 % level are observed for the male and special needs indicators within the exogenous pairing group. Although these differences are small in magnitude, baseline controls are incorporated in all outcome regressions to account

[6]It should be noticed that exogenous pairs formation shows slightly higher score in the standardized admission test (10% level).

for the imbalance.

5 Estimated Effect on Students' Performance

Table 4 reports the estimated treatment effects of each peer interaction condition on students' performance in both the practice exam and the final exam. We sequentially add controls across columns to test the robustness of our estimates. All specifications include homeroom fixed effects and robust standard errors.

Table 4 highlights three key findings. First, exogenous pairing, randomly assigned pairs without shared incentives, produces a positive and statistically significant increase in practice performance of approximately 0.15 σ. However, these gains do not translate into improvements in final exam performance. Within the conceptual framework, this pattern is consistent with peer interaction affecting contemporaneous practice outcomes, either through reduced effective effort costs or improved coordination, without generating inputs that persist into the final exam (i.e., a low effective pass-through term).

Table 4: Treatment Effects on Exam Test

	Practice Quiz			Final Quiz		
	No controls (1)	Initial score (2)	All controls (3)	No controls (4)	Initial score (5)	All controls (6)
Exogenous formation	0.169***	0.152***	0.149***	0.039	0.018	0.009
	[0.035]	[0.034]	[0.034]	[0.033]	[0.031]	[0.031]
Endogenous formation	–0.007	0.001	–0.003	–0.081*	–0.071*	–0.073*
	[0.044]	[0.042]	[0.042]	[0.044]	[0.041]	[0.041]
Shared grades	0.353***	0.339***	0.334***	0.136***	0.117***	0.113**
	[0.049]	[0.048]	[0.048]	[0.048]	[0.045]	[0.045]
Observations	4,606	4,606	4,606	4,606	4,606	4,606

Note: The table reports the estimated treatment effects following the specification outlined in Equation (1). The dependent variable is standardized to have mean zero and standard deviation one. Columns 1-3 correspond to the practice quiz and columns 4-6 to the final quiz. Within each group, columns 1 and 4 present estimates without controls, columns 2 and 5 include the students' baseline test score as a control, and columns 3 and 6 add the full set of baseline covariates, including an indicator for whether the baseline score was imputed and a dummy for students who failed four or more subjects in the previous year. All regressions include classroom fixed effects. Robust standard errors are used in all specifications. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Second, endogenous pairing, where students select their own partners, has no meaningful effect on practice performance but leads to a statistically significant decline in final exam

scores of roughly 0.07 σ. The absence of practice gains combined with negative final effects suggests that self-selection may introduce match-specific frictions or distractions that do not enhance learning during interaction and may even undermine the accumulation or persistence of knowledge. In the model, this is consistent with higher effective effort costs or adverse social responses that do not generate productive peer inputs.

Third, exogenous pairing with shared grades produces large gains in practice performance, approximately 0.33 σ, and these gains partially persist into the final exam, raising scores by about 0.11 σ relative to the control group. Because both exogenous pairing and shared grades involve random peer assignment, comparing these regimes isolates the effect of the grading rule. Under the benchmark model, shared grades mechanically reduce incentives and should therefore reduce performance. The fact that shared grades outperform exogenous pairing decisively rejects this prediction. Instead, the evidence indicates that linking outcomes within the pair alters behavior during practice in a way that increases both contemporaneous performance and the persistence of practice-stage inputs.

Taken together, these results indicate that peer interaction affects outcomes primarily through the practice stage, and that final exam effects depend on whether practice-stage gains translate into persistent inputs. Random pairing increases practice output X_{0i} but does not affect final performance X_{1i}, suggesting limited persistence. In contrast, when peer interaction is combined with shared incentives, practice gains are larger and partially persist, implying that incentives reshape behavior during interaction in ways that generate durable improvements in performance. These patterns reject the benchmark model in which shared grading merely weakens individual incentives and instead point to mechanisms operating through effort costs, social responses, or the productivity and persistence of peer inputs.

6 Peers' Ability when Working in Pairs

The conceptual framework highlights that peer interaction during the practice stage may operate through peer-related inputs that depend on partner characteristics, particularly peer ability. To examine this, we assess whether treatment effects vary with the ability composition of the pair.

Using the median of the baseline admission test, we classify students as high or low-ability and modify equation (1) to allow treatment effects to vary with peer ability. Specifically, we estimate the model separately for high and low-ability students, replacing the treatment indicators with interactions that distinguish whether a student is paired with a higher or lower-ability peer within each peer-interaction regime. In all specifications, the omitted category includes students of the respective ability type who worked alone in the control group.

Table 5 shows that treatment effects vary systematically with pair composition, consistent with a framework in which peer inputs $\psi_r(a_j, k_{ij})$ depend on partner ability.

For high-ability students paired with other high-ability peers, both exogenous and endogenous pairing produce modest gains in practice performance, although only the exogenous pairing effect is statistically significant. These gains do not persist into the final exam, suggesting that interaction lowers contemporaneous effort costs during practice but does not generate durable inputs (i.e., limited effective pass-through).

Under shared grades, however, high–high pairs exhibit substantially larger practice gains that partially persist into the final exam. Because peer composition is random in both exogenous pairing and shared grades, this contrast isolates the role of incentives. The improvement under shared grades, despite mechanically weaker individual incentives, rejects the benchmark model and indicates that linking outcomes within the pair reshapes practice-stage behavior in a way that enhances persistence.

Heterogeneity is sharper among mixed-ability pairs. For high-ability students paired with low-ability peers, exogenous pairing has little effect, whereas endogenous pairing produces

sizable declines in both practice and final performance. This pattern indicates that, without aligned incentives, interacting with lower-ability peers can raise the effective effort costs or cause social conflicts that hinder learning. Conversely, shared grades reduce these problems, aligning with the idea that incentives help prevent free-riding or coordination issues.

Low-ability students paired with high-ability peers experience large gains in practice performance under exogenous pairing, consistent with productive peer inputs operating through $\psi_r(a_j, k_{ij})$. These gains are amplified under shared grades, where aligned incentives strengthen engagement and partially translate into improved final outcomes. Although persistence remains incomplete, the shared grades regime generates the strongest and most robust improvements for mixed-ability pairs.

Finally, when both students are low ability, none of the peer interaction regimes produce meaningful gains in either practice or final performance. This pattern is consistent with limited scope for productive peer inputs when both partners face high marginal costs of effort.

Overall, the heterogeneity results reinforce the central implication of the model: peer interaction affects outcomes through its impact on practice-stage behavior and through the extent to which practice gains persist. Differences across regimes and match environments cannot be explained by grading incentives alone; instead, they reflect how peer ability and incentive structures jointly shape the productivity and persistence of interaction.

In the next section, we examine student behavior during practice to shed light on the mechanisms underlying these patterns.

Table 5: Heterogeneities by Pairs Ability

	High-High (1)	High-Low (2)	Low-High (3)	Low-Low (4)
Practice Test Score				
Exogenous pairs formation	0.149**	–0.015	0.376***	–0.007
	[0.059]	[0.058]	[0.059]	[0.070]
Endogenous pairs formation	0.115	–0.253***	0.126	–0.006
	[0.074]	[0.084]	[0.082]	[0.071]
Shared Grades	0.418***	0.135	0.564***	0.147
	[0.076]	[0.094]	[0.089]	[0.092]
Final Test Score				
Exogenous pairs formation	0.045	0.028	–0.010	–0.083
	[0.057]	[0.056]	[0.051]	[0.059]
Endogenous pairs formation	0.050	–0.111	–0.161**	–0.135**
	[0.074]	[0.087]	[0.073]	[0.068]
Shared Grades	0.106	0.121	0.003	0.114
	[0.076]	[0.093]	[0.081]	[0.081]

Note: The table reports heterogeneous treatment effects by pair type, classified according to students' initial ability: High–High, High–Low, Low–High and Low–Low pairs. All regressions control for students' baseline test scores and the full set of baseline covariates, including an indicator for whether the baseline score was imputed and a dummy for students who failed four or more subjects in the previous year. All specifications include classroom fixed effects. Robust standard errors are used in all specifications. * Significant at 10%; ** significant at 5%; *** significant at 1%.

7 Mechanisms

7.1 Social Proximity

Within our conceptual framework, peer interactions may generate social responses during the practice stage, captured by $\theta_r S_i(a_j, k_{ij})$. These responses depend not only on partner ability but also on relational proximity k_{ij}. If peer effects operate through reputational concerns or social pressure, treatment effects should vary systematically with pre-existing familiarity.

The results in Table 6 provide direct evidence on the role of relational proximity in mediating peer effects. We observe substantial heterogeneity by whether paired students knew each other prior to the intervention, consistent with social responses operating within socially meaningful relationships.

Under endogenous pairing (Panel B), the negative effects are entirely driven by pairs composed of students who did not previously know each other. Among unfamiliar peers ($k_{ij} = 0$), endogenous pairing reduces both practice performance X_{0i} and final performance X_{1i}. In contrast, when students are paired with a known peer ($k_{ij} = 1$), the negative effect disappears. This pattern is consistent with relational frictions arising when students work with

Table 6: Treatment Effects by Peer Familiarity (Known vs. Unknown)

	Practice Test Score (1)	Final Test Score (2)
Panel A: Exogenous pairing		
Unknown peer (U)	0.145***	0.015
	[0.037]	[0.034]
Known peer (K)	0.161***	-0.012
	[0.058]	[0.055]
Difference (K-U)	0.016	-0.028
	[0.063]	[0.058]
Panel B: Endogenous pairing		
Unknown peer (U)	-0.379***	-0.302***
	[0.083]	[0.081]
Known peer (K)	0.067	-0.031
	[0.044]	[0.043]
Difference (K-U)	0.446***	0.271***
	[0.086]	[0.084]
Panel C: Shared grades		
Unknown peer (U)	0.301***	0.090*
	[0.051]	[0.048]
Known peer (K)	0.490***	0.222**
	[0.095]	[0.090]
Difference (K-U)	0.190*	0.133
	[0.101]	[0.096]

Note: The table reports heterogeneous treatment effects by whether paired students knew each other prior to the intervention, which serves as a proxy for familiarity. *Unknown* refers to pairs composed of students from different homerooms, while *Known* refers to pairs from the same homeroom. Difference (K-U) reports the estimated interaction term. All regressions control for baseline test scores and student characteristics and include classroom fixed effects. Robust standard errors are reported in brackets. * Significant at 10%; ** significant at 5%; *** significant at 1%.

socially distant partners, leading to adverse social responses during practice (a negative $\theta_r S_i(\cdot)$). Because the final exam is taken individually, the persistence of this pattern into the final stage indicates that these relational dynamics affect the learning inputs generated during practice rather than merely contemporaneous effort[7].

The shared grades regime (Panel C) reveals the complementary side of this mechanism. Shared incentives increase performance for all pairs, but the effect is substantially larger when students are paired with a known peer. This amplification is consistent with incentives

[7]It is important to note that the meaning of an "unknown" peer differs across regimes. In the exogenous pairing arm, unfamiliar matches arise mechanically from random assignment and are therefore plausibly orthogonal to students' social position or preferences. By contrast, in the endogenous pairing arm, the small subset of pairs composed of students who did not previously know each other reflects an outcome of the choice process. These pairs might consist of students who either couldn't get their preferred partner or were socially marginalized within the classroom network. As a result, unfamiliar endogenous matches may elicit negative social responses, such as disappointment, exclusion, or weak coordination, that are absent under random assignment. This distinction helps reconcile why unfamiliar peers are not systematically harmful under exogenous pairing (Panel A) but are strongly detrimental under endogenous pairing (Panel B).

interacting with relational proximity: when grades are linked, reputational concerns and social accountability are more salient within socially proximate pairs, strengthening effort during practice. In terms of the model, $\theta_r S_i(a_j, k_{ij})$ becomes more positive when $k_{ij} = 1$, and part of the resulting increase in practice-stage inputs persists through the pass-through term λ_r. By contrast, exogenous pairing without shared incentives (Panel A) exhibits little systematic variation by familiarity, suggesting that relational proximity alone is insufficient to generate strong peer effects absent either coordination frictions or aligned incentives. Taken together, the evidence indicates that peer effects are mediated by social responses that depend on relational proximity and that these responses interact with the institutional structure governing peer interaction.

7.2 Student Behavior During Pair Work

Some patterns in Section 7.1 might result from selection rather than interaction. Since endogenous pairing allows students to select their own partners, it's possible that the small group who end up paired with an unfamiliar peer differs systematically from those paired with a familiar one. While this is a plausible explanation, we do not test this directly. Exogenous pairing and shared grades, however, assign partners at random. This means the Known/Unknown gap under shared grades (Table 6, Panel C) cannot reflect who was matched with whom; so if it is real, it must come from a genuine difference in how known and unknown pairs behave under a shared incentive. This subsection tests that directly, using self-reported behavior during the practice stage, in particular, cooperation, interest in the partner's grade and learning, and disruptions, to see whether the incentive regime changes the nature of interaction in ways consistent with the framework: social accountability under shared grades, relational friction under endogenous matching.

Because these questions were administered only to students working in pairs, the control group is not observed in these outcomes. We therefore compare the three paired regimes to each other, using exogenous pairing as the omitted category. The resulting coefficients

capture how endogenous pairing and shared grades alter the nature of interaction relative to this benchmark. Because the goal here is to establish whether regimes differ in the nature of interaction at all, we pool known and unknown pairs within each regime.

Results are shown in Table 7. Endogenous pairing increases reported cooperation relative to exogenous pairing (column 1), and students in this regime are more likely to report caring about their partner's learning (column 3). This pattern is consistent with greater pro-social behavior or altruistic motives within self-selected pairs. However, endogenous pairing also increases reported disruptions (column 4). The coexistence of higher cooperation and higher disruption suggests that greater social closeness does not necessarily translate into more effective learning inputs. In the language of the model, endogenous pairing may increase social interaction but does not guarantee that the resulting social responses $\theta_r S_i(a_j, k_{ij})$ are productive for learning; in some cases, they may even raise effective coordination costs.

Table 7: Student Behavior: Cooperation and Disruptions

	Cooperation	Interest in Peer's Grade	Interest in Peer's Learning	Disruptions
	(1)	(2)	(3)	(4)
Endogenous formation	0.088***	0.055**	0.069***	0.032**
	[0.017]	[0.025]	[0.025]	[0.015]
Shared grades	0.035*	0.102***	0.038	–0.026*
	[0.019]	[0.026]	[0.026]	[0.014]
Observations	2,965	2,965	2,964	2,965
Exog. formation mean	0.820	0.423	0.442	0.0885

Note: Sample of students in treatment arms working in pairs (i.e. not the control group). The base group is exogeneous pairing. All regressions include classroom fixed effects and the full set of controls. Robust standard errors are reported in brackets. * Significant at 10%; ** significant at 5%; *** significant at 1%.

The shared grades regime exhibits a different behavioral profile. While cooperation increases modestly, the most pronounced effect is on students' interest in their partner's grade, with no corresponding increase in concern for their partner's learning. At the same time, disruptions decline relative to exogenous pairing. This pattern is consistent with incentives reshaping social responses; when grades are linked, students appear to internalize the consequences of their partner's performance, generating accountability rather than purely altruistic motives. In terms of the framework, the shared grades treatment appears to increase the intensity of

the social response term $\theta_r S_i(a_j, k_{ij})$ in a way that aligns effort within the pair.

Table 8 splits these averages by whether the pair was known or unknown. One result stands out immediately; even under exogenous pairing, unknown pairs cooperate less than known ones. Being known or unknown affects cooperation on its own, with no incentive or selection story behind it.

Table 8: Heterogeneities by Peer's Homeroom

Peer relationship	Practice Test Score	Final Test Score	Cooperation	Interest in Peer's Grade	Interest in Peer's Learning	Disruptions
	(1)	(2)	(3)	(4)	(5)	(6)
Exogenous pairing x Known peer	0.161***	–0.012				
	[0.058]	[0.055]				
Exogenous pairing x Unknown peer	0.145***	0.015	–0.108***	–0.054	0.024	0.005
	[0.037]	[0.034]	[0.021]	[0.033]	[0.033]	[0.021]
Endogenous pairing x Known peer	0.067	–0.031	0.017	0.025	0.105***	0.020
	[0.044]	[0.043]	[0.022]	[0.037]	[0.037]	[0.022]
Endogenous pairing x Unknown peer	–0.379***	–0.302***	0.051	0.009	–0.036	0.115***
	[0.083]	[0.081]	[0.038]	[0.050]	[0.050]	[0.038]
Shared grades x Known peer	0.490***	0.222**	0.029	0.119**	0.083	0.004
	[0.095]	[0.090]	[0.033]	[0.057]	[0.057]	[0.033]
Shared grades x Unknown peer	0.301***	0.090*	0.043**	0.102***	0.027	–0.033**
	[0.051]	[0.048]	[0.021]	[0.029]	[0.029]	[0.016]
p-value equal exog. coeffs.	0.795	0.636	0.000	0.107	0.479	0.794
p-value equal endog. coeffs.	0.000	0.001	0.437	0.791	0.021	0.030
p-value equal shared grades coeffs.	0.061	0.166	0.714	0.784	0.372	0.309

Note: All regressions include classroom fixed effects and the full set of controls. Robust standard errors. The Practice and Final Test Score columns use the control group as the omitted baseline, exactly as in Table 6: all six regime x familiarity cells are estimated relative to students who worked alone. The Cooperation, Interest in Peer's Grade, Interest in Peer's Learning, and Disruptions columns were only administered to students working in pairs, so the control group cannot serve as a baseline for those columns; each cell there is instead benchmarked against Exogenous pairing with the same homeroom status (Known peers against Exogenous-Known, Unknown peers against Exogenous-Unknown), since Exogenous pairing is randomly assigned and therefore not subject to the selection concerns discussed for Endogenous pairing. Exogenous-Known is marked (ref.) as the omitted reference; Exogenous-Unknown is recovered from the same regression. This means the two sets of columns are not on the same baseline and should not be compared directly in levels. * significant at 10

In endogenous pairing, only disruptions vary between known and unknown pairs, with unknown pairs causing more disruption, supporting the coordination-friction explanation in Section 7.1. There is no difference in cooperation or interest in the partner's learning based on familiarity; the cooperation benefit from endogenous pairing appears to be a general feature of self-selected pairs rather than something unique to unknown pairs.

Shared grades show no difference in familiarity on any behavioral measure. While interest in the partner's grade increases, disruptions do not, consistently across both known and unknown pairs. Additionally, the score advantage observed in shared-grades pairs with

known scores, as shown in Table 6, is not reflected in any comparable measure.

Same-homeroom versus different-homeroom is still a coarse proxy for closeness. Two students from the same classroom could barely know each other or be best friends. We turn to that distinction next.

7.3 Selection and Match Quality

This section narrows in on endogenous pairing specifically to asks whether the benefits of familiarity keep growing with closeness, or level off. The same-homeroom category used so far pools together two very different relationships: students who share a homeroom but barely know each other, and students who are best friends. Table 9 splits it into three rungs, unknown, known but not a best friend, and best friend.

Moving from an unknown to a known not best friend peer accounts for essentially all of the benefit of familiarity. The gain is large and highly significant on both the practice and final quiz. Pairing with an actual best friend gives back part of that gain; performance falls significantly relative to the known not best friend group on the practice quiz, and the point estimate is negative, though smaller and imprecise, on the final exam. This is a hump shape rather than a straight line. Some familiarity resolves the coordination frictions that hurt unfamiliar pairs, but the closest pairs do not perform any better than moderately familiar ones, and by our practice-quiz estimates perform worse.

The Disruptions column speaks only partly to this pattern. Moving from an unknown to a known, not best friend peer significantly reduces disruptions, mirroring the gain in test scores at that same step and consistent with the coordination-friction story developed above. For best friends, however, the indicator is collinear with classroom fixed effects once the sample is restricted to students with non-missing controls and is dropped by the estimator; we cannot recover a disruption estimate for that group in this specification, so the data are silent on whether off-task socializing explains their lower academic performance. We therefore describe the best-friend result as an open feature of $S_i(a_j, k_{ij})$: closeness past a certain point

stops adding value, and possibly reverses it, but we do not have a behavioral marker in hand that pins down the channel.

Table 9: Peer Closeness in Endogenous Pairing: Performance and Behavior

Peer relationship	Practice Test Score	Final Test Score	Disruptions
	(1)	(2)	(3)
Unknown peer	–0.379***	–0.301***	0.087**
	[0.083]	[0.081]	[0.043]
Known peer, not best friend	0.125**	–0.010	–0.023
	[0.051]	[0.050]	[0.027]
Best friend	–0.054	–0.074	
	[0.068]	[0.065]	
Difference (Known - Unknown)	0.505***	0.291***	–0.110***
	[0.089]	[0.088]	[0.039]
Difference (Best friend - Known)	–0.179**	–0.063	
	[0.077]	[0.074]	

Note: The table extends Table 6's Known/Unknown split for Endogenous pairing into three categories, using classroom co-membership and an independent friendship-nomination measure (whether the assigned partner was already named as a friend, regardless of how the pair was formed): Unknown (different homeroom), Known but not best friend (same homeroom), and Best friend. This replaces an earlier version of this table based on students' self-reported reason for choosing their partner, which was available only for Endogenous pairing by construction and potentially subject to ex-post rationalization. The Disruptions column tests whether closeness affects off-task behavior during practice; as in Table 8, this outcome is observed only for students in a paired treatment arm. n.e. = not estimable: within this sample (N=1,553, once non-missing controls are required), the Best friend indicator is collinear with classroom fixed effects and is dropped by the estimator, so no coefficient exists for that cell or the difference row built from it. Only Endogenous pairing is refined this way; Exogenous pairing and Shared Grades keep the original Known/Unknown split from Table 6, since best-friend pairs are too rare in those arms to support a further split. All specifications include classroom fixed effects and the full set of baseline controls. Robust standard errors are used in all specifications. * significant at 10

From here since tomorrow

7.4 Rank Concerns

A natural concern with our results is that what we interpret as evidence of peer interaction may instead reflect a more general response to social comparison. The behavioral economics literature shows that effort in low-stakes settings responds not only to material incentives but to psychological ones: status, self-image, and relative-performance feedback have each been shown to move behavior independently of monetary or grade-based stakes (Levitt et al., 2016; Tran and Zeckhauser, 2012; Barankay, 2011). If simply knowing how one compares to others is sufficient to raise effort, then part of what we attribute to shared grades in Section 5 could reflect this broader comparison channel rather than the interaction-specific mechanisms outlined in our conceptual framework.

This possibility maps directly onto a limitation of our model as specified. The social-response term, $\theta_r S_i(a_j, k_{ij})$, is defined through interaction with a partner: it depends on partner ability a_j and relational proximity k_{ij}, and is mechanically absent for a student working alone. The model therefore has no channel through which a student who never interacts with a peer could nonetheless respond to comparison-based incentives. If such a channel exists empirically, our interaction-based interpretation of the shared-grades effect would be, at best, incomplete.

To assess this directly, we exploit an additional experimental arm in which students complete the practice activity entirely on their own, with no partner and no interaction, but are informed that they will learn their ranking among classmates during the practice and the final stage of the experiment (Rank Concerns). This arm isolates comparison from interaction: any effect it produces cannot operate through $\theta_r S_i(a_j, k_{ij})$, since there is no partner to generate ability-dependent or proximity-dependent social responses, and therefore provides a direct test of whether the accountability we document under shared grades requires genuine peer interaction, or whether it can be replicated by comparison alone. The arm was implemented as a separate experiment in 15 schools, comprising 45 classrooms and 1,428 students randomly assigned to control or treatment,[8] using the same specification as our main design: classroom fixed effects, the full set of baseline controls, and robust standard errors.

Table 10: Treatment Effect of Rank Concerns on Exam Test

	Practice Quiz			Final Quiz		
	No controls	Initial Score	All controls	No controls	Initial Score	All controls
	(1)	(2)	(3)	(4)	(5)	(6)
Rank Concerns	0.010	–0.008	–0.010	0.086*	0.064	0.062
	[0.049]	[0.048]	[0.048]	[0.050]	[0.048]	[0.047]
Observations	1,366	1,366	1,366	1,366	1,366	1,366

Note: The table reports the estimated treatment effect of the Rank Concerns arm, within its own separately randomized experiment (own control group, not pooled with the sample in Table 4). The dependent variable is standardized to have mean zero and standard deviation one. Columns 1-3 correspond to the practice quiz and columns 4-6 to the final quiz. Within each group, columns 1 and 4 present estimates without controls, columns 2 and 5 include the students' baseline test score as a control, and columns 3 and 6 add the full set of baseline covariates. All regressions include classroom fixed effects. Robust standard errors are used in all specifications. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table 10 reports the results. On the practice quiz, the treatment effect is close to zero,

[8]Students complete the practice and final exercises individually. They are told that, at the end of the exercise, they will be informed of their standing in the class relative to their classmates.

and none of these coefficients are significant. This is consistent with the structure of our conceptual framework: because students in this arm work entirely alone, the social-response term $\theta_r S_i(a_j, k_{ij})$ is mechanically zero, and no partner-generated inputs $\psi_r(a_j, k_{ij})$ can arise during practice. On the final exam, the estimated effect is not significant once we condition on the baseline test score and the full set of covariates. Both patterns suggest that informing students of their relative rank, without any peer interaction, does not replicate the gains we document under shared grades.

Table 11 formalizes this comparison, using the full-controls specification to test whether the Shared Grades and Rank Concerns effects are statistically distinguishable. On the practice quiz, the gap is large and highly significant: shared grades outperform Rank Concerns by 0.344σ, rejecting the hypothesis that comparison alone can account for the practice-stage gains we attribute to peer interaction. On the final exam, however, the estimated gap is not statistically distinguishable from zero, so we cannot reject equality between the two treatments.

Table 11: Shared Grades vs. Rank Concerns: Formal Comparison

	Practice Test Score	Final Test Score
	(1)	(2)
Shared grades	0.334***	0.113**
	[0.048]	[0.045]
Rank Concerns	–0.010	0.062
	[0.048]	[0.047]
Difference (SG - Rank)	0.344***	0.051
	[0.068]	[0.065]

Note: The table reports estimated treatment effects (relative to control) for Shared Grades and Rank Concerns, each estimated within its own separately randomized experiment using the same specification as elsewhere in the paper: classroom fixed effects, the full set of baseline controls, and robust standard errors. The difference row is computed analytically as the gap between these two coefficients; because the two experiments are independent (no shared students or classrooms), the standard error of the difference is the square root of the sum of the two estimates' squared standard errors, with no covariance term. * significant at 10

The practice-stage results provide the clearest evidence for an interaction-based interpretation. The gain we document under shared grades cannot be explained by comparison or rank concerns alone: Rank Concerns produces no discernible effect on practice performance, and the resulting gap with shared grades is significant.

The final-exam comparison is less sharp. The estimated gap between shared grades and

Rank Concerns is smaller and not statistically significant. We cannot reject that the two treatments have similar effects on final performance. Even so, the evidence still favors shared grades: its effect is precisely estimated and significant on its own, while Rank Concerns' is not, and the smaller sample underlying this arm makes the added imprecision unsurprising rather than informative. We therefore treat the final-exam comparison as suggestive rather than conclusive, and reserve the language of persistence for the shared-grades results, where practice-stage gains are clearly carried forward.

8 Conclusions

This paper examines how peer-to-peer learning operates in a short-run instructional setting. We implement two framed field experiments with seventh-grade students in Mexico City, varying both the structure of peer interaction during practice and the incentive scheme governing final grades. The design allows us to disentangle the role of peer composition, match formation, and incentive alignment in shaping academic performance.

Three central findings emerge. First, peer interaction alone is not sufficient to improve learning. Random pairing increases performance during the practice stage but does not translate into gains on the final exam. These results suggest that interaction can lower contemporaneous effort costs or facilitate coordination, yet fail to generate inputs that persist beyond the interaction stage.

Second, allowing students to choose their own partners does not enhance outcomes and, in fact, reduces final performance. The negative effects are concentrated among pairs lacking pre-existing social ties and among mixed-ability matches. The evidence indicates that endogenous pairing generates mostly familiar matches, but also a subset of low-quality pairings characterized by coordination problems and disruptions. These relational frictions appear to undermine the quality of practice-stage inputs and reduce learning persistence.

Third, and most strikingly, linking students' grades within randomly assigned pairs sub-

stantially improves both practice and final exam performance. Despite mechanically weaker individual incentives, the shared grades regime outperforms both individual work and standard exogenous pairing. This finding rejects a benchmark model in which students respond only to direct grading incentives. Instead, it points to the importance of social accountability and aligned incentives during peer interaction. Behavioral evidence indicates that shared grades increase students' concern for their partner's grade and reduce disruptions, suggesting that incentives reshape social responses within the pair.

Heterogeneity patterns further discipline interpretation. Low-ability students paired with high-ability peers experience large gains under shared grades, while high–high pairs exhibit improvements primarily when incentives are aligned. Familiarity amplifies positive effects under shared grades but does not independently generate gains in the absence of incentive alignment. Taken together, these patterns indicate that peer effects operate through the interaction of partner characteristics and institutional structure.

Our findings contribute to a growing literature emphasizing that peer effects are highly context-dependent and mediated by mechanisms beyond simple exposure to higher-ability peers. The results highlight three key lessons. First, interaction without aligned incentives may generate limited or even negative effects. Second, self-selection does not guarantee efficient matching, particularly in environments where social considerations dominate academic ones. Third, modest institutional changes in incentive design can substantially alter how peer interaction translates into learning.

From a policy perspective, the evidence suggests that peer-to-peer learning is not inherently beneficial; its effectiveness depends critically on how interactions are structured. Designing environments that align individual incentives within collaborative settings may enhance both engagement and learning persistence, even in short-run instructional contexts.

More broadly, the experiment underscores the importance of modeling peer interaction as a joint function of technology, incentives, and social responses. Understanding how these elements interact is central to explaining when and why peer-to-peer learning succeeds.

References

Barankay, I. (2011). Rankings and social tournaments: Evidence from a crowd-sourcing experiment.

Barrios-Fernandez, A. (2023). Peer effects in education. In *Oxford Research Encyclopedia of Economics and Finance*.

Berlinski, S., Busso, M., and Giannola, M. (2023). Helping struggling students and benefiting all: Peer effects in primary education. *Journal of Public Economics*, 224:104925.

Booij, A. S., Leuven, E., and Oosterbeek, H. (2017). Ability peer effects in university: Evidence from a randomized experiment. *Review of Economic Studies*, 84(2):547–578.

Busso, M. and Frisancho, V. (2025). Ability grouping in the classroom: Performance and behavioral responses. Technical report, Working paper.

Carrell, S. E., Sacerdote, B. I., and West, J. E. (2013). From natural variation to optimal policy? the importance of endogenous peer group formation. *Econometrica*, 81(3):855–882.

Coveney, M. and Oosterveen, M. (2021). What drives ability peer effects? *European Economic Review*, 136:103763.

Feld, J. and Zölitz, U. (2017). Understanding peer effects: On the nature, estimation, and channels of peer effects. *Journal of Labor Economics*, 35(2):387–428.

Gu, X. (2023). Classmates and friends matter! peer effects on cognitive ability formation. *China Economic Review*, 79:101910.

Hoxby, C. M. (2000). Peer effects in the classroom: Learning from gender and race variation.

Levitt, S. D., List, J. A., Neckermann, S., and Sadoff, S. (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, 8(4):183–219.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In *Handbook of the Economics of Education*, volume 3, pages 249–277. Elsevier.

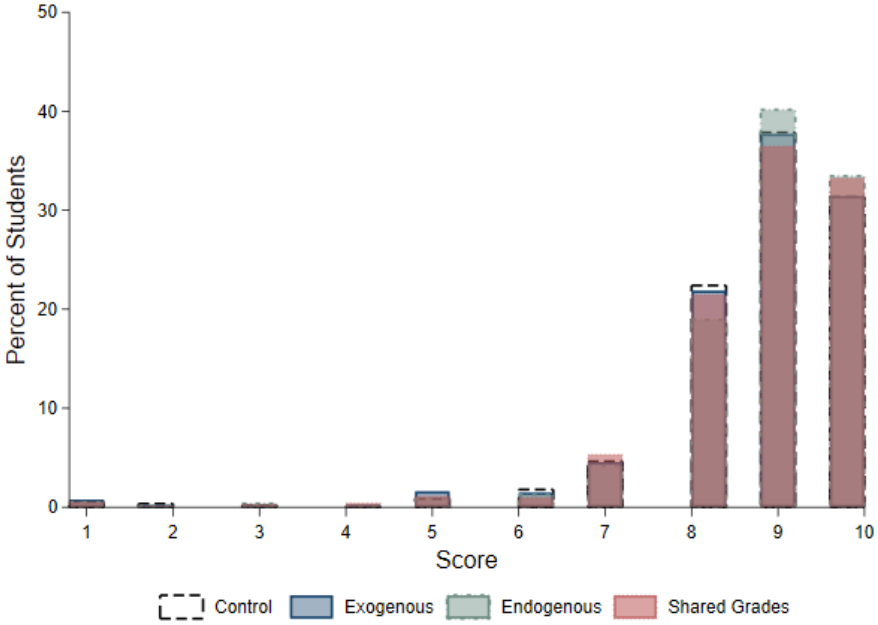
Sacerdote, B. (2014). Experimental and quasi-experimental analysis of peer effects: Two steps forward? *Annual Review of Economics*, 6(1):253–272.

Tran, A. and Zeckhauser, R. (2012). Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics*, 96(9-10):645–650.

Wu, J., Zhang, J., and Wang, C. (2023). Student performance, peer effects, and friend networks: Evidence from a randomized peer intervention. *American Economic Journal: Economic Policy*, 15(1):510–542.

A Appendix Tables and Figures

Figure A.1: Class quality



Notes: The figure shows the distribution of students' class evaluations (1–10) by treatment condition, where the y-axis indicates the percentage of students selecting each score. Question: "On a scale from 1 to 10, where 10 is the highest grade, what grade do you give the class you received on fractions and decimal numbers?"

Table A.1: Teacher's Quality

	Freq.	Percent	Cum.
Very good	2,419	50.96	50.96
Good	1,599	33.68	84.64
Regular	281	5.92	90.56
Bad or very bad	12	0.25	90.82
Did not answer questions	24	0.51	91.32
Nobody asked questions	412	8.68	100
Total	4,747	100.00	

Note: Question: "How did the substitute teacher answer questions that came up during class or practice?".

Table A.2: Quality Class and Tests

	Attention in class	Difficulty Practice Test	Difficulty Exam Test
	(1)	(2)	(3)
Exogenous formation	0.003	–0.063*	–0.070**
	[0.032]	[0.034]	[0.033]
Endogenous formation	–0.188***	0.136***	0.037
	[0.050]	[0.047]	[0.046]
Rank Concerns	0.053	–0.189***	–0.182***
	[0.062]	[0.052]	[0.053]
Shared grades	0.040	0.095*	0.063
	[0.042]	[0.050]	[0.047]
Observations	5,966	5,966	5,966

Note: All regressions include classroom fixed effects and the full set of controls. Robust standard errors. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.3: Balance: Experiment 1

Baseline variable	Mean control	Exogenous pairing	Endogenous pairing	Observations
Raw initial exam	31.628	0.061	–0.396	2477
	[10.803]	[0.387]	[0.392]	
Years of educ., mother	9.778	0.158	0.240	2404
	[3.390]	[0.157]	[0.158]	
Years of educ., father	10.246	–0.026	–0.027	2090
	[3.362]	[0.173]	[0.171]	
Living with both parents	0.699	–0.000	–0.021	2425
	[0.459]	[0.023]	[0.023]	
Household size	4.741	0.003	0.027	2425
	[1.568]	[0.078]	[0.082]	
Male	0.490	0.079	–0.005	2477
	[0.500]	[0.024]***	[0.025]	
Older than 12	0.039	0.010	0.006	2452
	[0.194]	[0.010]	[0.010]	
Primary school GPA	8.652	0.019	0.008	2432
	[0.640]	[0.028]	[0.029]	
Special needs	0.016	–0.005	–0.002	2425
	[0.127]	[0.006]	[0.006]	

Note: All regressions include classroom fixed effects and the full set of controls. Robust standard errors. * significant at 10

Table A.4: Balance: Experiment 2

Baseline variable	Mean control	Exogenous pairing	Shared grades	Observations
Raw initial exam	30.226	0.970	0.664	2256
	[10.500]	[0.418]**	[0.424]	
Years of educ., mother	10.329	0.144	–0.004	2203
	[3.012]	[0.154]	[0.150]	
Years of educ., father	10.536	0.208	0.008	1927
	[3.302]	[0.174]	[0.175]	
Living with both parents	0.665	0.039	0.024	2228
	[0.472]	[0.024]	[0.024]	
Household size	4.773	–0.079	–0.016	2228
	[1.631]	[0.083]	[0.080]	
Male	0.491	0.042	–0.007	2256
	[0.500]	[0.026]	[0.026]	
Older than 12	0.051	–0.020	–0.014	2244
	[0.220]	[0.010]**	[0.011]	
Primary school GPA	8.615	0.029	0.025	2234
	[0.602]	[0.029]	[0.029]	
Special needs	0.030	–0.012	–0.012	2227
	[0.171]	[0.008]	[0.008]	

Note: All regressions include classroom fixed effects and the full set of controls. Robust standard errors. * significant at 10

Table A.5: Balance: Rank Concerns Experiment

Baseline variable	Mean control	Rank Concerns	Observations
	(1)	(2)	(3)
Baseline ability	30.580	0.768	1366
	[11.644]	[0.473]	
Years of educ., mother	9.844	0.147	1366
	[3.400]	[0.166]	
Living with both parents	0.680	–0.005	1366
	[0.467]	[0.025]	
Household size	4.864	0.017	1366
	[1.776]	[0.096]	
Male	0.494	–0.017	1366
	[0.500]	[0.027]	
Older than 12	0.050	–0.004	1366
	[0.218]	[0.011]	
Primary school GPA	8.579	0.045	1366
	[0.680]	[0.034]	
Special needs	0.016	–0.002	1366
	[0.124]	[0.006]	

Note: The table reports the mean of each covariate in the control group and the estimated treatment effect for the Rank Concerns arm, within its own separately randomized experiment (own control group, not pooled with the sample in Table 3). All specifications include classroom fixed effects. Robust standard errors are used in all specifications. * significant at 10%; ** significant at 5%; *** significant at 1%.

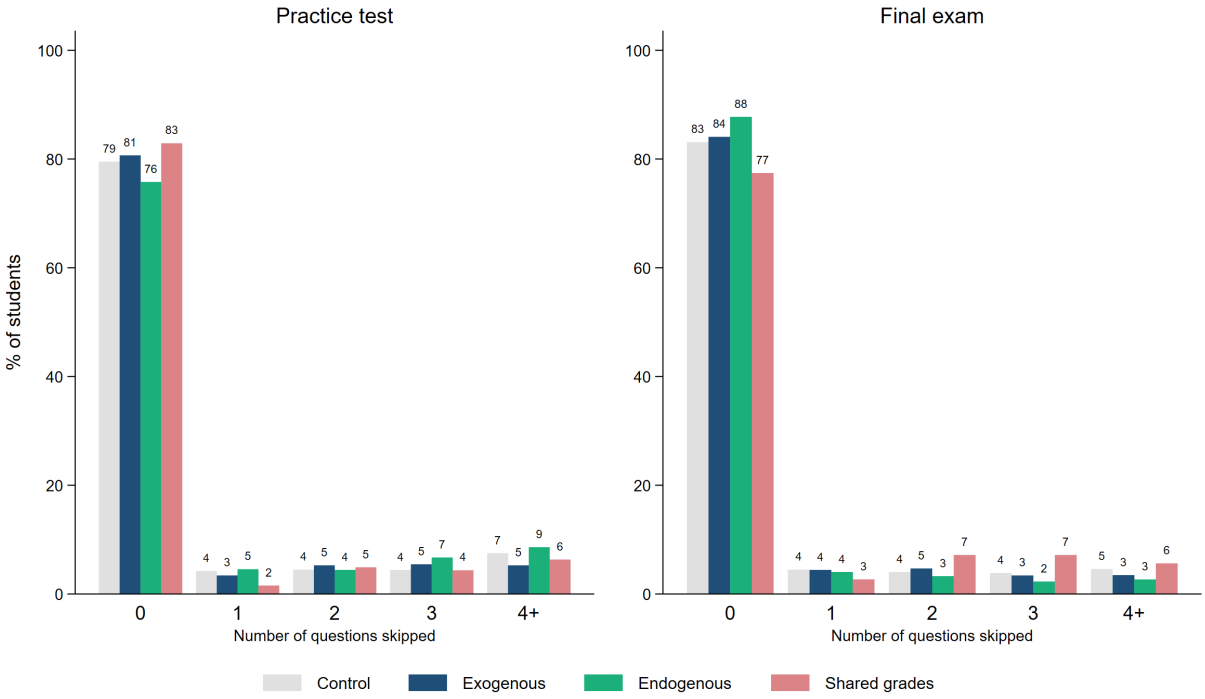


Figure A.2: Distribution of skipped questions: practice vs. final exam

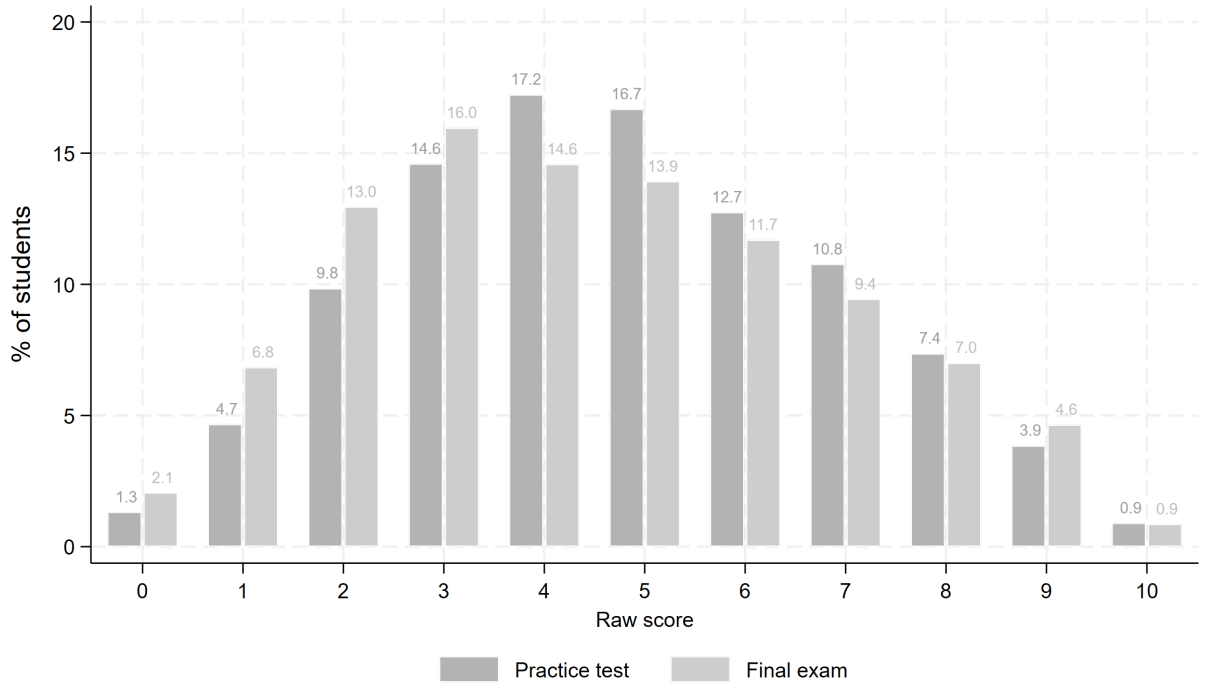


Figure A.3: Distribution of raw scores: practice vs. final exam