The Two Sides of Peer Feedback: Strategic Provision and Selective Use

Lorenzo Casaburi and Lea Heursen*

September 2, 2026

Abstract

Peer feedback is widely used in organizations, making effective design crucial. How do strategic considerations affect the feedback sent and how recipients use it? In a laboratory experiment (N=742), participants complete a complex task in a competitive environment and are randomly assigned to independent or reciprocal peer review. Reciprocity reduces feedback quantity and quality, as the first reviewer in the sequence withholds criticism. Recipients, however, incorporate only half the objectively useful feedback. Consequently, independent feedback produces no performance gains, though implementing it could have raised scores. Effective feedback design is two-sided: regimes must elicit and help recipients use valuable input.

Keywords: peer feedback, strategic communication, reciprocity, competition, organizational learning, experiment

JEL: D83, M54, D23, L23

*Casaburi: University of Zurich, Schönberggasse 1, 8001 Zurich, Switzerland (lorenzo.casaburi@econ.uzh.ch). Heursen: Technische Universität Berlin (TU Berlin), Straße des 17. Juni 135, 10623 Berlin, Germany (heursen@tu-berlin.de). This study was reviewed by the Institutional Review Board of the University of Zurich Faculty of Business, Economics and Informatics and determined to be exempt from further review. It was preregistered in the AEA RCT Registry under ID AEARCTR-0007718. We are grateful to participants at several conferences and seminars for helpful comments and suggestions. Timme Ariens, Tia Collins, Haoyu Cao and Yaqing Wang provided excellent research assistance. Lea Heursen thanks the Deutsche Forschungsgemeinschaft through CRC TRR 190 (project number 280092119) for financial support.

1. Introduction

Peer feedback plays a central role in organizations. When performance is multidimensional or difficult to observe, organizations rely on peer evaluations to aggregate dispersed information and guide decisions such as promotions, rewards, and task allocation. The importance of communication and information flows for organizational performance is well established (Garicano, 2000). Peer evaluations, ranging from formal 360-degree reviews to informal feedback platforms, have become prevalent in modern workplaces (Di Fiore and Souza, 2021), making the design of effective peer feedback regimes increasingly important.

Designing peer feedback regimes that elicit informative evaluations is challenging (Miller, Resnick and Zeckhauser, 2005). In many environments, feedback affects not only the recipient's learning and development but also the evaluator's own outcomes—either directly through competition for rewards or indirectly through reciprocity in future interactions—creating strategic incentives to distort reports (Carpenter, Matthews and Schirm, 2010; Bolton, Greiner and Ockenfels, 2013; Hussam, Rigol and Roth, 2022). Competition predicts excessive criticism or downgrading of others' work, whereas reciprocity concerns can either discipline such behavior or lead reviewers to withhold critical but useful information. Major firms have implemented different peer feedback regimes and have addressed this tension between competition and reciprocity in peer reviewing in different ways. Google's performance reviews, for example, are largely anonymized while Netflix relies on attributable, often public feedback.[1]

However, the effectiveness of different feedback regimes depends not only on the feedback supplied, but also on how recipients respond to the evaluations they receive, that is, the demand side of feedback. The predictions are ambiguous: feedback given under reciprocity concerns may be taken more seriously if it is perceived as less aggressively competitive, but it may also be perceived as less helpful if reviewers withhold critical information. Conversely, independent feedback may contain more valuable suggestions, but recipients may discount its helpfulness if they perceive it as overly critical or strategically motivated. How feedback design shapes what reviewers write, how recipients respond and how it ultimately affects subsequent performance remains largely an open question (Villeval, 2020; Drouvelis and Paiardini, 2022; Boudreau, 2026).

This paper provides, to the best of our knowledge, the first experimental evidence on how strategic incentives in peer feedback regimes affect feedback quantity and quality, as well as subsequent performance through the receiver's willingness to act on it. To identify these effects, we compare a feedback regime in which only competition may

[1] Trisca, L. (2025), Learn How to Run Employee Performance Reviews Like Google: Mission, Transparency, and Voice, *Deel Blog*, accessed February 17, 2026; McCord, P. (2024), How Netflix Reinvented HR, Harvard Business Review: 92(1).

distort feedback with one in which both competition and reciprocity may influence what information is provided and trace the complete pathway from feedback incentives to feedback content, its implementation, and ultimately performance.

We conduct a laboratory experiment in which participants (N=742) complete two rounds of a complex, creative task: designing a recruitment poster whose efficacy depends on multiple aspects of content and formatting. After the first round, half provide feedback to peers, who can then revise their posters. Roles subsequently reverse. First-round posters rank higher in a bonus tournament when they receive fewer remarks, creating competitive motives in peer review (Carpenter, Matthews and Schirm, 2010), while final posters are rewarded based on absolute performance.

Therefore, feedback can negatively affect first-round tournament outcomes but positively affect future earnings by helping recipients improve their final poster. These experimental incentives capture a broad tension in organizations, where peer feedback often serves both an evaluative and a developmental purpose. Critical feedback can harm the reviewed peer's position in a competition for bonuses, prestigious project assignments, or even promotions, while also helping their long-run development and performance. The incentives in the experiment model this tension in a tractable and transparent way. Reviewers' behavior will depend on which purpose—improving relative standing or fostering learning—they believe matters more to their recipients or themselves.

The key manipulation varies the strategic relationship between reviewer and recipient, holding everything else constant. Feedback is either independent (perfect-stranger matching when roles reverse) or reciprocal, with participants reviewing each other sequentially. Under reciprocity, the first reviewer may provide constructive feedback to elicit useful feedback in return, improving both peers' final performance and pay. Alternatively, they may withhold criticism to induce lenient feedback in return, improving both peers' first-round tournament standing (Klapper, Piezunka and Dahlander, 2024). Independent reviewing removes these reciprocal incentives, allowing more candid feedback but also potentially encouraging strategically harsh evaluations that disadvantage competitors (Balietti, Goldstone and Helbing, 2016).

These strategic incentives may also influence recipients' responses. Independently reviewed participants may discount feedback perceived as strategically motivated, whereas reciprocal reviewing may foster trust in the reviewers' intentions and the feedback's helpfulness, encouraging implementation. We examine these mechanisms using detailed measures of feedback content, first-round and final poster content and survey measures of feedback perceptions.

To gauge ex ante beliefs about how these competing strategic forces would influence feedback effectiveness, we surveyed 200 human resources (HR) professionals.[2] Their

[2]Recruited through Prolific Academic, 65% of HR professionals worked in organizations with more than 1,000 employees and 30% in organizations with 250–1,000 employees. On average, they were 39

beliefs are highly dispersed: half predict that reciprocal feedback is more effective for improving performance in our experiment, while a sizable share (41%) believes that independent feedback is more effective and only 8% expect no difference. Whether competition or potential reciprocity dominates in shaping feedback provision and recipients' response is an empirical question which our study is designed to answer.

We report three main findings. First, reciprocal feedback decreases both the quantity and quality of feedback. Relative to independent reviewing, reciprocity decreases the number of comments that the first set of evaluators provide by 15.6%. It also reduces the number of *implementable* comments—defined as actionable suggestions that would improve the recipient's performance if followed—by 15.1%. This pattern suggests that reciprocity concerns lead reviewers to withhold valuable information: they do not simply reduce unnecessary negative comments but also avoid providing useful feedback, arguably because it could be perceived as critical and provoke retaliation.

Second, participants incorporate only a fraction of the useful feedback they receive. This is despite a strong incentive to improve the final poster. On average, participants in the independent group implement 56.4% of the implementable comments and the implementation rate in the reciprocity group is strikingly similar (54.2%). Participants' subjective perceptions of feedback quality appear to play a central role. Approximately one-third of the subjects report that the feedback is at most somewhat helpful, even when its objectively assessed content suggests it could improve their performance. Accordingly, this subjective assessment of feedback quality is correlated with implementation rates ($\rho = 0.43$), while its correlation with objective quality is weaker ($\rho = 0.18$). Across feedback regimes, subjective feedback quality is the same even though objective quality differs. However, the feedback is perceived differently along several other dimensions: participants in the independent group rate the feedback they received as somewhat more aggressive and less friendly.

Third, although the performance distribution shows substantial learning from the initial to the final poster, we find no detectable effect of feedback regime on final performance, despite its effects on feedback provision. Performance on the final poster is statistically indistinguishable across treatments: the point estimate indicates a 0.80% reduction in average performance under reciprocal feedback relative to independent reviewing. The 95% confidence interval rules out reductions greater than 4% and improvements greater than 2.5%. A back-of-the-envelope calculation of best-response performance—had participants fully acted on the feedback they received—shows that this null effect is consistent with the limited implementation rate. Although independent feedback increases the supply of high-quality comments, recipients incorporate only part of this additional information, diluting the performance gains that improved feedback quality

years old, held a college degree, had eight years of HR experience and 71% had worked in performance management, training and development, or organizational development.

would otherwise generate.

Our findings offer several novel insights for designing peer feedback regimes, which have become increasingly relevant in firms, governments, and scientific institutions (Li and Agha, 2015; Li, 2017). Even in competitive environments, independent feedback, which precludes retaliation, need not induce outright sabotage, consistent with Dufwenberg, Görlitz and Gravert (2025). We find that independent feedback leads to more—and more useful—feedback than reciprocal review does. Reciprocity concerns appear to meaningfully distort the supply of high-quality feedback: reviewers withhold valuable suggestions when future interactions are expected. Because reciprocity concerns in practice often arise when reviewers' identities are known, the strategic withholding we document in an anonymous setting likely reinforces the documented reluctance to provide negative feedback when peer feedback is given openly rather than anonymously (e.g., Gneezy et al., 2017; Mickeler et al., 2025).

Importantly, the limited uptake of implementable feedback points to a demand-side bottleneck: learning from peer input depends not only on the quality of the feedback provided, but also on recipients' willingness to act on it. Workers may overlook or discount the potential value of peer suggestions, thereby muting the performance effects of improved feedback quality. We find no evidence that reciprocal reviewing helps overcome skepticism about the helpfulness of feedback. In line with this, we show that cooperative behavior in a public goods game played with a participant who was not involved in the feedback process is equally high under both feedback regimes.

These results shift the design problem in peer feedback from a purely supply-side question to a two-sided one. As emphasized by Crawford and Sobel (1982), communication outcomes depend on both the incentives of those who transmit information and the responses of those who receive it. Limiting reciprocity concerns in performance evaluations can improve feedback quality. But better feedback alone is not sufficient: organizations must also design regimes that help recipients recognize, trust, and incorporate valuable peer input. The central design challenge is therefore not only to elicit high-quality feedback, but to ensure that this feedback translates into learning and performance gains.

2. Experimental Design

The experiment was designed to investigate how strategic incentives in feedback regimes affect the production of information, its use and, ultimately, subsequent performance. The study was conducted in monitored online sessions with participants from the subject pool of the ETH Zurich Decision Science Laboratory, most of whom were university students from ETH Zurich or the University of Zurich. In this section, we describe the experimental tasks and incentives, the peer feedback process and the treatments, as well as the procedures in more detail.

2.1. Experimental Tasks and Incentives

The experiment centers on a multi-dimensional real-effort task developed for this project: designing a recruitment poster for a university decision science laboratory. Table 1 describes the poster grading scheme: performance is based on ten content and eight design & style items, with up to 80 points in each dimension. The criteria were developed with two lab managers and a communications expert. The multidimensional task permits substantive peer review and scope for improvement; pre-tests with exogenous feedback confirmed that feedback could improve performance.

At the beginning of the experiment, all subjects create an initial poster, with minimal guidance and a 15-minute time limit (Part 1).[3] See Figure A1 in the Online Appendix for an example poster. To discourage participants from seeking online tips for creating such posters, they are informed that a script records metadata on whether they leave the experimental webpage.[4] Then, participants are randomly assigned to a *revise-first* or a *revise-second* group.

In Part 2, revise-second participants review a randomly assigned peer's poster while revise-first participants wait (see Subsection 2.2).

In Part 3, revise-first participants have 15 minutes to revise their posters using the feedback; these final posters are our primary performance outcomes.

In Part 4, roles reverse and revise-first participants review revise-second participants.

Finally, all participants unexpectedly play a standard linear public-goods game with a stranger, allocating up to 20 tokens to a group project (MPCR=0.75; 1 token=CHF 0.20). This measures potential spillovers to cooperation. Revise-second participants then revise their poster while revise-first participants complete a survey.

Monetary Incentives. All participants within a session compete for a CHF 30 bonus (≈$30), awarded to the authors of the 25% of *first-round* posters receiving the fewest remarks during peer review.[5] Thus, the number of remarks acts as an inverse measure of poster quality. This mirrors organizational settings where peer feedback has evaluative consequences for bonuses, assignments, or promotions and creates scope for sabotage (Carpenter, Matthews and Schirm, 2010).

In addition, participants receive CHF 5, CHF 15, CHF 25, or CHF 35 based on their *final* poster's total score (0–160), with half the points for content and half for design and style (Table 1). Each poster was graded by one to three RAs after the session.[6]

[3] They use Quill, a free open-source WYSIWYG editor similar to Google Docs, integrated into our experimental software, which generates shareable HTML code that reviewers view as a formatted poster.

[4] At the time of data collection (2021-2022), conversational LLMs such as ChatGPT had not yet been widely adopted.

[5] During initial drafting, participants know that their chance of winning depends on another's review but are not aware of the later revision opportunity, discouraging external design search while waiting.

[6] In the first wave (≈ 1/3 of the data), two RAs graded posters independently, with a third resolving

Table 1: Grading Scheme for Posters

Content (0 or 8 points)	**Design & Style (0–10 points)**
1. Mentions the Decision Science Laboratory	11. Has a catchy title
2. Clarifies that recruitment is for academic or social science research	12. Attracts attention
3. Clarifies that participation is voluntary and non-obligatory	13. Nicely formatted (bold key points, bullet points, legible font size)
4. Leaves contact information	14. Easy to read from a distance
5. Provides info on how to sign up for the subject pool	15. Most important information is immediately visible
6. Mentions possibility to earn extra money	16. Not too wordy; appropriate amount of information
7. Gives an indication of expected payment	17. Content and presentation (fonts, colors) appropriate for university poster
8. Indicates target audience (students)	18. Correct grammar and punctuation
9. Indicates rough time commitment for a study	
10. Mentions participation requirements (e.g., skills, technical needs)	

Notes. This table presents the 18 criteria according to which each poster was graded by research assistants blind to the research question and design. These criteria were developed with experienced professionals (managers of university research labs and a communications expert). For each *content category*, posters received 8 points if they contained the information and 0 points otherwise. For each *design & style* category, the graders rated the extent to which the statement applies to the poster on a Likert scale from 0-"does not apply at all" to 10-"definitely applies". Thus, final performance can take values between 0 and 160 points.

2.2. Feedback Structure

In Parts 2 and 4, reviewers have 12 minutes to write feedback on the quality of their peer's poster. Reviewers can type it in up to seven entry fields with a character limit of 200 each. Each non-empty entry field counts as a remark for determining the competitive bonus. The instructions requested "specific points" but imposed no restrictions on whether comments should be positive or negative (see Online Appendix B for the instructions).[7] To avoid mechanical effects driven by irrelevant entries, we base our analysis of the feedback on what was actually written across all entry fields rather than relying on the entry-field count.

Three RAs who were blind to the research question independently classified the feedback into 26 categories. The entry-field count and the comment-count based on the RA classifications diverge when a peer entered two comments in a single entry field, e.g., the suggestions to change the font size of the title and also to fix a typo in it. Online Appendix Table A1 describes the classification categories. The first eighteen categories

discrepancies; the final score averages the independent evaluations. Later waves used one of the three wave-1 RAs, given the high inter-rater correlation in wave 1 (Pearson's $r = 0.850$).

[7]The instructions informed reviewers that their "[..] review should have the following format: a numbered list of comments on the quality of the poster you are reviewing. You can write up to 7 comments. Your comments should focus on specific points that you can include in the numbered list. You have to enter each point that you want to comment on in a new entry field with a character limit of 200."

correspond to the poster grading criteria. The table also reports inter-rater agreement, which is generally substantial.

Importantly, our finding regarding the quantity of feedback provided across feedback regimes is robust to comparing the number of remarks—that is, the number of non-empty entry fields filled during peer review—instead of the number of classified comments, as we show in the results section. While the classified content additionally allows us to study effects on the quality of feedback, our inference on the quantity of feedback is very similar regardless of how feedback content is counted.

Before providing feedback, reviewers learn the exact content grading criteria and that design and style would also be evaluated (see Table 1 for grading criteria). Thus, reviewers understand desirable poster characteristics before writing their review. Reviewers face the following tradeoff when they give feedback: writing a detailed review that identifies many shortcomings can lower their peer's chance of winning the bonus for the best first posters. At the same time, this type of review can help to improve their peer's final poster score and associated earnings. This captures that peer feedback often serves two purposes in organizations: evaluation and development. Since different feedback regimes imply different strategic relationships between reviewer and recipient, our treatments test whether these relationships affect how this tradeoff is resolved.

Given this feedback structure, our analysis focuses on the revise-first group's final poster performance and the feedback they received from the revise-second group. Before revising, revise-second members had seen their peers' posters and the grading criteria and had received peer feedback. Furthermore, in Part 4, revise-second group members received feedback from revise-first group members, who had already been reviewed in Part 2. Thus, the revise-first group's feedback could not have been motivated by the prospect of receiving reciprocal feedback later.

2.3. Feedback Treatments

Our experimental treatments induce exogenous variation in the feedback regime by changing the strategic relationship between reviewer and recipient. We randomize matching groups of four participants into one of two treatments, stratifying by session. The treatments vary the probability (p) that a reviewer from the revise-second group is later reviewed by the recipient of their feedback from the revise-first group. Otherwise, the two treatments are identical. For example, identities are never revealed.

In the *independent* treatment ($p = 0$), the probability of repeated interaction is zero, as we implement a perfect-stranger matching protocol. In the *reciprocity* treatment ($p = 1$), the probability of repeated interaction is one as feedback is exchanged in pairs. The matching protocol is common knowledge from Part 2 on, allowing participants to form expectations about whether their feedback may be reciprocated.

Under the independent treatment, there are no reciprocal concerns when writing feedback. The first reviewer knows that their feedback cannot directly influence the feedback they will later receive, as they are evaluated by an independent third peer. This third peer also understands that they are reviewing someone who did not review them. Consequently, the first reviewer need not consider how their comments in Part 2 might affect the feedback they receive in Part 4. By contrast, in the reciprocity treatment, the second peer can condition their feedback on the feedback they previously received and both peers are aware of this. This can influence the feedback strategy of the first peer in the sequence.

With these treatments, we test whether the feedback regime changes the content and learning value of peer feedback, as well as how recipients perceive and use it. Independent reviewing may encourage more candid and informative evaluations, but competitive motives in the first-poster tournament may also lead to overly critical feedback that recipients perceive as strategically motivated. In contrast, reciprocal reviewers may provide higher-quality feedback to elicit useful feedback in return, but they may also withhold pointed or critical comments if they fear retaliation. Recipients may, in turn, trust reciprocal feedback more because it is perceived as less competitively motivated, or derive less benefit from it if reviewers suppress valuable criticism. Which of these forces dominates in shaping feedback provision, implementation, and final performance is an empirical question.

2.4. Procedures

The online experiment was implemented in oTree (Chen, Schonger and Wickens, 2016) and conducted in four waves between summer 2021 and fall 2022. Participants were recruited from the ETH Zurich Decision Science Laboratory subject pool and took part in 20 monitored online sessions with 20–52 participants. Each 1.5-hour session was monitored via Zoom; participants used anonymous IDs with cameras off. The study was conducted in English. Initial instructions were read aloud to establish common knowledge, and detailed part-specific instructions appeared onscreen. Participants completed mandatory comprehension checks before Parts 1, 2, and 4 and could ask questions for clarification via Zoom. Average earnings were CHF 44.92, including a CHF 10 show-up fee, paid by bank transfer after posters were graded.

3. Results

We present results on feedback provision, recipients' responses to feedback, and final performance. Following the pre-registration, analyses are restricted to revise-first participants (N=367) and the feedback they received from revise-second reviewers (see Section 2.2).

3.1. Provision of Feedback

Table 2 reports treatment effects on feedback quantity and informational content, using the peer-feedback classifications described in Section 2.2.[8]

Feedback Quantity. Column 1 of Table 2 shows that reciprocal feedback significantly reduces the number of comments relative to independent feedback. The estimated coefficient of a linear regression implies that reviewers in the reciprocal treatment provide on average 0.88 fewer comments per poster than reviewers in the independent condition, with a 95% CI of [-1.28, -0.47].[9] This corresponds to a reduction of 15.6% relative to the independent mean (5.62 comments). The outcome counts all aspects of the grading mentioned in the feedback, including repeated mentions, non-helpful comments, and positive remarks. Results are similar using the number of non-empty entry fields ("remarks"), which does not rely on RA classifications: reciprocity reduces remarks by 1.14 on average [-1.50, -0.77].

Table 2: Quantity and Quality of Peer Feedback Sent

	(1) Comments	(2) Implementable Comments	(3) Implementable Comments: Content	(4) Implementable Comments: Design	(5) Non-Implementable Comments
Reciprocity	-0.88***	-0.70***	-0.24	-0.46***	-0.18
	(0.206)	(0.184)	(0.165)	(0.135)	(0.119)
	[-1.28, -0.47]	[-1.06, -0.33]	[-0.56, 0.09]	[-0.72, -0.19]	[-0.41, 0.05]
Control group mean	5.624	4.652	3.421	1.230	0.972
Observations	367	367	367	367	367

Notes. Revise-first participants (N=367). Table presents results from OLS regressions. *Comments*: number of comments received in the peer feedback as independently classified by three RAs blind to the research question, *Implementable Comments*: specific suggestions that can help the performance of the recipient given her performance on the first poster, *Implementable Comments Content/Design*: specific suggestions that can help the performance of the recipient along the content/design dimension, *Non-implementable Comments*: comments that cannot improve performance such as redundant or congratulatory statements. Each regression controls for performance on the first poster and includes session fixed effects. Heteroskedasticity-robust standard errors in parentheses and 95% CIs in square brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Feedback Quality. The reduction in feedback under reciprocity admits two interpretations. Independent reviewers may provide excessive or unhelpful comments to reduce peers' chances of winning the tournament. Alternatively, reciprocity may suppress critical but valuable feedback because reviewers fear retaliation when roles are reversed.

[8]Online Appendix Table A1 reports inter-rater agreement, which is generally substantial, and Table A2 provides examples of classified feedback.

[9]Throughout this section, numbers in square brackets show 95% CIs. In all tables, we report individual-level robust standard errors. Although treatment is assigned at the matching-group level, the two revise-first participants within a group never interact, so there is no behavioral link between their outcomes beyond session effects, which are absorbed by session fixed effects. Clustering at the matching-group level yields empirically indistinguishable standard errors for all regression results.

To distinguish between these interpretations, we examine treatment effects on the number of *implementable* comments. Implementable comments are defined as suggestions that, if followed, would increase the recipient's score on the final poster. This classification combines the content of the comment with the recipient's baseline performance on the first poster. A comment is considered implementable if it identifies a potential improvement in any of the 18 dimensions (see Table 1) for which the recipient did not receive full points in the initial draft.

Column 2 of Table 2 shows that reciprocal feedback significantly reduces the number of implementable comments relative to independent feedback. The estimated coefficient is -0.70 [-1.06, -0.33], corresponding to a reduction of approximately 15.1% relative to the independent mean. While column 5 of Table 2 shows that there are also slightly fewer non-implementable comments under reciprocal feedback (-0.18, [-0.41, 0.05], 18.5% less), the evidence does not support the interpretation that reciprocal feedback primarily reduces comments that would not help to improve performance.

The reduction in implementable comments under reciprocity reflects decreases in both content-related and design-related suggestions. It is informative to examine both dimensions separately because the content criteria are objective (e.g., whether or not information is in a poster) whereas the design criteria are more subjective (e.g., catchiness of the poster title). Moreover, because reviewers knew all the content criteria but not the design criteria, content-related suggestions were arguably safer to make. In absolute terms, the decline is nearly twice as large for design-related comments (-0.46) compared to content-related ones (-0.24). Because design-related comments are less frequent on average, the proportional reduction is even larger for design than for content.

These results indicate that reciprocity concerns primarily operate by suppressing the transmission of useful information. The stronger decline for design comments suggests that reviewers were particularly reluctant to provide more subjective or contestable criticism when they expected reciprocal evaluation. A plausible mechanism is fear of retaliation: reviewers anticipate that critical but informative feedback may be reciprocated when they themselves later receive feedback from the same peer. There is, indeed, a highly significant positive correlation between the number of remarks the first peer in the reviewing sequence sent and the number of remarks they receive in return when reviewing is reciprocal ($\beta = 0.40$ remarks received, robust SE=0.078, N=195 revise-second participants). By contrast, there is a negative correlation under independent reviewing (see Online Appendix Table A3 for the regression estimates).

When reciprocity concerns are removed under independent reviewing, reviewers provide more feedback overall and, crucially, more feedback with genuine learning value.[10]

[10]As an additional check, weaker first-round posters receive more comments: in a regression of comments on first-round performance, reciprocity, and their interaction, the performance coefficient is -0.028 (robust SE=0.008). The interaction with reciprocity is positive but imprecise (0.007, SE=0.012), suggestively consistent with reciprocity dampening feedback responsiveness to performance.

3.2. Response to Feedback

We next examine how recipients respond to the feedback they receive, either directly when deciding what to implement or indirectly by deciding how much to cooperate with another (unrelated) peer after experiencing the peer review. In this section, we also look at how the feedback regime affects perceptions of the feedback.

Table 3: Implementation of Peer Feedback Received and Final Performance

Panel A			
	(1)	(2)	(3)
	Implementation Rate	Implementation Rate: Content	Implementation Rate: Design
Reciprocity	-0.02	-0.02	-0.04
	(0.035)	(0.039)	(0.062)
	[-0.09, 0.05]	[-0.10, 0.06]	[-0.16, 0.08]
Control group mean	0.564	0.604	0.495
Observations	367	334	184
Panel B			
	Implemented Comments	Implemented Comments: Content	Implemented Comments: Design
Reciprocity	-0.32*	-0.10	-0.22***
	(0.172)	(0.160)	(0.076)
	[-0.66, 0.02]	[-0.42, 0.21]	[-0.37, -0.07]
Control group mean	2.596	2.028	0.567
Observations	367	367	367
Panel C			
	Final Poster: Total points	Final Poster: Content	Final Poster: Design
Reciprocity	-0.74	-0.93	0.19
	(1.541)	(1.347)	(0.729)
	[-3.77, 2.29]	[-3.58, 1.72]	[-1.25, 1.62]
Control group mean	92.64	56.95	35.69
Observations	367	367	367

Notes. Revise-first participants. Table presents results from OLS regressions. *Implementation Rate*: the ratio of implemented comments in the final draft to implementable comments received in the peer feedback. In the first row of column (1), the 10 missing values arising from participants who received no implementable comments are set to 0. In columns (2) and (3) the sample sizes are determined by the number of participants who received > 0 implementable content or design comments. *Implemented Comments*: number of comments that the recipient implemented in the final poster. *Final Poster*: total points achieved on the final poster (0–160 points). Each regression controls for performance on the first poster and includes session fixed effects. Heteroskedasticity-robust standard errors in parentheses and 95% CIs in square brackets.* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Feedback Implementation. Panels A and B of Table 3 report treatment effects on two sets of outcomes: the rate at which implementable comments are incorporated into the final poster (Panel A) and the number of implemented comments (Panel B).

Panel A of Table 3 shows that participants incorporate only a fraction of the implementable comments they receive.[11] In the independent group, the mean implementation

[11] An implementable comment identifies a potential improvement in one of the 18 grading dimensions

rate is 0.56. Importantly, this implementation rate is not significantly affected by the feedback regime. The estimated effect of reciprocal feedback on the overall implementation rate is -0.02 [-0.09, 0.05]. The corresponding estimates are -0.02 for content implementation [-0.10, 0.06] and -0.04 for design implementation [-0.16, 0.08].

A likely explanation for this limited adoption is that many recipients do not perceive the feedback as particularly useful, even when implementing it would improve performance. Overall, 34% of participants report that the feedback they received was only "a little" or somewhat helpful, with no significant difference in this subjective assessment of feedback quality across treatments (see Figure 2). Implementation rates are moderately positively correlated with subjective feedback quality (Spearman $\rho = 0.43, p < 0.001$).[12] By contrast, this subjective assessment correlates only weakly with an objective measure of feedback quality, namely, the number of implementable comments received ($\rho = 0.18, p < 0.001$). Online Appendix Figure A2 visualizes these patterns. Taken together, these results are consistent with perceived usefulness playing a central role in determining whether feedback is implemented.

Another possible explanation for the low implementation rate is that participants lacked sufficient time to process and incorporate feedback. However, Figure 1 presents two analyses that suggest otherwise: participants receiving more implementable comments are no more likely to report wanting additional time (Panel A), and the propensity to incorporate at least one implementable comment from an entry field does not decline with its position in the feedback list (Panel B).

Panel B of Table 3 turns to treatment effects on the number of *implemented* comments. Reciprocal feedback reduces the number of implemented comments by 0.32 relative to independent feedback [-0.66, 0.02]. This estimate corresponds to a 12% reduction relative to the independent mean of 2.60. Thus, although incomplete implementation attenuates the effect on implemented comments relative to received implementable comments, the effect remains sizable.

The decline in implemented comments is concentrated in design-related feedback. Reciprocal feedback reduces the number of implemented design comments by 0.22 [-0.37, -0.07], relative to the mean of 0.57 in the independent group. By contrast, the effect on implemented content-related comments is small and imprecisely estimated (-0.10 [-0.42, 0.21], control mean 2.03). These patterns mirror the results in Section 3.1, where reciprocity led to a larger reduction in design-related implementable comments.

Consistent with actual implementation, participants in the reciprocity group also

on which the final poster is graded. It is implemented if a recipient's final poster scored higher in that dimension compared to their first poster. The implementation rate is the number of implemented comments divided by the number of implementable comments.

[12]Participants answered the question, "To what extent was the review that you have received helpful for improving your second poster?", using a four-point scale: 1-"a little", 2-"some", 3-"pretty much", 4-"very much".

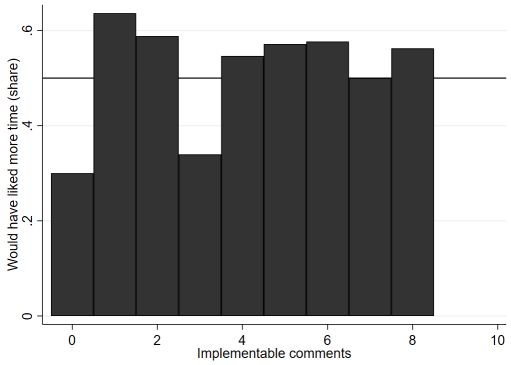
report having implemented 0.78 fewer comments on average (robust SE 0.183) compared to the 4.00 comments participants in the independent group report.[13]

Even in a setting that rewards performance gains, recipients only selectively act on the useful information they receive, regardless of the feedback regime. This is all the more striking because the feedback met several conditions known to enhance its effectiveness: it was timely, specific and task-focused, could help to close the gap between current and desired performance, and posed little threat to recipients' self-esteem (Hattie and Timperley, 2007).

[13]Participants were asked in the survey "How many comments did you end up incorporating in the draft of your final poster?". The estimated coefficient comes from a regression with our standard set of controls.

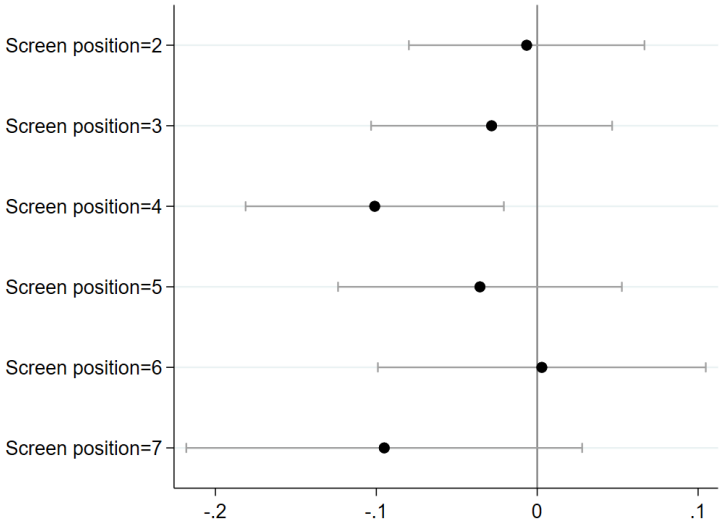
Figure 1: Time Constraints and Feedback Implementation

(a) Wanting More Time to Work on Poster and Number of Implementable Comments



Notes. Share of *Revise-first* participants (N=367) who said that they would have liked more time to create their poster as a function of how many implementable comments they received.

(b) Early and Late Comments: Feedback Screen Position and Propensity to Implement



Notes. Sample comprises 1,411 entry fields containing at least one implementable comment submitted to revise-first participants. Coefficient plot from a linear regression estimating the propensity to incorporate at least one implementable comment written in an entry field in the final poster based on the entry field's screen position (1 = top; 7 = bottom) with position 1 as the omitted category (mean probability = 0.64). The regression includes session fixed effects and standard errors are clustered at the participant level. Whiskers show 95% CIs.

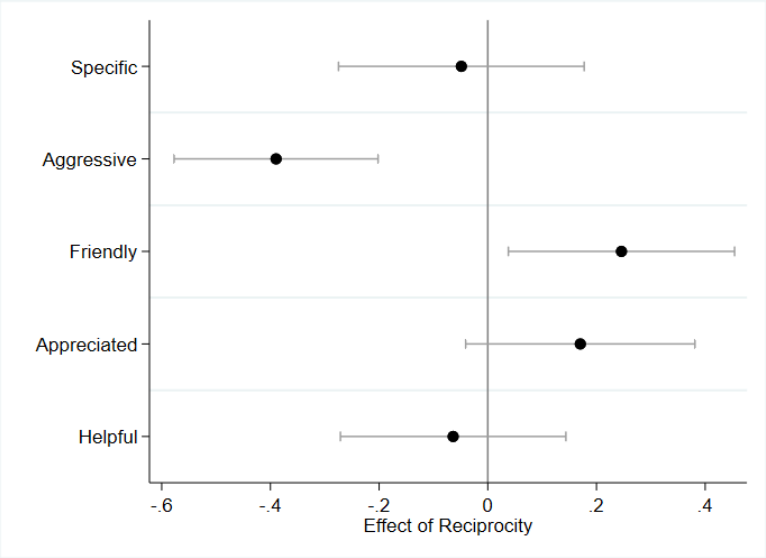
Feedback Perception and Cooperation. Next, we examine whether the feedback regime affects how recipients perceive the feedback they received and their willingness to cooperate with peers. Panel A of Figure 2 reports the effect of reciprocal feedback on several perception measures, expressed in standard deviations of the independent group. Reciprocity significantly reduces the perceived aggressiveness of feedback: recipients in the reciprocity treatment rate the comments as less aggressive by approximately 0.4 standard deviations. Reciprocity also increases the perceived friendliness of comments and appreciation of them, with effect sizes of roughly 0.2 standard deviations, although this latter estimate is not statistically significant at conventional levels. By contrast, we find no evidence that the feedback regime changes the perceived helpfulness of feedback.

In addition, we investigate whether the feedback regime affects cooperative behavior in the public goods game, which is played with another participant who was not previously encountered (see Section 2.1 for details). This design tests for potential spillover effects of the feedback experience on general cooperative behavior, rather than direct reciprocity or retaliation toward the feedback partner. As shown in Panel B of Figure 2, average contributions to the group project are the same across the two feedback regimes: 12.8 tokens out of 20. Overall, contribution levels are relatively high for a one-shot public goods game, but they do not vary meaningfully with the feedback regime.

Although reciprocal feedback is perceived as friendlier and less aggressive, the findings provide no evidence that the two feedback regimes affect how recipients respond to the feedback they receive or behave in subsequent social interactions.

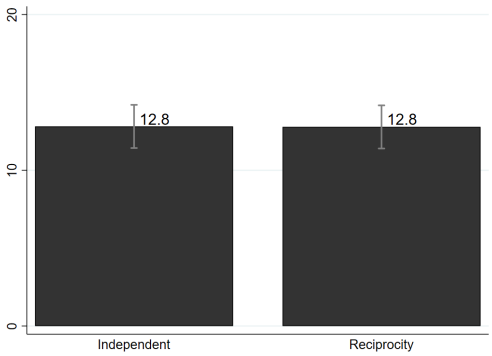
Figure 2: Feedback Perception and Cooperation

(a) Effect of Reciprocity on Feedback Perceptions



Notes. Revise-first participants (N=367). Dots show point estimates from regressions of the dependent variable (e.g., *specific*) on *reciprocity* treatment indicator when also controlling for session FEs. Whiskers show 95% CIs based on heteroskedasticity-robust standard errors. Effect sizes in terms of standard deviations of the independent group. Outcomes (1)-(4) were originally on a Likert scale: 1-"strongly disagree" to 7-"strongly agree". (1) *Specific*: "Overall, the comments I have received were specific." (2) *Aggressive*: "Overall, the comments I have received were written in an aggressive tone." (3) *Friendly*: "Overall, the comments I have received were written in a friendly tone." (4) *Appreciated*: "I appreciate the review that I have received." Outcome (5) *Helpful*, originally on Likert scale 1="a little", 2="some", 3="pretty much", 4="very much": "To what extent was the review that you have received helpful for improving your second poster?"

(b) Cooperation Levels by Treatment



Notes. Revise-first participants (N=367). Average number of tokens contributed to the group project out of 20. Whiskers show 95% CIs. Public Goods Game was played with someone not encountered before.

3.3. Performance after Feedback

In this final section, we consider how the feedback regime affects performance after feedback. Throughout this section, we again focus on outcomes for revise-first participants for whom peer feedback was the only new input before they worked on their final poster. Panel C of Table 3 shows the treatment effects of the feedback regime on the final poster score (0–160 points) and on the two subscores, content and design (0–80 points each).

Despite the reductions in both implementable and implemented comments under reciprocal feedback, we find no detectable effect on final poster performance. The estimated effect of reciprocity on total poster score is -0.74 points [-3.77, 2.29], relative to the independent-treatment mean of 92.64 points. The point estimate amounts to a 0.80% reduction in average performance and the confidence interval rules out reductions greater than 4% or improvements greater than 2.5%. The absence of significant performance effects is consistent across poster components: the objective content score and the more subjective design score. The effect on content scores is -0.93 [-3.58, 1.72], while the effect on design scores is 0.19 [-1.25, 1.62]; neither estimate is statistically different from zero. Figure A3 in the Online Appendix further shows that the similarity in performance across feedback regimes extends to the entire distribution, not just the mean.[14]

Thus, although independent feedback increases both the supply of implementable comments and the amount of implemented feedback, and although participants could substantially improve their posters, final performance is indistinguishable across feedback regimes.

Using a back-of-the-envelope calculation, we can provide suggestive evidence that the limited implementation rate (see Section 3.2) is driving this zero result on performance after feedback. We compare across the two feedback regimes the predicted best-response poster score, defined as the score a participant would have received had all implementable comments been fully implemented.[15] This benchmark represents the best possible improvement based on the received feedback, assuming perfect additivity of all comments and abstracting away from time constraints and substitution effects across dimensions. Relative to the independent group (mean = 104.0), the mean best-response score in the reciprocity group is 4.11 points lower on average (robust SE = 1.34; [-6.8, -1.5]).[16] The predicted best-response performance under reciprocal feedback is therefore about 4%

[14]The figure also shows substantial improvements from the first to the final poster under both feedback regimes. Evidence from the revise-second group—which learned the grading criteria before revising—further indicates scope for improvement: it improved by 14 points more on average than the revise-first group. Thus, the absence of a significant treatment effect on final poster performance is not due to a ceiling effect.

[15]For an implementable comment on content, the predicted final score increases from the first-round performance of 0 to 8, reflecting how these content categories are scored. For an implementable comment received on design aspects of the poster, the predicted final score increases from the first-round performance $\in \{0, 1, ..., 9\}$ to 10, i.e., the maximum score in a design category.

[16]Coefficient and heteroskedasticity-robust SE from a regression with our standard set of controls.

lower, on average, and we can confidently reject the null hypothesis of equal performance across treatments.

Looking at best-response scores on content and design categories separately, the predicted content scores are 2.35 points lower (-3.5%; [-4.9, 0.2]) and predicted design scores 2.02 points lower (-5.6%; [-3.3, -0.7]) under reciprocity.[17] These estimates show a comparably sized reduction on the objective content measure. Thus, the aggregate result appears not driven by noise in the more subjective design scoring.

These findings highlight the central contribution of our paper. The absence of performance effects does not imply that peer feedback lacks valuable information: independent feedback generated more implementable suggestions and full implementation would have led to higher performance. Rather, the results reveal an important demand-side bottleneck. Recipients incorporate only about half of the implementable comments they receive, limiting the extent to which higher-quality feedback translates into learning gains.

4. Conclusion

Peer feedback regimes may influence not only what reviewers communicate but also how recipients use the information they receive. Competition for rewards and expected future interactions create strategic incentives that may distort both feedback provision and use. In a laboratory experiment, we compare a regime in which competition alone may distort feedback with one in which both competition and reciprocity may do so. Participants complete two rounds of a complex, creative task, sequentially reviewing one another. Feedback serves both evaluative and developmental roles. More critical comments can reduce the likelihood of winning a competitive bonus but improve future performance and pay.

We find that an independent reviewing regime generates more informative feedback with greater learning value than a reciprocal one. Using granular data on the content of feedback and recipients' performance before and after feedback, we further show that recipients systematically underuse the feedback they receive, regardless of the feedback regime. Failure to recognize the full learning value of feedback emerges as one plausible explanation for limited implementation. This underutilization dampens the performance gains that higher-quality feedback generated by different regime design could have otherwise produced.

Our results point to two design implications for feedback regimes. First, feedback regimes that weaken strategic links between reviewers and recipients may elicit more informative evaluations. Second, even high-quality feedback may not be acted upon. Thus, effective feedback regimes must address two distinct challenges: creating incentives

[17] Same regression specification as the aggregate effect, run separately on each subscore and controlling for the first poster subscore.

for reviewers to provide informative evaluations and helping recipients act on valuable feedback.

References

Balietti, Stefano, Robert L Goldstone, and Dirk Helbing. 2016. "Peer review and competition in the Art Exhibition Game." *Proceedings of the National Academy of Sciences*, 113(30): 8414–8419.

Bolton, Gary, Ben Greiner, and Axel Ockenfels. 2013. "Engineering trust: reciprocity in the production of reputation information." *Management Science*, 59(2): 265–285.

Boudreau, Kevin J. 2026. "Field Experiments in the Science of Science: Lessons from Peer Review and the Evaluation of New Knowledge." *NBER Working Paper* 34811.

Carpenter, Jeffrey, Peter Hans Matthews, and John Schirm. 2010. "Tournaments and Office Politics: Evidence from a Real Effort Experiment." *American Economic Review*, 100(1): 504–517.

Chen, Daniel L, Martin Schonger, and Chris Wickens. 2016. "oTree—An Open-Source Platform for Laboratory, Online, and Field Experiments." *Journal of Behavioral and Experimental Finance*, 9: 88–97.

Crawford, Vincent P., and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica*, 50(6): 1431–1451.

Di Fiore, Alessandro, and Marcio Souza. 2021. "Are peer reviews the future of performance evaluations." *Harvard Business Review*.

Drouvelis, Michalis, and Paola Paiardini. 2022. "Feedback quality and performance in organisations." *The Leadership Quarterly*, 33(6): 101534.

Dufwenberg, Martin, Katja Görlitz, and Christina Gravert. 2025. "Peer Evaluation Tournaments." *CESifo Working Paper* 11720.

Garicano, Luis. 2000. "Hierarchies and the Organization of Knowledge in Production." *Journal of political economy*, 108(5): 874–904.

Gneezy, Uri, Christina Gravert, Silvia Saccardo, and Franziska Tausch. 2017. "A must lie situation – avoiding giving negative feedback." *Games and Economic Behavior*, 102: 445–454.

Hattie, John, and Helen Timperley. 2007. "The Power of Feedback." *Review of Educational Research*, 77(1): 81–112.

Hussam, Reshmaan, Natalia Rigol, and Benjamin N Roth. 2022. "Targeting high ability entrepreneurs using community information: Mechanism design in the field." *American Economic Review*, 112(3): 861–898.

Klapper, Helge, Henning Piezunka, and Linus Dahlander. 2024. "Peer evaluations: Evaluating and being evaluated." *Organization Science*, 35(4): 1363–1387.

Landis, J. Richard, and Gary G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics*, 33(1): 159–174.

Li, Danielle. 2017. "Expertise versus Bias in Evaluation: Evidence from the NIH." *American Economic Journal: Applied Economics*, 9(2): 60–92.

Li, Danielle, and Leila Agha. 2015. "Big names or big ideas: Do peer-review panels select the best science proposals?" *Science*, 348(6233): 434–438.

Mickeler, Maren, Diego Zunino, Tobias Kretschmer, and Marine Hadengue. 2025. "Anonymity and Peer-Feedback in Crowdsourcing: Evidence from a Field Experiment." ESSEC Business School. Available at SSRN: https://ssrn.com/abstract=5282432.

Miller, Nolan, Paul Resnick, and Richard Zeckhauser. 2005. "Eliciting Informative Feedback: The Peer-Prediction Method." *Management Science*, 51(9): 1359–1373.

Villeval, Marie Claire. 2020. "Performance feedback and peer effects." *Handbook of Labor, Human Resources and Population Economics*, 1–38.