Performance-Feedback

Jean-Pierre Benoît
London Business School, email: jpbenoit@london.edu

Ashley Perry
New York University Abu Dhabi, email: ashley.perry@nyu.edu

Ernesto Reuben
New York University Abu Dhabi, Center for Behavioral Institutional Design,
Luxembourg Institute of Socio-Economic Research, email: ereuben@nyu.edu

Abstract

Feedback plays a critical role in shaping beliefs, guiding decisions, and improving performance. We conduct an online experiment to study the nature and effectiveness of qualitative feedback. Although qualitative feedback is widely used, it has received little attention in experimental economics, where the primary focus has been on quantitative feedback. We find that subjects interpret the open-ended feedback they receive properly and update their beliefs appropriately. In particular, they do not incorporate the feedback in a way that is unduly favorable to themselves. This contrasts with previous work on quantitative feedback, which finds an upward bias in updating. The providers of written feedback exhibit a kindness bias: their feedback is more positive than their true opinions. However, recipients anticipate the bias. We find no difference in how feedback is given to men and women, unlike prior work. We identify two channels through which feedback influences decisions: a belief-updating channel and an encouragement channel. Women respond to both, while men are less responsive to encouragement. We find improved performance for individuals who receive feedback that includes more concrete advice.

This version: April 2026

JEL Codes: D83, D91, J16

Keywords: performance, feedback, qualitative data, beliefs, decision-making, gender

Acknowledgments: We thank Loukas Balafoutas, Juan Dubra, Roel van Veldhuizen, and various conference and seminar participants for their helpful comments and suggestions. We also thank Gabriel Møller for his research assistance. We are grateful to the London Business School's Wheeler Institute for Business and Development for supporting this research. We also gratefully acknowledge financial support from Tamkeen under the NYU Abu Dhabi Research Institute Award CG005. The project received IRB approval at London Business School (REC816-16062025). A link to the preregistration can be found here: https://aspredicted.org/LG8_JPK.

1. Introduction

Feedback is important for performance in a variety of settings. Employees receive periodic appraisals of their work; students are given grades and comments throughout their schooling.[1] In recent years, there has been a shift toward greater use of qualitative, rather than quantitative, feedback. For example, in 2016 General Electric introduced a qualitative feedback system for its 300,000 employees (Silverman, 2016), and in the United Kingdom, a 2015 education report advised schools to rely less on numerical assessments when providing student feedback (McIntosh, 2015). But is qualitative feedback effective? Despite its growing use in practice, qualitative feedback has received far less attention in the economics literature than its quantitative counterpart.

By qualitative feedback, we refer to textual descriptions of performance; by quantitative feedback, we mean numerical information. Quantitative feedback can vary in precision—for example, it may be a specific performance rating or an imprecise signal indicating that performance probably ranks in the top quartile. Qualitative feedback inherently involves a degree of vagueness, often a significant one. For instance, two people who observe the same performance and agree on its quality may nevertheless describe it using very different language. Conversely, two people may use similar language to describe performances of objectively different quality. Qualitative feedback requires recipients to decipher the meaning of the text, posing particular challenges to its usefulness.[2] The purpose of this paper is not to establish if one is optimal; there are many contexts in which one or both types of feedback may be of use. Rather, in light of the potential ambiguities in language, we seek to understand how qualitative feedback is used by those who give it and those who receive it.

Feedback and performance are related in a *performance-feedback sequence*: an individual undertakes a task, forms beliefs about their performance, has their performance evaluated, receives feedback, updates their beliefs, and takes subsequent actions, which may include steps to improve their performance. We study the entire sequence, since considering only some stages can lead to misleading conclusions. For example, a finding that evaluators' feedback is systematically biased could, by itself, suggest that feedback is unhelpful. Only by also studying how the recipients respond to the feedback can we determine whether they anticipate the biases

[1] There is growing evidence that management practices, of which feedback is one aspect, are important for improving performance of firms (Bloom and Van Reenen, 2007; Bloom et al., 2015, 2019). In schools, quantitative feedback has been shown to improve student performance (Bandiera et al., 2015; Andrabi et al., 2017).

[2] The challenges of qualitative feedback are also present in certain quantitative feedback settings, in which case the present study applies there as well. Moreover, whilst we study the particular problem of updating beliefs about performance with only access to qualitative feedback, there is complementary work which studies how to incorporate qualitative information, such as stories and memories, into models of belief formation (Bordalo et al., 2025; Graeber et al., 2024).

and correct for them.[3] To the best of our knowledge, we are the first to study qualitative feedback using an experiment that covers the entire performance-feedback sequence with each stage undertaken by participants.[4]

In practice, a worker may get feedback from their immediate supervisor, while their end-of-year bonus is determined by a committee on which that supervisor has just one vote; a professor may solicit comments from a colleague who has no direct influence on publication decisions. Throughout their lives, people receive feedback from a subset of the people who evaluate them. For qualitative feedback to be effective, the recipients must:

i. Correctly interpret the feedback. For instance, determine whether the phrase "good job" indicates that the evaluator believes performance is above average, average, or even below average.
ii. Assess how informative the feedback giver's opinion is about the views of other evaluators.
iii. Incorporate the feedback into their beliefs and subsequent decisions.

Our online experiment shares these features. The experiment centers on an essay-writing task. Participants are assigned to one of two roles: writer or evaluator. Each writer composes a short essay inspired by an image. The essay is then graded by a group of ten evaluators, each of whom assigns a number grade. Writers are not shown any of these grades nor provided with quantitative feedback. Instead, they receive written qualitative feedback from one randomly chosen evaluator. Writers report their beliefs about their average grade both before and after receiving the feedback. Even though the performance metric we use in this experiment is quantitative, the key feature of our design is that the feedback about performance *is* qualitative and so presents the writer with the issue of how to interpret this type of information given the performance metric used.

Previous research has found that recipients of *quantitative* feedback often update their beliefs about their performance in an upwardly biased manner, placing greater weight on favorable information (Eil and Rao, 2011; Möbius et al., 2022). We might expect that qualitative feedback, with its open-ended nature, is even more prone to bias. It can contain ambiguous messaging, and psychological phenomena, such as motivated reasoning, may allow for a variety of interpretations. Consider the following feedback, taken from our experiment:

[3]For example, Jampol and Zayas (2020) find women receive kinder feedback than men and conclude that this makes feedback less useful for women. However, their experimental design does not allow them to examine how the feedback is interpreted and whether women anticipate this effect and account for it. See Section 2. and Result 2 for more on their study.

[4]Prior work in psychology and economics has examined one or two stages. This work includes experiments that focus on biases at the evaluation stage (Goldberg, 1968; Mechtenberg, 2009), in the way individuals form beliefs about their performance (Exley and Kessler, 2022), in the feedback given (Bohren et al., 2018; Jampol and Zayas, 2020; Jampol et al., 2022), in how individuals update their beliefs after feedback (Eil and Rao, 2011; Ertac, 2011; Zimmermann, 2020; Möbius et al., 2022), and the impact of feedback on choices (Wozniak et al., 2014; Brandts et al., 2015; Shastry et al., 2020; Abel, 2024; Abel and Buchman, 2024).

> "I think this was a good attempt. You've explored the different parts of the picture, while also delving deeper into Josh's thoughts and emotions, providing a context to the scene. The flow does seem to be a bit muddled at times, for example I think the description of the other people could have been incorporated into the story in a slightly neater way. Some more creative use of language would have been nice also. The grammar and spelling is accurate though. All in all, it was enjoyable!"

This feedback corresponds to an essay for which the evaluator gave a grade of 3 on a 1-to-5 scale, although the writer only saw the text, not the numerical score. To us, the content of the feedback appears consistent with the grade. However, the recipient could selectively attend to different parts of the text. An optimistic writer might focus on the positive elements— "You've explored the different parts of the picture … The grammar and spelling is accurate … All in all, it was enjoyable!"—and infer an above average grade, perhaps estimating a 4. In contrast, a pessimistic writer might focus on the negatives—"The flow does seem to be a bit muddled … Some more creative use of language would have been nice also"—and conclude they received a below-average grade of 2.

On this accounting, qualitative feedback might be particularly ineffective. On the other hand, people have a lifetime of experience with qualitative feedback and may have a good intuitive understanding of it. Perhaps they integrate imprecise qualitative feedback into their beliefs better than the somewhat artificial quantitative signals they receive in some experiments.

Of course, the artificiality of the signals in some experiments is purposeful: a well-chosen quantitative signalling structure has the advantage of enabling researchers to make precise Bayesian calculations, which can form a benchmark from which to judge the beliefs and behaviour of subjects. Our experiment is purposely less structured, to better reflect real-world qualitative feedback environments. The feedback provided by the evaluators is open-ended and written in their own words. This provides a more natural setting, and it is important to investigate belief-updating in natural settings both because those are, ultimately, the domain of interest, and because there is evidence that more natural environments may improve information processing.

For example, the early experimental evidence for confirmation bias was based on artificial settings (Wason, 1960, 1968). Wason (1968) showed subjects four cards, each with a letter on one side and a number on the other. The visible faces were *D, 3, B,* and *7*. Subjects were asked which cards should be turned over to falsify the rule that a card with a D on one side has a 3 on the other. Only three percent correctly identified *D* and *7*. However, Griggs and Cox (1982) showed that people did not make this type of error in a more realistic setting. In their study, the four cards had a person's age on one side and what they were drinking on the other. The rule to falsify was that a person drinking beer must be over 19 years old. Participants saw cards displaying *coke, age 16, beer*, or *age 22*, and seventy-three percent correctly identified *age 16* and *beer* as the cards to be turned over.

In our design, the content of the feedback is unique – evaluators use language of their own choosing – making the probability of receiving particular feedback difficult, or impossible, to calculate. Nevertheless, by collecting writers' beliefs about their performance before and after they receive qualitative feedback, and taking into account the unseen grade that accompanies the feedback, we can determine if the feedback is interpreted appropriately.

Beyond beliefs, we study how feedback influences decisions and how the content of feedback affects its usefulness. In one set of treatments, writers are given the option to compete for a bonus payment based on their average grade, while in another treatment, they are given the opportunity to revise their essays and have them regraded. By combining the participants' decisions with their beliefs, we are able to examine the motivational and informational channels through which feedback can shape behavior.

We also explore the nature of the feedback itself by analyzing its textual content using a Generative Pre-trained Transformer (GPT); we compare comments that evaluators knew would be given as feedback to writers to assessments of the same essays written in a confidential setting.

Finally, motivated by prior evidence on gender disparities in self-assessment and responsiveness to feedback, we examine several gender-related questions: Do women and men differ in their initial beliefs about their performance? Do they receive systematically different feedback for equivalent performance? Do they respond differently to the feedback in belief-updating and subsequent decisions?

We note that we ran the online experiment before the widespread availability of ChatGPT, so we can be sure that participants did not use it to write essays or provide feedback.

Overview of the findings

Below, we provide an overview of our main findings.

Qualitative feedback is interpreted appropriately. Despite the open-ended nature of qualitative feedback and the inherent subjectivity in its interpretation, writers interpret the feedback in a manner consistent with the (unseen) grade that accompanies it: they revise their beliefs upward when the grade is above their prior and downward when it is below.

Belief-updating is suboptimal. On average, the magnitude of belief-updating is less than optimal.

There is an upwards kindness effect in feedback. When the evaluators' comments will be seen by the writer, the feedback is more positive than when the comments are confidential.

For example, feedback accompanying an essay graded 2 is, on average, as positive as confidential comments written for an essay graded 3. However, writers anticipate the kindness effect.

There is no gender bias in feedback or belief-updating. Contrary to some previous findings, female and male writers receive equally positive feedback (for essays with similar grades). Moreover, conditional on having the same prior belief, men and women update their beliefs similarly in response to the feedback.

Feedback should, arguably, be gender specific. While feedback is equally positive and interpreted similarly by men and women, differences in the accuracy of prior beliefs imply that optimal updating requires different revisions across genders. This suggests that feedback may need to be tailored to address underlying differences in priors.

Feedback and behavior. In the choice to compete, there are two channels through which qualitative feedback affects behavior: a belief-updating channel and an encouragement channel. When it comes to revising their essay, feedback improves essay quality, with more concrete feedback leading to larger improvements.

2. Relation to the literature

We contribute to the experimental literature on performance, feedback, beliefs, and decision-making, and how these relate to gender. We discuss the previous literature on these issues below.

In psychology, some studies find that women receive systematically more positive feedback than men (Jampol and Zayas, 2020; Correll et al., 2020; Jampol et al., 2022; Sheppard et al., 2025). Although this literature has not explored the effects within the performance-feedback sequence, making it hard to appraise the total impact. For example, Jampol and Zayas (2020), participants are given a poorly written essay and either told it was written by a woman or by a man. When asked to provide written feedback to the purported (fictional) writer, participants give more positive feedback when they believe the writer is a woman. While this experiment controls for the content of the essay, allowing them to identify gender biases in feedback provision, it cannot examine how recipients interpret and react to feedback, and therefore cannot speak to its effectiveness.[5]

In experimental economics, a growing body of work examines quantitative feedback. Several studies explore belief-updating in response to noisy signals about performance (Eil and Rao,

[5]There is a large body of work studying feedback in psychology and management for a review of this literature see Kluger and DeNisi (1996) and Heine et al. (2026)

2011; Ertac, 2011; Zimmermann, 2020), while others investigate how feedback influences outcomes (Ertac and Szentes, 2011; Wozniak et al., 2014; Shastry et al., 2020; Kessel et al., 2021) or both beliefs and outcomes (Brandts et al., 2015; Buser et al., 2018; Möbius et al., 2022; Coffman et al., 2024; Coutts et al., 2026). These studies typically provide quantitative feedback based on a well-defined signal structure, which allows the authors to compare updating to a Bayesian benchmark but abstracts from the ambiguity and richness of qualitative feedback. In contrast, our experiment uses open-ended, text-based feedback, where the information content must be inferred by the participant, introducing distinct types of challenges.

A separate line of research examines how feedback affects economic decision-making across genders—particularly in the context of choosing between tournament and piece-rate compensation (Niederle and Vesterlund, 2007). Several papers find that feedback can reduce or eliminate gender gaps in willingness to compete (Ertac and Szentes, 2011; Wozniak et al., 2014; Brandts et al., 2015; Shastry et al., 2020; Kessel et al., 2021). These studies rely on quantitative feedback and do not explore how such effects operate through the interpretation of textual feedback.

More recently, some experimental studies in economics have incorporated qualitative feedback based on performance (Bohren et al., 2018; Abel, 2024; Abel and Buchman, 2024). However, Bohren et al. (2018) focus primarily on discriminatory behavior in evaluation and do not examine how feedback affects recipients' beliefs or decisions, nor do they trace its effects across the full performance-feedback sequence. The studies of Abel and Buchman (2024) and Abel (2024) do not capture the entire performance-feedback sequence as they do not measure feedback recipients' performance beliefs.

Finally, we contribute to an emerging literature on learning from qualitative information, which examines a variety of issues, including the role selective memory plays in belief formation (Bordalo et al., 2025; Graeber et al., 2024), how senders use words to strategically bias receivers beliefs (Thaler et al., 2025), how senders use arguments to persuade voters' beliefs on a policy issue (Huning et al., 2022), how face-to-face communication impacts financial decision-making (Ambuehl et al., 2025), and the impact of word-of-mouth transmission of economic forecasts on beliefs (Graeber et al., 2026). Our study is complementary to this literature as we focus on a particular type of qualitative information, performance-feedback.

3. Experimental design

We ran an experiment using participants from the UK recruited with Prolific, an online research platform with a diverse pool of participants for academic and behavioral studies. The experiment consisted of three parts that took place within a three-week window. The parts were conducted in order from one to three, with a new part only commencing once the prior one had been completed. Appendix E provides a complete description of the study. Here, we limit our-

selves to describing the aspects of the study that are analyzed in the current paper. The study comprises a number of treatments to which participants were randomly assigned. As most of the study structure is common to all treatments, we first describe the two baseline treatments used in the initial analysis, *No-Feedback* and *Feedback*. We describe the other treatments in detail later on.

3.1. Baseline treatments

Participants were assigned to one of two roles: *writers* or *evaluators*. Writers participated in Parts 1 and 3, while evaluators participated in Part 2. All participants received a participation fee of £4 and a bonus payment based on performance.

Part 1: Writers

In Part 1, writers were given 15 minutes to write an essay inspired by an image (the same image was used for all writers and is available in Appendix E). Essays were required to be between 100 and 1000 words. Writers were informed that their essay would be graded by ten evaluators on an integer scale from 1 to 5. Their final grade would be the average of the ten grades. Writers were told that evaluators were recruited through the same platform and were instructed to assess the essays based on four criteria: accuracy and detail, flow and structure, creativity and engagement, and spelling and grammar. Writers were also told that, upon returning for Part 3, they might receive written feedback on their essay.

Writers received a bonus based on their final grade. Specifically, each writer's final grade was compared with those of nine other randomly selected writers. A writer earned £4 if their grade ranked among the top three and £1 otherwise. To minimize attrition, participants were informed that payment would only be made if they completed both Part 1 and Part 3.

After submitting their essay, writers were asked to indicate their expected final grade using a slider ranging from 1 to 5, with increments of one decimal point.[6] We chose not to incentivize belief elicitation. Work by Danz et al. (2022) suggests that incentivized belief elicitation with proper scoring rules can be cognitively demanding and confuse participants, potentially distorting the elicited beliefs. In addition, incentivized belief elicitation creates opportunities for hedging across tasks (Blanco et al., 2010). Consistent with these concerns, Charness et al. (2021) find that incentive-compatible methods do not outperform simply asking participants to state their beliefs.[7]

[6] Participants' point predictions are typically interpreted as the mean of their belief distribution (e.g., Eil and Rao, 2011; Möbius et al., 2022). While some participants may report other summary statistics (e.g., the median or the mode), our primary interest lies in the direction of belief-updating, which is likely to be robust across different summary statistics.

[7] In any case, our positive findings indicate that writers were not answering carelessly.

Finally, writers were asked to select an alias from a list of gender-congruent names.[8] These aliases were displayed to evaluators in place of participant names. The use of aliases served two purposes. First, some real names may be gender ambiguous (e.g., one of the authors of this paper, Ashley, has such a name). In contrast, the aliases we used were unambiguously gendered. Second, to control for potential ethnicity effects, we restricted the aliases to typically white names commonly used in the UK. Section A.1. in the Appendix describes in detail how the aliases were selected.

Part 2: Evaluators

In Part 2, evaluators were randomly assigned to ten essays. They graded each essay on an integer scale from 1 to 5, using the criteria described above. Evaluators knew that multiple evaluators would grade each essay and that writers would not see individual grades but would learn whether their final grade placed them among the top 30%, which would determine their bonus payment. Evaluators were shown the image that inspired the essays, below which they saw the phrase "Written by [writer's alias]," followed by the essay text.

To encourage careful grading, an evaluator's grade was compared to the grades given by nine other evaluators to the same essay. Evaluators earned £0.50 per essay for which their assigned grade matched the modal grade given by the other nine evaluators.[9] Since evaluators graded ten essays, the maximum bonus possible was £5.

After completing the grading task, each evaluator was asked to write between 50 and 1000 words about one of their essays, randomly chosen. In the *No-Feedback* treatment, they were asked to describe the reasoning behind their grade and told that their comments would not be shared with the writer. In the *Feedback* treatment, they were asked to provide feedback directly to the writer on how well they thought the writer had done. Evaluators knew that each writer would receive feedback from only one evaluator.[10] Evaluators were explicitly instructed not to mention the numeric grade they had assigned. We refer to this grade as the (unseen) grade accompanying feedback.

We chose not to incentivize the written feedback. Since evaluators were already paid for their grading we expected them to take the task seriously. Our results, which we discuss below, along with the overall quality of the written comments, give us confidence that this was indeed the

[8]Participants self-identified their gender. Fewer than 1% selected "Other," rather than "Female" or "Male.", when given the option.

[9]Technically, the evaluators played a so-called Keynesian beauty contest. A plausible first step in this game is to form a judgement and assume that the judgement of others is positively correlated with yours, so that a higher grade belief on your part would lead you to give a higher grade. An evaluator might simply provide their sincere grade, or might engage in intricate higher order reasoning. The particularities of this aspect were unimportant for us; rather, our primary goal was to give evaluators the incentive to properly consider the essays.

[10]Evaluators were reminded of the writer's gender in the screen on which they wrote their feedback, which began with the phrase "Dear [writer's alias]."

case (The feedback example in the introduction is fairly representative of evaluators' feedback).

Part 3: Writers

Writers from Part 1 were invited to return for Part 3. Those in the *No-Feedback* treatment were shown their essay. Those in the *Feedback* treatment were shown their essay along with the written feedback provided by one randomly selected evaluator. In both treatments, writers were then once again asked to report their expected final grade.[11]

3.2. Additional Treatments

This study contains several treatment variations. We summarize them here and provide more details later, in the sections where they are relevant to the analysis. The *Feedback* treatment has three sub-treatments, all randomly assigned: (i) *Feedback-Only*, which follows the exact structure described in the previous section, (ii) *Feedback-Compete*, where writers were given a choice between a lottery payment and the competitive payment scheme after receiving their feedback, (iii) *Feedback-Compete-Hidden*, which mirrors *Feedback-Compete* but without the disclosure of the writers' gender to the evaluators, and (iv) *Feedback-Edit*, where writers could choose whether to edit their essay and have it regraded. Note that writers were not assigned to treatments in Part 1. They learned the details of their treatment when they came back for Part 3.

In parts of our analysis we aggregate the data across the sub-treatments, for example when comparing the sentiment of feedback shared versus not shared with the writer or when examining belief-updating in response to feedback. For our main outcome variables, such as grade beliefs, we find no differences across these sub-treatments (see Tables A1 and A2 in the Appendix). In our analysis we indicate where we have aggregated the data.

3.3. Implementation

We recruited a gender-balanced sample of evaluators and writers using the platform Prolific. Recruitment was open to Prolific participants who were at least 18 years old, were based in the United Kingdom, and had a 96% or higher approval rating. All the studies were conducted in August 2022 and programmed in Qualtrics.

We recruited 900 writers. Of these, we have complete submissions for 847 writers who completed Parts 1 and 3, of which 417 were female and 430 male.[12] The large majority of writers

[11] We did not also ask subjects for their guess of the grade that accompanied the feedback as we did not want to influence their updating procedure.

[12] Of the 53 missing writers, 22 did not return for Part 3, and 31 received invalid feedback. Although evaluators were told not to mention the grade they assigned in their feedback, 31 did. Hence, we drop these observations. Attrition was not significantly different by gender (6.9% for women and 4.9% for men; χ^2 test, $p = 0.75$).

Table 1. Treatment sample sizes

Treatment	Writers Parts 1 & 3	Evaluators Part 2
No-Feedback	98	123
Feedback-Only	184	241
Feedback-Compete	192	421
Feedback-Compete-Hidden	185	436
Feedback-Edit	188	339

Note: Number of writers and evaluators with complete submissions for the various treatments.

identified as white (85%), grew up in the UK (90%), and considered English as their mother tongue (91%). Sample characteristics do not differ by gender or treatment assignment (see Tables B1 and B2 in the Appendix).

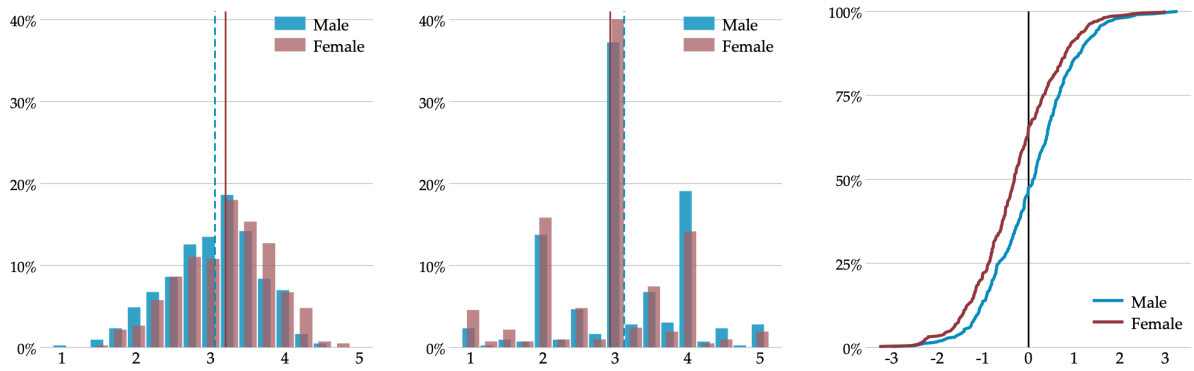
We have 1560 completed submissions from evaluators in Part 2, of which 785 identify as female, 765 as male, and 10 who selected "Other." Similar to the writers, 85% identified as white, 92% grew up in the UK, and 93% considered English as their mother tongue. Again, sample characteristics do not differ by gender or treatment assignment (see Tables B3 and B4 in the Appendix). The overwhelming majority of participants should have been familiar with British English spelling and the stereotyped gender associated with the alias of the writers.[13] Table 1 gives an overview of the sample size of writers and evaluators for each treatment condition at each part of the study.

4. Results

4.1. Final grades and prior beliefs

Figure 1 presents the distribution of final grades, prior grade beliefs, and the cumulative distribution of grade overestimation—defined as the difference between prior beliefs and final grades—for all writers by their gender. Vertical lines indicate group means, with solid lines for female writers and dotted lines for male writers. On average, female writers receive significantly higher final grades than male writers (3.20 vs. 3.06; t-test, $p < 0.01$). Despite their stronger performance, female writers report significantly lower prior grade beliefs than male writers (2.93 vs. 3.12; t-test, $p < 0.01$). As a result, female writers underestimate their grade by an average of

[13]We found that sometimes the computer displayed the essay and feedback texts without the correct spacing between a few words, which might have been perceived as a spelling mistake. We corrected for this in the essays for later participants. Moreover, if we test whether this bug impacted grading, we find no effect (see the subsection on spacing display errors in Section A.2. in the Appendix for details). Nonetheless, we control for it in the subsequent analysis.



(a) Final grade **(b) Prior grade belief** **(c) Grade overestimation**

Figure 1. Distributions of writers' final grades and prior grade beliefs

Note: Panel (a) shows the histograms of the writers' final grade by gender. Panel (b) shows the histograms of the writers' prior grade beliefs by gender. In Panels (a) and (b), means are depicted by the vertical lines, with female writers corresponding to the solid blue line and male writers to the dashed red line. Panel (c) plots the cumulative distribution of grade overestimation: the difference between writers' prior grade beliefs and their final grades. The vertical solid line corresponds to a gap of zero. The sample comprises writers from all treatments ($N = 847$).

0.27 points (t-test, $p < 0.01$), while male writers slightly overestimate theirs by 0.06 points (t-test, $p = 0.16$). Figure 1c shows that this gender gap in grade overestimation is present across the distribution: the distribution of overestimation for male writers first-order stochastically dominates that for female writers (Kolmogorov-Smirnov test, $p < 0.01$). These findings are consistent with previous work on gender differences in overconfidence (see, e.g., Niederle and Vesterlund, 2007; Reuben et al., 2014).

To examine whether revealing a writer's gender influences grading, we compare outcomes across the *Feedback-Compete* and *Feedback-Complete-Hidden* treatments. In both treatments, evaluators graded the same set of essays; the only difference is that writer aliases were shown in the former but not in the latter.[14] We find no evidence that revealing gender affects grading. When aliases are disclosed, the average grade is 0.04 points lower for female writers and 0.07 points lower for male writers—neither difference is statistically significant at the 5% level (see Table C1 in the Appendix).

4.2. Feedback characteristics

Feedback and grades

To update beliefs about their average grade, each writer had to interpret the qualitative feedback they received. Moreover, since this grade was the average of ten evaluators' scores, but feedback

[14] As described in Section 3.1., aliases were disclosed with the phrase "Written by [writer's alias]" when presenting the essay and "Dear [writer's alias]" when prompting the evaluator to write feedback. In *Feedback-Compete-Hidden*, neither phrase was shown. When reading their feedback in Part 3, writers saw a screenshot of what the evaluator saw, including whether their alias was disclosed.

was provided by only one of them, the writer also needed to form expectations about the nine grades for which they did not get feedback.

Suppose a writer successfully infers the grade associated with the feedback they received. What should they infer about the other evaluators' grades? Intuitively, a high grade from one evaluator suggests that the remaining grades are also likely to be high. A strong version of this intuition is that the information follows first-order stochastic dominance. That is, for any grade x, look at all essays that were graded as x by at least one evaluator and plot the distribution of the other grades of those essays. Repeat the procedure for a grade y. If $x > y$ and the distribution associated with grade x first-order stochastically dominates the one associated with grade y, then higher grades from a single evaluator systematically signal higher grades overall.

Figure 2 shows that first-order stochastic dominance holds in our data. This implies that if writers can accurately infer the (unseen) grade that accompanies their feedback, they should update their beliefs about their average grade more positively the higher the accompanying grade.[15] We note that the finding of first-order stochastic dominance reassures us that evaluators approached the grading task seriously, and did not, for instance, assign grades haphazardly.[16]

Feedback sentiment

Since feedback is an important factor in improving performance, how this feedback is communicated is a first-order concern for determining its impact. The emotional tone or valence of feedback has been shown to be an important characteristic when considering performance (Belschak and Den Hartog, 2009; Lechermeier and Fassnacht, 2018).[17] As stated in our preregistration, we focus on the tone of the feedback text and in this section, we apply sentiment analysis—a natural language processing technique—to analyze the emotional tone of the text written by the evaluators. Specifically, we use the OpenAI API for GPT-3.5, a large language model with a neural network architecture that has demonstrated strong performance across a range of human-like tasks, including passing the bar exam (Katz et al., 2024) and constructing

[15] A more positive updating is also what we would intuitively expect, even without a finding of first order stochastic dominance.

[16] We can also assess the grading consistency using the intra-class correlation coefficient. A two-way random effects model yields an average intra-class correlation of 0.80 across essay groups, which is generally considered a high level of inter-rater agreement.

[17] Text data is inherently multi-dimensional, making its analysis complex. There is an emerging literature proposing particular methodologies for extracting meaningful information, such as using human coders to give human interpretable concepts to differences identified through machine learning approaches (Ludwig and Mullainathan, 2024) or using hidden layers of large language models to provide interpretable concepts for relationships between text data and a target variable (Movva et al., 2025). These approaches require a large number of text data and an associated target variable to apply machine-learning approaches to mine for hidden relationships. Although in total we have 1560 feedback texts, per treatment group we do not have enough observations to reliably apply these techniques.

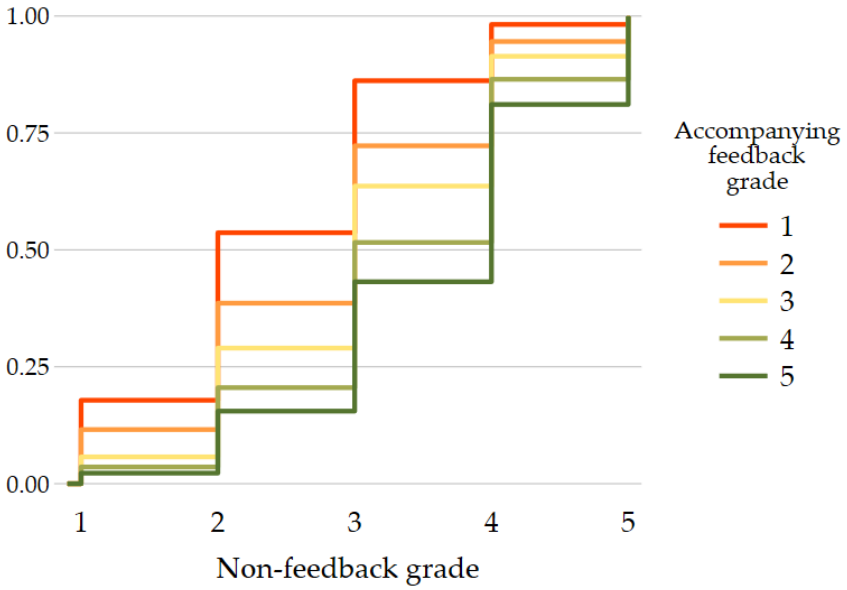


Figure 2. Cumulative distribution of the non-feedback grades depending on the grade accompanying the feedback text

Note: Since a writer's essay was graded by multiple evaluators but only one was selected at random to provide feedback, the figure plots the cumulative distribution of the non-feedback grades depending on the grade accompanying the feedback text. For example, the red line corresponds to writers whose accompanying grade was 1 and plots the distribution of the other remaining grades. The sample comprises essays from all feedback treatments ($N = 749$).

psychological measures (Rathje et al., 2024). For each text, we generate a sentiment score on a continuous scale from -1 (most negative) to $+1$ (most positive), where the score reflects the overall emotional leaning of the writing.[18] See Section D in the Appendix for more details.

As expected, a strong positive relationship exists between the sentiment score of the text written by the evaluators and the grade they assigned to the essay, with a correlation coefficient of 0.62 ($p < 0.01$). This confirms that the sentiment scores are meaningful and provides evidence that evaluators reflected their thoughts about the essay's quality in their writing. As a robustness check, we also replicate the sentiment scoring using Google Natural Language (GNL), which yields qualitatively similar results. Descriptive statistics for the evaluators' writings, such as sentiment and feedback length, are presented in Table C2 in the Appendix.

Next, we examine how the sentiment of the evaluators' writing depends on whether the writer will see it as feedback or not. To visualize the results, we divide feedback into three groups based on the (unseen) grade that accompanies the text: grades of 1 or 2 form the *low* group, grade 3 the *medium* group, and grades of 4 or 5 the *high* group.

Figure 3 illustrates the average GPT sentiment scores of the text written by the evaluators across the three grade groups: *Low*, *Medium*, and *High*.[19] The data is shown separately for

[18]The exact prompt was: "What is the sentiment of this text? Answer with a continuous numerical variable that ranges from minus 1.0 (negative) to plus 1.0 (positive) and corresponds to the overall emotional leaning of the text. Only respond with a continuous numerical variable. Here is the text."

[19]Figure C1 in the Appendix presents box plots that confirm the upward trend in sentiment across grade groups

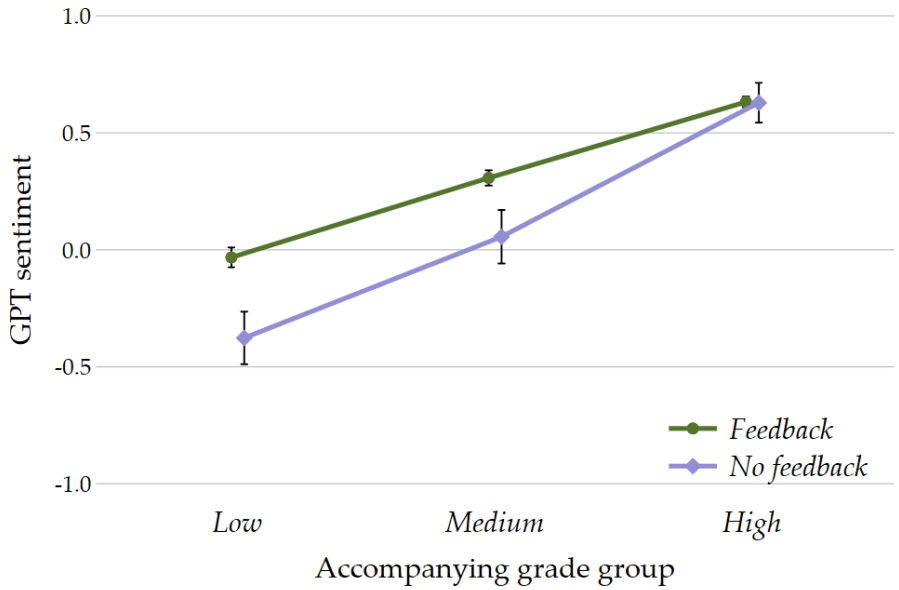


Figure 3. GPT sentiment of the evaluators' text depending on the accompanying grade and whether the text would be shared with writers as feedback

Note: Mean GPT sentiment score of the evaluators' written text depending on the accompanying grade group. The data is shown separately for evaluators who knew that their assessment would be shared with writers (*Feedback*) and those who knew it would not (*No-Feedback*). The accompanying grade groups are: *Low* for grades 1 or 2, *Medium* for grade 3, and *High* for grades 4 or 5. The GPT sentiment score ranges from −1 (negative sentiment) to +1 (positive sentiment). Error bars indicate 95% confidence intervals. The sample consists of evaluators from all treatments ($N = 1437$ for *Feedback* and $N = 123$ for *No-Feedback*).

evaluators who knew that their assessment would be shared with the writers (*Feedback*) and those who knew it would not (*No-Feedback*)

The figure reveals a clear *kindness effect*: evaluators are more positive when the writer will see their comments. The effect is most pronounced in the *Low* grade group. In *No-Feedback*, the average sentiment score is approximately -0.4; in *Feedback*, the average rises to around 0.0. Alternatively, the sentiment score associated with a low-grade essay in *Feedback* is as positive as the sentiment score associated with a medium-grade essay in *No-Feedback*. A similar, though smaller, effect is observed in the medium-grade group. The kindness effect disappears in the high-grade group, where sentiment scores are similarly high across treatments, suggesting that evaluators felt no need to soften their remarks for top-performing essays. These results are robust to the alternative sentiment scoring using Google Natural Language (see Figure C2 in the Appendix).

Table 2 presents linear regressions of the GPT sentiment score of the evaluators' text depending on the treatment, the accompanying grade, and the gender of the writer. To facilitate interpretation of the coefficients, we standardized the sentiment scores and the accompanying grades. Column (1) shows that, controlling for the accompanying grade, text that is not shown to the writer is, on average, 0.36 standard deviations less positive than feedback that is shared

while illustrating the variation within groups.

($p < 0.01$). Column (2) includes an interaction between the *No-Feedback* treatment and the accompanying grade. At the mean grade, sentiment is 0.38 standard deviations less positive when it is not shared. However, for each one-standard-deviation increase in the accompanying grade, the kindness effect diminishes by 0.33 standard deviations. In other words, the difference in sentiment between *Feedback* and *No-Feedback* narrows to just 0.05 standard deviations at grades one standard deviation above the mean but grows to 0.71 standard deviations at grades one standard deviation below the mean.[20] These results are robust to using GNL sentiment scores (see Table C3 in the Appendix). Result 1 summarizes these findings.

Result 1 *For a given grade, evaluators write systematically more positive comments when they know their remarks will be shared with the writer as feedback. This effect diminishes as the grade increases and effectively disappears for the highest grade essays.*

Does the kindness effect vary by the gender of the writer? To investigate this, we focus on treatments in which the writer's alias—and thus their gender—was disclosed to evaluators. Figure 4 plots the average GPT sentiment scores by the writers' gender. The kindness effect is present for both female and male writers. This pattern is also evident when sentiment is measured using the alternative GNL score (see Figure C3 and Table C3 in the Appendix).

In Table 2, we use linear regressions of the GPT sentiment score of the evaluators' text to evaluate whether the kindness effect varies with the writers' gender. Columns (3) and (4) replicate the specifications from columns (1) and (2) but include interactions with the writers' gender. We find no evidence of a significant gender difference in the overall sentiment or the impact of the *No-Feedback* treatment. In column (5), we further control for a range of evaluator and essay characteristics, including the GPT sentiment of the essay itself.[21] The sentiment of the essay is included to check whether the tone used by evaluators reflects the tone of the essay. The results are robust to the inclusion of these controls.

We further test whether the writer's gender affects feedback by comparing sentiment scores for the same essays, depending on whether the writer's gender was disclosed. As noted earlier, in *Feedback-Compete*, evaluators saw the writer's alias, whereas in *Feedback-Compete-Hidden*, the same essays were presented without any gender-identifying information. Table 3 reports linear regressions of the GPT sentiment score on treatment indicators, the accompanying grade, and

[20] Also, the sentiment of feedback does not vary by if it is instrumental to decision-making e.g. the recipient can take an action based on it, results are available on request. This is in contrast to work by Coutts et al. (2026), although they use a quantitative feedback environment. This underlines the need to study qualitative feedback environments.

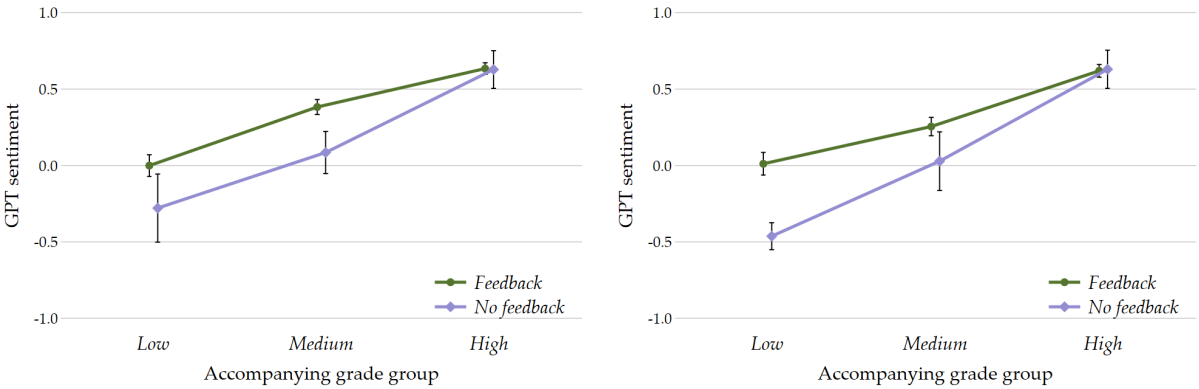
[21] The evaluator controls include their age, ethnic identity, gender, level of education, whether English is their native language, whether they grew up in the UK, the number of characters in the feedback, and their treatment assignment. The essay controls include the number of characters, whether there were spacing display errors, and the GPT sentiment of the essay. See Appendix B and C descriptive statistics and more details of these variables.

Table 2. GPT sentiment of the evaluators' text

	(1)	(2)	(3)	(4)	(5)
Constant	0.03	0.03	0.02	0.02	0.02
	(0.02)	(0.02)	(0.04)	(0.04)	(0.05)
Accompanying grade	0.63**	0.60**	0.60**	0.56**	0.54**
	(0.02)	(0.02)	(0.02)	(0.04)	(0.04)
No-Feedback	−0.36**	−0.38**	−0.41**	−0.42**	−0.46**
	(0.07)	(0.07)	(0.11)	(0.10)	(0.10)
No-Feedback × Accompanying grade		0.33**		0.40**	0.39**
		(0.06)		(0.07)	(0.07)
Female			0.08	0.08	0.05
			(0.05)	(0.05)	(0.05)
No-Feedback × Female			0.05	0.04	0.07
			(0.15)	(0.14)	(0.14)
Accompanying grade × Female				0.00	0.01
				(0.05)	(0.05)
No-Feedback × Accompanying grade				−0.06	−0.08
× Female				(0.12)	(0.12)
Essay GPT sentiment					0.02
					(0.02)
Controls	-	-	-	-	✓
N	1560	1560	1124	1124	1122
adj. R^2	0.399	0.406	0.377	0.389	0.402

Note: Linear regressions of the GPT sentiment score of the evaluators' text as the dependent variable. *No-Feedback* is a dummy variable indicating the evaluator's comments would not be shared with the writer. Female is a dummy variable indicating the writer was female. The accompanying grade is the grade assigned by the evaluator who wrote the comments. Essay GPT sentiment is the GPT sentiment score of the essay's text. Columns (1) and (2) utilize the entire sample of evaluators. In columns (3)-(5), observations from the *Feedback-Compete-Hidden* treatment were dropped since gender was not disclosed to the evaluators. In column (5), two observations were dropped as the GPT sentiment score of the essay returned a non-numeric value. All continuous variables—the GPT sentiment score, the accompanying grade, and the essay GPT sentiment score—are standardized to have a mean of zero and a standard deviation of one. Controls include the evaluators' age, ethnic identity, gender, level of education, whether English is their native language, whether they grew up in the UK, their treatment assignment, the presence of spacing display errors in the essay, the number of characters in the feedback and the number of characters in the essay. Robust standard errors in parentheses and statistical significance of non-zero coefficients is indicated by * $p < 0.05$ and ** $p < 0.01$.

their interactions with the writer's gender. Because multiple evaluators assessed the same essays across treatments, we can include essay fixed effects to control for idiosyncratic essay characteristics. Once again, we standardized the sentiment scores and the accompanying grades. Column (2) further controls for evaluator characteristics (see footnote 21). We find no statistically significant differences in the sentiment of feedback when the writer's gender is disclosed, neither for male nor for female writers. This result also holds when sentiment is measured using the alternative GNL sentiment score (see Table C4 in the Appendix). The next result summarizes



(a) Female writers **(b) Male writers**

Figure 4. GPT sentiment of the evaluators' text depending on the writers' gender, the accompanying grade, and whether the text would be shared with writers as feedback

Note: Mean GPT sentiment score of the evaluators' written text depending on the accompanying grade group and the writers' gender. The data is shown separately for evaluators who knew that their assessment would be shared with writers (*Feedback*) and those who knew it would not (*No-Feedback*). The accompanying grade groups are: *Low* for grades 1 or 2, *Medium* for grade 3, and *High* for grades 4 or 5. The GPT sentiment score ranges from -1 (negative sentiment) to $+1$ (positive sentiment). Error bars indicate 95% confidence intervals. The sample consists of evaluators from all treatments where writer aliases were disclosed ($N = 1001$ for *Feedback* and $N = 123$ for *No-Feedback*).

these findings.

Result 2 *There is no difference in the positivity of feedback given to female and male writers. For the same essay, feedback sentiment does not vary with the disclosure of the writer's gender.*

Our finding of no gender difference in the sentiment of feedback contrasts with some studies in psychology that report a female positivity bias. Jampol and Zayas (2020) find that women receive more positive feedback than men for the same essay. In their design, the gender of (supposed) writers is revealed through a name, just as in our study. However, their evaluators believed they had engaged in a live, back-and-forth chat with the writer, whereas our evaluators provided feedback to writers without interacting with them and with some time delay.[22] It is possible that gender differences in feedback are attenuated when it is not delivered 'in the moment.' Jampol et al. (2022) report a gender bias in the sentiment of 360-degree feedback received by MBA students from their former colleagues. In that setting, evaluators had prior relationships with the recipients of the feedback. In contrast, our evaluators and writers were anonymous to each other. The absence of preexisting relationships may account for the gender-based differences in our study. These potential explanations are speculative, however, as our experiment was not designed to test them.

[22] As is common in many studies in psychology, Jampol and Zayas (2020) use deception. They do not use human participants as writers; instead, evaluators believe a man or a woman wrote the essay and their live-chat is with this person, when actually the writer's responses have been pre-programmed.

Table 3. GPT sentiment depending on whether the writers' gender is disclosed

	(1)	(2)
Constant	0.04	0.03
	(0.04)	(0.04)
Feedback-Compete-Hidden	−0.06	−0.03
	(0.08)	(0.08)
Accompanying grade	0.61**	0.58**
	(0.06)	(0.06)
Feedback-Compete-Hidden × Female	−0.02	−0.07
	(0.12)	(0.12)
Accompanying grade × Female	−0.08	−0.08
	(0.08)	(0.08)
Essay fixed effects	✓	✓
Controls	-	✓
N	857	857
adj. R^2	0.462	0.469

Note: Linear regressions of the GPT sentiment score of the feedback text as the dependent variable in treatments *Feedback-Compete* and *Feedback-Compete-Hidden*. *Feedback-Compete-Hidden* is a dummy variable indicating the writer's gender was not disclosed to the evaluator. The accompanying grade is the grade assigned by the evaluator who wrote the feedback. Female is a dummy variable indicating the writer was female. Since the same essays were used across treatments, we control for essay characteristics by including essay fixed effects. All continuous variables—the GPT sentiment score and the accompanying grade—are standardized to have a mean of zero and a standard deviation of one. Controls include the evaluators' age, ethnic identity, gender, level of education, whether English is their native language, the number of characters in the feedback, and whether they grew up in the UK. Robust standard errors in parentheses and statistical significance of non-zero coefficients is indicated by * $p < 0.05$ and ** $p < 0.01$.

In summary, we find that sentiment scores capture a meaningful component of feedback. There is a clear kindness effect whereby evaluators are more positive when they know their comments will be seen by the writer. We find no evidence of gender bias in the sentiment of feedback. This result can be viewed as identifying boundary conditions for such a bias or as casting doubt on its existence. In the next section, we examine whether the kindness effect undermines a writer's ability to interpret the feedback they receive accurately.

Feedback and beliefs

Is qualitative feedback effective? Do writers understand the feedback they receive and properly incorporate it into their beliefs? To address these questions, we first present a Bayesian model of belief-updating.

A Simple Updating Model. A writer composes an essay which is graded by ten evaluators, each of whom assigns a number grade $g \in \{1, 2, 3, 4, 5\}$. An evaluator's grade is an i.i.d. draw

from a probability distribution θ over $\{1, 2, 3, 4, 5\}$. We can think of θ as describing the quality of the essay. For example, $\theta = (0.07, 0.08, 0.15, 0.60, 0.10)$ indicates a high-quality essay that will most likely be given a grade of 4 but with elements that could result in a grade of 1 with a 7% chance, 2 with an 8% chance, and so forth. Alternatively, the grade dispersion could be due to idiosyncrasies of the graders or errors on their part.

The writer is uncertain of the quality of their essay and has a prior belief π over possible θ's. A standard approach in a setting like this is to model the writer's prior π as a Dirichilet distribution. In this instance, the Dirichilet distribution is characterized by a five-dimensional vector $x \in \mathbf{R}^5_+$ with the feature that the mean belief is given by

$$\mathbb{E}_{\pi(x)}(\theta) = \left(\frac{x_1}{\sum_{i=1}^5 x_i}, \cdots, \frac{x_5}{\sum_{i=1}^5 x_i} \right).$$

That is, the expected value of θ is a multinomial distribution in which the probability of observing a draw of j is $x_j / \sum x_i$. The writer's initial expectation of their average grade is

$$\mathbb{E}\left(\frac{1}{10} \sum_{j=1}^{10} g_j \right) = \frac{\sum_{i=1}^5 i x_i}{\sum_{i=1}^5 x_i} \equiv z.$$

The writer receives written feedback from, say, the first evaluator. We suppose that the writer updates their prior by proceeding in two steps. First, they "decode" the feedback to decide what grade accompanies the feedback; then they update based on that grade.

Suppose the writer determines that the accompanying grade was $g_1 = k$. The Dirichilet has the property that, following a draw of k, the posterior $\pi(x|k)$ is characterized by the vector

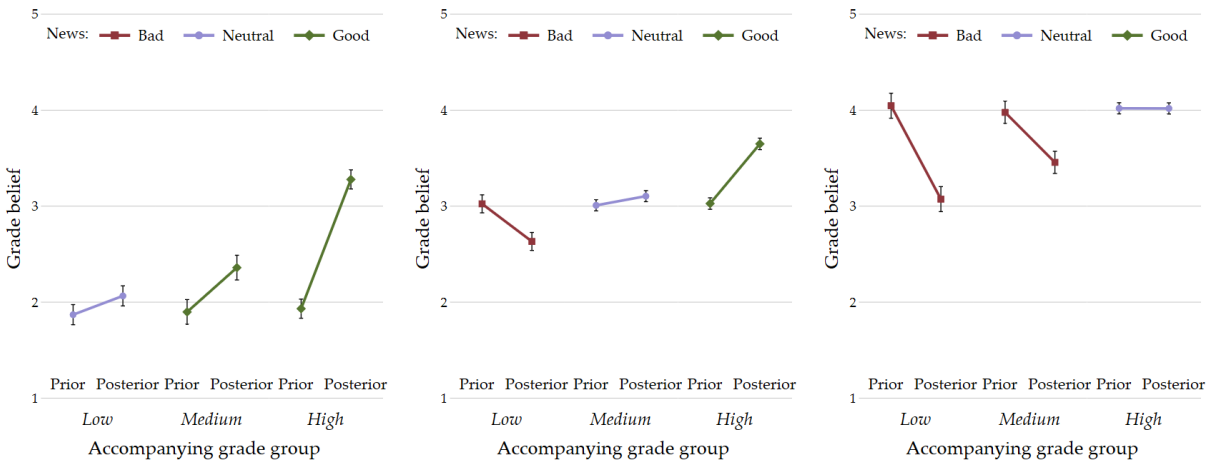
$$x|k = (x_1, \cdots, x_k + 1, \cdots, x_5).$$

Hence, the writer's updated expected belief of their average grade is now

$$\mathbb{E}\left(\frac{1}{10} \left(k + \sum_{j=2}^{10} g_j \right) \right) = \frac{1}{10} \left(k + 9 \frac{x_1 + \cdots + k(x_k + 1) + \cdots + 5x_5}{1 + \sum_{i=1}^5 x_i} \right).$$

It is easy to verify that the writer updates their expected belief upward if and only if $k > z$. Moreover, if $(k' - z) > (k - z) > 0$, the writer's updated mean belief is larger following k' than k (and similarly if the inequalities are reversed).

Thus, our test of whether subjects properly understand and incorporate the qualitative feedback is whether i) they revise their expectations upward (downward) when the actual grade accompanying the feedback is larger (smaller) than their prior and ii) the greater (lesser) the accompanying grade is, the more they revise.



(a) Low prior beliefs **(b) Medium prior beliefs** **(c) High prior beliefs**

Figure 5. Mean prior and posterior beliefs depending on the (unseen) accompanying grade

Note: Writers' mean prior and posterior beliefs depending on the accompanying grade group: *Low* (grades 1 and 2), *Medium* (grade 3), and *High* (grades 4 and 5). Panel (a) shows the beliefs for writers in the *Low* prior-belief group (priors in the range $[1, 2.5]$), Panel (b) for those in the *Medium* prior-belief group (priors in the range $(2.5, 3.5)$), and Panel (c) for those in the *High* prior-belief group (priors in the range $[3.5, 5]$). Beliefs are labeled as good news (in green) if the accompanying grade group is above the prior-belief group, as bad news (in red) if it is below, and as neutral news (in lavender) if it is equal. Error bars indicate 95% Cousineau-Morey confidence intervals calculated with within-subject variability (Cousineau, 2005; Morey, 2008). The sample consists of writers in the *Feedback-Only*, *Feedback-Compete*, and *Feedback-Compete-Hidden* treatments ($N = 561$).

The Data. We now consider our findings. We analyze within-subject belief-updating for the 561 writers who reported their expected grade before receiving feedback (Part 1) and again afterward (Part 3). This corresponds to writers in the *Feedback-Only*, *Feedback-Compete*, and *Feedback-Compete-Hidden* treatments.

To visualize how beliefs are updated, we divide writers into three prior-belief groups: *Low* for prior beliefs in the range $[1, 2.5]$, *Medium* for beliefs in the range $(2.5, 3.5)$, and *High* for beliefs in the range $[3.5, 5]$, mirroring the three accompanying grade groups used in the sentiment analysis of Section 4.2.. We refer to the grade accompanying the feedback as good news if it is in a grade group above the prior-belief group, as bad news if it is below, and as neutral news if it is in the same grade group. If writers interpret the feedback correctly, we expect them to revise their beliefs upward in response to good news and downward in response to bad news, with larger adjustments the greater the gap between the grade and the prior.

For each prior-belief group, Figure 5 depicts the writers' mean prior and posterior beliefs depending on whether the feedback's accompanying grade was *Low*, *Medium*, or *High*. Across all groups, writers revise beliefs upward in response to good news, downward in response to bad news, and make little adjustment to neutral news. Moreover, they adjust more the larger the gap. This can be seen in Figure 5a, where writers whose accompanying grade was *High* adjusted upwards more than those with the lower accompanying grade of *Medium*.[23] If we

[23]These patterns are robust to changes in the groups' cutoff values. Figure C4 in the Appendix plots individual-

inspect the sign of the belief updates, we find that in response to good news the majority of updates are positive and vice versa for bad news (see Table C5). This suggests that writers are able to see through the kindness effect identified in the feedback text (Result 1) and correctly infer its informational content. For instance, writers with medium priors who receive a low accompanying grade revise their beliefs downward, even though the sentiment of the feedback matches that of unshared evaluations for medium-grade essays.

To more formally assess whether writers correctly interpret qualitative feedback, we estimate a linear regression in which the dependent variable is the writer's posterior grade belief (i.e., their expected grade after receiving feedback). As independent variables, we use the writer's prior grade belief and their grade-prior gap, defined as the difference between the grade accompanying the writer's feedback and their prior grade belief.[24] Note that a positive grade-prior gap indicates good news, and a negative one indicates bad news. If writers correctly distinguish good from bad news and update their beliefs in the correct direction, the coefficient on the grade-prior gap should be positive, reflecting that writers increase (decrease) their belief when receiving good (bad) news and this increase (decrease) depends on the size of the gap. In addition, if writers correctly recognize when they receive neutral news, meaning the accompanying grade matches their prior belief, then the coefficient on the prior grade belief should equal 1, indicating that their belief remains unchanged in the absence of good or bad information.

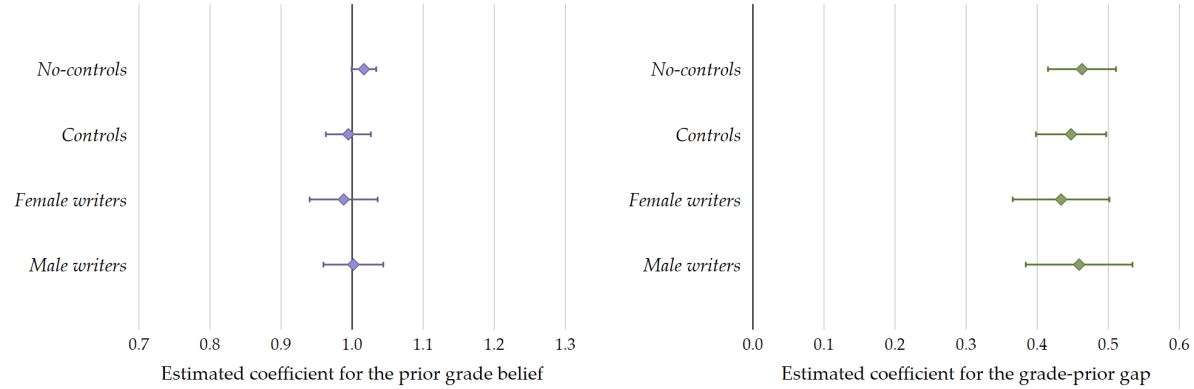
Figure 6 displays the estimated coefficients, while Table C6 in the Appendix contains the regressions' output. Coefficients labeled as *No-controls* in the figure correspond to the regression described above (column (1) in Table C6). Coefficients labeled as *Controls* correspond to regressions that control for a range of writer, feedback, and essay characteristics[25] (column (2) in Table C6). In both regressions, the coefficient of the grade-prior gap is positive and statistically significant (around 0.45; $p < 0.01$), indicating writers correctly incorporate good and bad news into their beliefs and update more the larger the gap, while the coefficient of the prior grade belief is very close to 1, consistent with beliefs being unaffected by neutral news.

So far, our analysis has assumed that writers respond symmetrically to good and bad news. However, prior research using quantitative feedback suggests that this may not always be the case. Some studies report positive asymmetries, where individuals respond more strongly to

level prior and posterior beliefs, showing that the belief-updating pattern holds quite generally.

[24] Specifically, we estimate the regression $\mu_i^1 = \beta_1 \mu_i^0 + \beta_2(g_i - \mu_i^0) + \gamma X_i + \epsilon_i$, where μ_i^1 denotes writer i's posterior grade belief, μ_i^0 their prior grade belief, g_i the grade accompanying their feedback, and X_i is the vector of controls. Hence, $\beta_1 = 1$ implies the posterior belief equals the prior when news is neutral (i.e., $g_i = \mu_i^0$), and β_2 captures how strongly writers update their belief about their final grade after receiving feedback with an accompanying grade of g_i.

[25] The writer controls include their age, ethnic identity, gender, level of education, whether English is their native language, whether they grew up in the UK, and their treatment assignment. The feedback controls include the number of characters and if there was a spacing display error. The essay controls include if there was a spacing display error. See Appendix B and C descriptive statistics and more details of these variables. Note that the controls include the same as the evaluator ones in footnote 21 where relevant.



(a) Prior grade belief **(b) Grade-prior gap**

Figure 6. Grade belief-updating depending on prior beliefs and the grade-prior gap (accompanying grade − prior grade belief)

Note: Estimated coefficients from linear regressions of the writer's posterior grade belief as the dependent variable. Panel (a) plots the estimated coefficient of the first dependent variable: the writers' prior grade belief. Panel (b) plots the estimated coefficient of the second dependent variable: the grade-prior gap (i.e., the difference between the feedback's accompanying grade and the prior grade belief). The precise specification is described in footnote 24. *No-controls* corresponds to the regression without additional variables. *Controls* further controls for essay, feedback, and writer characteristics (see footnote 25). *Female writers* restricts the sample to only female writers and *Male writers* to only male writers. Table C6 contains the regression results. Error bars indicate 95% confidence intervals calculated with robust standard errors. The sample consists of writers in the *Feedback-Only*, *Feedback-Compete*, and *Feedback-Compete-Hidden* treatments ($N = 561$).

good news than to bad news (Eil and Rao, 2011; Zimmermann, 2020; Möbius et al., 2022), while others report negative asymmetries, with stronger responses to bad news (Ertac, 2011).[26] In Table C7 in the Appendix, we test for such asymmetries. We find that, directionally, writers update more in response to good news than to bad news, but the difference is not statistically significant.

We now turn to whether women and men differ in how they update their beliefs in response to qualitative feedback. Figure 6 depicts the estimated belief-updating coefficients for female and male writers (for corresponding regressions see columns (3) and (4) of Table C6). As the figure indicates, there are no substantial gender differences. The effect of prior grade belief is nearly identical for both genders (0.99 for women vs. 1.00 for men; Wald test, $p = 0.83$), and the response to the feedback signal—the gap between the feedback's accompanying grade and the prior—is also statistically indistinguishable (0.43 for women vs. 0.46 for men; Wald test, $p = 0.35$). This finding contrasts with previous work that reports gender differences in belief-updating (Ertac, 2011; Möbius et al., 2022). However, these studies differ from ours in both the type of feedback (quantitative vs. qualitative) and the nature of the task (solving word puzzles vs. writing an essay). It remains an open question whether the feedback format or the task domain drives these differences. Result 3 summarizes our findings.

[26] Barron (2021) finds no evidence of asymmetric belief updating and Zimmermann (2020) shows that with substantially large financial incentives this positive asymmetry can be eliminated.

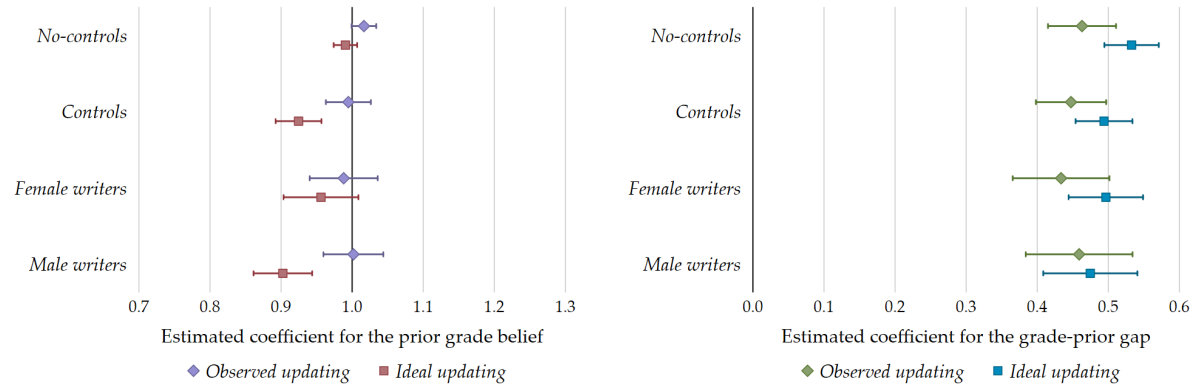
Result 3 *Upon receiving feedback, writers revise their beliefs in the appropriate direction: upward in response to good news, downward in response to bad news, and minimally in response to neutral news. The magnitude of belief revision increases with the size of the gap between the feedback-giver's evaluation and the writer's prior belief. We find no gender differences in belief-updating.*

We have thus far examined how writers update their beliefs in response to feedback, focusing on both the direction and the magnitude of their revisions. We now study how these updates compare to an ideal benchmark: adjusting one's belief to match the final grade exactly. To do this, we re-estimate the regressions reported earlier (see footnote 24), but use the writer's final grade as the dependent variable instead of their posterior belief. Figure 7 summarizes the results (the regression estimates are reported in Table C6 of the Appendix). To facilitate comparisons, we include the coefficients reported in Figure 6 for actual belief-updating, labeled as *Observed updating* in Figure 7. The newly estimated coefficients are labeled as *Ideal updating* and represent the average belief adjustment required for writers to match their final grade. Panel (a) shows the estimated coefficients on the writer's prior grade belief, while Panel (b) shows the coefficients for the grade-prior gap. As before, *No-controls* refers to regressions without additional covariates; *Controls* includes controls for essay, feedback, and writer characteristics; and *Female writers* and *Male writers* indicate regressions run separately by gender.

In the *No-controls* specification, the coefficient on the grade-prior gap is significantly smaller for observed updating than for ideal updating (0.46 vs. 0.53; Wald test, $p = 0.02$).[27] This suggests that writers underreact to the feedback they receive, revising their beliefs less than would be needed to match their final grade. This result mirrors previous findings of conservative belief-updating in response to quantitative feedback (Eil and Rao, 2011; Möbius et al., 2022). When we include controls, the gap between observed and ideal updating decreases but remains in the same direction and is close to statistical significance (0.45 vs. 0.49; Wald test, $p = 0.10$). In contrast, the coefficient on prior grade belief tends to be larger for observed updating than for ideal updating, especially with controls (0.99 vs. 0.92; Wald test, $p < 0.01$), suggesting that writers overestimate their performance such that when they receive neutral news, their ideal response would involve a slight downward revision.

Figure 7 further breaks down these findings by gender. Female writers react to feedback slightly under their ideal updating, though the difference is not statistically significant (0.43 vs. 0.50; Wald test, $p = 0.08$). At the same time, they place the appropriate weight on their prior grade beliefs (0.99 vs. 0.96; Wald test, $p = 0.30$). In contrast, while male writers' response to feedback is close to ideal updating (0.46 vs. 0.47; Wald test, $p = 0.71$), they place too much weight on their prior grade beliefs than is ideal (1.00 vs. 0.90; Wald test, $p < 0.01$),

[27] To compare coefficients across regressions, we use seemingly unrelated estimation to combine parameter estimates into a single vector and compute a joint (co)variance matrix (White, 1994).



(a) Prior grade belief

(b) Grade-prior gap (accompanying grade – prior grade belief)

Figure 7. Ideal and observed belief-updating depending on prior beliefs and the grade-prior gap (accompanying grade – prior grade belief)

Note: Estimated coefficients from linear regressions. In the regressions labeled as *Observed updating*, the dependent variable is the writer's posterior grade belief (also seen in Figure 6). In the regressions labeled as *Ideal updating*, the dependent variable is the writer's final grade. Panel (a) plots the estimated coefficient of the first dependent variable: the writers' prior grade belief. Panel (b) plots the estimated coefficient of the second dependent variable: the grade-prior gap (i.e., the difference between the feedback's accompanying grade and the prior grade belief). The precise specification is described in footnote 24. *No-controls* corresponds to the regression without additional variables. *Controls* further controls for essay, feedback, and writer characteristics (see footnote 25). *Female writers* restricts the sample to only female writers and *Male writers* to only male writers. Table C6 contains the regression results. Error bars indicate 95% confidence intervals calculated with robust standard errors. The sample consists of writers in the *Feedback-Only*, *Feedback-Compete*, and *Feedback-Compete-Hidden* treatments ($N = 561$).

consistent with men being more overconfident than women. These results suggest that while belief-updating is directionally appropriate, there are systematic deviations from the ideal, which differ slightly by gender. The evidence on belief-updating presented here is summarized in the following result.

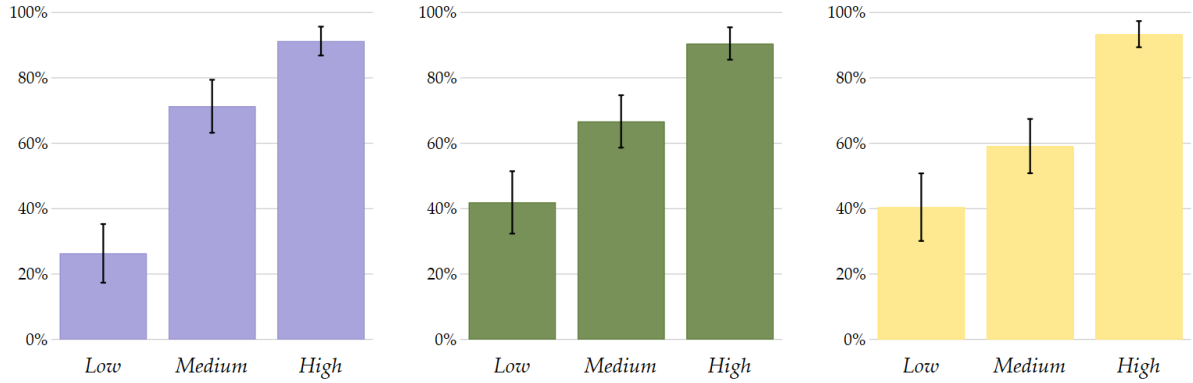
Result 4 *When updating their grade beliefs, writers deviate from the ideal response. Overall, they slightly overweigh their prior grade beliefs and underreact to feedback. Male writers tend to place excessive weight on their priors.*

4.3. Feedback and choices

We have shown that qualitative feedback shapes writers' grade beliefs. We next study how qualitative feedback influences decision-making and how it is used by writers.

Competition

We begin by analyzing whether feedback affects the decision to compete. As described in Section 3.1., writers in Part 1 were informed that their final grade would be compared to those of nine other randomly selected writers and they would earn a £4 bonus if their essay ranked in



(a) Posterior grade belief **(b) Accompanying grade** **(c) GPT sentiment**

Figure 8. Percentage of writers choosing to compete depending on their posterior grade beliefs, the grade accompanying their feedback, and their feedback's sentiment score

Note: Percentage of writers who chose the competitive payment scheme. Panel (a) shows the proportion competing depending on the writers' posterior grade beliefs, where *Low* corresponds to beliefs in the range $[1, 2.5]$, *Medium* to beliefs in the range $(2.5, 3.5)$), and *High* to beliefs in the range $[3.5, 5]$. Panel (b) shows the proportion competing depending on the (unseen) grade accompanying the feedback, where *Low* corresponds to grades 1 and 2, *Medium* to grade 3, and *High* to grades 4 and 5. Panel (c) shows the proportion competing depending on the feedback's GPT sentiment score, where *Low* corresponds to the lowest tercile, *Medium* to the middle tercile, and *High* to the highest tercile. Error bars indicate 95% confidence intervals. The sample consists of writers in the *Feedback-Compete* and *Feedback-Compete-Hidden* treatments ($N = 377$).

the top three (with ties broken randomly) and £1 otherwise. In treatments *Feedback-Compete* and *Feedback-Compete-Hidden*, when writers returned for Part 3, they were given a choice: stay with the competitive bonus scheme or opt for a lottery that paid £4 with a 30% chance and £1 otherwise. (Evaluators in these treatments were informed that writers would be given this choice before providing feedback.) A natural hypothesis is that writers who receive more positive feedback, resulting in higher updated grade beliefs, should be more likely to stick with the competitive payment scheme.

We begin by visualizing how beliefs and different aspects of the feedback relate to a writer's decision to compete. Figure 8 presents the proportion of writers who chose the competitive payment scheme depending on: (a) their posterior grade belief, grouped as *Low* ($[1, 2.5]$), *Medium* ($(2.5, 3.5)$), or *High* ($[3.5, 5]$); (b) the (unseen) grade accompanying their feedback, grouped as *Low* (grades 1 and 2), *Medium* (grade 3), or *High* (grades 4 and 5), and (c) the feedback's GPT sentiment score divided into the lowest, middle, or highest tercile. In all three cases, we see a clear positive relationship with the decision to compete. This shows that feedback influences not only beliefs but also consequential behavior.

Table 4 formally analyzes these patterns using linear probability models. The dependent variable is a binary indicator equal to one if the writer chose the competitive payment scheme. The results confirm the three relationships shown in Figure 8. In column (1), a one standard deviation increase in the posterior grade belief is associated with a 25 percentage point increase

Table 4. Effects of feedback on the choice to compete

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	0.68**	0.68**	0.68**	0.68**	0.68**	0.68**	0.68***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Posterior grade belief	0.25**			0.20**	0.19**	0.21**	0.19**
	(0.02)			(0.02)	(0.03)	(0.02)	(0.02)
Accompanying grade		0.19**		0.09**		0.07**	
		(0.02)		(0.02)		(0.03)	
GPT sentiment			0.21**		0.11**		0.10***
			(0.02)		(0.02)		(0.02)
Final grade						0.02	0.02
						(0.03)	(0.02)
Controls	-	-	-	-	-	✓	✓
N	377	377	377	377	377	377	377
adj. R^2	0.279	0.157	0.192	0.306	0.320	0.312	0.327

Note: Linear regressions where the dependent variable equals one if the writer chose the competitive payment scheme and zero otherwise. The posterior grade belief corresponds to writers' expected final grade after receiving feedback. The accompanying grade is the grade assigned by the evaluator who wrote the feedback. GPT sentiment is the GPT sentiment score of the feedback's text. Final grade is the average grade given to the writer by all evaluators. All dependent variables are standardized to have a mean of zero and a standard deviation of one. Controls include the writers' age, ethnic identity, gender, level of education, whether English is their native language, whether they grew up in the UK, their treatment assignment, the presence of spacing display errors in both the essay and feedback, and the number of characters in the feedback. The sample consists of writers in the *Feedback-Compete* and *Feedback-Compete-Hidden* treatments. Robust standard errors in parentheses and statistical significance of non-zero coefficients is indicated by * $p < 0.05$ and ** $p < 0.01$.

in the likelihood of competing.[28] Similarly, column (2) shows that a one standard deviation increase in the (unseen) grade accompanying the feedback increases the likelihood of competing by 19 percentage points, and column (3) shows that the same increase in feedback sentiment raises the likelihood by 21 percentage points.

Does feedback influence the choice to compete solely through its effect on beliefs? Columns (4) and (5) include both the posterior grade belief and one of the two feedback variables to isolate distinct channels through which feedback may operate. In both cases, the posterior belief and the feedback variable remain positive and statistically significant.

Since feedback is endogenous, in that it is written in response to an essay, it is possible that better writers are inherently more competitive, hold high beliefs, and write essays that elicit feedback with higher sentiment scores. To address this, columns (6) and (7) include the writers' final grade as a proxy for their ability, along with additional controls for essay, feedback, and writer characteristics (see footnote 25). Even with these controls, both the posterior grade belief and the feedback variables remain positive and statistically significant. In the Appendix, we

[28] When we split the posterior into the prior and the change in the grade belief (posterior – prior), we find positive and significant effects for both the change in the grade belief and the prior grade belief, demonstrating that feedback matters for this choice, not just priors.

show that the results are robust to using the alternative GNL sentiment score (Table C8) and that the encouragement effect persists when we flexibly control for posterior beliefs, suggesting that it is not simply the result of model misspecification or measurement error in the elicitation of posterior grade beliefs (Table C9).

These findings suggest that feedback affects the choice to compete through two distinct channels. The first is a *belief channel*, in which writers incorporate information from the feedback into their grade expectations, thereby informing their decision to compete. The second we call an *encouragement channel*, in which the tone or content of the feedback motivates writers to compete beyond their impact on beliefs. We believe this interpretation is conceptually plausible: qualitative feedback may provide encouragement, express confidence, or convey interpersonal warmth in ways that influence writers' motivation independently of their updated beliefs. While this encouragement effect is not randomly assigned and thus should be interpreted with caution, the fact that the sentiment of feedback remains predictive of competition decisions after controlling for posterior beliefs supports the idea that qualitative aspects of communication can shape behavior through both motivational and informational pathways. This interpretation is consistent with a broader literature showing that how information is conveyed through its tone, framing, or interpersonal valence can shape behavior in ways not captured by belief-updating (see Kamenica, 2012).

Broken down by gender, 74.3% of women and 62.6% of men chose to compete (χ^2 test, $p = 0.01$). This pattern contrasts with the common finding that women are less likely to compete than men (Niederle and Vesterlund, 2011). However, prior work has shown that the gender gap in willingness to compete is context-dependent and can be attenuated or even reversed when tasks are perceived as stereotypically female (Dreber et al., 2011; Cárdenas et al., 2012; Dreber et al., 2014; Grosse et al., 2014; Apicella and Dreber, 2015; Flory et al., 2015). Consistent with this explanation, in our setting both female and male participants correctly expect female writers to perform better than male writers (see Figure C5 in the Appendix). Another aspect of our study is that writers are in the competitive payment scheme by default, which has been shown to reduce the gender gap in competition (Erkal et al., 2022).

Table 5 examines whether the feedback's belief and encouragement channels differ by gender. We estimate linear probability models, using the decision to compete as the outcome variable. In column (1), we include the posterior grade belief, which we interact with either a Female or a Male dummy. In columns (2) and (3), we introduce the encouragement channel by including either the accompanying grade or the GPT sentiment score. Lastly, columns (4) and (5) add controls for the writer's final grade, as well as essay, feedback, and writer characteristics.

We find that feedback affects the competitive choices of male and female writers through both the belief and encouragement channels. However, there are noticeable gender differences. Female writers' decision to compete is less sensitive to their posterior grade beliefs than male

Table 5. Effects of feedback on the choice to compete by gender

	(1)	(2)	(3)	(4)	(5)
Constant	0.61**	0.62**	0.63**	0.61**	0.62**
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Female	0.14**	0.13**	0.10*	0.14**	0.11**
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Posterior grade belief × Female	0.18**	0.13**	0.12**	0.14**	0.13**
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)
Posterior grade belief × Male	0.31**	0.28**	0.27**	0.28**	0.27**
	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)
Accompanying grade × Female		0.11**		0.10**	
		(0.03)		(0.03)	
Accompanying grade × Male		0.07*		0.05	
		(0.03)		(0.04)	
GPT sentiment × Female			0.13**		0.12**
			(0.04)		(0.04)
GPT sentiment × Male			0.06*		0.06
			(0.03)		(0.03)
Final grade				0.01	0.02
				(0.03)	(0.02)
Controls	-	-	-	✓	✓
N	377	377	377	377	377
adj. R^2	0.315	0.343	0.347	0.346	0.352

Note: Linear regressions where the dependent variable equals one if the writer chose the competitive payment scheme and zero otherwise. Female and Male are dummy variables indicating the writer's gender. The posterior grade belief corresponds to writers' expected final grade after receiving feedback. The accompanying grade is the grade assigned by the evaluator who wrote the feedback. GPT sentiment is the GPT sentiment score of the feedback's text. Final grade is the average grade given to the writer by all evaluators. All dependent variables are standardized to have a mean of zero and a standard deviation of one. Controls include the writers' age, ethnic identity, gender, level of education, whether English is their native language, whether they grew up in the UK, their treatment assignment, the presence of spacing display errors in both the essay and feedback, and the number of characters in the feedback. The sample consists of writers in the *Feedback-Compete* and *Feedback-Compete-Hidden* treatments. Robust standard errors in parentheses and statistical significance of non-zero coefficients is indicated by * $p < 0.05$ and ** $p < 0.01$.

writers' (Wald test, $p < 0.01$). Conversely, the encouragement channel, captured by the coefficients of the associated grade and sentiment scores, is directionally stronger for women, though not significantly different (Wald test, $p > 0.17$). Taken together, these findings suggest that belief and encouragement channels are similarly important in shaping women's choices, while for men, feedback operates more strongly through an informational pathway.

Result 5 *Qualitative feedback influences the choice to compete through two distinct channels: a belief channel, whereby feedback affects expectations about performance, and an encouragement channel, whereby the tone or content of feedback motivates action beyond belief-updating. Female*

writers respond similarly to both channels, whereas male writers respond more strongly to the belief channel and less to the encouragement channel.

Overall, 68.4% of writers chose the competitive payment scheme, rather than the lottery. At first glance, this figure may seem to indicate that writers are overconfident, given that only 30% will rank in the top three of their group. This reasoning is erroneous, as up to 99% can rationally prefer betting on themselves finishing in the top three (Benoît and Dubra, 2011). Put differently, 99% could have over a 30% chance of placing in the top three of their respective groups.[29]

The compete choice alone does not reveal whether writers made good decisions. To assess decision quality, we define two types of errors based on monetary outcomes: false positives, where writers choose to compete despite having a low chance of winning, and false negatives, where writers avoid competition despite having a high chance of winning. To estimate these errors, we calculate each writer's probability of winning a competition by simulating 10,000 random tournaments for each essay, calculating how often the essay ranks in the top three. We find that 37.9% of essays have at least a 30% chance of winning, which is substantially lower than the 68.4% of writers who chose to compete. Among those who competed, 53.9% would have had higher expected earnings by not competing (false positives). Among those who did not compete, 20.2% would have been better off competing (false negatives). Gender differences in error rates are small and not statistically significant: 54.7% of women and 52.9% of men make false positive errors; 20.8% of women and 19.7% of men make false negative errors (χ^2 tests, $p > 0.24$).

Although we cannot directly observe how writers would behave in the absence of feedback, we can use our data to evaluate whether feedback, particularly its motivational component, improves decision quality. Since feedback predicts final grades and writers incorporate it into their beliefs, the belief channel should reduce decision errors. In contrast, the effect of the encouragement channel is less clear: it may help or hinder earnings-maximizing decisions. To evaluate its role, we construct two counterfactual predictions of the decision to compete. First, we use the regression from column (1) of Table 4 to predict each writer's probability of competing based solely on their posterior grade belief and gender. Second, we predict the same probability using the regression from column (5), which includes the sentiment score, to capture the added impact of the encouragement channel. For each prediction, we compute the mean probability of competing conditioning on whether the writer maximizes their earnings by competing or not competing. For writers who earn more by competing, we use the predicted probability of

[29]Excessive looking entry into tournaments is a common finding in the experimental literature (e.g., see Niederle and Vesterlund, 2011; Dechenaux et al., 2015). "Non-rational" explanations for this include overconfidence, preferences for competition (Niederle and Vesterlund, 2007; Lozano and Reuben, 2025) and preferences for control (Benoît et al., 2022).

competing to calculate the chance they make false negative errors. Similarly, for writers who earn more by not competing, we use their predicted probability of competing to obtain the chance they make false positive errors. This allows us to understand the impact on these two error types when incorporating the encouragement channel of qualitative feedback.

We find that incorporating the encouragement effect into the predicted probability of competing significantly reduces both types of decision errors for men and women. Consistent with the encouragement channel playing a greater role for female writers, the reduction in error rates is larger for women. These results suggest that the motivational component of feedback, even when not reflected in belief-updating, enhances decision quality. Full details and robustness checks are reported in Section C.4. of the Appendix (see Tables C10 and C11).

In conclusion, we identify two distinct channels through which qualitative feedback influences the decision to compete: a belief channel, which operates through updated grade beliefs, and an encouragement channel, which captures other aspects of the feedback, such as tone, that affect the decision to compete beyond belief-updating. The encouragement channel appears more important for female writers. Using each writer's probability of winning to benchmark optimal choices, we find that both women and men make false positive errors (competing when they shouldn't) and false negative errors (not competing when they should). We find suggestive evidence that qualitative feedback helps reduce both types of errors.

Editing

We now turn to examining how qualitative feedback influences writers' willingness to revise their work and whether this leads to improved performance.

In the *Feedback-Edit* treatment, writers could edit their essay after receiving feedback. Evaluators were informed writers would have this option before writing their feedback. Writers who chose not to edit were paid based on how their unedited essay ranked relative to nine randomly selected unedited essays. These writers were asked to indicate their final grade belief for the unedited essay. Writers who opted to edit were given five minutes to revise their essay, with both the original essay and the feedback visible during the editing process. A new set of evaluators graded the revised essays, and payment was based on how the new final grade ranked against those of nine randomly selected unedited essays. We used unedited essays for the rankings so that writers' incentive to edit did not depend on whether others choose to revise their work. After submitting their revised essay, these writers were asked to indicate their final grade belief for the edited essay.[30] All writers were paid after all the edited essays were graded.

We recruited a new pool of 200 evaluators to grade the edited essays. They also gave grades to the essays that were not edited. They were paid £0.50 bonus per essay for which their grade

[30]To avoid overburdening participants and potential anchoring effects, writers who chose to edit were not asked to report beliefs about their original, unedited essay.

matched the modal grade given by other evaluators for that essay. In addition, they received a participation fee of £3 – the participation fee was only £3 because they did not write any feedback.

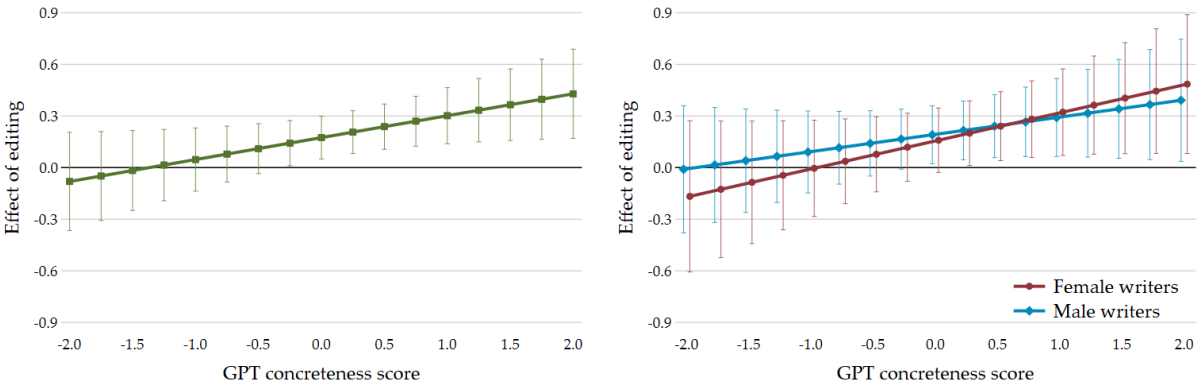
Unlike the decision to compete, which becomes more attractive as (expected) performance increases, the relationship between performance and the decision to edit is not straightforward. Because editing requires effort, writers should choose to edit only if they believe it will meaningfully improve their chances of winning. Writers who believe they performed well may see little value in editing, while those who believe they performed poorly may feel that even with revisions, they are unlikely to win. Additionally, editing does not guarantee a higher grade, so writers who feel they already performed to the best of their ability may see little benefit in revising, regardless of their expected performance.

Overall, 37.2% of writers chose to edit their essay. There is no statistically significant gender difference in editing rates: 35.4% of male writers and 39.1% of female writers chose to edit (χ^2 test, $p = 0.60$). Consistent with the idea that the decision to edit is not systematically related to performance, we find that neither unedited final grades, prior grade beliefs, nor the positivity of feedback, as measured by the accompanying grade or GPT sentiment, significantly predict the decision to edit or vary by gender (for details, see Tables C12 and C13 in Section C.5. of the Appendix). Given this, we now turn to examining whether feedback influences the impact of editing.

Since the new evaluators graded both edited and unedited essays, we can examine whether the choice to edit leads to improvements in final grades. The average grade assigned to unedited essays by the new evaluators was 3.11, which is statistically indistinguishable from the original average of 3.10 (paired t-test, $p = 0.81$). In contrast, the average grade of essays that were edited improves from of 3.09 to 3.27 or 0.18 grade points, a statistically significant effect corresponding to approximately 0.28 standard deviations (paired t-test, $p < 0.01$). If we look at this improvement by gender, we find that essays edited by male writers improve by 0.19 grade points, while those edited by female writers improve by 0.18 grade points (paired t-tests, $p < 0.02$). The improvement of male and female writers is not significantly different (t-test, $p = 0.87$).

Prior research suggests that feedback is more effective when it provides concrete advice (see Yeomans, 2021). To test this idea, we utilized GPT-3.5 to generate a concreteness score for each feedback text, where higher values indicate more concrete advice.[31] We then estimated linear regressions where the dependent variable is the change in the final grade: the difference between the new (regraded) and original grades. The key independent variables are the feedback's concreteness score, a dummy for whether the writer edited their essay, and their interaction.

[31] The GPT-3.5 prompt used to generate concreteness scores was: "How concrete is the advice in this text? Answer with a continuous numerical variable on a scale from 0 to 100, where 0 indicates no concrete advice, 50 indicates some concrete advice, and 100 indicates a lot of concrete advice. The advice should be on how to improve an essay. Only respond with a continuous numerical variable. Here is the text: ..."



(a) All writers **(b) Writers by gender**

Figure 9. Estimated change in the final grade due to the editing choice depending on the GPT concreteness of the feedback text and the writers' gender

Note: Predicted impact of editing on final grades, estimated using linear regressions where the dependent variable is the difference between the new (regraded) and original final grade. Independent variables include the feedback's GPT concreteness score (see footnote 31), a dummy variable for whether the writer edited their essay, and their interaction. The GPT concreteness score is standardized to have a mean of zero and a standard deviation of one. Panel (a) plots the estimated effect of editing—the coefficient on the editing dummy plus the coefficient of its interaction with GPT concreteness—for all writers. Panel (b) shows the same estimated effect separately for female and male writers. Regression results are reported in Table C14. Error bars indicate 95% confidence intervals calculated with robust standard errors. The sample consists of writers in the *Feedback-Edit* treatment ($N = 188$).

Full regression results are reported in Table C14 in the Appendix. We find that among writers who edited their essay, a one-standard-deviation increase in feedback concreteness is associated with a 0.13-point improvement in the final grade. In other words, the effectiveness of editing depends on how concrete the feedback is. Figure 9 illustrates the estimated effect of editing on changes in final grade depending on the standardized concreteness score of the feedback. Panel (a) shows that editing significantly improves grades only when the concreteness score is above the mean. Panel (b) shows that the benefits of concrete feedback do not differ by gender. These results are robust to the inclusion of controls of the essay, feedback, and writer characteristics (see Table C14). The findings from this section are summarized in the following result.

Result 6 *Writers who edit their essay after receiving feedback improve their final grade, with greater improvements when feedback is more concrete.*

5. Discussion and Concluding Remarks

Despite its widespread use, qualitative feedback remains relatively understudied in the economics literature. This paper demonstrates that qualitative feedback, despite its inherent subjectivity and lack of structure, can function as an effective and interpretable signal that shapes beliefs and influences behavior in economically meaningful ways. Through a controlled experiment, we examine the entire feedback-performance sequence, observing how feedback is given,

how it is interpreted and integrated into beliefs, and how it impacts consequential choices, such as whether to compete, as well as whether it enhances performance by motivating and informing revisions.

Our experimental setting presents meaningful challenges for feedback to be effective. Writers complete a relatively familiar task—writing a short essay—but under unusual conditions: the essay is based on an unfamiliar image, the task is one-shot, and evaluation comes from anonymous individuals with whom they have no interaction. Feedback is open-ended, qualitative, and composed in an evaluator's own words without standardization.

We believe this setting reflects many real-world environments, where evaluation is subject to a significant degree of subjectivity and feedback is loosely structured. At each stage of the feedback-performance sequence, there is potential for bias: feedback givers may soften criticism due to norms of politeness; writers may misinterpret the tone or intended message; and even when feedback is correctly understood, it may be over- or under-weighted in belief-updating or decision-making. These features make our setting a demanding test of the effectiveness of qualitative feedback.

Nevertheless, we find that the qualitative feedback is well-understood and meaningfully interpreted by writers. Although the feedback does not explicitly mention a grade, writers revise their beliefs upward when the feedback was written by an evaluator who assigned a grade higher than the writer's prior grade belief, and downward when the grade was lower. Remarkably, they do this despite the presence of a *kindness effect*: evaluators write much more positive comments when the writer will see the feedback than when they will not. Writers appear to effectively unravel this kindness effect and adjust their beliefs accordingly. This suggests that individuals, perhaps due to their extensive experience with qualitative feedback in everyday life, are capable of interpreting such messages accurately and updating appropriately, even in unfamiliar environments.[32]

In contrast to some earlier studies, we find no systematic differences in how feedback is given to male and female writers. Moreover, conditional on their prior beliefs, men and women update their beliefs similarly in response to feedback. However, because women's prior beliefs tend to underestimate their performance relative to men, the absence of gender differences in feedback giving and belief-updating means that feedback did not correct this initial gender difference. This finding suggests a potential policy implication: qualitative feedback might be more effective if tailored to the recipient's gender. Of course, the effectiveness of such tailoring would depend on whether recipients are aware of it and how they respond to it when they are. While our study is not designed to evaluate such interventions, this remains a promising direction for future experimental research.

[32]In contrast, several studies find that subjects do not update appropriately when given quantitative feedback Eil and Rao (2011); Ertac (2011); Zimmermann (2020); Möbius et al. (2022).

We present evidence that qualitative feedback shapes decision-making through two channels. A *belief channel*, where writers extract information from the feedback to update their expectations, and an *encouragement channel*, in which features of the feedback, such as tone, confidence, or warmth, motivate writers to act beyond what their beliefs imply. By combining the writers' belief data with sentiment analysis of the feedback text, we find evidence that both channels shape the decision to compete. Women appear to respond equally to both channels, whereas men rely more heavily on the belief channel and are less affected by the encouragement channel.

In addition to the choice to compete, we also examine how feedback influences writers' decisions to edit their work and the impact of this editing on performance. We find that writers who edit after reading their feedback improve their grades, but this improvement depends on how the feedback is written. More concrete feedback leads to greater improvements, indicating that the content of qualitative feedback plays a crucial role in its effectiveness.

Taken together, the findings on competing and editing reveal that qualitative feedback influences behavior beyond belief-updating. The way feedback is communicated—its tone, specificity, and content—can affect motivation and effort. These findings point to a promising direction for future research: identifying how to structure and deliver qualitative feedback to maximize its impact.

There are several limitations to our experimental design that should be considered when interpreting our findings. First, feedback in our study was delivered in written form rather than face-to-face. The mode of communication may shape how qualitative feedback is expressed through a phenomenon known in psychology as the disinhibition effect (Joinson, 2007). The impact of face-to-face delivery on qualitative feedback is not immediately apparent. On the one hand, written formats may promote greater honesty—patients, for example, are more likely to under-report alcohol consumption when speaking to a doctor than when interacting with a computer (Lind et al., 2013). On the other hand, digital communication has been found to reduce civility, particularly toward certain groups such as women (Coe et al., 2014; Wu, 2018; Ederer et al., 2024). Understanding how the kindness effect varies across communication modes and whether recipients can still anticipate and unravel it, is an important avenue for future research.

Second, while our experiment focuses only on receiving qualitative feedback, many environments involve both qualitative and quantitative feedback, such as in product reviews, academic evaluations, and workplace assessments. It remains an open question whether individuals interpret qualitative feedback differently when quantitative feedback is also present, and how these two types of feedback might interact to shape beliefs and behavior.

Third, our experiment captures only a single iteration of the feedback-performance sequence. In many situations, feedback is embedded in repeated interactions, where feedback givers and receivers can both adjust their behavior over time. Such repetition may influence how feedback

is formulated, how it is interpreted, and how it shapes subsequent performance, as individuals learn about each other's expectations, communication styles, and responsiveness.

Finally, the effectiveness of qualitative feedback is likely influenced by cultural norms. Communication styles vary across cultures, such as the degree of directness or indirectness (Meyer, 2015), which may affect how feedback is expressed and interpreted. To mitigate the impact of cultural differences in our study, we restricted participation to individuals residing in the UK. Investigating how qualitative feedback functions in cross-cultural contexts remains a promising area for future research.

References

Abel, M. (2024). Do workers discriminate against female bosses? *Journal of Human Resources*, 59(2):470–501.

Abel, M. and Buchman, D. (2024). The effect of manager gender and performance feedback: Experimental evidence from india. *Economic Development and Cultural Change*, 73(1):307–338.

Ambuehl, S., Bernheim, B. D., Ersoy, F., and Harris, D. (2025). Peer advice on financial decisions: A case of the blind leading the blind? *Review of Economics and Statistics*, 107(1):240–255.

Andrabi, T., Das, J., and Khwaja, A. I. (2017). Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review*, 107(6):1535–1563.

Apicella, C. L. and Dreber, A. (2015). Sex differences in competitiveness: Hunter-gatherer women and girls compete less in gender-neutral and male-centric tasks. *Adaptive Human Behavior and Physiology*, 1(3):247–269.

Bandiera, O., Larcinese, V., and Rasul, I. (2015). Blissful ignorance? a natural experiment on the effect of feedback on students' performance. *Labour Economics*, 34:13–25.

Barron, K. (2021). Belief updating: Does the 'good-news, bad-news' asymmetry extend to purely financial domains? *Experimental Economics*, 24(1):31–58.

Belschak, F. D. and Den Hartog, D. N. (2009). Consequences of positive and negative feedback: The impact on emotions and extra-role behaviors. *Applied Psychology*, 58(2):274–303.

Benoît, J.-P. and Dubra, J. (2011). Apparent overconfidence. *Econometrica*, 79(5):1591–1625.

Benoît, J.-P., Dubra, J., and Romagnoli, G. (2022). Belief elicitation when more than money matters: Controlling for "control"'. *American Economic Journal: Microeconomics*, 14(3):837–888.

Blanco, M., Engelmann, D., Koch, A. K., and Normann, H.-T. (2010). Belief elicitation in experiments: Is there a hedging problem? *Experimental Economics*, 13(4):412–438.

Bloom, N., Brynjolfsson, E., Foster, L., Jarmin, R., Patnaik, M., Saporta-Eksten, I., and Van Reenen, J. (2019). What drives differences in management practices? *American Economic Review*, 109(5):1648–1683.

Bloom, N., Propper, C., Seiler, S., and Van Reenen, J. (2015). The impact of competition on management quality: Evidence from public hospitals. *Review of Economic Studies*, 82(2):457–489.

Bloom, N. and Van Reenen, J. (2007). Measuring and explaining management practices across firms and countries. *Quarterly Journal of Economics*, 122(4):1351–1408.

Bohren, A., Imas, A., and Rosenberg, M. (2018). The language of discrimination: Using experimental versus observational data. *AEA Papers and Proceedings*, 108:169–174.

Bordalo, P., Burro, G., Coffman, K., Gennaioli, N., and Shleifer, A. (2025). Imagining the future: Memory, simulation, and beliefs. *Review of Economic Studies*, 92(3):1532–1563.

Brandts, J., Groenert, V., and Rott, C. (2015). The impact of advice on women's and men's selection into competition. *Management Science*, 61(5):1018–1035.

Buser, T., Gerhards, L., and Van Der Weele, J. (2018). Responsiveness to feedback as a personal trait. *Journal of Risk and Uncertainty*, 56(2):165–192.

Cárdenas, J.-C., Dreber, A., Von Essen, E., and Ranehill, E. (2012). Gender differences in competitiveness and risk taking: Comparing children in colombia and sweden. *Journal of Economic Behavior & Organization*, 83(1):11–23.

Charness, G., Gneezy, U., and Rasocha, V. (2021). Experimental methods: Eliciting beliefs. *Journal of Economic Behavior & Organization*, 189:234–256.

Coe, K., Kenski, K., and Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, 64(4):658–679.

Coffman, K., Ugalde Araya, M. P., and Zafar, B. (2024). A (dynamic) investigation of stereotypes, belief-updating, and behavior. *Economic Inquiry*, 62(3):957–983.

Correll, S. J., Weisshaar, K. R., Wynn, A. T., and Wehner, J. D. (2020). Inside the Black Box of Organizational Life: The Gendered Language of Performance Assessment. *American Sociological Review*, 85(6):1022–1050.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, 1(1):42–45.

Coutts, A., Koh, B. H., and Murad, Z. (2026). The Signals we Give: Performance Feedback, Gender, and Competition. *Management Science*. Forthcoming.

Danz, D., Vesterlund, L., and Wilson, A. (2022). Belief elicitation and behavioral incentive compatibility. *American Economic Review*, 112(9):2851–2883.

Dechenaux, E., Kovenock, D., and Sheremeta, R. M. (2015). A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics*, 18(4):609–669.

Dreber, A., Von Essen, E., and Ranehill, E. (2011). Outrunning the gender gap—boys and girls compete equally. *Experimental Economics*, 14(4):567–582.

Dreber, A., Von Essen, E., and Ranehill, E. (2014). Gender and competition in adolescence: Task matters. *Experimental Economics*, 17(1):154–172.

Ederer, F., Goldsmith-Pinkham, P., and Jensen, K. (2024). Anonymity and identity online. Working paper, Yale School of Management.

Eil, D. and Rao, J. M. (2011). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–138.

Erkal, N., Gangadharan, L., and Xiao, E. (2022). Leadership selection: Can changing the default break the glass ceiling? *The Leadership Quarterly*, 33(2):101563.

Ertac, S. (2011). Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization*, 80(3):532–545.

Ertac, S. and Szentes, B. (2011). The effect of information on gender differences in competitiveness: Experimental evidence. Working Paper.

Exley, C. L. and Kessler, J. B. (2022). The gender gap in self-promotion. *Quarterly Journal of Economics*, 137(3):1345–1381.

Flory, J. A., Leibbrandt, A., and List, J. A. (2015). Do competitive workplaces deter female workers? A large-scale natural field experiment on job entry decisions. *Review of Economic Studies*, 82(1):122–155.

Goldberg, P. (1968). Are women prejudiced against women? *Trans-action*, 5(5):28–30.

Graeber, T., Noy, S., and Roth, C. (2026). The Transmission of Reliable and Unreliable Information. *Working Paper*.

Graeber, T., Roth, C., and Zimmermann, F. (2024). Stories, statistics, and memory. *The Quarterly Journal of Economics*, 139(4):2181–2225.

Griggs, R. A. and Cox, J. R. (1982). The elusive thematic-materials effect in wason's selection task. *British JOurnal of Psychology*, 73(3):407–420.

Grosse, N., Riener, G., and Dertwinkel-Kalt, M. (2014). Explaining gender differences in competitiveness: Testing a theory on gender-task stereotypes. SSRN Working Paper 2551206.

Heine, E. C. E., Stouten, J., and Liden, R. C. (2026). Performance Feedback: A Critical Systematic Review. *Journal of Organizational Behavior*, 47(2):312–361.

Huning, H., Mechtenberg, L., and Wang, S. W. (2022). Using arugments to persuade: Experimental evidence. *Working Paper*.

Jampol, L., Rattan, A., and Wolf, E. B. (2022). A bias toward kindness goals in performance feedback to women (vs. men). *Personality and Social Psychology Bulletin*, 49(10):1–16.

Jampol, L. and Zayas, V. (2020). Gendered white lies: Women are given inflated performance feedback compared with men. *Personality and Social Psychology Bulletin*, 47(1):57–69.

Joinson, A. N. (2007). Disinhibition and the internet. In Gackenbach, J., editor, *Psychology and the Internet*, pages 75–92. Academic Press, Burlington, MA.

Kamenica, E. (2012). Behavioral economics and psychology of incentives. *Annual Review of Economics*, 4:427–452.

Katz, D. M., Bommarito, M. J., Gao, S., and Arredondo, P. (2024). GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 382(2270):20230254.

Kessel, D., Mollerstrom, J., and van Veldhuizen, R. (2021). Can simple advice eliminate the gender gap in willingness to compete? *European Economic Review*, 138:103777.

Kluger, A. N. and DeNisi, A. (1996). The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory. *Psychological Bulletin*, 119(2):254–284.

Lechermeier, J. and Fassnacht, M. (2018). How do performance feedback characteristics influence recipients' reactions? A state-of-the-art review on feedback source, timing, and valence effects. *Management Review Quarterly*, 68(2):145–193.

Lind, L. H., Schober, M. F., Conrad, F. G., and Reichert, H. (2013). Why do survey respondents disclose more when computers ask the questions? *Public Opinion Quarterly*, 77(4):888–935.

Lozano, L. and Reuben, E. (2025). (re)measuring preferences for competition. NYUAD working paper.

Ludwig, J. and Mullainathan, S. (2024). Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827.

McIntosh, J. (2015). Final report of the commission on assessment without levels. UK Department for Education and Standards and Testing Agency.

Mechtenberg, L. (2009). Cheap talk in the classroom: How biased grading at school explains gender differences in achievements, career choices and wages. *Review of Economic Studies*, 76(4):1431–1459.

Meyer, E. (2015). *The culture map: Decoding how people think, lead, and get things done across cultures*. PublicAffairs.

Möbius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2022). Managing self-confidence: Theory and experimental evidence. *Management Science*, 68(11):7793–7817.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, 4(2):61–64.

Movva, R., Peng, K., Garg, N., Kleinberg, J., and Pierson, E. (2025). Sparse autoencoders for hypothesis generation. *arXiv*.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics*, 122(3):1067–1101.

Niederle, M. and Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, 3(1):601–630.

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., and Bavel, J. J. V. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34):e2308950121.

Reuben, E., Sapienza, P., and Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 111(12):4403–4408.

Shastry, G. K., Shurchkov, O., and Xia, L. L. (2020). Luck or skill: How women and men react to noisy feedback. *Journal of Behavioral and Experimental Economics*, 88:101592.

Sheppard, L. D., Trzebiatowski, T. M., and Prasad, J. J. (2025). Paternalism in the Performance Context: Evaluators Who Feel Social Pressure to Avoid Exhibiting Prejudice Deliver More Inflated Performance Feedback to Women. *Journal of Business and Psychology*, 40(2):439–454.

Silverman, R. E. (2016). GE does away with employee ratings. The Wall Street Journal.

Thaler, M., Toma, M., and Wang, V. Y. (2025). Numbers tell, words sell. *Working Paper*.

Wason, P. C. (1960). On the Failure to Eliminate Hypotheses in a Conceptual Task. *Quarterly Journal of Experimental Psychology*, 12(3):129–140.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly journal of Experimental Psychology*, 20(3):273–281.

White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge University Press.

Wozniak, D., Harbaugh, W. T., and Mayr, U. (2014). The menstrual cycle and performance feedback alter gender differences in competitive choices. *Journal of Labor Economics*, 32(1):161–198.

Wu, A. H. (2018). Gendered language on the Economics Job Market Rumors Forum. *AEA Papers and Proceedings*, 108:175–179.

Yeomans, M. (2021). A concrete example of construct construction in natural language. *Organizational Behavior and Human Decision Processes*, 162:81–94.

Zimmermann, F. (2020). The dynamics of motivated beliefs. *American Economic Review*, 110(2):337–363.