Performance-Feedback

JEAN-PIERRE BENOÎT
London Business School, e-mail: jpbenoit@london.edu

ASHLEY PERRY
New York University Abu Dhabi, e-mail: ashley.perry@nyu.edu

ERNESTO REUBEN
New York University Abu Dhabi, Center for Behavioral Institutional Design, Luxembourg Institute of Socio-Economic Research, e-mail: 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 focus has been primarily on quantitative feedback. Our design captures the full performance-feedback sequence: participants complete an essay-writing task, assess their performance, receive feedback from an evaluator, and then update their beliefs and make choices. Despite the presence of an upwards kindness bias in how feedback is given, we find that qualitative feedback is effective: beliefs are updated appropriately. We find no difference in how feedback is given to men and women. 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. The more concrete feedback is, the more useful.

This version: September 2025

ACKNOWLEDGEMENTS

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. We gratefully acknowledge financial support from Tamkeen under the NYU Abu Dhabi Research Institute Award CG005.

جامعة نيويورك أبوظبي
NYU | ABU DHABI

مركز التصميم السلوكي المؤسساتي
CENTER for BEHAVIORAL INSTITUTIONAL DESIGN

1. Introduction

Feedback is important for performance in a variety of settings. Employees receive periodic appraisals of their work, and 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.[2] 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.[3]

To understand how qualitative feedback affects beliefs and performance, we study the entire *performance-feedback sequence*: an individual completes 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. Considering only some stages of the sequence 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 interpret and respond to the feedback can we determine whether they anticipate the biases and correct for them.[4] To the best of our knowledge, we are the first to study qualitative feedback using an experiment that covers the entire

[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] More accurately, quantitative feedback is isomorphic to numeric feedback. Thus, a grading system consisting of good, satisfactory, and unsatisfactory is quantitative feedback, as it corresponds to 3, 2, and 1.

[3] The challenges of qualitative feedback are also present in certain quantitative feedback settings, in which case the present study applies there as well.

[4] 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 Sections 2. and 4.2. for more on their study.

performance-feedback sequence, with each stage of the sequence undertaken by participants.[5]

Note that people typically do not receive feedback from everyone who evaluates them. For example, 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 or 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.

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). Qualitative feedback, by its open-ended nature, may be even more prone to bias. It can contain mixed messaging and psychological phenomena, such as motivated reasoning, which allow for a variety of interpretations. Consider the following feedback, taken from our experiment:

> "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!"

[5]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).

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, motivated reasoning could lead the recipient to 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.

Given the ubiquity of qualitative feedback, understanding how it shapes beliefs is essential. From a research perspective, a well-specified quantitative feedback structure has the advantage of providing a precise Bayesian benchmark for evaluating participants' belief-updating. Our experiment is purposely less structured to better reflect real-world qualitative feedback environments, where such calculations are infeasible. The feedback provided by evaluators is open-ended and written in their own words, making it difficult to assign probabilities to specific formulations. Nevertheless, by eliciting writers' beliefs about their performance both before and after receiving feedback, we can assess whether participants interpret qualitative feedback in a manner consistent with the underlying (but unseen) grade.

Beyond beliefs, we study how feedback influences decisions and how the content of the feedback affects its usefulness. In one set of treatments, writers are given the option to compete for a bonus payment that will be based on their average grade; in another treatment, writers 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), and by comparing 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? And do they respond differently to qualitative feedback in belief-updating and subsequent decisions?

Note that we ran the 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.

There is an upwards kindness effect in feedback. When evaluators know their 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.

Qualitative feedback is interpreted appropriately, but belief-updating is suboptimal. Despite the open-end nature of qualitative feedback and the inherent subjectivity in its interpretation, writers anticipate the kindness effect and 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. However, on average, the magnitude of belief-updating is less than optimal.

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.

Some studies in psychology find that women receive systematically more positive feedback than men. In 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.

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, 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). 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. Whereas, Abel and Buchman (2024) and Abel and Buchman (2024) do not capture the entire performance-feedback sequence as they do not measure feedback recipients' performance beliefs.

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. Section E of the Appendix provides a complete description of the study. Here, we limit ourselves 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 ran-

domly 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 the Appendix). 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. Recent 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.

Finally, writers were asked to select an alias from a list of gender-congruent names.[7] These aliases were displayed to evaluators in place of participant names. The use of aliases served

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

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. Since evaluators graded ten essays, their maximum bonus 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.[8] 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 case (The feedback example given 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

[8]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]."

the written feedback provided by one randomly selected evaluator. In both treatments, writers were then once again asked to report their expected final grade.

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 regarded. 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.[9] The large majority of writers 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

[9] 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.

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.[10] Table 1 gives an overview of the sample size of writers and evaluators for each treatment condition at each part of the study.

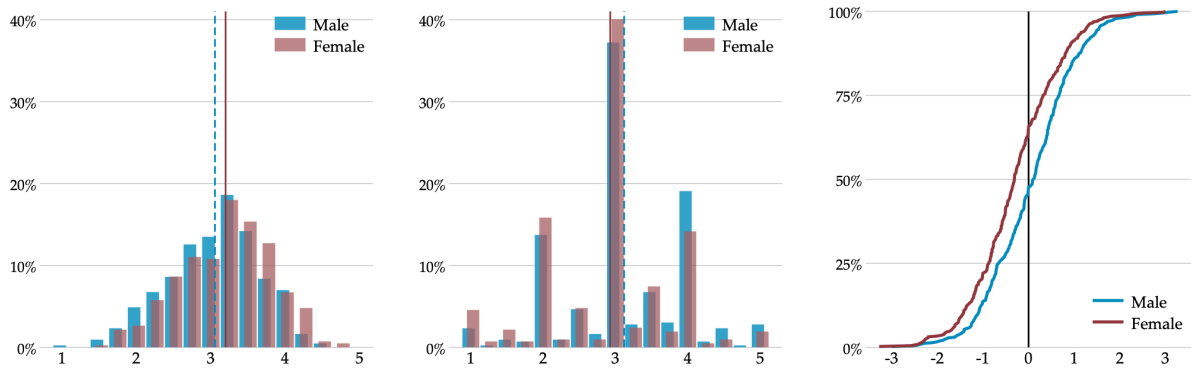
4. Results

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 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

[10] In a few instances, we found that 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 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 ($N = 847$).

across the *Feedback-Complete* 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.[11] 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 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

[11] 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.

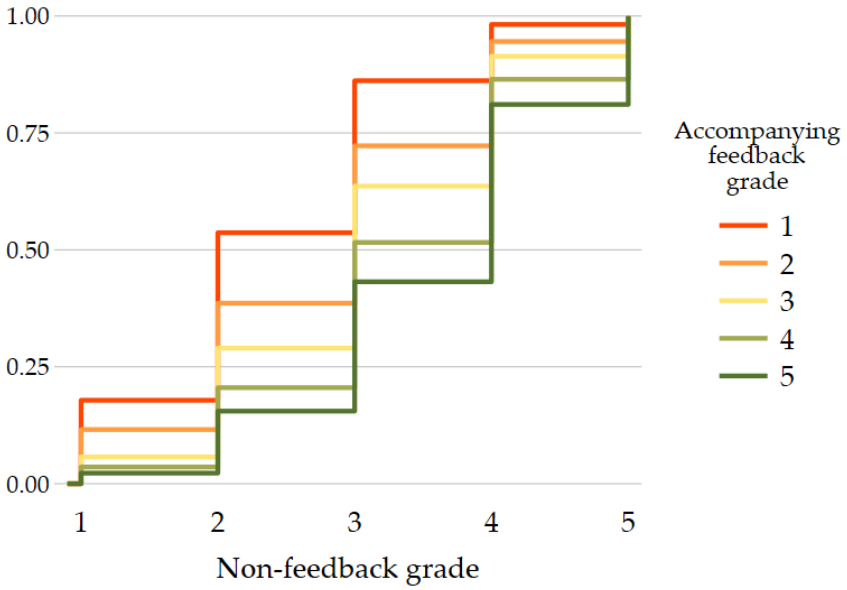


Figure 2. Cumulative distribution of the non-feedback grades depending on the grade accompanying the feedback text

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 ($N = 749$).

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.[12]

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.[13]

Feedback sentiment

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 demon-

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

[13] 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.

strated strong performance across a range of human-like tasks, including passing the bar exam (Katz et al., 2024) and constructing 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.[14] 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. Descriptive statistics for the evaluators' writings are presented in Table C2 in the Appendix. As a robustness check, we also replicate the sentiment scoring using Google Natural Language (GNL), which yields qualitatively similar results.

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*.[15] The data is shown separately for 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

[14] 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."

[15] Figure C1 in the Appendix presents box plots that confirm the upward trend in sentiment across grade groups while illustrating the variation within groups.

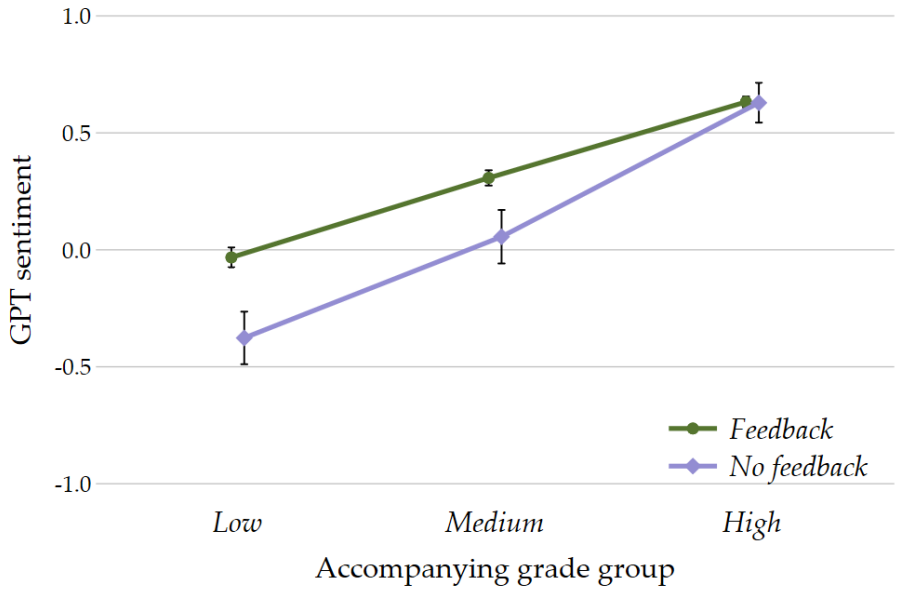


Figure 3. GPT sentiment of the evaluators' text depending on the accompanying grade and whether the text would be shared with writers as feedback

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*).

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 ($p < 0.01$). 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. These results are robust to using GNL sentiment scores (see Table C4 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

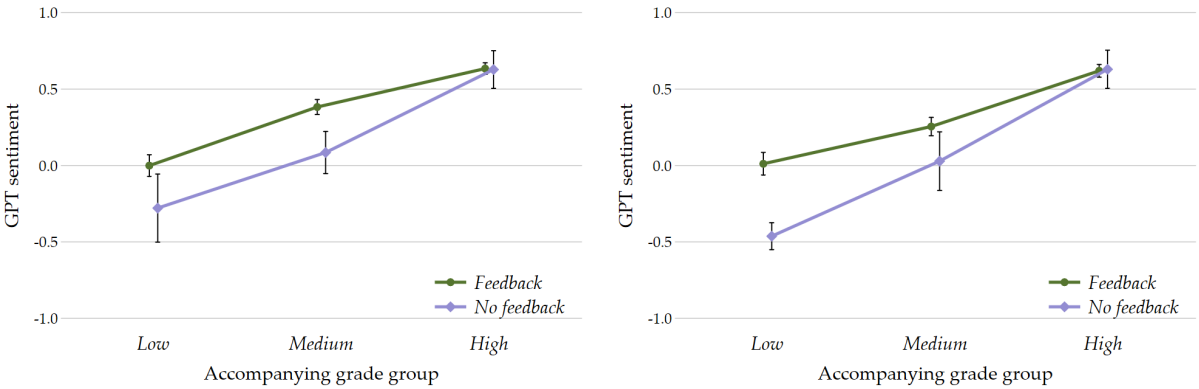
Table 2. GPT sentiment of the evaluators' text

	(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**	−0.43**	−0.45**
	(0.07)	(0.07)	(0.11)	(0.11)	(0.10)
No-Feedback × Accompanying grade		0.33**		0.40**	0.40**
		(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	0.401

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 errors in the essay, 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 ** $p < 0.01$.

measured using the alternative GNL score (see Figure C3 and Table C4 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



(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*).

and essay characteristics,[16] including the GPT sentiment of the essay itself. This last control 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 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 16). 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 C3 in the Appendix). The next result summarizes these findings.

[16]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, and their treatment assignment. The essay controls include the number of characters, whether there were spacing errors, and the GPT sentiment of the essay. See Appendix B and C descriptive statistics and more details of these variables.

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**	0.59**
	(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, 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 ** $p < 0.01$.

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 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.[17] 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

[17] As is common in many studies in psychology, Jampol and Zayas (2020) use deception. They do not use human participants as writers; instead, they deceive evaluators into believing a man or a woman wrote the same essay and programmed the writer's responses for the chat with the evaluator.

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.

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.

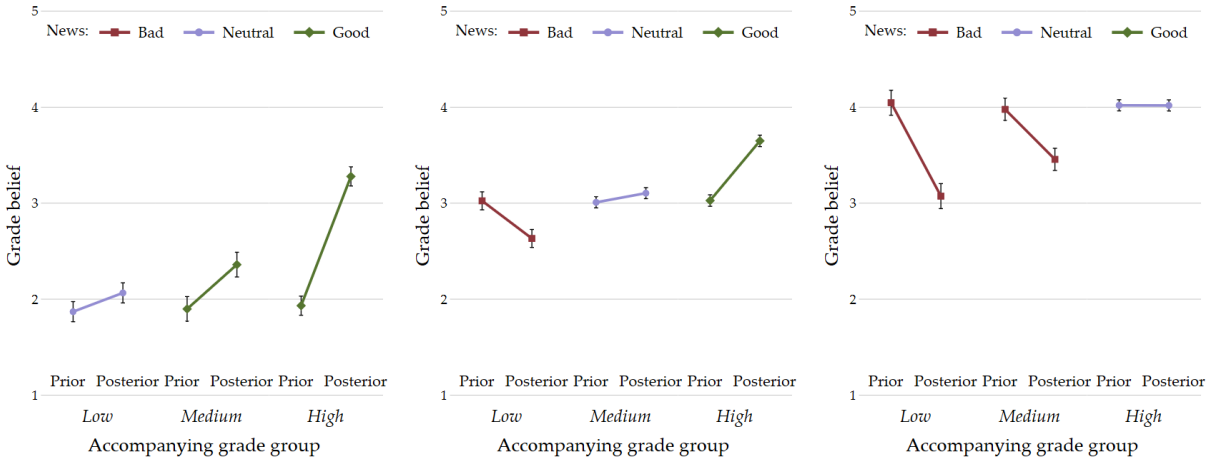
Feedback and beliefs

Is qualitative feedback effective? Do writers accurately interpret the feedback they receive and incorporate it into their beliefs? To address these questions, 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.

Writers received qualitative feedback from one evaluator, based on which they could attempt to infer the accompanying (unseen) grade. 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 $[2.5, 3.5]$, mirroring the three accompanying grade groups used in 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. Section C.3. of the Appendix presents a formal Bayesian model with this updating property.

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.[18] 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

[18]These patterns are robust to changes in the groups' cutoff values. Figure C4 in the Appendix plots individual-level prior and posterior beliefs, showing that the belief-updating pattern holds quite generally.



(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

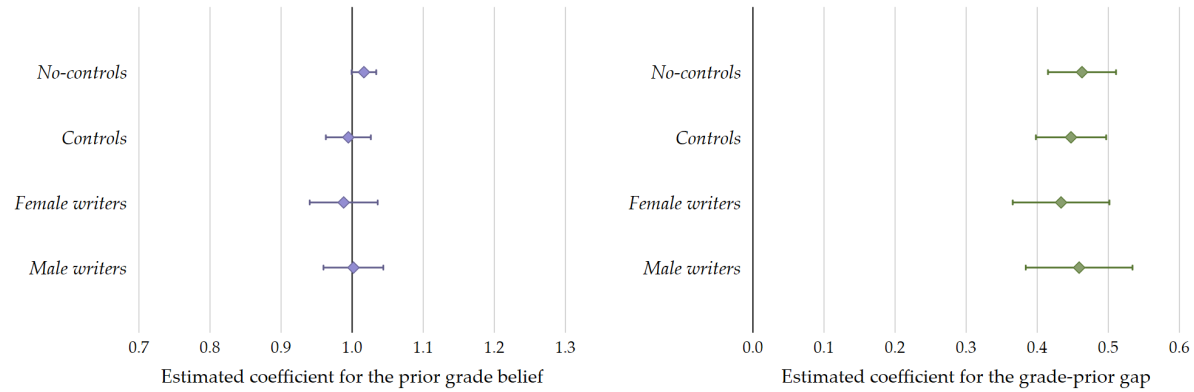
Note: Writers' mean prior and posterior beliefs depending on the accompanying grade group: *Low* (grades 1 and 2), *Medium* (grade 3), and *High* (grades 4 and 5). Panel (a) shows the beliefs for writers in the *Low* prior-belief group (priors in the range $[1, 2.5]$), Panel (b) for those in the *Medium* 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 ($N = 561$).

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.[19] 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. 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 C5 in the Appendix contains the regressions' output. Coefficients labeled as *No-controls* in the figure correspond to the regression

[19]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.



(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 19. *No-controls* corresponds to the regression without additional variables. *Controls* further controls for writer and essay characteristics (see footnote 20). *Female writers* restricts the sample to only female writers and *Male writers* to only male writers. Table C5 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 ($N = 561$).

described above (column (1) in Table C5). Coefficients labeled as *Controls* correspond to regressions that control for a range of writer and essay characteristics[20] (column (2) in Table C5). 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, 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 good news than to bad news (Eil and Rao, 2011; Möbius et al., 2022), while others report negative asymmetries, with stronger responses to bad news (Ertac, 2011). In Table C6 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

[20]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 essay controls include the number of characters and whether there were spacing errors. See Appendix B and C descriptive statistics and more details of these variables. Note that the writer and essay controls are the same as the evaluator ones in footnote 16 except that the essay controls do not include the sentiment of the essay.

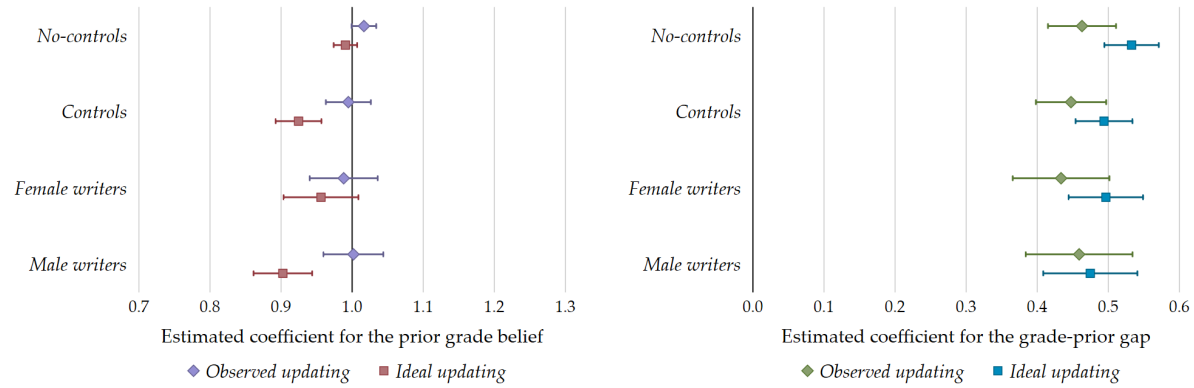
and male writers (for corresponding regressions see columns (3) and (4) of Table C5). As the figure indicates, there are no substantial gender differences. The effect of prior grade belief is nearly identical for both genders (0.99 for women vs. 1.00 for men; Wald test, $p = 0.83$), and the response 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.

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 also footnote 19), 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 C5 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 writer and essay 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, $p = 0.02$).[21] 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

[21] 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 19. *No-controls* corresponds to the regression without additional variables. *Controls* further controls for writer and essay characteristics (see footnote 20). *Female writers* restricts the sample to only female writers and *Male writers* to only male writers. Table C5 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 ($N = 561$).

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 tend to underreact to feedback relative to 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, $p < 0.01$), 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, while female writers tend to underreact to feedback.

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

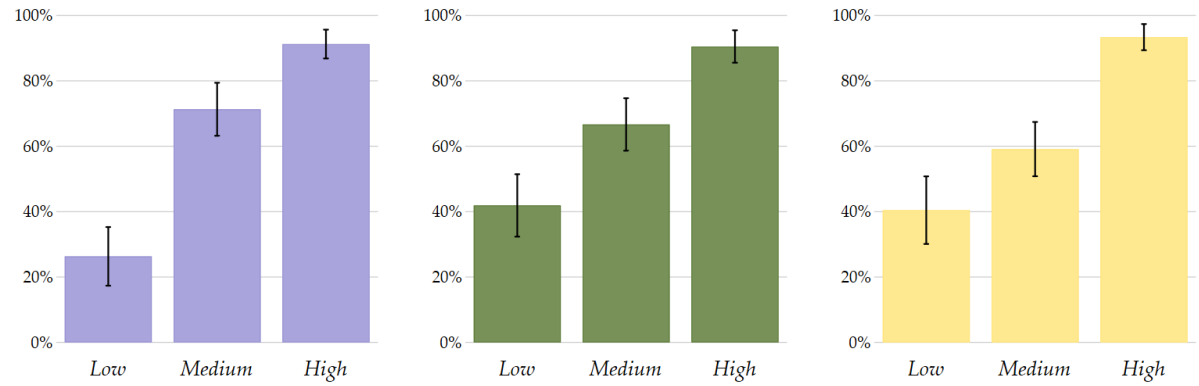
We begin by analyzing whether feedback affects the decision to compete. As described in Section 3.1., writers in Part 1 were informed that their final grade would be compared to those of nine other randomly selected writers and they would earn a £4 bonus if their essay ranked in the top three (with ties 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 in the likelihood of competing.[22] 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

[22] 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.



(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 ($N = 377$).

(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 and writer characteristics (see footnote 20). Even with these controls, both the posterior grade belief and the feedback variables remain positive and statistically significant. In the Appendix, we show that the results are robust to using the alternative GNL sentiment score (Table C7) 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 C8).

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

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 errors in the essay, and the number of characters in the essay. 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 ** $p < 0.01$.

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

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 errors in the essay, and the number of characters in the essay. 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 ** $p < 0.01$.

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 and evaluator 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 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.[23]

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

[23] 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).

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 5 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 (3), 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 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.5. of the Appendix.

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.[24] 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 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.

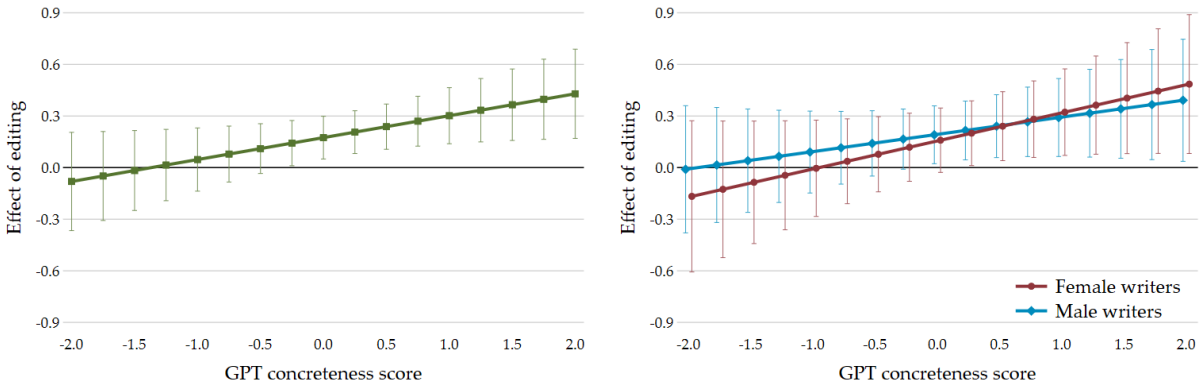
Unlike the decision to compete, which becomes more attractive as (expected) performance increases, the relationship between performance and the decision to edit is not straightforward. Because editing requires effort, writers should choose to edit only if they believe it will meaningfully improve their 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 (for details, see Section C.6. in 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

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



(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 25), 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 C13. Error bars indicate 95% confidence intervals calculated with robust standard errors. The sample consists of writers in the *Feedback-Edit* treatment ($N = 188$).

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.[25] 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. Full regression results are reported in Table C13 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 for evaluator and essay characteristics (see Table C13). 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.*

[25]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: . . ."

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, even in unfamiliar environments.

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.

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 solely on 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 United Kingdom. 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.

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.

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.

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.

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.

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 predjudiced against women? *Trans-action*, 5(5):28–30.

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.

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.

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.

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.

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.

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

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.

Online appendices for: Performance-Feedback

by Jean-Pierre Benoît, Ashley Perry, and Ernesto Reuben

This document contains supplementary materials for the paper (Benoît et al., 2025). Appendix A contains additional details on the experimental design and implementation that were mentioned in the paper but were not fully described due to space constraints. Appendix B contains descriptive statistics of the study participants. Appendix C contains more details of the data analysis, a model of belief-updating and numerous robustness tests, referenced in the paper. Appendix D contains the details of the methods used for textual analyses.

Appendix A. Additional information about the experiment

A.1. Gendered alias

To create the lists of highly gendered UK names from which the writers selected their aliases, we used the 200 most common female and male birth names, one hundred for each gender, registered in 1994 with the Office for National Statistics. This year was chosen to ensure that the names are common today and so likely to be known to the participants of our study. Common names from a more recent list are not necessarily very common among adults today (e.g., Ayla).

To determine which names are highly gendered, we used the web-based service Gender API, which performed well compared to similar services (Santamaría and Mihaljević, 2018). The API integrates data from multiple sources, including publicly available government records and social media sites. Names must be present in multiple sources to be considered valid. For each name, the service will return a gender assignment (female, male, or unknown), a probability that the assigned gender is correct, and a count of sources in the database that match the name. We restricted the search to names associated with sources derived from the UK. For our 200 male and female names, we retained names that had a source count of at least 2000 and a probability of correct classification of at least 98%. This ensured that they were common and highly gendered. All the names fulfilling these criteria are typically white names from the UK. For each gender, we randomly generated three lists of ten names. In the instructions for Part 1, writers were instructed to select an alias to maintain anonymity. Based on their stated gender, they were randomly shown one of three lists. The order of the names in a list was randomized across writers.

A.2. Implementation details

As specified in our preregistration, we recruited 900 writers. The link to the preregistration can be found here: https://aspredicted.org/LG8_JPK. This was balanced across gender with 49.8% females. The average completion time of Part 1 was approximately 16 minutes. In Part 1, 906 writers had been invited to the study, six of which were rejected. Of these six writers, three did not consent to the study and three failed to answer the understanding questions following the instructions. They had multiple attempts to answer understanding questions.

For Part 2, the essays written in Part 1 were randomly allocated to our five treatments. Before doing this, we performed computer-based checks that the submissions were valid and that the essays were written in English (we used python package langdetect). Within each treatment, the essays were randomly allocated to groups of ten essays and balanced by gender. We assigned 100 essays to the *No-Feedback* treatment, 200 to both treatment *Feedback-Only* and *Feedback-Edit*. The remaining 400 were allocated to *both* treatments *Feedback-Compete* and *Feedback-Compete-Hidden*, the difference being that the alias of the writer was not revealed in *Feedback-Compete-Hidden*. This allows us to identify the effect of a writer's gender being disclosed to the evaluator. For each writer we have two observations of their feedback but we only showed them one piece of feedback, which was decided randomly. From the writer's perspective, they were in one of two treatments (each of which had 200 writers). For *Feedback-Edit*, we aimed to collect 400 feedback observations, double the number of writers, to create a larger sample for text analysis. This meant that we aimed to collect 1500 feedback observations, as stated in our pre-registration. However, we could not predict which evaluators would complete the study once they started, which would have meant that some writers would not have received feedback. Hence, to ensure that we met the minimum requirement of one written feedback per writer, we randomly over-sampled. During data collection in this part, 91 evaluators were shown the wrong alias during the feedback stage. Since the alias was correct in the prior grading stage, we were able to retain the grade data, but we do not use the feedback data during our text analysis.

In total, we collected 1,651 submissions from evaluators in Part 2, which includes the 91 evaluators who only graded essays and did not provide feedback and so we only use their data for determining a writer's final grade. In total, 1685 evaluators were invited to the study, 34 of whom were rejected. Of these 34 evaluators, one did not consent to the study, two had a malfunction which meant no grade data was collected as they did not see the study materials and so were dropped, and 31 failed to answer multiple attempts at the understanding questions. For the analysis at the evaluator level, such as the sentiment analysis, we have 1560 complete submissions with feedback from evaluators. Part 2 began a few days after Part 1 had ended, and the average completion time of Part 2 was approximately 25 minutes.

This study has a number of different *Feedback* conditions: *Feedback-Only*, *Feedback-Compete*, *Feedback-Compete-Hidden*, and *Feedback-Edit*. The analysis corresponding to the feedback data is done at the evaluator level. Table A1 shows that there are no treatment differences in the unseen grade accompanying the feedback text or the sentiment of the feedback text.

Table A1. Check of treatment differences for evaluator outcome variables

	Feedback $N = 241$		Compete $N = 421$		Compete-Hidden $N = 436$		Edit $N = 339$		
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	p-value
Accompanying grade	3.16	1.12	3.11	1.05	3.10	1.07	3.14	1.10	0.545
GPT sentiment	0.37	0.43	0.34	0.45	0.30	0.47	0.35	0.43	0.258
GNL sentiment	0.16	0.39	0.16	0.41	0.11	0.42	0.15	0.39	0.194

Note: All evaluators who took part in Part 2 of the study. For the unseen grade the p-value is derived from a chi-squared test of independence between the given group categories, integer grades from 1 to 5, across the treatment groups. For the two sentiment scores, the p-value is derived from an analysis of variance (ANOVA) as they are continuous variables.

Part 3 began a few days after Part 2 had ended. We invited the 900 writers to return and complete the study. In addition to the initial invitation, we sent reminders to those who had not yet completed this part of the study. In total, 878 writers returned, 433 women, and 445 men. Attrition by gender was around the same for women and men (3.3% vs. 1.5%;χ^2 test, $p = 0.79$). There are no significant differences in the rates of attrition across treatment groups (χ^2 test, $p = 0.29$). During Part 2 evaluators were instructed not mention the grade they had given in their feedback. After the data had been collected, we checked if this rule was followed and found that a small minority had deviated. In the feedback seen by the writers,[A1] the evaluator explicitly stated the grade in 31 cases. Hence, for the analysis pertaining to writers (sections 4.2. and 4.3.) we drop these observations. However, including them has little effect on the results. After dropping these observations, we are left with 417 women and 430 men. Attrition in this sample by gender also similar and not significantly different (6.9% for women and 4.9% for men; χ^2 test, $p = 0.75$), as well as attrition across treatment groups (χ^2 test, $p = 0.29$). The average completion time of Part 3 was approximately seven minutes.

We recruited 200 new evaluators to evaluate the edited essays from *Feedback-Edit*. They passed all understanding questions. We used the original 200 essays that had been assigned to the *Feedback-Edit* treatment and swapped the original essay to the edited essays if a writer had chosen to edit. The essays were randomly assigned to groups of ten essays and balanced by gender. The average completion time of this re-evaluation was approximately 14 minutes.

[A1] This excludes all observations in the *No-Feedback* treatment and any observations from the other treatments that were not shown to the writers; the corresponding number of observations is 780.

The analysis that corresponds to belief-updating is done at the writer level. From the *Feedback* conditions we use the following conditions: *Feedback-Only*, *Feedback-Compete*, and *Feedback-Compete-Hidden*. Table A2 shows that for the outcome variables, prior and posterior grade beliefs, there are no treatment difference. We exclude *Feedback-Edit* because writers who edited their essay were not asked for their current grade belief, but instead of their grade belief about their edited essays. This was done this to minimize the number of questions they were asked and to avoid any anchoring effects.

Table A2. Check of treatment differences for writer outcome variables

	Feedback Only $N = 184$		*Feedback Compete* $N = 192$		*Feedback Compete-Hidden* $N = 185$		
	mean	s.d.	mean	s.d.	mean	s.d.	p-value
Prior grade belief	3.09	0.88	2.98	0.83	3.03	0.81	0.424
Posterior grade belief	3.17	0.95	3.08	0.85	3.19	0.84	0.391

Note: All writers who had a complete submission for Part 3. The p-value is derived from an analysis of variance (ANOVA) as they are continuous variables.

Spacing errors

During the data collection for Part 3, we identified a coding error in the presentation of the essay and feedback text to participants. The error caused a few words to be combined, creating what could be interpreted as a spelling mistake (e.g., the words "this" and "essay" would appear as "thisessay"). Of the total words written in the essays the spacing error affected 1.3% of the words and was present in 84% of the 900 essays. In Part 3, given the *No-Feedback* treatment a total of 780 writers saw their feedback. Of the total number of words written in the corresponding feedback, the spacing error affected 3.0% of the words and was present in 83% of the feedback. We believe that computer-generated spacing errors are not a concern for our analysis for the following three reasons. First, we corrected the code when essays were reevaluated in the *Feedback-Edit* treatment, ensuring that there were no computer-generated spacing errors. Therefore, for writers who did not edit their essay, we have an observation with the spacing error and one without. We find no significant difference between the original and new final grades for these essays (paired t-test, $p = 0.37$). We also find no difference if we restrict the test to only male or female writers (paired t-tests, $p > 0.44$). Second, the presence of the computer-generated spacing errors did not differ significantly by gender of the writer for both the essays or the feedback text (t-tests, $p > 0.18$). Third, although spelling and grammar were part of the grading criteria, they were only one out of four criteria, the other being accuracy and detail, flow and structure, and creativity and engagement.

Appendix B. Descriptive statistics

This section provides descriptive statistics and tests whether there are significant differences between genders and across treatments. Table B1 shows the descriptive statistics of writers who completed Parts 1 and 3. The sample is more diverse than typical samples in experimental laboratories at universities (e.g., 29% had high school as their highest level of education, and 71% are 31 years or older). The table also shows statistics by the writers' gender. For each variable, the table displays the p-value obtained when testing whether there is a significant gender difference using χ^2 tests. There are no significant gender differences.

Table B1. Descriptive statistics for writers by gender

		All $N = 847$		Female $N = 433$		Male $N = 445$		diff. in	p-
		mean	s.d.	mean	s.d.	mean	s.d.	means	value
Gender	Female	0.49	0.50						
	Male	0.51	0.50						
Age	18-30	0.29	0.45	0.30	0.46	0.28	0.45	−0.02	0.796
	31-50	0.49	0.50	0.48	0.50	0.49	0.50	0.01	
	51-84	0.22	0.42	0.22	0.41	0.23	0.42	0.01	
Ethnicity	Arab	0.00	0.05	0.00	0.07	0.00	0.00	0.00	0.127
	Asian	0.08	0.26	0.06	0.23	0.10	0.29	0.04	
	Black	0.03	0.17	0.03	0.17	0.03	0.16	0.00	
	White	0.85	0.35	0.87	0.34	0.84	0.37	−0.02	
	Mixed heritage	0.02	0.15	0.02	0.15	0.02	0.15	0.00	
	Other	0.02	0.13	0.02	0.15	0.01	0.11	−0.01	
Education	School	0.12	0.32	0.12	0.32	0.11	0.32	0.00	0.910
	Sixth form	0.17	0.38	0.18	0.38	0.16	0.37	−0.01	
	Some university	0.10	0.30	0.10	0.30	0.10	0.30	0.00	
	Undergraduate degree	0.39	0.49	0.40	0.49	0.39	0.49	−0.01	
	Graduate degree	0.22	0.41	0.20	0.40	0.23	0.42	0.03	
English mother tongue		0.91	0.29	0.91	0.28	0.91	0.29	0.00	1.000
Grew up in UK		0.90	0.31	0.89	0.32	0.90	0.29	0.02	0.474
Feedback by male evaluator		0.46	0.50	0.47	0.50	0.46	0.50	−0.01	0.729
Spacing error in the essay		0.85	0.36	0.83	0.38	0.86	0.35	0.03	0.254
Spacing error in the feedback		0.83	0.37	0.82	0.39	0.85	0.36	0.03	0.304

Note: All writers who had complete submissions for Parts 1 and 3. Means and standard deviations are calculated overall and separately by gender. Spacing errors in the text written by evaluators in the *No-Feedback* treatment are not included since those assessments were not shared with them. The p-values are derived from χ^2 tests of the variable categories and the writers' gender.

Table B2 shows that the writers' variables are almost all balanced across treatments. The

only exceptions are the variables indicating if English was their mother tongue, if they grew up in the UK, and the presence of a computer-generated spacing error in their essay. If we adjust p-values with the Benjamini-Hochberg method to account for multiple comparisons (Benjamini and Hochberg, 1995), then we find a statistically significant difference only for the presence of spacing errors. There are more of these errors in the *Feedback-Compete* and *Feedback-Edit* treatments than in the others. We control for this and other essay characteristics in our analysis and find that it does not affect our results.

Table B2. Treatment balance of writers

		No-Feedback $N = 98$		*Feedback* *Only* $N = 184$		*Compete* $N = 192$		*C-Hidden* $N = 185$		*Edit* $N = 188$		
		mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	p-value
Gender	Female	0.48	0.50	0.49	0.50	0.49	0.50	0.50	0.50	0.49	0.50	0.999
	Male	0.52	0.50	0.51	0.50	0.51	0.50	0.50	0.50	0.51	0.50	
Age	18-30	0.29	0.45	0.28	0.45	0.27	0.44	0.28	0.45	0.35	0.48	0.446
	31-50	0.52	0.50	0.45	0.50	0.51	0.50	0.52	0.50	0.45	0.50	
	51-84	0.19	0.40	0.27	0.45	0.23	0.42	0.21	0.41	0.20	0.40	
Ethnicity	Arab	0.01	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.07	0.475
	Asian	0.10	0.30	0.06	0.24	0.06	0.23	0.06	0.24	0.11	0.32	
	Black	0.04	0.20	0.03	0.18	0.02	0.12	0.03	0.18	0.03	0.16	
	White	0.79	0.41	0.88	0.33	0.89	0.31	0.86	0.35	0.82	0.39	
	Mixed heritage	0.03	0.17	0.02	0.13	0.03	0.17	0.02	0.15	0.02	0.13	
	Other	0.03	0.17	0.01	0.10	0.01	0.07	0.03	0.16	0.02	0.14	
Education	School	0.15	0.36	0.10	0.31	0.12	0.33	0.11	0.32	0.11	0.31	0.844
	Sixth form	0.14	0.35	0.18	0.38	0.19	0.39	0.16	0.37	0.16	0.37	
	Some university	0.10	0.30	0.15	0.36	0.09	0.28	0.08	0.27	0.10	0.30	
	Undergraduate degree	0.39	0.49	0.38	0.49	0.39	0.49	0.41	0.49	0.41	0.49	
	Graduate degree	0.21	0.41	0.19	0.39	0.22	0.41	0.24	0.43	0.22	0.41	
English mother tongue		0.89	0.32	0.95	0.22	0.94	0.24	0.90	0.30	0.87	0.34	0.029
Grew up in UK		0.86	0.35	0.93	0.25	0.92	0.28	0.89	0.31	0.86	0.35	0.097
Spacing error in the feedback				0.81	0.39	0.85	0.36	0.82	0.39	0.85	0.36	0.601
Spacing error in the essay		0.76	0.43	0.82	0.39	0.90	0.30	0.81	0.40	0.90	0.30	0.001
Feedback from male evaluator		0.43	0.50	0.46	0.50	0.49	0.50	0.46	0.50	0.45	0.50	0.671

Note: All writers who had complete submissions for Parts 1 and 3. Means and standard deviations are calculated separately by treatment. Spacing errors in the text written by evaluators in the *No-Feedback* treatment are not included since those assessments were not shared with them. The p-values are derived from χ^2 tests of the variable categories and the writers' assigned treatment.

Table B3 shows the descriptive statistics of all evaluators who completed Part 2. Similarly to the writers, 30% finished high school and 69% are 31 years or older. The table also shows

statistics by the evaluators' gender. For these statistics, since those who selected "Other" as their gender make up less than 1% of the sample, we considered only those who indicated their gender as female or male. For each variable, the table displays the p-value obtained when testing whether there is a significant gender difference using χ^2 tests. The evaluators' variables are almost all balanced across genders. The only variable showing a significant gender difference is growing up in the UK, although this is no longer the case if we adjust p-values with the Benjamini-Hochberg method for multiple comparisons.

Table B3. Descriptive statistics of evaluators by gender

		All $N = 1560$		Female $N = 785$		Male $N = 765$		diff in.	p-
		mean	s.d.	mean	s.d.	mean	s.d.	means	value
Gender	Female	0.50	0.50						
	Male	0.49	0.50						
	Other	0.01	0.08						
Age	18-30	0.30	0.46	0.30	0.46	0.30	0.46	0.00	0.987
	31-50	0.47	0.50	0.47	0.50	0.47	0.50	0.00	
	51-83	0.23	0.42	0.23	0.42	0.23	0.42	0.00	
Ethnicity	Arab	0.00	0.06	0.00	0.05	0.00	0.06	0.00	0.431
	Asian	0.08	0.28	0.08	0.27	0.09	0.29	0.02	
	Black	0.03	0.16	0.03	0.17	0.02	0.15	−0.01	
	White	0.85	0.36	0.86	0.35	0.84	0.37	−0.03	
	Mixed heritage	0.02	0.12	0.01	0.12	0.02	0.13	0.00	
	Other	0.02	0.14	0.02	0.12	0.03	0.16	0.01	
Education	School	0.11	0.31	0.11	0.31	0.11	0.31	0.00	0.411
	Sixth form	0.19	0.39	0.19	0.40	0.17	0.38	−0.02	
	Some university	0.10	0.30	0.09	0.28	0.11	0.32	0.03	
	Undergraduate degree	0.41	0.49	0.41	0.49	0.42	0.49	0.02	
	Graduate degree	0.20	0.40	0.20	0.40	0.19	0.39	−0.02	
English mother tongue		0.92	0.27	0.92	0.27	0.94	0.24	0.02	0.179
Grew up in UK		0.91	0.29	0.90	0.30	0.93	0.25	0.03	0.033
Gave feedback to a female writer		0.51	0.50	0.52	0.50	0.50	0.50	−0.02	0.546
Spacing error in the essay		0.85	0.36	0.84	0.36	0.85	0.36	0.01	0.838

Note: All evaluators who took part in Part 2. Columns for Female and Male evaluators exclude ten evaluators who indicated "Other" as their gender. Means and standard deviations are calculated overall and separately by gender. The p-values are derived from χ^2 tests of the variable categories and the evaluators' gender.

Table B4 shows that variables are almost all balanced across treatments. The only variable showing a significant difference across treatments is the presence of computer-generated spacing errors in the essay they graded. This difference remains significant after adjusting p-values with

the Benjamini-Hochberg method for multiple comparisons. There are fewer of these errors in the *No-Feedback* treatment than in the feedback treatments. This treatment is not used in the analysis of belief-updating and decision-making, but it is used in sentiment analysis. We find that controlling for the presence of spacing errors and other essay characteristics does not affect our results.

Table B4. Treatment balance of evaluators

		No-Feedback $N = 123$		*Feedback*								
				Only $N = 241$		*Compete* $N = 421$		*C-Hidden* $N = 436$		*Edit* $N = 339$		
		mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	p-value
Gender	Female	0.51	0.50	0.51	0.50	0.49	0.50	0.50	0.50	0.51	0.50	
	Male	0.49	0.50	0.49	0.50	0.50	0.50	0.49	0.50	0.48	0.50	0.989
	Other	0.00	0.00	0.00	0.06	0.01	0.08	0.01	0.08	0.01	0.09	
Age	18-30	0.28	0.45	0.34	0.47	0.30	0.46	0.29	0.45	0.32	0.47	
	31-50	0.53	0.50	0.44	0.50	0.47	0.50	0.49	0.50	0.43	0.50	0.592
	51-83	0.19	0.39	0.22	0.42	0.23	0.42	0.22	0.41	0.25	0.43	
Ethnicity	Arab	0.01	0.09	0.00	0.00	0.00	0.05	0.00	0.05	0.01	0.08	
	Asian	0.09	0.29	0.06	0.24	0.09	0.28	0.07	0.26	0.11	0.31	0.195
	Black	0.01	0.09	0.05	0.22	0.03	0.16	0.02	0.13	0.03	0.18	
	White	0.85	0.35	0.87	0.34	0.84	0.37	0.87	0.33	0.82	0.39	
	Mixed heritage	0.02	0.13	0.01	0.11	0.02	0.14	0.02	0.14	0.01	0.08	
	Other	0.02	0.15	0.00	0.06	0.02	0.15	0.02	0.13	0.03	0.17	
Education	School	0.12	0.33	0.10	0.30	0.10	0.30	0.09	0.29	0.13	0.34	
	Sixth form	0.15	0.35	0.20	0.40	0.18	0.38	0.20	0.40	0.18	0.38	0.458
	Some university	0.13	0.34	0.11	0.32	0.08	0.28	0.10	0.30	0.11	0.31	
	Undergraduate degree	0.36	0.48	0.42	0.50	0.45	0.50	0.39	0.49	0.40	0.49	
	Graduate degree	0.24	0.43	0.17	0.38	0.18	0.39	0.22	0.41	0.18	0.38	
English mother tongue		0.96	0.20	0.92	0.27	0.93	0.25	0.93	0.26	0.92	0.27	0.651
Grew up in UK		0.95	0.22	0.90	0.29	0.93	0.26	0.91	0.28	0.90	0.30	0.342
Gave feedback to a female writer		0.48	0.50	0.50	0.50	0.52	0.50	0.50	0.50	0.51	0.50	0.927
Spacing error in the essay		0.74	0.44	0.82	0.39	0.86	0.35	0.84	0.36	0.89	0.31	0.001

Note: All evaluators who took part in Part 2. Means and standard deviations are calculated separately by treatment. The p-values are derived from χ^2 tests of the variable categories and the evaluators' assigned treatment.

Appendix C. Supplementary data analysis

This section contains robustness checks and additional analysis for results reported in Sections 4.1., 4.2., and 4.3. of the main body of the paper.

C.1. Final grades and prior beliefs

Evaluators in treatments *Feedback-Compete* and *Feedback-Compete-Hidden* saw the same essays. However, in *Feedback-Compete-Hidden*, they did not see the gendered alias. This allows us to isolate the effect of disclosing the writers' gender to evaluators. The number of evaluators we can use for this analysis is 893, which includes the 857 evaluators who provided valid feedback and 36 evaluators who graded the essay but were unable to provide feedback due to a computer error (as explained in Section A.2.).

We check whether there is a gender difference in grading. Table C1 presents the results of linear regressions with final grades as the dependent variable. Each evaluator graded ten essays, which gives us 8,920 observations. Since multiple evaluators saw the same essays in both treatments, we use essay fixed effects. Column (1) controls for the treatment and its interaction with the writers' gender. Column (2) also controls for the evaluators' characteristics described in footnote 16. Grades of female writers with disclosed aliases are 0.03 grade points lower than those with undisclosed aliases. Similarly, grades of male writers with disclosed aliases are around 0.07 grade points lower than those with undisclosed aliases. Since these differences are small, we consider that there is no meaningful difference in the grading.

Table C1. Predicting grades

	(1)	(2)
Constant	3.13**	3.20**
	(0.02)	(0.09)
Feedback-Compete	−0.06	−0.07
	(0.04)	(0.04)
Feedback-Compete × Female	0.03	0.03
	(0.03)	(0.03)
Essay fixed effects	✓	✓
Evaluator controls	-	✓
Observations	8930	8930
Evaluators	893	893
adj. R^2	0.001	0.003

Note: Linear regressions with the essay grades as the dependent variable. *Feedback-Compete* is a dummy variable that equals one if the writer's alias is disclosed to the evaluator grading the essay and zero otherwise. Female is a dummy variable indicating that the writer was female. Each evaluator graded ten essays, and each essay had between 10 and 15 grades. The sample is restricted to essays seen by evaluators in both the *Feedback-Compete* and *Feedback-Compete-Hidden* treatments. Controls include the evaluators' age, level of education, ethnic identity, gender, whether English is their native language, and whether they grew up in the UK. Robust standard errors clustered on evaluators in parentheses and statistical significance of non-zero coefficients indicated by * $p < 0.05$ and ** $p < 0.01$.

C.2. Characteristics of feedback

Table C2 summarizes the sentiment variables generated with NLP methods (see Section D).

Table C2. Feedback statistics of evaluators

Participant	Variable	N	mean	s.d.
Evaluators	GPT sentiment	1560	0.32	0.46
	GNL sentiment	1560	0.13	0.41

Note: The data corresponds to evaluators from all treatments with a complete submission. GPT and GNL sentiment are on a scale from −1 (negative sentiment) to +1 (positive sentiment).

To visualize the sentiment data, 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 C1 shows the box plot of the GPT sentiment score within the accompanying grade groups. Despite substantial variation within groups, a clear positive relationship exists with the GPT sentiment score.

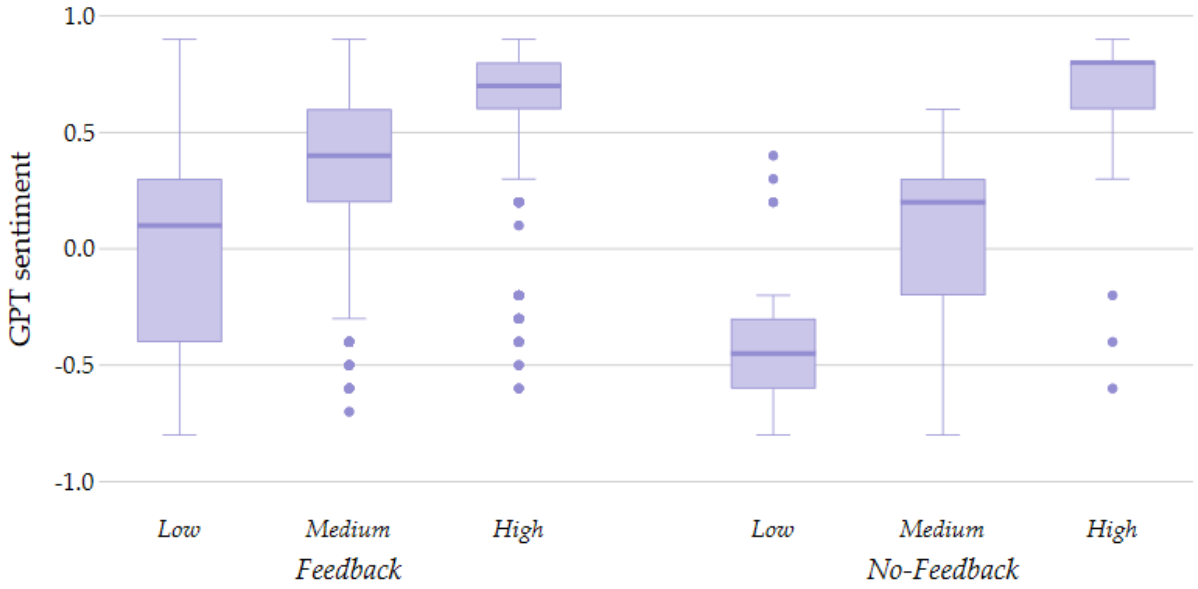


Figure C1. Box plot of the GPT sentiment of the evaluators' text depending on the accompanying grade and whether the text would be shared with writers as feedback

Note: Box plots of the 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). The lower and upper bounds of the box correspond to the first and third quartiles. The points correspond to outliers that exceed 1.5 times the inter-quartile range. The sample consists of evaluators from all treatments ($N = 1437$ for *Feedback* and $N = 123$ for *No-Feedback*).

We next address the concern that the findings of our sentiment analysis may be specific to

the tool we used, OpenAI's GPT. We use an alternative sentiment score from Google Natural Language (GNL) and replicate our main findings. Figure C2 uses the GNL sentiment measure and visually confirms the "kindness" effect (Result 1).

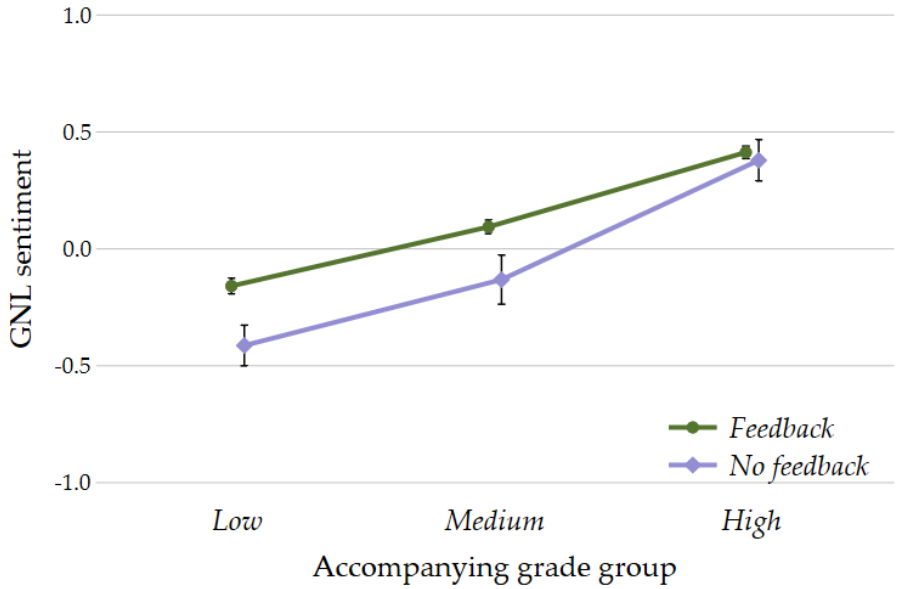
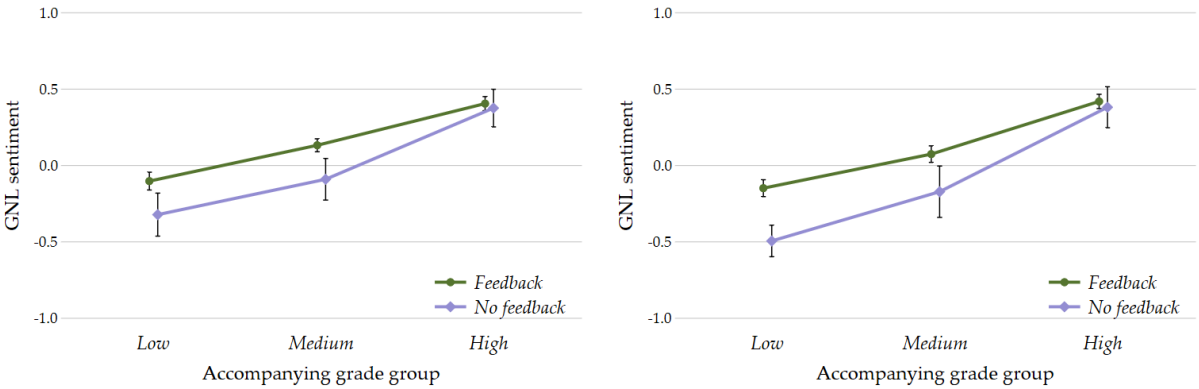


Figure C2. GNL sentiment of the evaluators' text depending on the accompanying grade and whether the text would be shared with writers as feedback

Note: Mean GNL 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 GNL 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*).

Table C4 contains linear regressions of the evaluators' GNL sentiment score on the accompanying grade, the gender of the writer, and whether the evaluator was in the *No-Feedback* or one of the *Feedback* treatments. We replicate the findings of Table 2: namely, GNL sentiment scores of evaluators in *No-Feedback* are more negative than those of evaluators in the *Feedback* treatments, but the gap narrows for higher accompanying grades.

Next, we examine the finding of no gender differences in the feedback given to writers (Result 2). Figure C3 illustrates the mean GNL sentiment scores depending on the writers' gender, the accompanying grade, and whether the alias was visible to evaluators. The kindness effect is seen for both genders. Columns (3) to (5) of Table C4 replicate the same findings in Table 2. There is no statistically significant gender difference in the sentiment of text or the effect of the *No-Feedback* treatment. As seen in column (5), the results are robust to the inclusion of controls for evaluator and essay characteristics and the GNL sentiment of the writer's essay.



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

Figure C3. GNL 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 GNL 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 GNL 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*).

Table C3. GNL sentiment depending on whether the writers' gender is disclosed

	(1)	(2)
Constant	0.05	0.05
	(0.04)	(0.04)
Feedback-Compete-Hidden	−0.08	−0.06
	(0.09)	(0.09)
Accompanying grade	0.60**	0.58**
	(0.06)	(0.06)
Feedback-Compete-Hidden × Female	−0.04	−0.09
	(0.12)	(0.13)
Accompanying grade × Female	−0.11	−0.11
	(0.08)	(0.08)
Essay fixed effects	✓	✓
Controls	-	✓
N	857	857
adj. R^2	0.418	0.415

Note: Linear regressions of the GNL 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 GNL 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, 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 ** $p < 0.01$.

Table C4. GNL sentiment of the evaluators' text

	(1)	(2)	(3)	(4)	(5)
Constant	0.03	0.03	0.04	0.04	0.07
	(0.02)	(0.02)	(0.04)	(0.04)	(0.08)
Accompanying grade	0.60**	0.58**	0.57**	0.58**	0.57**
	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
No-Feedback	−0.36**	−0.38**	−0.44**	−0.46**	−0.50**
	(0.07)	(0.07)	(0.11)	(0.10)	(0.11)
No-Feedback × Accompanying grade		0.24**		0.31**	0.32**
		(0.06)		(0.08)	(0.08)
Female			0.05	0.05	0.02
			(0.05)	(0.05)	(0.05)
No-Feedback × Female			0.10	0.10	0.10
			(0.15)	(0.14)	(0.14)
Accompanying grade × Female				−0.07	−0.06
				(0.05)	(0.05)
No-Feedback × Accompanying grade				−0.08	−0.09
× Female				(0.13)	(0.13)
Essay GPT sentiment					0.04
					(0.03)
Controls	-	-	-	-	✓
N	1560	1560	1124	1124	1124
adj. R^2	0.368	0.372	0.347	0.354	0.360

Note: Linear regressions of the GNL 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 GNL sentiment is the GNL 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. All continuous variables—the GNL sentiment score, the accompanying grade, and the essay GNL 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 errors in the essay, 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 ** $p < 0.01$.

As mentioned in the paper, we can utilize a feature of our experimental design that enables us to isolate the effect of a writer's gender being disclosed to the evaluators. In *Feedback-Compete*, we disclosed the writer's alias to the evaluators, whereas in *Feedback-Compete-Hidden*, the same essays were shown to evaluators without the alias disclosed. Hence, we can control for the essay and estimate the effect of the alias being disclosed. Table C3 contains the linear regressions of the GNL sentiment score on treatment indicators, the accompanying grade, and their interactions with the writer's gender. We include essay fixed effects and standardize both the sentiment scores and the accompanying grades. Column (2) also controls for evaluator characteristics (see footnote 16). Regressions are restricted to writers in *Feedback-Compete* and *Feedback-Compete-*

Hidden treatments. We replicate the findings from Table 3. Namely, there are no statistically significant differences in the sentiment of feedback when the writer's gender is disclosed.

C.3. A simple model of belief-updating

In this section, we describe a simple model of belief-updating in which an individual responds positively to good news, negatively to bad news, and neutrally to neutral news.

A writer composed an essay that was graded by ten evaluators, each of whom assigned 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. Thus, $\theta = (0.07, 0.08, 0.15, 0.60, 0.10)$ indicates a high-quality essay that will most likely be graded a 4 but with elements that could result in a grade 1 with a 7% chance, 2 with an 8% chance, and so forth. Alternatively, grade dispersion could be due to idiosyncrasies of the graders.

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

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

Thus, 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

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

Suppose a writer receives written feedback from the first evaluator and correctly infers that $g_1 = k$. The Dirichilet has the property that, after observing a grade draw of k, the posterior of $\pi(x)$ is

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

Hence, the writer's updated mean belief is

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

and the writer's updated belief of their average grade is

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

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 is greater following k' than k.

C.4. Reactions to Feedback

Figure C4 shows the grade belief adjustments of individual writers depending on their accompanying grade group. Writers are labeled according to the gap between their accompanying grade group and prior-belief group. Red lines correspond to writers who got bad news: their accompanying grade group is below their prior-belief group. Lavender lines correspond to writers who got neutral news: their accompanying grade group equals their prior-belief group. Green lines correspond to writers who got good news: their accompanying grade group is above their prior-belief group. Note that the figure does not convey the density of writers with the same accompanying grade group, prior, and posterior. The majority of belief adjustments align with the news received. For example, 66.9% of writers who received bad news adjust their belief downward, while 75.6% of writers who received good news adjust their belief upward.

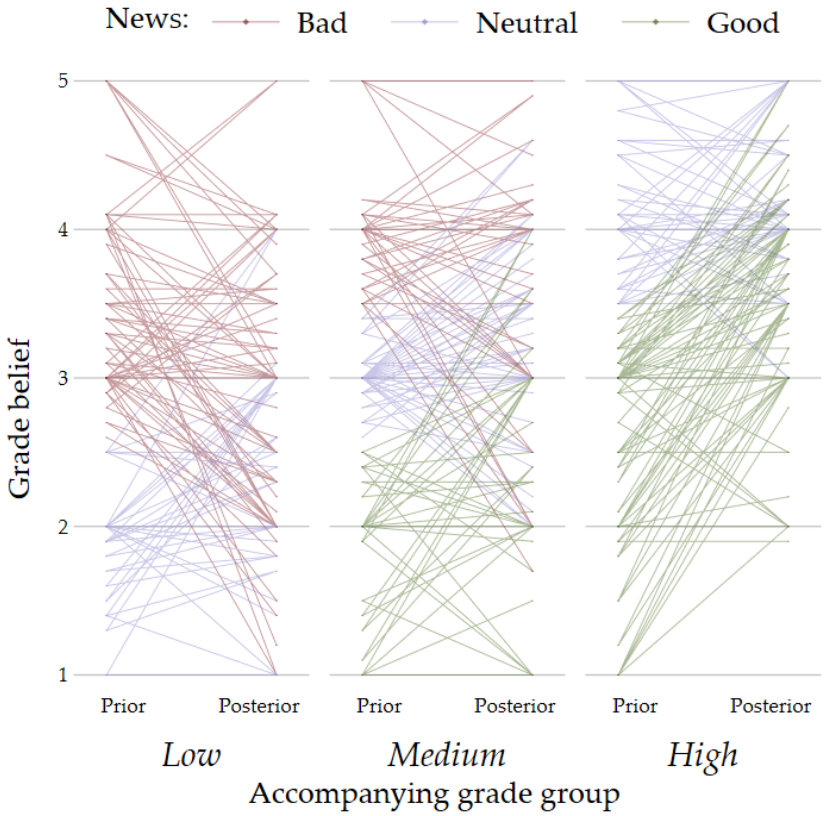


Figure C4. Individual belief adjustment in response to different types of feedback

Note: Individual writers' prior and posterior beliefs depending on the accompanying grade group: *Low* (grades 1 and 2), *Medium* (grade 3), and *High* (grades 4 and 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. The prior-belief group is *Low* for priors in the range $[1, 2.5]$, *Medium* for those in the range $(2.5, 3.5)$), and *High* for those in the range $[3.5, 5]$. The sample consists of writers in the *Feedback-Only*, *Feedback-Compete*, and *Feedback-Compete-Hidden* treatments ($N = 561$).

As mentioned in the paper, we examine whether writers correctly interpret qualitative feedback by estimating regression of the form $\mu_i^1 = \beta_1 \mu_i^0 + \beta_2(g_i - \mu_i^0) + \gamma X_i + \epsilon_i$, 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. Note that $g_i - \mu_i^0 > 0$ indicates good news, $g_i - \mu_i^0 < 0$ bad news, and $g_i - \mu_i^0 = 0$ neutral news. Hence, if writers correctly identify good from bad news and update beliefs in the right direction, then β_2 should be positive. Moreover, if writers recognize when they receive neutral news, then their posterior belief should equal their prior belief, implying $\beta_1 = 1$. Table C5 contains the regression results. Column (1) corresponds to the regression described above, estimated with all writers from the *Feedback-Only*, *Feedback-Compete*, and *Feedback-Compete-Hidden* treatments. Column (2) additionally controls for writer and essay characteristics (see footnote 20). Columns (3) and (4) restrict the sample to solely female or male writers, respectively. The coefficients of these regressions are depicted graphically in Figure 6. In all regressions, the coefficient of the grade-prior gap (β_2) is positive and statistically significant ($p < 0.01$), and the coefficient of the prior grade belief (β_1) is very close to 1. Moreover, when we estimate the regressions separately for women and men, we find very similar coefficients. If we use seemingly unrelated estimation to compare these coefficients across regressions (White, 1994), we find they are statistically indistinguishable across genders ($p = 0.83$ for β_1 and $p = 0.35$ for β_2).

Table C5 also contains regressions to evaluate what would be the ideal belief-updating: namely, adjusting beliefs such that the posterior matches the actual final grade. To do this, we re-estimate the same regressions but we use the writer's final grade as the dependent variable instead of their posterior belief. Mirroring the previous regressions, column (5) corresponds to the regression without controls, column (6) adds controls for writer and essay characteristics, column (7) restricts the sample to female writers, and column (8) to male writers. These coefficients are depicted graphically in Figure 7.

Comparing the coefficients of the grade-prior gap across regressions suggests that, on average, writers underreact to feedback and fail to correct for their initial overestimation of their performance. Using seemingly unrelated estimation to test coefficients across regressions, we find that the coefficient of the grade-prior gap is significantly smaller in column (1) compared to column (5) ($p = 0.02$) and is close to being statistically smaller in column (2) compared to column (6) ($p = 0.10$). Conversely, the coefficients of the prior grade belief tend to be smaller for ideal updating compared to observed updating (with and without controls, $p < 0.01$), and significantly lower than 1 when controls are included ($p < 0.01$ with controls and $p = 0.26$ without). Comparing columns (3) and (7) suggests that female writers underreact to feedback relative to the ideal, with the difference being close to statistical significance ($p = 0.08$), but they place the appropriate weight on their prior grade beliefs ($p = 0.30$). Comparing columns (4) and (8) suggests that male writers' reaction to feedback is close to the ideal ($p = 0.71$), but

they place too much weight on their prior grade beliefs ($p < 0.01$).

Table C5. Observed and ideal grade belief-updating

	Observed				Ideal			
	All		Female	Male	All		Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Grade-prior gap (accompanying	0.46**	0.45**	0.43**	0.46**	0.53**	0.49**	0.50**	0.47**
grade – prior grade belief)	(0.02)	(0.03)	(0.03)	(0.04)	(0.02)	(0.02)	(0.03)	(0.03)
Prior grade belief	1.02**	0.99**	0.99**	1.00**	0.99**	0.92**	0.96**	0.90**
	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.03)	(0.02)
Controls	-	✓	✓	✓	-	✓	✓	✓
N	561	561	278	283	561	561	278	283
adj. R^2	0.957	0.957	0.953	0.960	0.960	0.965	0.964	0.963

Note: Estimated coefficients from linear regressions. In columns (1) to (4), the dependent variable is the writer's posterior grade belief. In columns (5) to (6), the dependent variable is the writer's final grade. 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. The precise specification is described in footnote 19. Columns (1) and (5) do not include additional independent variables. All other columns include controls for writer and essay characteristics. Columns (3) and (7) restrict the sample to only female writers, and columns (4) and (8) to only male writers. The sample is restricted to writers in the *Feedback-Only*, *Feedback-Compete*, and *Feedback-Compete-Hidden* treatments. Controls include the writers' age, level of education, ethnic identity, gender, whether English is their native language, whether they grew up in the UK, their treatment assignment, the presence of spacing errors in their essay or feedback, and the number of characters in their feedback. Robust standard errors in parentheses and statistical significance of non-zero coefficients indicated by * $p < 0.05$ and ** $p < 0.01$.

Next, we relax the assumption that writers respond symmetrically to good and bad news. Table C6 presents regression results where the dependent variable is the writer's posterior grade belief. As before, we include the writer's prior grade belief and the grade-prior gap as independent variables. To test for asymmetric updating, we introduce two dummy variables: 'Bad news,' which equals one when the accompanying grade is lower than the writer's prior belief, and 'Good news,' which equals one when the accompanying grade is higher. We interact each dummy variable with the grade-prior gap to allow belief-updating to differ based on the direction of the news. The regression in column (1) does not include other covariates, while that in column (2) includes controls for writer and essay characteristics.

We find that writers' response to good news is somewhat stronger than to bad news. Directionally, these findings are consistent with papers that uncover a positive asymmetry when updating to quantitative feedback (Möbius et al., 2022; Zimmermann, 2020; Eil and Rao, 2011). However, in our case, the difference between the two coefficients is not statistically significant (Wald tests, $p = 0.22$ without controls and $p = 0.21$ with controls). Hence, we cannot reject that belief-updating is symmetric, at least in the context of qualitative feedback

Table C6. Observed grade belief-updating with differential responses to good and bad news

	(1)	(2)
Bad news	0.09	0.08
	(0.10)	(0.10)
Good news	0.10	0.09
	(0.11)	(0.11)
Bad news × Grade-prior gap (accompanying grade – prior grade belief)	0.39**	0.37**
	(0.06)	(0.06)
Good news × Grade-prior gap (accompanying grade – prior grade belief)	0.50**	0.49**
	(0.07)	(0.07)
Prior grade belief	0.98**	0.96**
	(0.02)	(0.02)
Controls	-	✓
N	561	561
adj. R^2	0.957	0.957

Note: Estimated coefficients from linear regressions with the writer's posterior grade belief as the dependent variable. As independent variables, we use the writer's prior grade belief, a dummy variable called 'Bad news' indicating that the feedback's accompanying grade is lower than the prior grade belief, a dummy variable called 'Good news' indicating that the feedback's accompanying grade is higher than the prior grade belief, and the interaction of these dummies with the writer's grade-prior gap (i.e., the difference between the feedback's accompanying grade and the writer's prior grade belief). Column (1) does not include additional covariates, while column (2) includes controls for writer and essay characteristics. The sample is restricted to writers in the *Feedback-Only*, *Feedback-Compete*, and *Feedback-Compete-Hidden* treatments. Controls include the writers' age, level of education, ethnic identity, gender, whether English is their native language, whether they grew up in the UK, their treatment assignment, the presence of spacing errors in their essay or feedback, and the number of characters in their feedback. Robust standard errors in parentheses and statistical significance of non-zero coefficients indicated by * $p < 0.05$ and ** $p < 0.01$.

C.5. Competition

First, we evaluate whether the results concerning the encouragement channel are sensitive to the sentiment score used to identify it. Table C7 reproduces the regressions used in Table 4 using GNL sentiment scores instead of the GPT sentiment scores. In all regressions, the dependent variable is a binary indicator equal to one if the writer chose to compete. Column (1) reproduces the regression in column (3) of Table 4 using the GNL sentiment score as the only independent variable. Column (2) reproduces the regression in column (5) of Table Table 4, which includes the writers' posterior grade belief along with the GNL sentiment score as independent variables. Finally, column (3) reproduces the regression in column (7) of Table 4, which adds controls for writer and essay characteristics. The GNL sentiment score very closely replicates the estimates of the GPT sentiment score. In particular, both the posterior belief and the sentiment score are positive and statistically significant when considered together, suggesting that feedback tone has an impact on the decision to compete beyond its impact on beliefs.

Table C7. Effects of feedback on the choice to compete with GNL sentiment

	(1)	(2)	(3)
Constant	0.68**	0.68**	0.68***
	(0.02)	(0.02)	(0.02)
Posterior grade belief		0.19**	0.19**
		(0.02)	(0.02)
GNL sentiment	0.21**	0.12**	0.11***
	(0.02)	(0.02)	(0.02)
Final grade			0.02
			(0.02)
Controls	-	-	✓
N	377	377	377
adj. R^2	0.201	0.325	0.333

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. GNL sentiment is the GNL 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 errors in the essay, and the number of characters in the essay. 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 Table C8 we test the robustness of the encouragement channel by allowing for nonlinearity in the belief channel. If the linear specification does not capture the relationship between posterior beliefs and the decision to compete well, then the feedback variables may simply be capturing these non-linear effects. A similar argument can be made for the feedback variables picking up error in the measurement of posterior beliefs (Gillen et al., 2019). Specifically, instead of including the posterior belief as a continuous variable, we include dummy variables for each possible posterior grade belief, which ranged from 1 to 5 in increments of one decimal place. In addition to the posterior belief, we include the feedback's accompanying grade in column (1), the GPT sentiment score in column (2), and the GNL sentiment score in column (3). All regressions include controls for writer and essay characteristics. Table C8 shows that controlling flexibly for the posterior grade belief does not affect the magnitude or statistical significance of the coefficients of the accompanying grade or the GPT and GNL sentiment scores. These results suggest that the encouragement channel is not the result of misspecification or measurement error in the posterior grade beliefs.

Table C8. Effects of feedback on the choice to compete controlling flexibly for beliefs

	(1)	(2)	(3)
Constant	0.42**	0.48**	0.47**
	(0.15)	(0.15)	(0.14)
Accompanying grade	0.08**		
	(0.03)		
GPT sentiment		0.11**	
		(0.03)	
GNL sentiment			0.11**
			(0.02)
Final grade	0.04	0.04	0.04
	(0.03)	(0.03)	(0.03)
Posterior belief fixed effects	✓	✓	✓
Controls	✓	✓	✓
N	377	377	377
adj. R^2	0.358	0.376	0.380

Note: Linear regressions where the dependent variable equals one if the writer chose the competitive payment scheme and zero otherwise. The accompanying grade is the grade assigned by the evaluator who wrote the feedback. GPT and GNL sentiment refer to the sentiment scores of the feedback's text, as determined by the GPT and GNL APIs. 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. The posterior belief fixed effects correspond to dummy variables for each possible posterior grade belief, which ranged from 1 to 5 in increments of one decimal place. 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 errors in the essay, and the number of characters in the essay. 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 ** $p < 0.01$.

Gendered beliefs about performance

In the final questionnaire of the *Feedback-Compete* and *Feedback-Compete-Hidden* treatments, we asked writers to predict whether women or men performed better in the essay task by asking them "On average, do you think men or women obtained a better final grade?" The possible answers , which are abbreviated in the figure, were "Women obtained a much better final grade than men," "Women obtained a slightly better final grade than men," "Women and men obtain equal final grades," and "Men obtained a slightly better final grade than women," and "Men obtained a much better final grade than women." Figure C5 plots the distribution of answers depending on the writers' gender. The two most common answers were "Women and men obtain equal final grades" and "Women obtained a slightly better final grade than men." In other words, both female and male writers think women perform better in this task.

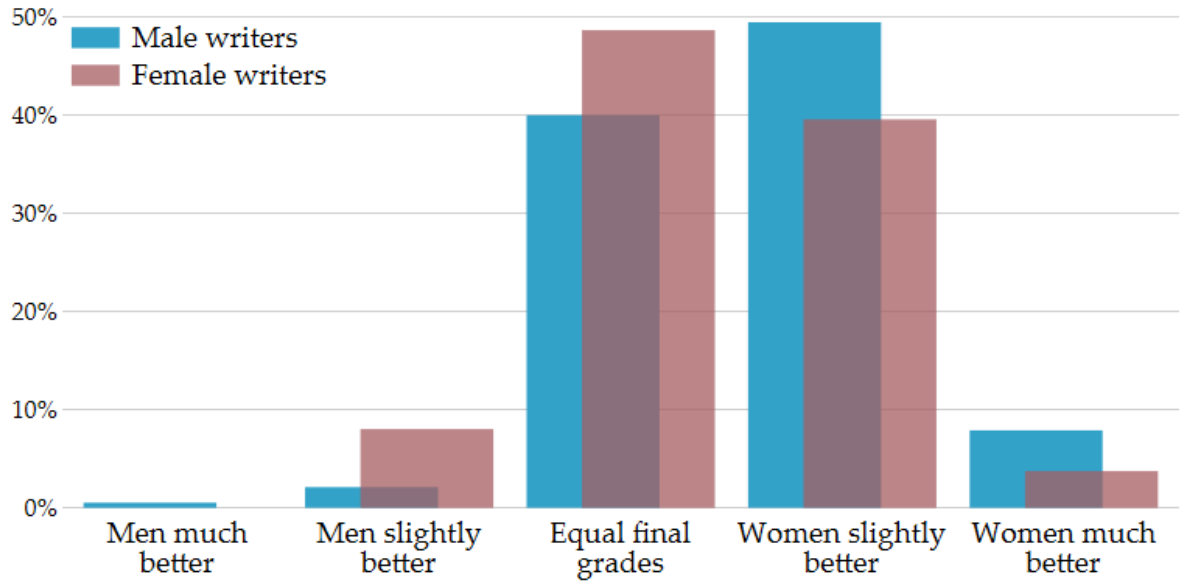


Figure C5. Beliefs about which gender performs better by the writers' gender

Note: Histogram of the writers' responses to the question "On average, do you think men or women obtained a better final grade?" depending on the respondents' gender. The sample consists of writers from the *Feedback-Compete* and *Feedback-Compete-Hidden* treatments ($N = 377$).

Competition error rates

As discussed in Section 4.3. of the paper, writers can make two errors in their choice to compete: competing when they should not, a false positive, and not competing when they should, a false negative. We determine these error rates by estimating the probability that any particular essay would end up in the top three. Given an essay, we randomly draw nine other essays from the sample of 900 and rank them by their final grade. We repeat this procedure, drawing with replacement 10,000 times to arrive at the probability of a top-three placement.

To estimate the impact of the encouragement channel, we construct two counterfactual predictions of the decision to compete. We regress the choice to compete on the writers' posterior grade belief and the GPT sentiment of their feedback text (i.e., column (5) in Table 4). With this regression, we can predict each writer's probability of competing given both the belief and encouragement channels. Next, we estimate this same probability but using only the posterior grade belief as the independent variable (i.e., column (1) in Table 4), which accounts only for the belief channel. Finally, with the estimated probability of a top-three placement, we construct an indicator variable Γ that equals 1 if a writer's probability of a top-three placement is greater than 30% and 0 if it is less than 30%.[A2] For each prediction of competing p we calculate the conditional mean probability of competing when you should not, i.e. $\mathbb{E}[p|\Gamma = 0]$, and the conditional mean probability of not competing when you should, $\mathbb{E}[1 - p|\Gamma = 1]$. Then, for a

[A2] A risk-neutral writer is indifferent between competing or not with a probability of exactly %30. In our sample, none of the estimated probabilities equaled exactly %30.

false positive error (competing when you should not), we compute the difference in the mean probability of competing p, both with and without the impact of the encouragement channel. If the mean probability of competing is lower with the encouragement channel than without, this suggests that the encouragement channel helps reduce this type of error. We repeat the procedure for a false negative error (not competing when you should) and the mean probability of not competing $1-p$. As robustness checks, we utilize other variables that capture the encouragement channel: the GNL sentiment of the feedback text and the unseen grade accompanying the feedback. Table C9 contains the results of this analysis.

The presence of the encouragement channel helps reduce the likelihood of committing both types of errors. For example, row (A) contains the mean probability of competing for those who are better off not competing when we exclude the encouragement channel, 62.3%. If we consider the encouragement channel as captured by the GPT sentiment, row (B), we see a statistically significant reduction in the likelihood of making a false positive error of 1.3 percentage points ($p < 0.05$). For the false negative error, we find that the encouragement channel significantly reduces it by 2.4 percentage points ($p < 0.01$). These results suggest that the content of qualitative feedback is useful to writers in reducing the likelihood of these two types of errors.

Table C9. The effect of feedback on error types for the competition choice

			Difference with the probability in (A)	
	False Positive (1)	False Negative (2)	False Positive (3)	False Negative (4)
(A) Posterior grade belief	62.3	22.0		
(B) GPT sentiment	61.0	19.6	1.3*	2.4**
(C) Accompanying grade	60.1	19.0	2.2**	3.0**
(D) GNL sentiment	60.6	20.0	1.7**	2.0**

Note: The effect of the encouragement channel on the likelihood of making false positive and false negative errors with respect to the choice to compete. Column (1) contains estimates of the mean probability of competing for writers who commit a false positive error (competing with a less than 30% chance of placing in the top three). The estimate of row (A) is based on the regression in column (1) of Table 4. The estimates of rows (B), (C), and (D) are based on the regressions in columns (5), (6), and (7) of Table 4. Column (2) contains the mean probability of not competing for writers who commit a false negative error (not competing with a greater than 30% chance of placing in the top three). Column (3) contains the difference in the mean probability of competing between row (A) column (1) and individually each row (B) to (D). Column (4) contains the difference in the mean probability of not competing between row (A) column (2) and individually each of row (B) to (D). For columns (3) and (4), a positive value indicates that including the encouragement channel reduces the likelihood of committing a particular error. 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 (GNL) sentiment is the GPT (GNL) sentiment score of the feedback's text. Statistical significance of non-zero coefficients is indicated by * $p < 0.05$ and ** $p < 0.01$. The sample consists of writers from the *Feedback-Compete* and *Feedback-Compete-Hidden* treatments ($N = 377$).

We also look at this by gender. We follow the same procedure as above, but we use the regressions in Table 4, which estimate separate coefficients by gender. Table C10 contains the results split by gender. We find that for all three measures, female and male writers make fewer errors of both types with the inclusion of the encouragement channel. However, for some of the male estimates, we can not rule out that there is no statistically significant effect. Furthermore, when comparing the gender difference in the benefit for the two types of errors, female writers benefit more than male writers. This is consistent with the encouragement channel playing a greater role for female writers.

Table C10. The effect of feedback on error types for the competition choice by gender

				Difference with (A)		Gender difference	
	Gender	False Positive (1)	False Negative (2)	False Positive (3)	False Negative (4)	False Positive (5)	False Negative (6)
(A) Posterior grade belief	Female	70.0	19.4				
	Male	55.1	24.8				
(B) GPT sentiment	Female	67.6	15.7	2.5*	3.6**	+2.2	+2.6
	Male	54.8	23.7	0.2	1.1		
(C) Accompanying grade	Female	66.5	15.1	3.5**	4.2**	+2.5	+2.4
	Male	54.1	22.9	1.0*	1.8**		
(D) GNL sentiment	Female	66.9	16.3	3.1*	3.0*	+2.7	+2.0
	Male	54.6	23.7	0.4	1.0		

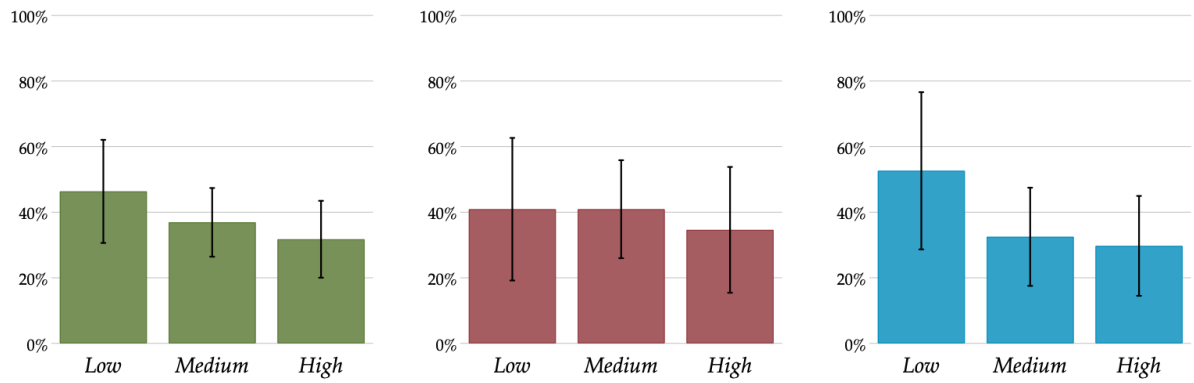
Note: The effect of the encouragement channel on the likelihood of making false positive and false negative errors with respect to the choice to compete by writer gender. Column (1) contains estimates of the mean probability of competing for writers who commit a false positive error (competing with a less than 30% chance of placing in the top three). The estimate of row (A) is based on the regression in column (1) of Table 5. The estimates of rows (B) and (C) are based on the regressions in columns (3) and (2) of Table 5, and those of (D) of an equivalent regression using the GNL sentiment score. Column (2) contains the mean probability of not competing for writers who commit a false negative error (not competing with a greater than 30% chance of placing in the top three), split by writer gender. Column (3) contains the difference in the mean probability of competing between row (A) column (1) and individually each row (B) to (D). Column (4) contains the difference in the mean probability of not competing between row (A) column (2) and individually each of row (B) to (D). For columns (3) and (4), a positive value indicates that including the encouragement channel reduces the likelihood of committing a particular error. Columns (5) and (6) indicate the gender differences in the differences of columns (3) and (4) respectively, calculated as the female difference minus the male difference, with positive values indicating that females are predicted to make fewer errors with the inclusion of the encouragement channel. 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 (GNL) sentiment is the GPT (GNL) sentiment score of the feedback's text. Statistical significance of non-zero coefficients is indicated by * $p < 0.05$ and ** $p < 0.01$. The sample consists of writers from the *Feedback-Compete* and *Feedback-Compete-Hidden* treatments ($N = 377$).

C.6. Editing

In all the tables of this section, any variables generated with NLP methods have been applied to the clean feedback text (see Section D for details).

In this section, we analyze the choice to edit. For the choice to compete, the worse a writer believes they performed, the less likely they are to compete. For the editing decision, the comparative static is not clear-cut. Suppose a writer believes they performed badly. On the one hand, they may want to improve their essay by editing; on the other hand, they might believe they are simply poor writers, so that editing would not help. There is no natural hypothesis to make regarding the relationship between how a writer believed they performed and their editing choice.

We cannot directly empirically examine this relationship, since those who chose to edit were asked to predict the grade of their edited essay, but not the grade of their original essay. However, we can use the regression described in footnote 19 to infer the writers' posterior grade beliefs about their original essays, based on their prior grade beliefs and the gap between the accompanying grade and their prior grade.[A3] Figure C6 plots the percentage of those who edited against the inferred posterior grade belief buckets, both overall and by gender. We see a downward trend, but the error bands indicate that the relationship is not statistically significant.



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

Figure C6. The percentage of writers who chose to edit depending on their inferred posterior grade belief

Note: Bar graphs of the percentage of writers who choose to edit their essay depending on their inferred posterior grade belief. Inferred posterior grades are estimated using the coefficients of the belief-updating regression (see footnote 19) using writer observations from *Feedback-Only*, *Feedback-Compete*, and *Feedback-Compete-Hidden*. For each writer in *Feedback-Edit*, we predict their posterior based on these coefficients and the observed values of their prior grade belief and the difference between the accompanying grade and this prior. Each bar plots the fraction of writers who edit when their inferred posterior grade belief is *Low*, in the range $[1, 2.5]$, *Medium*, in the range $(2.5, 3.5)$, or *High*, in the range $[3.5, 5]$. Error bars indicate 95% confidence intervals. The sample consists of writers in the *Feedback-Edit* treatment ($N = 188$).

[A3] The coefficients are estimated from the writer data of treatments: *Feedback-Only*, *Feedback-Compete*, and *Feedback-Compete-Hidden*.

Table C11. Possible determinants of the choice to edit

	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.37**	0.37**	0.37**	0.37**	0.37**	0.37**
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Inferred posterior grade belief	−0.04					
	(0.03)					
GPT sentiment		0.00				
		(0.03)				
GNL sentiment			−0.06			
			(0.03)			
Prior grade belief				−0.03		
				(0.03)		
Final grade					−0.01	
					(0.04)	
Accompamying grade						−0.03
						(0.03)
N	188	188	188	188	188	188
adj. R^2	0.002	−0.005	0.008	−0.002	−0.005	−0.002

Note: Linear regressions where the dependent variable equals one if the writer chose to edit their essay. Inferred posterior grades are estimated using the coefficients of the belief-updating regression (see footnote 19) using writer observations from *Feedback-Only*, *Feedback-Compete*, and *Feedback-Compete-Hidden*. For each writer in *Feedback-Edit*, we predict their posterior based on these coefficients and the observed values of their prior grade belief and the difference between the accompanying grade and this prior. GPT and GNL sentiment refer to the sentiment scores of the feedback's text, as determined by the GPT and GNL APIs. Prior grade beliefs are the writers' prior beliefs. Final grade is the average grade given to the writer by all evaluators. The accompanying grade is the grade assigned by the evaluator who wrote the feedback. All dependent variables are standardized to have a mean of zero and a standard deviation of one. The sample consists of writers in the *Feedback-Edit* treatment. Robust standard errors in parentheses and statistical significance of non-zero coefficients is indicated by * $p < 0.05$ and ** $p < 0.01$.

In Table C11, we use a linear probability model to analyze the edit decision. The independent variables are standardized with a mean of zero and a standard deviation of one. In column (1), we use only the inferred posterior grade belief. The estimate is negative, in line with Figure C6a, but the magnitude is small and the coefficient is not statistically significant. In column (2), we use only the GPT sentiment of the feedback text. We find no statistically significant relationship between the sentiment and the choice to edit. Given this null result, could it be that we are underpowered to detect any effects? Since there was no prior literature to inform power calculations, we employ an ex-post power analysis for the minimum detectable effect size (Dupont and Plummer, 1998), assuming particular parameter values for $n = 188$. We use the standard of detecting 80% of true effects and the default value of 1 for standard deviation. The effect size δ for a linear regression is defined as the difference between the alternative and null values of the slope multiplied by the ratio of the standard deviations of the covariate to the error term. With assumptions, we estimate a $\delta = 0.21$. Given our coefficient estimates in Table

C11, it is possible that our study is underpowered, suggesting that further research is needed to determine how the content of the feedback affects this particular decision.

Table C12. Possible determinants of the choice to edit by writer gender

	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.36**	0.35**	0.36**	0.36**	0.35**	0.35**
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Female	0.03	0.04	0.03	0.03	0.04	0.04
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
Inferred posterior grade belief	−0.07					
	(0.05)					
Inferred posterior grade belief	0.06					
× Female	(0.07)					
GPT sentiment		0.00				
		(0.05)				
GPT sentiment × Female		−0.01				
		(0.07)				
GNL sentiment			−0.03			
			(0.05)			
GNL sentiment × Female			−0.05			
			(0.07)			
Prior grade belief				−0.04		
				(0.05)		
Prior grade belief × Female				0.02		
				(0.07)		
Final grade					−0.02	
					(0.05)	
Finale grade × Female					0.03	
					(0.07)	
Accompanying grade						−0.06
						(0.05)
Accompanying grade × Female						0.08
						(0.07)
N	188	188	188	188	188	188
adj. R^2	−0.004	−0.015	0.001	−0.011	−0.014	−0.005

Note: Linear regressions where the dependent variable equals one if the writer chose to edit their essay. Female is a dummy taking the value one if the writer was female. Inferred posterior grades are estimated using the coefficients of the belief-updating regression (see footnote 19) using writer observations from *Feedback-Only*, *Feedback-Compete*, and *Feedback-Compete-Hidden*. For each writer in *Feedback-Edit*, we predict their posterior based on these coefficients and the observed values of their prior grade belief and the difference between the accompanying grade and this prior. GPT and GNL sentiment refer to the sentiment scores of the feedback's text, as determined by the GPT and GNL APIs. Prior grade beliefs are the writers' prior beliefs. Final grade is the average grade given to the writer by all evaluators. The accompanying grade is the grade assigned by the evaluator who wrote the feedback. All continuous dependent variables are standardized to have a mean of zero and a standard deviation of one. The sample consists of writers in the *Feedback-Edit* treatment. Robust standard errors in parentheses and statistical significance of non-zero coefficients is indicated by * $p < 0.05$ and ** $p < 0.01$.

Table C12 shows the results of the same regressions including a gender dummy interacted with all other dependent variables. There are no significant gender differences. These findings contrast sharply with our results on the choice to compete, where several statistically significant coefficients are observed. This contrast is perhaps unsurprising, given that a higher grade makes competing more attractive but has no clear implication for editing.

Table C13 presents results from linear regressions where the dependent variable is the change in final grade: the difference between the new (regraded) and original final grades. Since previous work has found that feedback is more effective when it is more concrete (see Yeomans, 2021), we used GPT-3.5 to generate a concreteness score for each feedback. The precise prompt is available in footnote 25. Column (1) includes this concreteness score and its interaction with the editing decision. The coefficient of the interaction between GPT Concreteness and Edited indicates that, among those who edited, each standard deviation increase in the concreteness

Table C13. Relationship between grade performance, editing, and feedback

	(1)	(2)	(3)	(3)
Constant	0.01	0.00	0.01	0.01
	(0.04)	(0.05)	(0.04)	(0.05)
Edited	0.17**	0.19*	0.16*	0.19*
	(0.06)	(0.09)	(0.06)	(0.09)
GPT concreteness	−0.06	−0.04	−0.07	−0.05
	(0.04)	(0.05)	(0.04)	(0.05)
GPT concreteness × Edited	0.13*	0.10	0.15*	0.11
	(0.06)	(0.08)	(0.07)	(0.08)
Female		0.01		0.01
		(0.08)		(0.08)
Edited × Female		−0.03		−0.04
		(0.13)		(0.13)
GPT concreteness × Female		−0.05		−0.07
		(0.09)		(0.09)
GPT concreteness × Edited × Female		0.06		0.10
		(0.13)		(0.14)
Controls	-	-	✓	✓
N	188	188	188	188
adj. R^2	0.043	0.024	0.059	0.042

Note: Linear regressions where the dependent variable is the difference between the new (regraded) and original final grades. Edited is a dummy variable indicating the writer chose to edit their essay. GPT concreteness is generated by asking GPT-3.5 "How concrete is the advice in this text?" in reference to the feedback's text (see footnote 25 for the detailed prompt). Concreteness scores are standardized to have a mean of zero and a standard deviation of one. Female is a dummy variable indicating the writer's gender is female. Controls include the writers' age, ethnic identity, gender, level of education, whether English is their native language, whether they grew up in the UK, the presence of spacing errors in the essay, and the number of characters in the essay. The sample consists of writers in the *Feedback-Edit* treatment. Robust standard errors in parentheses and statistical significance of non-zero coefficients is indicated by * $p < 0.05$ and ** $p < 0.01$.

Appendix D. Text Analysis

To run the sentiment text analysis we used the feedback text data without any of the unforced spelling errors (see Section 3.3. for details). We pre-processed the text data before conducting the sentiment analysis with the following steps. We normalized hyphenated words such as *miss-spelled* to *misspelled*. We converted numerical digits to string characters e.g. 1 to one. In the feedback text we often find that evaluators, to aid the point they were making or to indicate grammatical errors, quoted a passage directly from the essay they were grading. To ensure the sentiment analysis is only capturing the sentiment of evaluators own words and not that of the writer, we removed all text between quotation marks in the feedback text. For the same reason, we also removed a word if it was misspelled and present in the essay and feedback text. We also analysed the sentiment of the essay text. This allows us to control for the sentiment of the essay text which could influence the sentiment of the feedback text.

D.1. OpenAI GPT: Sentiment analysis

We analysed the sentiment of the text using a GPT of OpenAI. With the introduction of the high performing GPT-3.5 in 2022 the ability to generate bespoke machine learning text analysis has become accessible to social scientists. GPT is a large language model with a neural network architecture. Previously, to use such a model for text analysis required specialized knowledge to build the neural network architecture and vast quantities of data to train the neural network. GPT version 3.5 and 4 have been shown to work well on a number of human-like tasks e.g. the bar exam (Katz et al., 2024) and constructing psychological measures (Rathje et al., 2024). For each feedback text, GPT-3.5 to construct a sentiment measure of the text. For each text we used GPT-3.5 to generate a sentiment score $\in [-1, 1]$, where negative scores indicate negative sentiment and positive scores indicate positive sentiment. We refer to this sentiment score as GPT sentiment. Since OpenAI are continuously updating their model, for ease of replication we used a snapshot of GPT-3.5 taken on the 1st of March 2023. In the documentation this is referred to as *gpt-3.5-turbo-0301*.

D.2. Google Natural Language: Sentiment analysis

Google Natural Language API is a pre-trained machine learning model with a neural network architecture, which allows users to run NLP tasks such as sentiment analysis or entity detection. For each feedback text the model generates a sentiment score $\in [-1, 1]$, where negative scores indicate negative sentiment and positive scores indicate positive sentiment. The absolute value of the score indicates the strength of the sentiment. We used Google cloud version 2.8.1.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Benoît, J.-P., Perry, A., and Reuben, E. (2025). Performance-feedback. Working Paper.

Dupont, W. D. and Plummer, W. D. (1998). Power and sample size calculations for studies involving linear regression. *Controlled Clinical Trials*, 19(6):589–601.

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.

Gillen, B., Snowberg, E., and Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the Caltech Cohort Study. *Journal of Political Economy*, 127(4):1826–1863.

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.

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.

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.

Santamaría, L. and Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, 4:e156.

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

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.