Online appendices for: Performance-Feedback

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

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 feedback conditions for 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 7 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 display 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 display 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 display error affected 3.0% of the words and was present in 83% of the feedback. We believe that computer-generated spacing display 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 display errors. Therefore, for writers who did not edit their essay, we have an observation with the spacing display 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 display 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	
	31-50	0.49	0.50	0.48	0.50	0.49	0.50	0.01	0.796
	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	
	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	0.127
	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	
	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	0.910
	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 display 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 display 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 display 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 display errors. 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 display error in the feedback				0.81	0.39	0.85	0.36	0.82	0.39	0.85	0.36	0.601
Spacing display 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 display 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 display 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 display errors in the essay they graded. This difference remains significant after adjusting p-

values with the Benjamini-Hochberg method for multiple comparisons. We find that controlling for the presence of spacing display 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 display 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 21. 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) and shows that the length of the feedback by treatment is quite similar.

Table C2. Feedback statistics of evaluators

Treatment	Variable	N	mean	s.d.
All	GPT sentiment	1560	0.32	0.46
All	GNL sentiment	1560	0.13	0.41
No-Feedback		123	523.10	250.38
Feedback-Only		241	511.23	239.12
Feedback-Compete	Feedback Length	421	501.04	303.22
Feedback-Compete-Hidden		436	511.57	322.39
Feedback-Edit		339	545.26	334.58

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). Feedback length represents the number of characters of the text.

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.

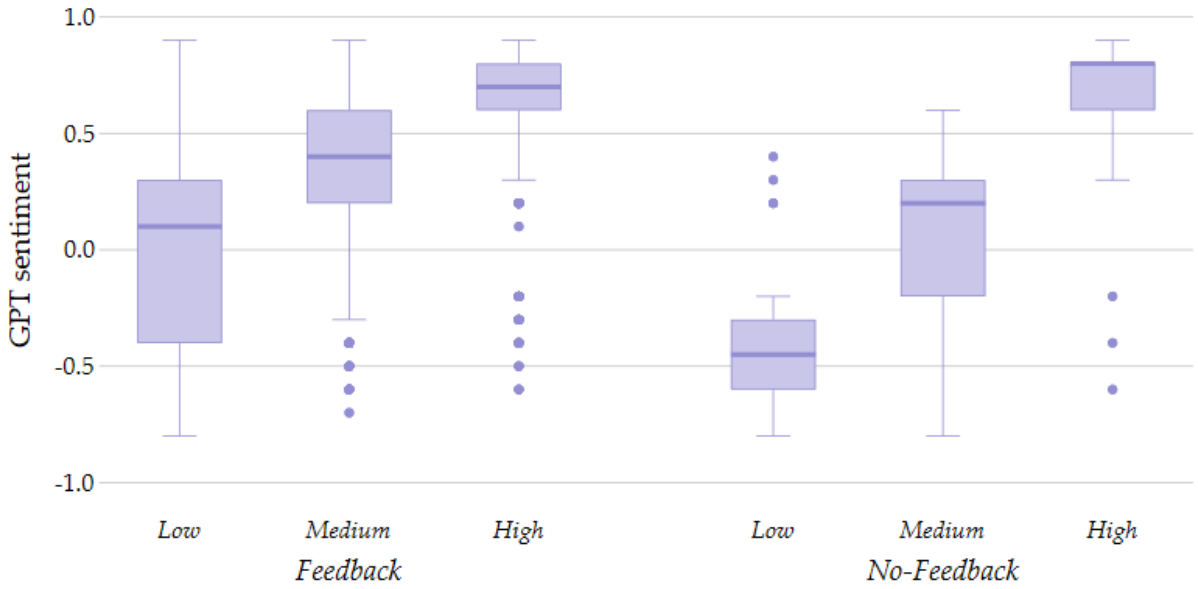


Figure C1. Box plot of the GPT sentiment of the evaluators' text depending on the accompanying grade and whether the text would be shared with writers as feedback

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).

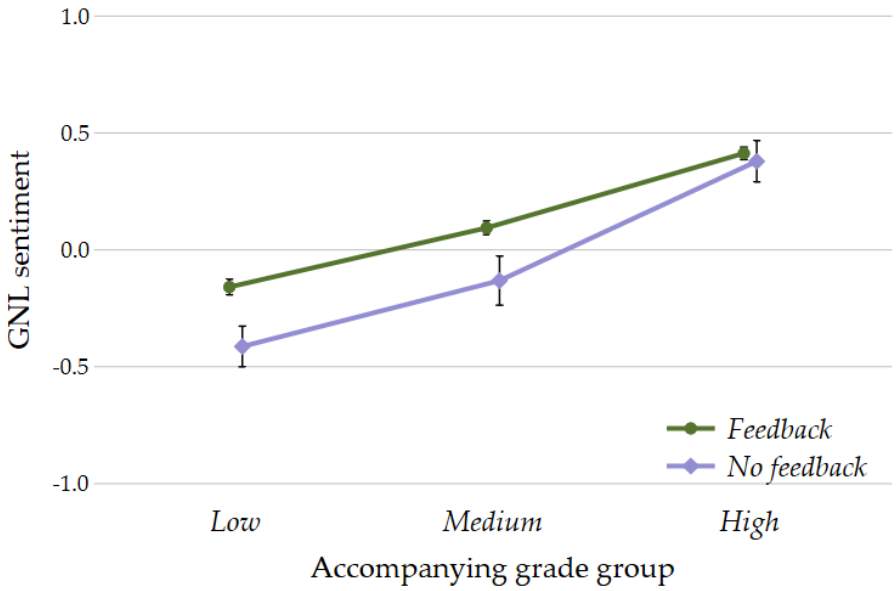
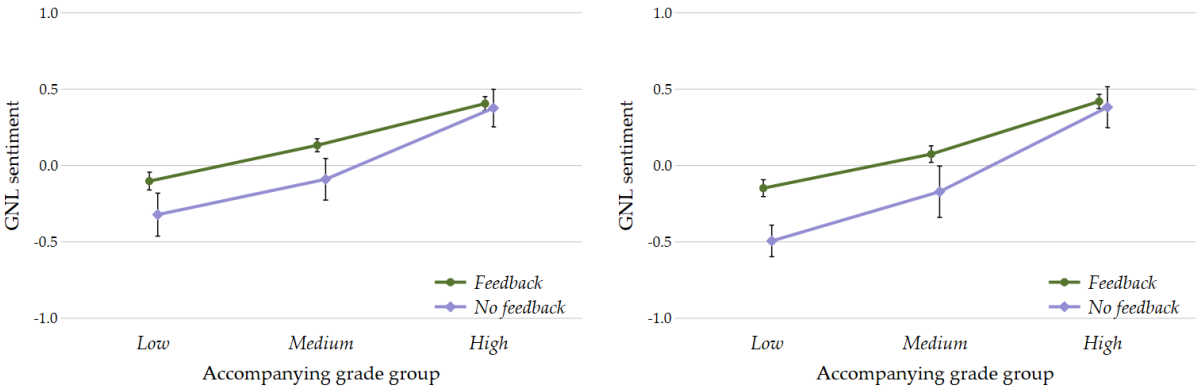


Figure C2. GNL sentiment of the evaluators' text depending on the accompanying grade and whether the text would be shared with writers as feedback

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 C3 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 C3 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 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)	(0.05)
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	0.03
			(0.05)	(0.05)	(0.05)
No-Feedback × Female			0.10	0.10	0.12
			(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 display errors in the essay, the number of characters in the feedback, and the number of characters in the essay. Robust standard errors in parentheses and statistical significance of non-zero coefficients is indicated by * $p < 0.05$ and ** $p < 0.01$.

Table C4. 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	−0.05
	(0.09)	(0.09)
Accompanying grade	0.60**	0.57**
	(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	0.417

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, the number of characters in the feedback, and whether they grew up in the UK. Robust standard errors in parentheses and statistical significance of non-zero coefficients is indicated by * $p < 0.05$ and ** $p < 0.01$.

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 C4 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 21). 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. 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.

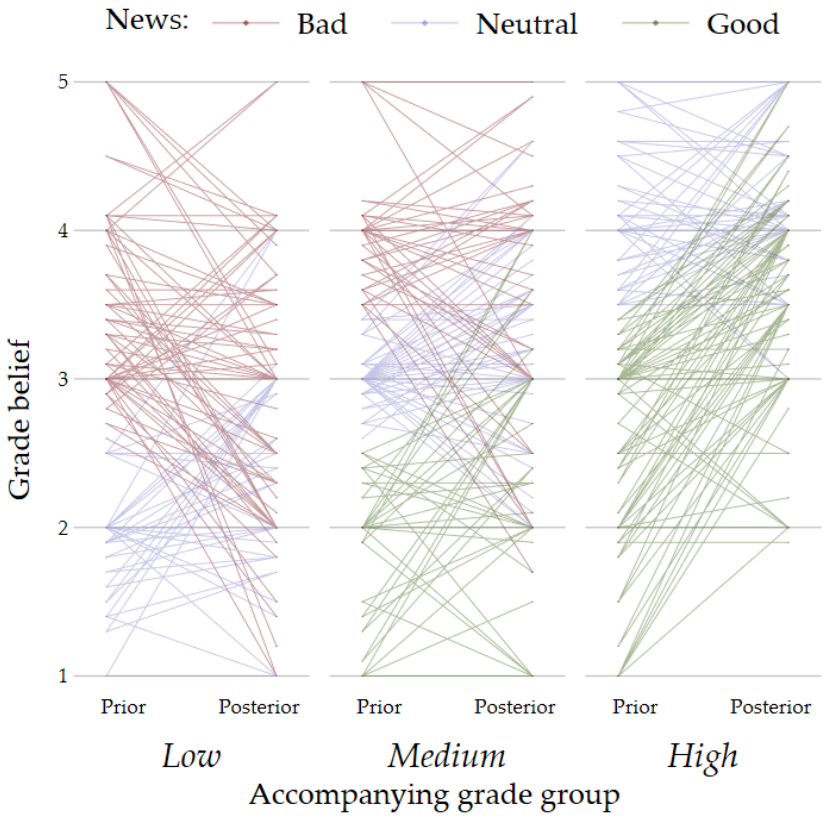


Figure C4. Individual belief adjustment in response to different types of feedback

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 ($N = 561$).

Table C5 contains the distribution of the sign of belief updating, for all nine possible combinations of the prior-belief grade groups (*Low, Medium, High*) and accompanying grade groups (*Low, Medium, High*). We can see that for good news (Accompanying grade group is above the prior grade group) that the majority of updates are positive. For example, for those in the *Low* prior group 62.22% update positively if they are in the *Medium* accompanying grade group and

Table C5. Distribution of belief-updating sign

	Accompanying grade group					
Prior group	*Low*		*Medium*		*High*	
Low		%	N	%	N	%	N
	Downward	32.65	16	22.22	10	0.00	0
	None	24.49	12	15.56	7	10.64	5
	Upward	42.86	21	62.22	28	89.36	42
Medium							
	Downward	58.11	43	28.26	26	5.68	5
	None	20.27	15	30.43	28	19.32	17
	Upward	21.62	16	41.30	38	75.00	66
High							
	Downward	87.80	36	62.50	30	32.47	25
	None	7.32	3	16.67	8	16.88	13
	Upward	4.88	2	20.83	10	50.65	39

Note: Each cell reports the distribution of sign of belief updates (Downward, None, or Upward) within a prior group × accompanying grade group cell. The accompanying grade group: *Low* (grades 1 and 2), *Medium* (grade 3), and *High* (grades 4 and 5). The prior grade groups: *Low* prior-belief group (priors in the range $[1, 2.5]$), *Medium* prior-belief group (priors in the range $(2.5, 3.5)$), and *High* prior-belief group (priors in the range $[3.5, 5]$). Percentages are within-cell percentages. N is the number of observations within group × accompanying grade group cell. The sample consists of writers in the *Feedback-Only*, *Feedback-Compete*, and *Feedback-Compete-Hidden* treatments ($N = 561$).

89.36% if they are in the *High* accompanying grade group.

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 C6 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 25). 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 C6 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 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 C6. 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 24. 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 display 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 C7 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 (Eil and Rao, 2011; Ertac, 2011; Möbius et al., 2022; Zimmermann, 2020). 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 C7. 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 display 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.4. Competition

First, we evaluate whether the results concerning the encouragement channel are sensitive to the sentiment score used to identify it. Table C8 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 C8. 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 display errors in both the essay and feedback, and the number of characters in the feedback. The sample consists of writers in the *Feedback-Compete* and *Feedback-Compete-Hidden* treatments. Robust standard errors in parentheses and statistical significance of non-zero coefficients is indicated by * $p < 0.05$ and ** $p < 0.01$.

In Table C9 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 C9 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 C9. 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 display errors in both the essay and feedback, and the number of characters in the feedback. The sample consists of writers in the *Feedback-Compete* and *Feedback-Compete-Hidden* treatments. Robust standard errors in parentheses and statistical significance of non-zero coefficients is indicated by * $p < 0.05$ and ** $p < 0.01$.

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.

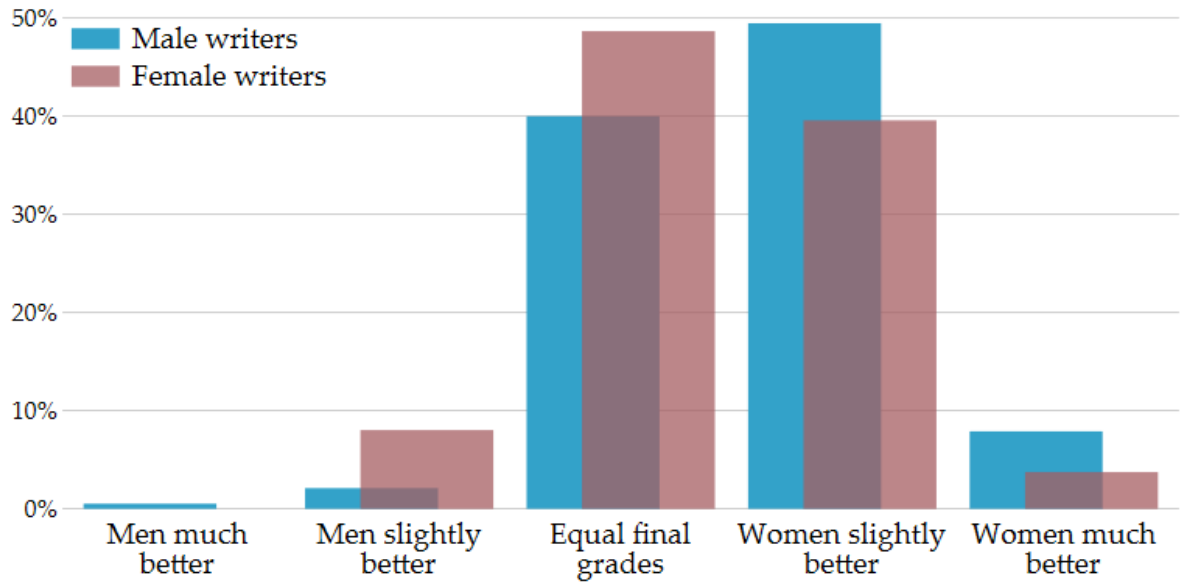


Figure C5. Beliefs about which gender performs better by the writers' gender

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

[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.

conditional mean probability of not competing when you should, $\mathbb{E}[1-p|\Gamma=1]$. Then, for a 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 C10 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 C10. 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 C11 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 C11. 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		

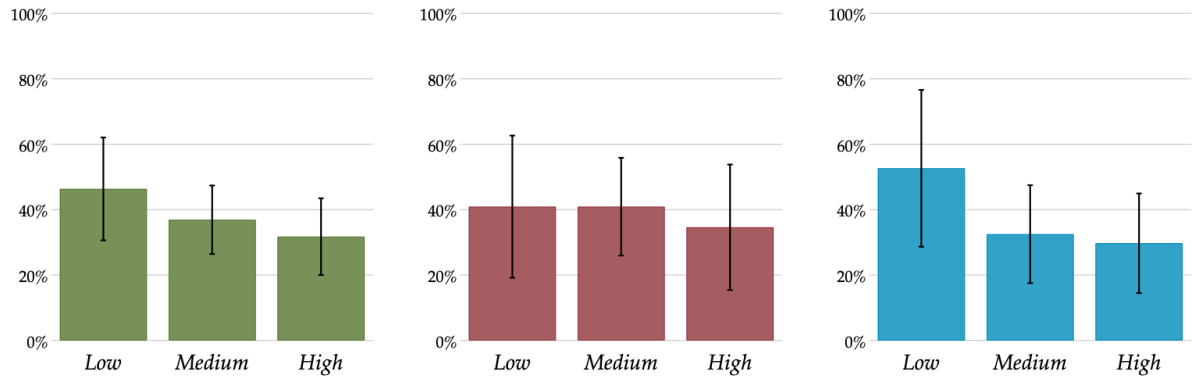
Note: The effect of the encouragement channel on the likelihood of making false positive and false negative errors with respect to the choice to compete by writer gender. Column (1) contains estimates of the mean probability of competing for writers who commit a false positive error (competing with a 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.5. Editing

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 24 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 24) 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 C12. 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 24) 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 C12, 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

C12, 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 C13. 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 24) 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 C13 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 C14 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 31. 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 C14. 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 31 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 display errors in both the essay and feedback, and the number of characters in the feedback. 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.

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.

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.

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.