Overcoming interpersonal barriers to improve performance feedback through expected candor

Abstract

Most organizations depend on effective feedback for improving employee performance. However, it can be difficult to give high quality feedback. In particular, feedback givers have competing goals of wanting to be honest, but also maintain relationships and appear benevolent. In response, we develop the construct of *"expected candor,"* as a way to elicit more concrete feedback. We theorize that expected candor licenses the feedback giver to provide direct and honest input because it sets out the recipient's expectations, thereby mitigating relational concerns that might prevent more concrete feedback. To test this theory, we draw on data from two years of a public sector organization's end-of-year 360 degree feedback reviews. Across two studies (Study 1: N=26,482 reviewers; Study 2A: N = 23,404 reviewers) we demonstrate feedback-givers' reluctance to give concrete developmental feedback, even when directly asked to do so. We also present results showing that feedback becomes more concrete when feedback givers receive a short intervention that prompts *expected candor*. Finally, in a third study (N=4,139 subjects) we find no evidence for relational backlash effects resulting from increased concreteness, nor to feedback elicited by the *expected candor* prompt. We conclude that the *expected candor* intervention gives organizations an easy way to promote better feedback-giving. Our findings also extend existing theoretical models on the perceived trade-off between honesty and benevolence, and add to our understanding of the broader set of relational concerns in feedback processes.

Keywords: feedback, concreteness, natural language processing, expected candor, field experiment

In organizations across industries, managers want their employees to be motivated to perform well and improve at their jobs. As the marketplace becomes increasingly competitive, enhancing employee productivity is critical to staying ahead of the competition. Performance feedback can provide employees information about the effectiveness of their own work behavior, improving organizational learning, and can also keep employees engaged in their jobs (Murphy & Cleveland, 1995). Accordingly, performance feedback is widely used in organizations, and the design and maintenance of feedback systems are essential for both individual and organizational performance (Ashford & Cummings, 1983; Bonacio & Dalal, 2006; Fedor, 1991; Kluger & DeNisi 1996; Taylor, Fisher, & Ilgen, 1984). However, the precise ways through which feedback influences employee motivation and performance is, to date, still unclear. In fact, as Latham and Locke (1991, p. 224) noted, "few concepts in psychology have been written about more uncritically and incorrectly than that of feedback."

Moreover, the effectiveness of feedback itself is not entirely clear. On the one hand, feedback is a central concept in most theories of learning - to improve future behavior, a person must be able to critically evaluate their past behavior. 85% of Fortune 500 companies have 360 degree feedback processes in place (Zenger, 2016). In their extensive review of the feedback literature, Kluger and DeNisi (1996) conclude that feedback improves performance about one-third of the time, has a negative impact on about one-third of the time, and has no apparent effect the remaining third of a time. Therefore, it is important to understand how that feedback can be made more constructive and useful.

As a channel for improving performance, feedback is often theorized to be more effective when it includes specific, actionable suggestions that can be followed, rather than abstract evaluations (Baron, 1988; Goodman, Wood & Hendrickx, 2004; Ilgen et al, 1979; Kraft &

Rogers, 2015). More concrete feedback is particularly useful for providing specific details on how a particular end state in the immediate future (e.g. in the coming work days, weeks or months) can be reached (Wiesenfeld, Reyt & Trope, 2017). Despite research advances in what makes good developmental feedback, there is still widespread dissatisfaction with the quality of feedback received in workplaces (Mishra & Farooqi, 2013; Wigert & Harter, 2017; Zenger & Folkman, 2014).

Though workers may want feedback to help them develop (Gong et al., 2017; Li et al., 2011; Wigert & Harter, 2017), it can be difficult for people to provide high-quality developmental feedback to their colleagues (Bies, 2013; Dibble & Levine, 2010; Kluger & DeNisi, 1996; Schaerer et al., 2018; Yeager et al., 2014). The previous literature suggests three alternative mechanisms for poor feedback provision. First, for some employees, it may simply be an issue of lack of attention (Blunden & Gino, 2018; Gino & Grant, 2015). Employees may not be focused on organizational development when providing feedback, especially when answering open-ended text questions. To them, the goal of providing concrete constructive input may not be salient. For others, providing good feedback may require a concerted effort that may be hard to muster without a timely reminder prompt to bring their attention to the goal of concrete feedback-giving.

Previous research also suggests that feedback-givers perceive a trade-off between honesty and benevolence (Levine & Cohen, 2018; Levine & Gomez, 2019; Levine & Schweitzer, 2014; Levine, Roberts & Cohen, 2019). Benevolence, in this literature has been defined as the motivation to improve someone's long-term welfare, which is consistent with giving specific feedback, but can come at the cost of short-term emotional harm (Mayer, Davis & Schoorman, 1995; Levine, Roberts & Cohen, 2019; Moore, Munguia Gomez & Levine, 2019). Levine and

colleagues (2019) argue that people may counterproductively think they are being benevolent by withholding such feedback because they (a) overestimate the magnitude of harm caused by the feedback, and (b) focus more on the short-term harm rather than the long-term benefits of the tough feedback. Furthermore, when people perceive a conflict between being benevolent (i.e., caring) or being honest (i.e., truthful), benevolence is often valued more than honesty (Levine & Schweitzer, 2014). This fear of interpersonal conflict thereby leads to inflated positive performance feedback (Waung & Highhouse, 1997).

To address this trade-off, we develop a potential remedy and introduce the construct of "expected candor". Expected candor describes the belief that feedback givers have about their recipients' preference for benevolence vs. honesty in their developmental feedback. We propose that feedback often fails to be concrete because feedback givers have low expected candor; that is, they believe their recipient will be discouraged by (and thus prefer not to receive) honest feedback about how they can improve. Accordingly, we develop a simple intervention to increase givers' expected candor. This intervention, using a short on-screen prompt as givers write their feedback, encourages feedback givers to recognize that the recipient expects direct and honest developmental input. We propose that this intervention helps givers see that concrete feedback is integral to, rather than at odds with, being a helpful provider. Our results demonstrate that this intervention does indeed improve the concreteness of feedback - and, further, that this improvement does not come with a cost of discouraging the recipient.

Across three field experiments (N = 54,025 respondents) we examine the barriers to giving concrete feedback, effective interventions to improve concreteness, and recipients' reactions to feedback from their managers. In doing so, we make three contributions to the feedback literature. First, we show that an attention-based intervention that explicitly asks for

specific examples, and a temporal framing intervention that focuses on future performance are each insufficient for improving the concreteness of feedback. Second, we provide empirical evidence of a relational-intervention that successfully encourages concrete feedback-giving, and demonstrate the role of relational concerns in workplace feedback. Finally, we provide some evidence suggesting this concern can be misplaced, and that concrete feedback does not necessarily induce a backlash from recipients.

1.1. Theory and Hypotheses

In this paper we focus on the concreteness of feedback. The literature refers to concrete feedback as that which "produces both learning and tangible, appropriate results, such as increasing effectiveness and improving performance on the job" (Cannon & Witherspoon, 2005, p. 120). While task-specific concrete feedback provided immediately has been shown to produce greater improvement in the short term on a particular task, in the long run, delayed and less frequent feedback has been shown to lead to better learning outcomes (Schooler & Anderson, 1990; Goodman, Wood & Hendrickx, 2004). Annual three-hundred sixty degree reviews are the focus of this research and they are by nature delayed and less frequent, however there is scope for it to be more effective (Zenger & Folkman, 2014).

As Fulham, Krueger and Cohen (2022) set out, "for feedback to be effective, recipients must be receptive and accurately understand the meaning and veracity of the feedback (i.e., discern the truth in feedback)." Concrete feedback is by definition, less ambiguous, and therefore more likely to be understood (Cannon & Witherspoon, 2005; Kopelman, 1986). Our theoretical contribution focuses on illuminating the barriers to providing concrete feedback.

One simple explanation for the dearth of concrete feedback is that it is attention-driven. Even if givers endorse concrete feedback as a goal in principle, they may fail to deliver it in

practice because that goal is not salient when they are prompted for feedback. Concrete feedback may not be generated by default, as givers who want to be concrete must recognize the opportunity to do so, and search through their memory for specific examples, and formulate actionable plans for improvement. According to Fishbach and Ferguson (2007), "as a memory construct, a goal necessarily fluctuates in accessibility (i.e., its activation potential; Higgins, 1996). This means that the likelihood of the goal being activated will vary across time and situations according to its accessibility at the moment."

Many papers have documented similar “failure[s] to enact one’s intentions,” (Rogers & Milkman, 2016) that prevent people from achieving their goals (Fife-Schaw, Sheeran, & Norman, 2007; Kristal & Whillans, 2020; Rhodes & de Brujin, 2013; Rhodes & Dickau, 2012; Sheeran & Webb, 2016; Webb & Sheeran, 2006). Research has demonstrated how reminders can be an effective way of increasing goal salience, which in turn, increase goal achievement in various domains ranging from medicine (Shea, DuMouchel, & Bahamonde, 1996) to savings (Karlan, Ratan, & Zinman, 2014) to shopping (Rogers & Milkman, 2016). Delivering the reminder at the right time is essential for aligning behavior with intentions (Austin, Sigurdsson, & Rubin, 2006).

Employees and managers have various demands on their time. Oftentimes providing feedback for their colleagues is an additional task employees need to undertake on top of their daily work requirements. Busy workers rarely have time for reflection (Bruch & Ghoshal, 2002). Feedback givers may be inattentive when filling out feedback surveys, and this may hinder good feedback provision (Blunden & Gino, 2018; Gino & Grant, 2015). Therefore, if providing concrete input is a goal the feedback giver holds, then making the goal of concreteness more accessible and salient at the time of feedback-giving should help to activate that goal.

Hypothesis 1. Activating the goal of giving concrete feedback at the time of feedback giving will lead to improved concreteness of developmental comments.

Hypothesis 1 posits that employees fail to provide useful feedback because they intend to be concrete but are inattentive to their goal in the moment. An alternative hypothesis is that when feedback is being generated, that may not be givers' primary intention. Instead, givers may be trying to balance two competing goals: giving concrete feedback and preserving the relationship. These goal conflicts are common in mixed-motive conversations like feedback-giving (Yeomans, Schweitzer & Brooks, 2022). If there are interpersonal relational concerns causing the lack of concrete feedback, a timely reminder to give more concrete feedback (i.e., addressing the attentional barrier), will be insufficient for closing the intention-action feedback gap. As such, we put forward an alternative hypothesis based on the following theory.

There are two models that can help us think through the ideal way of framing feedback that accounts for these competing goals. According to Levine, Roberts and Cohen (2018), approaching feedback that is high in honesty and high in benevolence requires an integrative strategy. Another, complementary approach involves the "conversational circumplex" developed by Yeomans, Schweitzer and Brooks (2022) which is a framework that classifies conversational motives along two dimensions: informational and relational. When considering feedback in an organizational context, these conversations may satisfy one or both of the informational (corresponding to honesty) and relational (corresponding to benevolence) dimensions.

Preserving the relationship can form a barrier to providing concrete feedback: People generally focus more on the immediate immediate social harm of providing tough feedback than on the long-term benefits of providing that honest feedback, often leading to an aversion to the seemingly harmful action, leaving out information, or engaging in prosocial/paternalistic lying

(Cushman et al., 2012; DePaulo et al., 1996; Fulham, Krueger & Cohen, 2022; Jampol & Zayas, 2021; Levine et al., 2018; Levine, Roberts & Cohen, 2019; Lupoli, Jampol & Oveis, 2017; Lupoli, Levine & Greenberg, 2018; Sun & Slepian, 2020). This avoidance of communicating undesirable information is known as the MUM effect (Bond & Anderson, 1987; Dibble & Levine, 2010, Rosen & Tesser, 1970; Tesser, Rosen & Tesser, 1971). Managers, in particular, may have a vested interest in providing overly positive feedback because it decreases employee stress, disagreements with the manager and employee dissatisfaction, thus enabling managers and employees to have better working relationships (Murphy & Cleveland, 1995). As such, merely asking individuals to provide more concrete feedback would be insufficient for overcoming the psychological barriers preventing them from providing this input.

Surprisingly, when people are forced (or, as in an experiment, assigned to a condition) to be honest in conversation with close others, the experience is more positive for them and their social relationships than they (and others) would have predicted (Levine & Cohen, 2018). Furthermore, the literature demonstrates that when recipients of tough feedback believe in the good intentions of the communicator, the recipient does not experience emotional harm (Finkelstein, Fishbach & Tu, 2017; Yeager et al., 2014). Previous research has demonstrated how explicitly stating the benevolent intentions of the feedback giver makes recipients more accepting of the critical feedback (Yeager et al., 2014), However, there is less evidence to support the effectiveness of emphasizing to the feedback giver how giving critical feedback is expected and benevolent. However, we hypothesize that making these good intentions and recipients' expectations clear changes beliefs about how the recipient will react, leading to better feedback giving.

In this research we propose that feedback can be elicited in a way that allows givers to attain these multiple goals simultaneously. Specifically, we suggest that both the informational and the relational aspects of feedback-giving can be made salient, by reminding the feedback giver of the expectations their recipient has for candid feedback. By framing the prompt this way, we license the feedback-giver to provide critical, constructive feedback while still allowing the recipient to "save face" (Goffman, 1967). When feedback givers are prompted to be more honest, they may also feel less responsible for any relational harms that result. We test an "expected candor" intervention that decouples the association of being direct/honest with damaging the relationship.

Hypothesis 2. Asking people to give more concrete feedback in a way that promotes "expected candor" will lead to improved concreteness of developmental comments.

A competing reason for poor feedback-giving could be their natural tendency to conflate the word "feedback" with the request for a backwards-looking evaluative approach (Blunden et al., working paper). Focusing on the past has been shown to constrain one's ability to think about alternative possibilities (Linsey, Tseng, Fu, Cagan, Wood, & Schunn, 2010; Youmans & Arciszewski, 2014). The language of many feedback prompts leads feedback givers to adopt an evaluative backwards-looking approach on what was done, as opposed to a developmental future-oriented approach on what that person could do better in the future. However, when people are asked to give advice, they tend to think more about possible future actions (Brooks, Gino, & Schweitzer, 2015; Levari, Gilbert & Wilson, 2022). In fact, Bluden et al., (working paper) find that using the word "advice" evokes a greater future-focused mindset and in turn leads to more actionable, developmental input when people are asked to provide input on how someone performs at a specific task.

While replacing "feedback" with "advice" is hypothesized to change mindsets, it also may play a role in addressing relational concerns. It may be uncomfortable for people to discuss their colleagues' weaknesses when looking back as it is something that is immutable and happened in the past. However, when looking forward, it may be less uncomfortable to discuss areas for development because the potential to change is more salient.

Hypothesis 3. Asking people to give more concrete feedback in a future-oriented way, framed as "advice" will improve concreteness of developmental comments.

1.2. Present research

The present research documents three studies. In Study 1, we describe a field experiment with two arms (control and treatment) that attempted (without success) to increase concreteness in 360-degree feedback. Study 1 tests Hypothesis 1 where we directly remind feedback-givers to give more specific input at the moment of the feedback provision, without addressing feedback-givers relational concerns.

In Study 2A, we co-designed a field experiment to facilitate better feedback giving while maintaining relational trust. We test a control condition against three experimental conditions: (i) a "expected candor" that sets out the recipient's expectation for direct and honest feedback (testing Hypothesis 2); (ii) a "future focus" condition that reframes feedback as advice (testing Hypothesis 3); and (iii) a condition that combines the two previous conditions to explore additive or multiplicative effects. The "expected candor" manipulation proved to be most effective.

Finally, in Study 2B, we seek to explore whether there are backlash effects of increasing the concreteness of feedback. As such, we delivered a three-item survey to feedback recipients asking about the helpfulness, accuracy and how motivating they find their manager's feedback.

2. Experimental Design

2.1. Research Collaboration. The Behavioural Insights Team (BIT) is a former government organization, partly still owned by the UK Cabinet Office, with a mandate to generate and apply behavioral insights to inform policy, improve public services and deliver results for citizens and society. In 2018, the Behavioural Insights Team was asked to work on a project examining gender differences in feedback at the highest levels of a particular UK public sector organization. In particular, they were concerned about the concreteness of the feedback provided to some employees, which informed our theoretical model of effective feedback in this context, and our empirical strategy for program evaluation. These data and studies are part of a research collaboration with BIT and their collaboration with the UK Government Equalities Office. The treatment effects here do not vary by gender, so we do not address that research question in this paper. We plan to write a different manuscript detailing the general findings about benevolent sexism in workplace feedback language using these data.

The interventions were developed with sensitivity to the constraints of the context. Among all the possible ways to improve feedback delivery, we were only allowed the ability to change the prompts for the open-ended feedback in the survey form. This light touch intervention, embedded into existing processes, enabled us to study the impact of an individual-centered intervention in an organizational context (Lambert, Caza, Trinh & Ashford, 2022). Although these limitations perhaps put a ceiling on the potential effect size, they also forced us to design an intervention that would impose minimal costs on the organization and the employees.

2.2. Data Management. The data from Study 1 was collected between January and March 2019, and the data from Study 2 was collected between January and June 2021. The pre-registration for Study 1 was written after (and for Study 2, while) the data was collected by

the organization, but before the researchers had any access to the data. Furthermore, the researchers only received anonymized data, after identifiers were stripped. For more information on the variables we received see Appendix A.

Given the sensitive nature of the data, the raw data will not be posted publicly. However, we provide our pre-registrations, data dictionaries, all cleaning and analysis code, and an anonymized data set (which can be used to reproduce all our analyses) through the Open Science Framework at https://osf.io/jtzg6/?view_only=ce1e5245e466498982943c9e1f9c49f6.

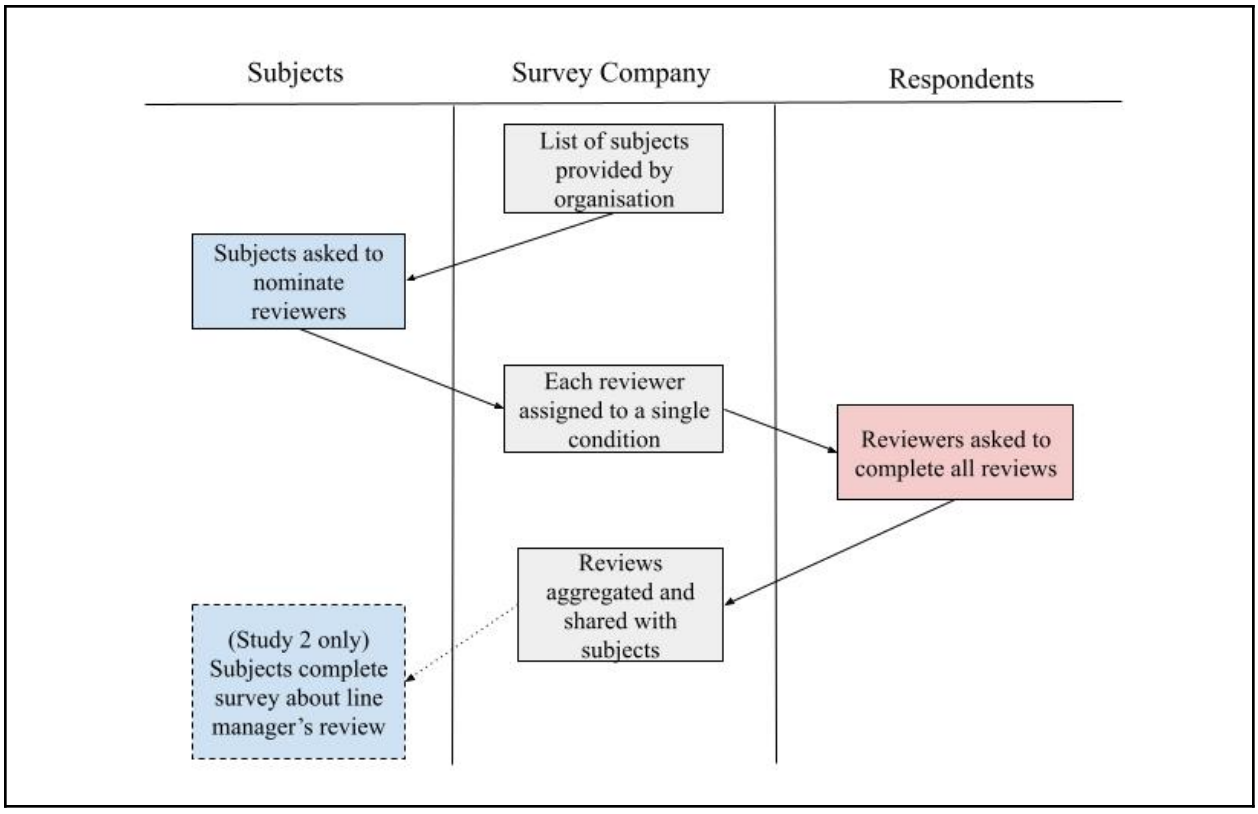
2.3. 360 Degree Review Experimental Protocol. The data in these studies were collected as part of an annual 360 degree review process conducted by a UK public sector organization. These reviews do not have any direct impact on bonus, compensation, or promotions. Instead, they are encouraged instead as a tool for reflection and learning within the organization - accordingly, all surveys were sent out (and were expected to be completed) during normal work hours. These 360 reviews were being collected and distributed as an organizational development exercise for several years before they were considered as a potential data source for research. We left most of the review protocol intact from previous years, for several reasons: to preserve the existing organizational value of the survey; to minimize implementation costs imposed on the organization; and to produce a scalable and modular intervention that could easily be adopted by others with a similar feedback procedure.

The review process starts when the public sector organization provides the survey company with a list of subjects (and their email addresses) to be reviewed. The survey company then contacts the subjects and asks them to nominate reviewers (and provide their email addresses) - including their manager, their direct reports, and up to ten peers. The survey company then randomly assigns each reviewer to one of two conditions in Study 1 or one of four

conditions in Study 2A. The reviewers are then given a 360 degree review survey to fill out about the subject. Reviewers may be nominated by multiple subjects, and since random assignment is at the reviewer level, all of the reviews they provide will be using the same condition questions. Conversely, since a given subject may nominate multiple reviewers, those reviewers will not necessarily all be assigned to the same condition; therefore, subjects may receive reviews from reviewers in different conditions. When all reviews are completed, the survey company aggregates the reviews and shares anonymized findings with the subjects. Finally, in Study 2B, we introduce a follow-up survey, where subjects are asked to provide feedback on the feedback they received from their managers. See Figure 1 for a diagram of this review process protocol.

Figure 1.

360 Degree Review Experimental Protocol



2.4. Measuring Concreteness in Feedback

As discussed above, we are interested in concreteness as a theoretically-motivated outcome. Empirically, we wanted to measure concreteness directly from the text. Algorithmic measures provide several advantages over human annotators. First, they are scalable - they can be applied to large datasets at very low cost, including (as in our case) situations where the raw text cannot be shared due to privacy constraints. Second, they allow us to interpret the contents of the text that are driving the results (whereas human ratings are holistic and usually do not come with explanations). However, algorithmic measures also come with concerns about generalizability - that performance suffers if the domain in which they are trained differs from the domain in which they are tested.

To mitigate this trade-off, we rely on a recently published paper that develops a domain - specific model of concreteness in advice-giving (Yeomans, 2021). This model was trained using 3,289 documents from six different advice datasets that had human annotations of advice specificity. This model used ngram features, in a supervised LASSO regression. The resulting model outperformed (i.e. predicted the human ratings better than) every other domain-general concreteness dictionary, and generalized well across the different advice contexts (e.g. parent-teacher feedback, 360 peer reviews, teaching darts). In sum, we contribute to the growing literature applying natural language processing methods to texts generated in organizational contexts to gain insight into organizational cognition and to leverage this understanding to improve performance (Lix, Goldberg, Srivastava & Valentine, 2022; Srivastava, Goldberg, Manian & Potts, 2018).

3. Study 1

We worked with a large public sector organization to help improve the quality of the comments written for high-level employees during their end-of-year 360-degree feedback reviews. Reviewers (peers, direct reports, line managers, or others) were nominated by subjects to provide numeric responses on 15 Likert scales (henceforth referred to as "structured measures") regarding the subject's performance and written feedback in the open-ended text boxes.

3.1. Study 1 Methods

3.1.1. Study Design

We employed a two-condition between-subjects design. The primary differences between these two conditions were the prompts for the three open-ended questions at the end of the survey. See Table 1 for the exact text prompts.

Table 1

Study 1 Question Prompts by Condition

Study 1	**Q1: Strengths**	**Q2: Development**	**Q3: Overall**
Condition 1: Control	What are their main strengths as a leader and why? Please include up to three examples.	What are their main areas to develop as a leader? Please include up to three examples.	Overall, please state if you feel they provide a good role model as a (org name) leader and how they demonstrate this.
Condition 2: Direct appeal	What are their main strengths as a leader? Please include examples of both how they relate to others and their leadership of their team's objectives.	What are their main areas to develop as a leader? Please include concrete examples of what they could do to achieve this.	Overall, are they a good role model as a (org name) leader? Please provide specific examples to explain your answer and if needed what actions

			they could take to improve.

3.1.2. Participants

The dataset we received contains 51,223 unique responses to a survey ("reviews"). Of these, 5,037 responses were "self-reviews," where the respondent evaluated themselves as a subject. We removed these self-reviewers from our formal analyses (except to identify the demographics of respondents in the other surveys). We also found that ten people conducted a self-review twice, so we only keep the first of these for each person.

These responses were written by 26,482 unique respondents, and targeted 5,037 unique subjects. However, 709 subjects were only present in self-reviews. When these are removed, we are left with a sample of 4,328 unique subjects who received at least one review from another person, and 25,627 unique respondents who wrote at least one review for another person, for a total of 46,176 unique survey responses.

The treatment was assigned at the level of respondent - that is, each respondent saw the same survey condition for all of their reviews - at the time the reviewee nominated the reviewer by providing their email address. This left us with 12,754 respondents assigned to the control group, 12,731 assigned to the treatment group, and 142 who were labeled as having been included in both conditions (due to two email addresses). Following our pre-registration, we dropped the respondents who were exposed to both conditions, leaving a focal sample of 25,485 respondents, 4,328 subjects, and 45,693 responses.

3.1.2.1. Invitation Rates. Subjects nominated potential respondents, who were then invited to respond. However, participation was voluntary, and the individuals in the focal sample display a wide range of response rates. Although every senior person within the organization was

invited to be a subject, we do not have exact data on who accepted those invitations. However, we observe every invitation sent to respondents. Excluding self-reviews, each respondent was invited to write an average of 1.79 reviews (SD = 1.96, min = 1, max = 31) and each subject received an average of 10.6 reviews (SD = 3.61, min = 4, max = 49). Within this focal sample, we observe that respondents assigned to the treatment condition were sent the same number of survey invitations (m= 1.79, SE=0.018) as respondents in the control condition (m= 1.80, SE=0.017, t(25483) = 0.6, p = .574).

3.1.2.2. Sample Characteristics. Out of all the people in the focal sample, 43.2% were invited to evaluate female employees, 49.6% to male employees, and the remaining 7.2% to employees who either did not disclose their gender identity or identify as "other." When a subject suggested potential respondents, they also described their relationship to the respondent. In the focal sample, 40.7% of responses involved a direct report reviewing their line manager, 26.2% of responses involved a respondent reviewing one of their peers, 10.0% involved a line manager reviewing one of their direct reports, while the remaining 23.1% were classified as an "other" relationship. Characteristics of the focal sample are described in Table 2.

Table 2

Sample characteristics for Study 1

	Control	Treatment	Total
Unique Reviews	22,955	22,738	45,693
Respondent % Female	38.7	38.8	
Subject % Female	43.7	42.7	
Relationship			
% Line Managers	10.1	9.9	
% Direct Reports	40.8	40.6	
% Peers	26.1	26.4	
% Others	23.0	23.2	
Unique respondents	**12,754**	**12,731**	**25,485**

% Female Reviews - Mean (SD)	39.0% 1.80 (1.98)	38.9% 1.79 (1.94)	
Unique subjects % Female Reviews - Mean (SD)	 5.31 (2.40)	 5.27 (2.46)	4,328 42.8% 10.6 (3.6)

3.1.2.3. Respondent Characteristics. We can only identify the respondent characteristics of the 5,037 subjects in this sample who also reviewed themselves. This means we have self-reported gender for 18,487 out of the 45,693 responses in our sample (40.5%). Within this subsample, we have 9,019 responses that can be attributed to male employees and 7,161 responses that can be attributed to female employees. We also found that female respondents composed 33.6% of responses from line managers, 40.8% of the responses from peers, 40.6% of responses from direct reports, and 37.8% of responses from others.

3.1.2.4. Balance Checks. We conducted balance checks to confirm that there were no systematic differences in the make-up of the control and treatment group. We found no significant differences in the gender make-up of subjects between the two groups ($X^2(2) = 0.252$, $p = .882$), the distribution of cohorts between the two groups ($X^2(256) = 272$, $p = 0.235$) or the distribution of relationships between the two groups ($X^2(3) = 3.11$, $p = .376$). While randomization was assigned at the *reviewer* level, we only collected demographic information for the subjects, meaning that we can only derive the reviewer characteristics based on the 5,037 who reviewed themselves.

3.1.3. Measures

The primary outcomes in this study were the three open-ended text boxes at the end of the survey. Each box asked a different question about the subject: the first, “Strengths”, asked for positive qualities; the second, “Development”, asked for things to improve on; and the third,

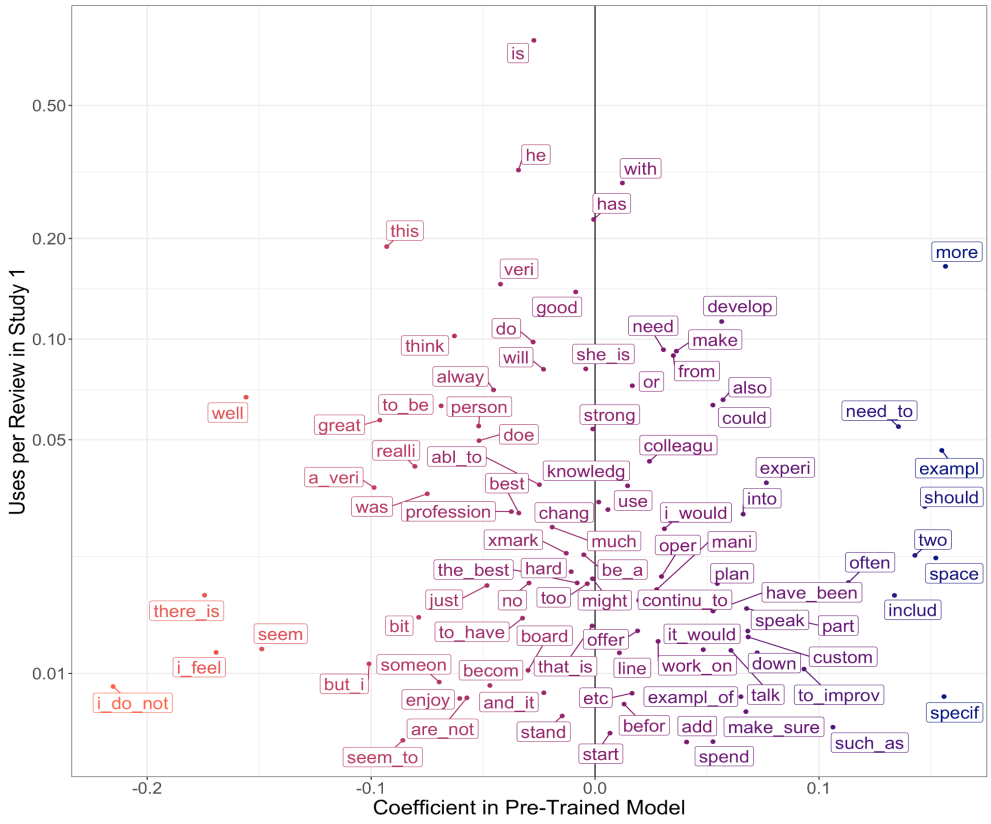
"Overall", asked for a more general assessment (see Table 1 for exact text). As planned, we analyzed the results of each text box separately, and we estimated all regressions using standard errors clustered at the subject and respondent levels. Given that we are most interested in improving developmental feedback, we will be focusing our discussion on the second text box.

3.1.3.1. Concreteness. In our pre-registrations we defined our primary outcome measure as the concreteness of the advice people wrote. We measured this using the pre-trained advice model from the doc2concrete package (Yeomans, 2021). That model was trained on human-annotated data from six different datasets, primarily involving workplace feedback, but also data from people giving advice on how to lead a good life or play darts, as well as advice from middle school teachers to their students' parents. This model was trained to accurately identify the domain-general markers of concrete advice across all of these data sources. Figure 2 provides a rough illustration of the contents of the final concreteness model from the package. Although there are 287 total coefficients in the model, most of the variance in the predictions is driven by phrases with large coefficients, that are common in the data, or both.

The model was applied to each of the three final text boxes separately, to produce three concreteness scores for each survey response. All of these scores were then standardized to a common mean and variance, to compare the relative effects of our interventions on the concreteness of the written responses.

Figure 2

Most Important Words and Phrases in the Concreteness Model



The X axis represents the coefficient in the pre-trained model (positive is more concrete, negative is more abstract). The Y axis shows the prevalence of each word and phrase in the data from Study 1. The model scores a particular review for concreteness by multiplying these coefficients and the counts of each phrase used by a review writer.

3.1.3.2. Performance. The survey asked respondents to provide 15 structured ratings of their subjects' job performance before writing in the text boxes (see Appendix A for exact text). Although the questions covered many distinct aspects of job performance, in practice these numeric ratings correlated highly with one another (cronbach's alpha = 0.96). Based on this we decided to create a single index for our analysis, after imputing the missing values (described below). As the main treatment occurred after the structured ratings, we did not expect ratings to differ by condition. We empirically verified that almost all of the numeric ratings did not differ by condition, with one exception, which was the only question where the wording changed

across conditions ($\beta = .119$, SE = .010, $t(45691) = 12.5$, $p < .001$). We removed that question from all further analysis, though our results do not change meaningfully if we include it instead.

Non-response on the numeric questions range between 24-49% (median 31%). We imputed all missing values for each question, using the mice package in R (Van Buuren & Groothuis-Oudshoorn, 2011). Using those imputed values, we then performed a principal components analysis (PCA) across the remaining set of numeric questions. From that PCA analysis, we extracted the first principal component, reflecting the main axis of variation in the data, as a single abstracted measure of "subjective performance", and there was no difference across conditions in that composite measure ($\beta = .001$, SE = .011, $t(45691) = 0.0$, $p = .974$).

Our measure of subjective performance is meant to capture the job performance of the subject, in the eyes of the reviewer. However, we did not have any truly objective measures of performance we could benchmark against. However we could at least measure the degree to which a respondent's subjective performance ratings track "consensus" performance ratings, by calculating the average of the ratings that all the other reviewers gave to each subject. The individual respondents' ratings correlated with these consensus ratings ($r = .229$, $t(45691) = 50.3$, $p < .001$) Furthermore, respondents' ratings correlated with the subjects' own self-ratings ($r = .110$, $t(45691) = 23.6$, $p < .001$). These correlations are not very high, as surely there are idiosyncrasies in people's thresholds and criteria that lead to divergence (especially when comparing self- vs. other-assessment). But perfect correlations are not the goal, as this index is meant to capture a reviewer's unique perspective about the subject. Still, these results suggest that this constructed performance index captures some generally-agreed aspects of a subject's performance.

Table 3

Summary statistics and bivariate correlations for Study 1

	Summary Stats			Bivariate Correlations						
Variable name	N	Mean	St. Dev	Gender	Performance	Concreteness (Strength)	Concreteness (Development)	Concreteness (Overall)	Word Count (Strengths)	Word Count (Development)
Gender (Male = 1)	45,693	0.50	0.50							
Performance (standardized)	45,693	0.01	1.02	-0.05***						
Concreteness (Strengths)	30,465	-0.10	0.32	-0.02***	-0.01					
Concreteness (Development)	27,227	0.02	0.38	0.01*	-0.11***	0.09***				
Concreteness (Overall)	28,885	-0.14	0.25	0	-0.03***	0.12***	0.08***			
Word Count (Strengths)	45,693	35.2	45.2	-0.03***	0.16***	-0.06***	0	-0.07***		
Word Count (Development)	45,693	28.8	40.9	-0.02***	-0.07***	-0.06***	0.06***	-0.08***	0.49***	
Word Count (Overall)	45,693	21.7	29.5	-0.02***	0.08***	-0.01	0	-0.11***	0.43***	0.35***

*Note.*St. Dev., standard deviation. **p<0.05; **p<0.01; ***p<0.001*

3.2. Study 1 Results

For all of our analyses in both studies, we used a common statistical model from our pre-registration. We estimated linear effects throughout (even for binary outcomes) and do not standardize any variables, with the exception of our concreteness measure. Our data are in a hierarchical design, with most respondents reviewing multiple subjects. However, because the treatment was assigned at the level of the respondent, we cannot control for respondent fixed effects. Instead, we estimate our model at the level of individual reviews, with standard errors adjusted for clustering at the subject and the respondent level (Graham, Arai & Hagstroemer, 2016). Throughout the paper, we report the primary coefficient and hypothesis tests in the main text, however we also provide full regression tables in Appendix B for each test for interested readers (and the full data and code are posted on OSF for further investigation).

Most of our primary hypothesis tests are presented as simple regressions, without any control variables. However, for robustness we also include models that include a small number of control variables motivated by the literature. Specifically, in our preregistration we mention that we anticipated differences in concrete feedback - unrelated to condition assignment - due to natural variation in performance (i.e. worse performers would get less concrete feedback) and gender differences (Correll et al., 2020), as well as differences in relationships (i.e. managers or direct reports would give more concrete feedback because they have more experience working with the recipient than peers) and department, and these expectations were borne out (see section 2.2.4 below). So we report models with those controls included in the manuscript, and compare the coefficients in full regression tables in Appendix B. Additionally, although we did not pre-register it, we also tested all our models with subject fixed effects. In all cases, every

combination of controls we considered produced similar results, and had no meaningful impact on the conclusions we drew from the data.

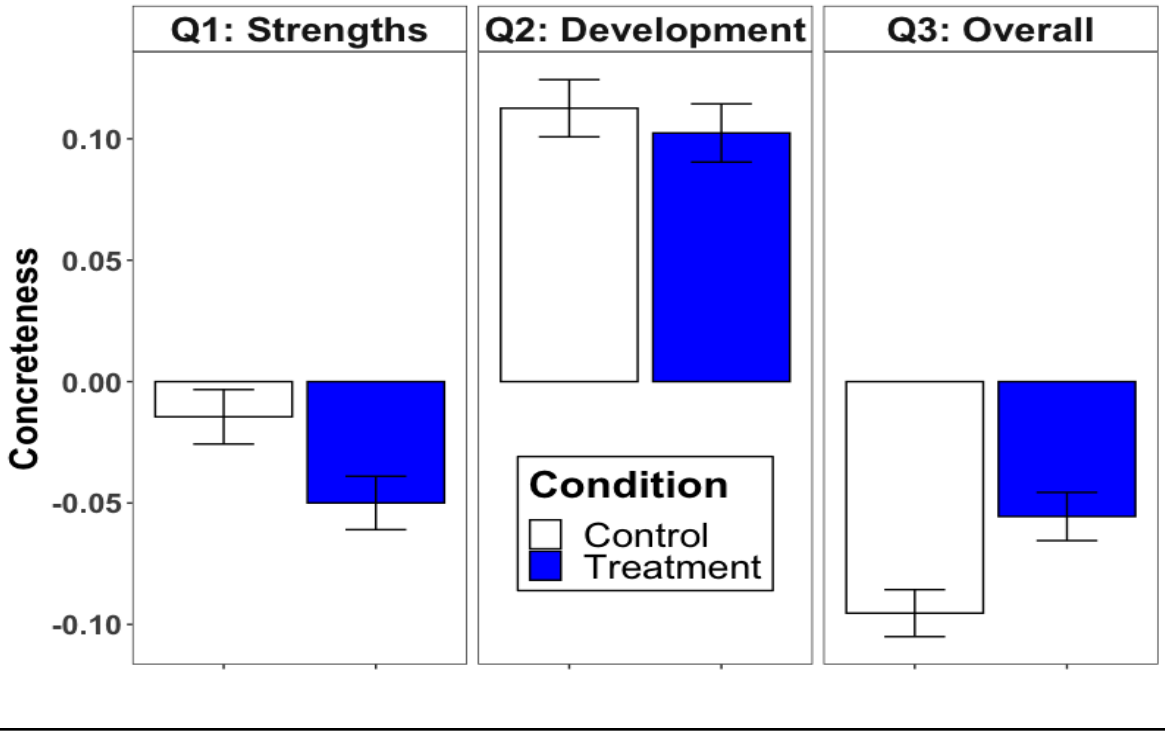
3.2.1. Non-Response. Respondents were not required to answer any of the questions on the survey, and many did not do so, even after accepting an invitation. Roughly a third of participants did not write a single word in one text box (Q1: 33%; Q2: 40%; Q3: 37%), with 32% not writing a single word in any of the text boxes. We wanted to focus on the most substantive text for analysis, so to remove empty and low-word-count responses, we only analyzed responses that were longer than five words (see Appendix C for details). Across different cut-offs, we tend to find that participants in the treatment condition were more likely to leave text responses blank, including for our chosen cut-off (Q1: $t(45691) = 4.6$, $p < .001$; Q2: $t(45691) = 4.8$, $p < .001$; Q3: $t(45691) = 9.4$, $p < .001$). However, our results are robust to other cut-off strategies, including quantile cut-offs that remove a similar number of texts from each condition (thus warding off selection bias). Our chosen cut-off results in samples of 23,103 responses to question 1, 23,053 responses to question 2, and 22,967 responses to question 3, with a total of 16,979 respondents and 4,325 subjects.

3.2.2. Response Length. After removing texts less than five words long, the remaining texts were all substantial (Strengths: mean = 65.8 words, SD = 45.9 words; Development: mean = 55.7 words, SD = 42.9 words; Overall: mean = 41.7 words, SD = 30.2 words). On top of the results in 2.2.1. showing differences in non-response by condition, we also find differences in length among the remaining texts. The direct appeal condition resulted in shorter feedback in both Strengths ($\beta = -2.90$, $SE = .83$, $t(23101) = 3.5$, $p < .001$) and Overall ($\beta = -1.75$, $SE = .50$, $t(22965) = 3.5$, $p < .001$), with no difference in the Development box ($\beta = .82$, $SE = .73$, $t(23051) = 3.5$, $p < .001$).

3.2.3. Concreteness. Our first hypothesis concerned the concreteness of the subjects' reviews. Specifically, we thought that we could increase the concreteness of feedback by nudging one half of reviewers to write more concrete feedback for their review subjects. In Figure 3, we display the means for each text box by condition.

Figure 3

Concreteness Results for Study 1



Average standardized concreteness scores for all three text responses, by treatment assignment. Each point represents a group mean and 95% CI.

3.2.3.1. Development. The intervention had no significant (and directionally negative) impact on the concreteness of the Development review text (β = -.025, SE = .015, t(23051) = 1.7, p = .091). When controlling for performance, we still do not see an effect of treatment on concreteness (β = -.024, SE = .015, t(23050) = -1.6, p = .102), nor when we control for

performance along with gender, relationship, and department (β = -.024, SE = .015, t(23029) = 1.6, p = .103).

3.2.3.2. *Strengths.* The intervention had a significant negative effect on the concreteness of the Strengths review text (β = -.099, SE = .015, t(23101) = -6.7, p < .001). The negative effect persists when controlling for performance (β = -.099, SE = .015, t(23100) = -6.8, p < .001), as well as when we control for performance along with gender, relationship and department (β = -.099, SE = .015, t(23079) = -6.7, p < .001).

3.2.3.3. *Overall.* The intervention did have a positive effect on the concreteness of the Overall review text (β = .148, SE = .015, t(22965) = 10, p < .001). This effect persists when controlling for performance (β = .147, SE = .015, t(22964) = 10, p < .001), as well as when we control for performance along with gender, relationship, and department (β = .147, SE =.015, t(22943) = 10, p < .001).

3.2.4. Outcome Validation. Given the surprising direction of our treatment effect estimates, we wanted to evaluate our chosen outcome variable with respect to several other variables as validation checks, to confirm the measure identifies expected differences in other variables. First, we show that there are average differences across questions, which can be seen in Figure 2. The Development question elicited feedback that was much more concrete than the Strengths question (β = .370, SE =.010, t(46154) = 37, p < .001). Furthermore, we found that employees who were rated as having higher subjective performance (as determined by our index described above) were also given less concrete feedback on the Development question (β = -.098, SE =.007, t(23051) = 14, p < .001) and the Overall question (β = -.037, SE = .007, t(22965) = 5.1, p < .001), though not for the Strengths question (β = -.006, SE =.007, t(23101) = 0.9, p = .366). Finally, we tested the effect of respondent relationship - managers gave more

concrete feedback to their reports than vice versa, on the Development question (β = -.093, SE =.024, t(23049) = 3.9, p < .001) and the Overall question (β = -.047, SE =.022, t(22963) = 2.1, p = .035), with no difference on the Strengths question (β = .028, SE =.023, t(23099) = 1.2, p = .231). All three of these results - showing that feedback is more concrete when it is focused on Development areas; when it is given to poorer-performing employees; and when it is given by managers - are consistent with previous theory, and add further validation for our measure of concreteness as a component of feedback quality.

3.3. Study 1 Discussion

In Study 1, we evaluated an intervention designed to increase concreteness of developmental feedback by directly asking people to provide more specific feedback. However, this intervention seems to mostly have failed in this goal. The treatment caused no significant changes in concreteness of developmental feedback, but it was associated with less concrete descriptions of the subject's strengths, and more concrete descriptions of the subject's overall leadership ability. We also investigated treatment effect heterogeneity and did not find any significant interactions with respect to gender, relationship, subjective performance or respondent characteristics.

In addition, we found across all three questions that prompting individuals to provide more specific feedback led to *lower* response rates on all three questions. And when respondents did write answers, they were shorter in the treatment condition than in the control condition. We speculate that the increased salience of the goal may have created a backlash, whereby people reduced the amount of feedback they gave, rather than expend the extra effort to provide feedback that met the expectations of the question.

We find little evidence in support of Hypothesis 1, suggesting the lack of concrete feedback is not caused by a lack of goal salience at the time of feedback giving. Thus, we hypothesized that to elicit concrete feedback we would need to address the givers' relational concerns. Consequently, in Study 2, we move to test Hypothesis 2 and Hypothesis 3.

4. Study 2

Study 2 was conducted with a very similar design to Study 1 (as depicted in Figure 1), with the following changes. First, we tested three new treatment groups (testing Hypotheses 2 and 3), in comparison to our control condition. We also standardized the structured questions across treatments. Reviewers also reported their own demographics. Finally, we added a new "feedback feedback" survey that allowed subjects to evaluate the feedback they received from their line managers. Our results are thus structured as Study 2A (the initial review survey where feedback is written) and Study 2B (the subjects' evaluations of their line managers' feedback).

4.1. Study 2A Methods

4.1.1 Study Design

We followed the same procedure as described in Study 1 (randomly assigning reviewers to different conditions); however, we included three different treatment arms. See Table 4 for information on the text prompts. We also used the same measures as described in Study 1; however, this time we were able to collect a more complete profile of the demographics of the survey respondent (i.e. the reviewer).

Table 4

Study 2 Question Prompts by Condition

	Q1: Strengths	**Q2: Development**	**Q3: Overall**

Condition 1: Past focused feedback	What have their main leadership strengths been? Please give your feedback, including up to three examples.	What have their main leadership areas to develop been? Please give your feedback, including up to three examples.	Overall, do you have any other feedback for them?
Condition 2: Expected candor feedback reminder	What have their main leadership strengths been? Please give your feedback, including up to three examples. REMEMBER: This person is expecting your candid responses. Please give as direct and honest feedback as you can.	What have their main leadership areas to develop been? Please give your feedback, including up to three examples. REMEMBER: This person is expecting your candid responses. Please give as direct and honest feedback as you can.	Overall, do you have any other feedback for them? REMEMBER: This person is expecting your candid responses. Please give as direct and honest feedback as you can.
Condition 3: Future focused advice reminder	What are this person's greatest strengths that they should keep demonstrating in future? Please give your advice, and include up to three examples.	What are their main leadership areas to develop in future? Please give your advice, and include up to three examples.	Overall, do you have any other advice for them?
Condition 4: Expected candor + future focused advice reminder	What are this person's greatest strengths that they should keep demonstrating in future? Please give your advice, and include up to three examples. REMEMBER: This person is expecting your candid responses. Please give as direct and honest advice as you can.	What are their main leadership areas to develop in future? Please give your advice, and include up to three examples. REMEMBER: This person is expecting your candid responses. Please give as direct and honest advice as you can.	Overall, do you have any other advice for them? REMEMBER: This person is expecting your candid responses. Please give as direct and honest advice as you can.

4.1.2. Participants

The dataset we received contains 48,496 unique responses to a survey (“reviews”). Of these, 5,150 responses were “self-reviews,” where the respondent evaluated themselves as a subject. These responses were written by 23,404 unique respondents, and targeted 5,150 unique

subjects. However, 757 subjects were only present in self-reviews. When these are removed, we are left with a sample of 4,393 unique subjects who received at least one review from another person, and 22,679 unique respondents who wrote at least one review for another person, for a total of 43,346 unique survey responses. See Table 5 for some descriptive statistics about this sample.

The treatment was assigned at the level of respondent - that is, each respondent saw the same survey condition for all of their reviews - at the time the reviewee nominated the reviewer by providing their email address. This left us with 5,898 respondents assigned to the control group, 5,844 assigned to the expected candor group, 5,868 respondents assigned to the future condition, and 5,794 respondents assigned to the expected candor + future condition.

Table 5

Study 2 Sample characteristics

	Control	Future	Benevolence	Both	Total
Unique Reviews	**11,903**	**12,185**	**12,236**	**12,172**	**48,496**
Respondent % Female	33.4	33.1	33.3	33.1	
Subject % Female	41.9	42.2	42.7	42.4	
Relationship					
% Line Managers	10.2	11.0	11.0	11.1	
% Direct Reports	39.2	38.5	38.8	38.1	
% Peers	25.7	26.3	25.1	25.6	
% Others	24.9	24.2	25.1	25.2	
Unique respondents	**5,714**	**5,674**	**5,671**	**5,620**	**22,679**
% Female	32.0%	31.7%	31.6%	31.0%	31.6%
Reviews - Mean (SD)	1.87 (1.97)	1.91 (2.02)	1.93 (2.09)	1.93 (2.03)	1.91 (2.03s)
Unique subjects	**3,987**	**4,007**	**4,023**	**4,019**	**5,150**
% Female	42.8%	42.9%	42.9%	43.1%	42.8%
Reviews - Mean (SD)	2.68 (1.65)	2.71 (1.54)	2.72 (1.61)	2.70 (1.58)	9.87 (4.13)

4.1.2.1. Invitation Rates: Subjects nominated potential respondents, who were then invited to respond. However, participation was voluntary, and the individuals in the focal sample display a wide range of response rates. Although every senior person within the organization was invited to be a subject, we do not have exact data on who accepted those invitations. However, we observe every invitation sent to respondents.

Excluding self-reviews, each respondent was invited to write an average of 1.91 reviews (SD = 2.03, min = 1, max = 23) and each subject received an average of 9.87 reviews (SD = 4.13, min = 6, max = 54). We observe that respondents across conditions were sent the same number of survey invitations ($m_{control}$ = 1.87, $SD_{control}$ = 1.97; $m_{expected}$ = 1.93, $SD_{expected}$ = 2.09; m_{future} = 1.91, SD_{future}= 2.02; m_{both} = 1.93, SD_{both} = 2.03; F(3, 23,400)=1.29, p = .277).

4.1.2.2. Subject Characteristics: Out of all the people invited to respond to the survey, 42.3% were invited to evaluate female employees, 47.6% to male employees, and the remaining 10.1% to employees who either did not disclose their gender identity or identify as “other.” The invited respondents in our data had a variety of relationships with respect to their subjects. In the focal sample, 38.7% of responses involved a direct report reviewing their line manager, 25.7% of responses involved a respondent reviewing one of their peers, 10.8% involved a line manager reviewing one of their direct reports, while the remaining 24.9% were classified as an “other” relationship.

4.1.2.3. Respondent Characteristics: Unlike in Study 1, in Study 2, we can now identify characteristics of all respondents. 33.2% of responses were written by a female respondent, 37.3% were written by a male respondent, and the remaining 29% did not disclose their gender, or reported another gender.

4.1.2.4. Balance Checks: We conducted balance checks to confirm that there were no systematic differences in the make-up of the treatment groups. We found no significant differences in the gender make-up of subjects between the groups ($X^2(6) = 5.0$, $p = .544$), the distribution of cohorts between the groups ($X^2(256) = 272$, $p = .235$), the distribution of relationships between the groups ($X^2(9) = 11.55$, $p = .240$) or the distribution of gender make-up of the respondents ($X^2(6) = 4.49$, $p = .611$).

4.1.3. Measures. The primary outcomes in this study were again the three open-ended text boxes at the end of the survey (see Table 3 for exact text). As planned, we analyze the results of each text box separately, and we estimate all regressions using standard errors clustered at the subject and respondent levels. We again measured the concreteness of the advice people received using the pre-trained advice model from the doc2concrete package (Yeomans, 2021).

With the numeric ratings, we found a similar level of non-response (ranging from 21% - 46%, median = 26%) and the answers were once again quite correlated with one another (cronbach's $\alpha = .96$). We again imputed all missing values, using the mice package in R (Van Buuren & Groothuis-Oudshoorn, 2011). Using those imputed values, we then performed a principal components analysis (PCA) across the remaining set of numeric questions. From that PCA analysis, we extracted the first principal component, reflecting the main axis of variation in the data, as a single abstracted measure of "subjective performance", and there was no difference across conditions in that composite measure ($F(3, 43{,}342) = 1.2$, $p = .293$). We again found that a respondent's subjective performance ratings correlates with "consensus" performance ratings, from the other reviewers ($r = .253$, $t(43344) = 54.6$, $p < .001$) and with the subjects' own self-ratings ($r = .128$, $t(43384) = 27.0$, $p < .001$). These checks again give us some confidence that this constructed performance index is a decent proxy for job performance.

Table 6

Summary statistics and bivariate correlations for Study 2

	Summary Stats			Bivariate Correlations							
Variable name	N	Mean	St. Dev	Gender	Performance	Concrete (Strength)	Concrete (Development)	Concrete (Overall)	Word Count (Strengths)	Word Count (Development)	Word Count (Overall)
Gender (Male = 1)	48,496	0.48	0.50								
Performance (standardized)	48,496	0	1	-0.03***							
Concreteness (Strengths)	36,023	-0.10	0.40	-0.01	-0.02***						
Concreteness (Development)	32,502	0.01	0.43	0.01*	-0.09***	0.16***					
Concreteness (Overall)	26,693	-0.21	0.34	0.03***	-0.07***	0.12***	0.11***				
Word Count (Strengths)	48,496	53.5	65.1	-0.04***	0.13***	0.05***	0.02***	-0.08***			
Word Count (Development)	48,496	41.4	56.9	-0.02***	-0.11***	-0.02***	0.11***	-0.07***	0.51***		
Word Count (Overall)	48,496	23.8	37.6	-0.02***	-0.03***	0.05***	0.06***	0	0.31***	0.35***	
Feedback Index (Study 2B ONLY) (standardized)	1,345	0	1	0.03***	0.06***	-0.02	-0.01	-0.01	0.02	0.01	0

*Note.*St. Dev., standard deviation. *$p<0.05$; **$p<0.01$; ***$p<0.001$

4.2. Study 2A Results

For all of our analyses, we use the same common statistical model from our pre-registration and from Study 1. Once again, we report the focal estimates and hypothesis tests in the main text, and include full regression tables in Appendix B. Additionally, for clarity, we report only the two main effects (expected candor and future focus) in the main text, and compile the analyses of interactions between the two interventions in Appendix D. In general, we found little evidence for any interaction between the interventions.

4.2.1. Non-Response. Out of all the invitations sent, 50.9% completed every question, 28.8% answered some of the questions, and 20.4% answered no questions. Roughly a third of participants did not write a single word in one text box (Q1: 25.7%; Q2: 33.0%; Q3: 45.0%), with 25.1% not writing a single word in any of the text boxes. We again focused our analysis on responses that were longer than five words (see Appendix C for details). Again, we confirm our results are robust to other cut-off strategies, including quantile cut-offs that remove a similar number of texts from each condition (thus warding off selection bias). Our chosen cut-off results in samples of 30,164 responses to question 1, 25,774 responses to question 2, and 20,650 responses to question 3, with a total of 17,160 respondents and 4,389 subjects.

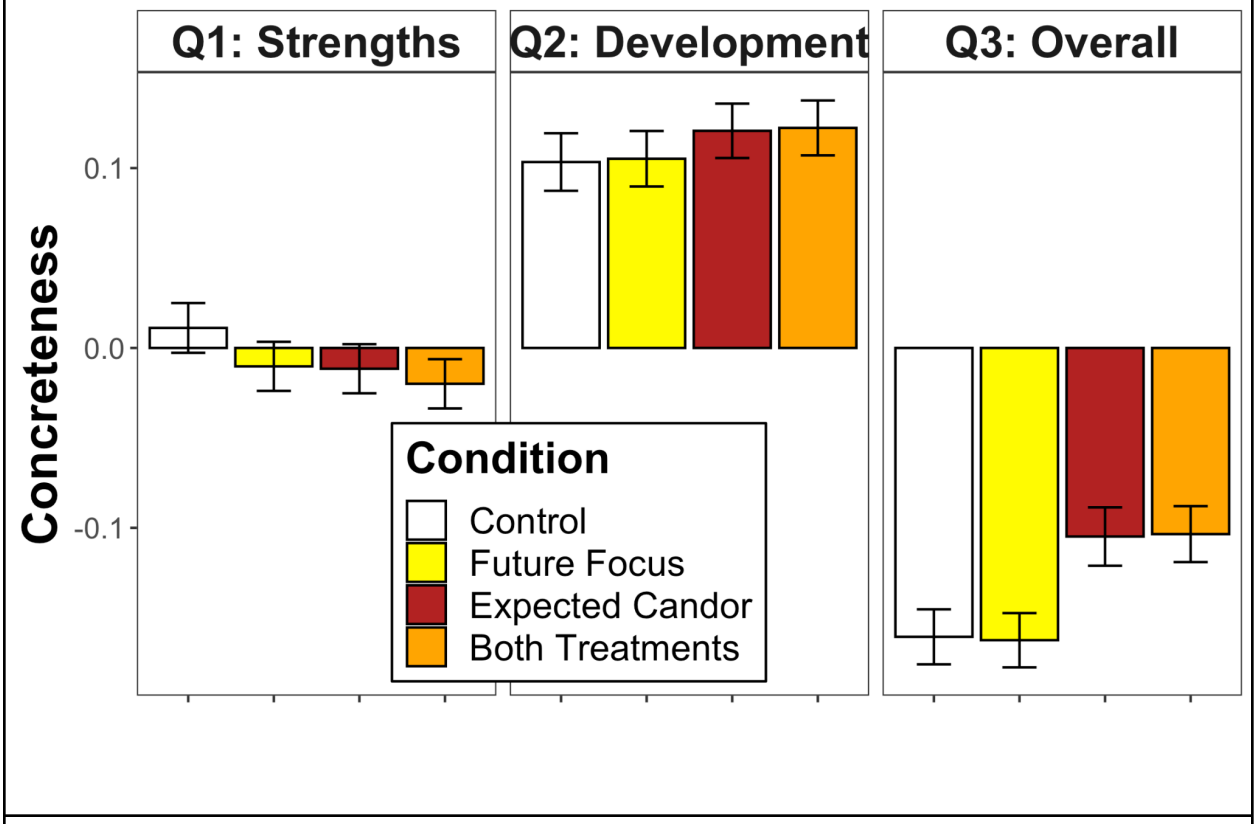
At our chosen cut-off, our interventions had no effect on whether the Strengths question was answered (future focus: $\beta = .009$, $SE = .007$, $t(43343) = 1.4$, $p = .169$; expected candor: $\beta = -.003$, $SE = .007$, $t(43343) = 0.4$, $p = .705$). However, both interventions increased the completion rate of the Development question (future focus: $\beta = .026$, $SE = .007$, $t(43343) = 3.7$, $p < .001$; expected candor: $\beta = .023$, $SE = .007$, $t(43343) = 3.2$, $p = .001$), and had opposite effects on the Overall question (future focus: $\beta = .033$, $SE = .007$, $t(43343) = 4.5$, $p < .001$; expected candor: $\beta = -.046$, $SE = .007$, $t(43343) = -6.3$, $p < .001$).

4.2.2. Response Length. After removing texts less than five words long, the remaining texts were all substantial (Strengths: mean = 77.3 words, SD = 64.2 words; Development: mean = 69.8 words, SD = 60.1 words; Overall: mean = 48.8 words, SD = 37.2 words). We also find some differences in length among the remaining texts. The future focus condition increased the length of feedback in all three text boxes (Strengths: $\beta = 2.24$, SE = 1.04, $t(30161) = 2.1$, $p = .032$; Development: $\beta = 4.74$, SE = 1.01, $t(25771) = 4.7$, $p < .001$; Overall: $\beta = 3.59$, SE = .69, $t(20647) = 5.2$, $p < .001$). The expected candor condition decreased the length of feedback in two text boxes (Strengths: $\beta = -4.96$, SE = 1.06, $t(30161) = 4.7$, $p < .001$; Overall: $\beta = -2.31$, SE = .68, $t(20647) = 3.4$, $p < .001$) but had no effect on the length of the Development feedback ($\beta = -.78$, SE = 1.0, $t(25771) = 0.8$, $p = .442$).

4.2.3. Concreteness. Our primary hypothesis concerned the concreteness of the subjects' reviews. In Figure 4, we display the means for each text box by condition.

Figure 4

Concreteness Results for Study 2A



Average standardized concreteness scores for all three text responses, by treatment assignment. Each point represents a group mean and 95% CI.

4.2.3.1. Development. The future focus intervention had no significant impact on the concreteness of the Development review text (β = -.002, SE = .014, t(25771) = .11, p = .912), while the expected candor intervention increased the concreteness of the Development text (β = .040, SE = .014, t(25771) = 2.8, p = .005). These results held in a model where we also control for performance, gender, and relationship (future focus: β = -.003, SE = .014, t(25763) = .23, p = .818; expected candor: β = .041, SE = .014, t(25763) = 2.9, p = .004). We can also use these control variables to benchmark the effect size of the candor intervention. The impact of the candor intervention on concreteness was approximately 29% of the difference in concreteness between line managers' developmental feedback to their direct reports, and reports' feedback

back to their managers. Alternatively, the effect of the intervention was equivalent to the increase in concreteness to be expected from a 0.51 standard deviation decrease in the subjective performance of the recipient (i.e. as judged by the reviewer).

4.2.3.2. Strengths. The future focus intervention reduced the concreteness of the Strengths review text (β = -.046, SE = .014, t(30161) = 3.2, p = .001), as did the expected candor intervention (β = -.046, SE = .014, t(30161) = 3.2, p = .001). These results held in a model where we also control for performance, gender, and relationship (future focus: β = -.044, SE = .014, t(30155) = 3.2, p = .002; expected candor: β = -.045, SE = .014, t(30155) = 3.2, p = .002). We do not benchmark these results against subjective performance, as it was not actually correlated with concreteness of the strengths text (see 3.2.4. below). Further, line managers gave *less* concrete strengths feedback than their reports - still, the effect of both of these interventions was roughly 44% of the difference between managers and reports.

4.2.3.3. Overall. The future focus intervention had no significant impact on the concreteness of the Overall review text (β = .003, SE = .016, t(20647) = .19, p = .848), while the expected candor intervention increased the concreteness of that text (β = .174, SE = .016, t(20647) = 11, p < .001). These results held in a model where we also control for performance, gender and relationship (future focus: β = -.000, SE = .016, t(20641) = 0.0, p = .991; expected candor: β = .171, SE = .016, t(20641) = 11, p < .001). This was arguably the largest treatment effect we measured. As a benchmark, the effect size of the candor intervention was roughly equivalent to a 2.2 standard deviation decrease in subjective performance, or 259% of the difference in concreteness between line managers and their reports.

4.2.4. Outcome Validation. We again conducted validation checks of our measure of concreteness. We again saw differences across questions. The Development question elicited

feedback that was much more concrete than the Strengths question (β = .279, SE = .009, t(62369) = 33, p < .001). Furthermore, we found that employees who were rated as having higher subjective performance were also given less concrete feedback on the Development question (β = -.079, SE =.007, t(25772) = 11.4, p < .001) and the Overall question (β = -.082, SE =.008, t(20648) = 10.8, p < .001), though not for the Strengths question (β = -.007, SE =.006, t(30162) = 1.1, p = .255). Finally, we again found that managers gave more concrete feedback to their reports than vice versa, on the Development question (β = .142, SE =.021, t(29057) = 6.6, p < .001) and the Overall question (β = .083, SE =.025, t(22581) = 3.2, p = .001), although they gave less concrete feedback on the Strengths question (β = -.099, SE =.022, t(33574) = 4.5, p < .001). These results are consistent with Study 1, adding further validation to our measure.

4.3. Study 2B Methods

4.3.1. Study Design. Subjects were sent a follow-up survey after receiving the feedback. The email containing the survey said, "*We would like you to take a 2-question survey to tell us what you think of the 360 degree feedback you have received. We are trying to help managers deliver better quality feedback. We will only conduct our analyses in combination with everyone else's responses so your response will be kept confidential and anonymous. Thank you for helping us improve the 360 degree feedback process."*

Once subjects opened the survey, they read "*We would like to know what you think of the open-ended feedback you have received from your manager. For these questions, only focus on your line manager's response (and ignore the feedback you received from your reports and peers)*."

They were then asked to answer the extent to which they agreed with the following statements (on a 1-7 Likert Scale from Extremely Disagree to Extremely Agree): "*My manager's*

feedback is an accurate reflection of my past performance," *"My manager's feedback is helpful for my future performance,"* and "*My manager's feedback makes me feel motivated to perform my job.*"

4.3.2. Participants. All 4,139 subjects who received written feedback from their line managers were invited to participate in Study 2B. Of those sent the survey, 1,345 (32.5%) responded to all three questions. The best predictor of whether someone participated in the Study 2B was whether their line manager had written feedback ($\beta = .043$, $SE = .015$, $t(4137) = 2.9$, $p = .004$). Accordingly, we exclude responses from people whose manager did not write any feedback, leaving 2,514 responses to the follow-up survey.

Among those whose managers did write them feedback, response rates did not differ by treatment condition ($F(3, 2510) = 0.3$, $p = .797$). Response rates were also not predicted by subjective performance ratings ($\beta = .010$, $SE = .011$, $t(2512) = 0.9$, $p = .350$) and were slightly higher among male subjects than female subjects ($\beta = .036$, $SE = .020$, $t(2511) = 1.8$, $p = .069$). Based on these results, we then decided to focus our analyses in Study 2B on the 859 participants who completed the followup survey, and whose managers had given them written feedback.

4.4. Study 2B Results

For all of our analyses, we use the same common statistical model from our pre-registration and from Study 1 and Study 2A. Once again, we report the focal estimates and hypothesis tests in the main text, and include full regression tables in Appendix B. Additionally, we report only the two main effects and compile the analyses of interactions between the two interventions in Appendix D (we found no meaningful interactions between the two interventions).

4.4.1. Treatment Effects

4.4.1.1. Accuracy. Neither treatment significantly impacted ratings of perceived accuracy of feedback (future focus: β = -.124, SE = .080, t(856) = 1.6, p = .120; expected candor: β = -.015, SE = .079, t(856) = 0.2, p = .855). These results held in a model where we also control for gender, performance rating by the manager, gender and the average of how other people rated the participant (future focus: β = -.125, SE = .076, t(852) = 1.7, p = .098; expected candor: β = -.007, SE = .075, t(852) = 0.1, p = .927).

4.4.1.2. Helpfulness. Neither treatment significantly impacted ratings of helpfulness (future focus: β = -.117, SE = .074, t(856) = 1.6, p = .117; expected candor: β = -.001, SE = .074, t(856) = 0.1, p = .929). These results held in a model where we also control for gender, performance rating by the manager, gender and the average of how other people rated the participant (future focus: β = .121, SE = .072, t(852) = 1.7, p = .096; expected candor: β = -.003, SE = .072, t(852) = 0.0, p = .965).

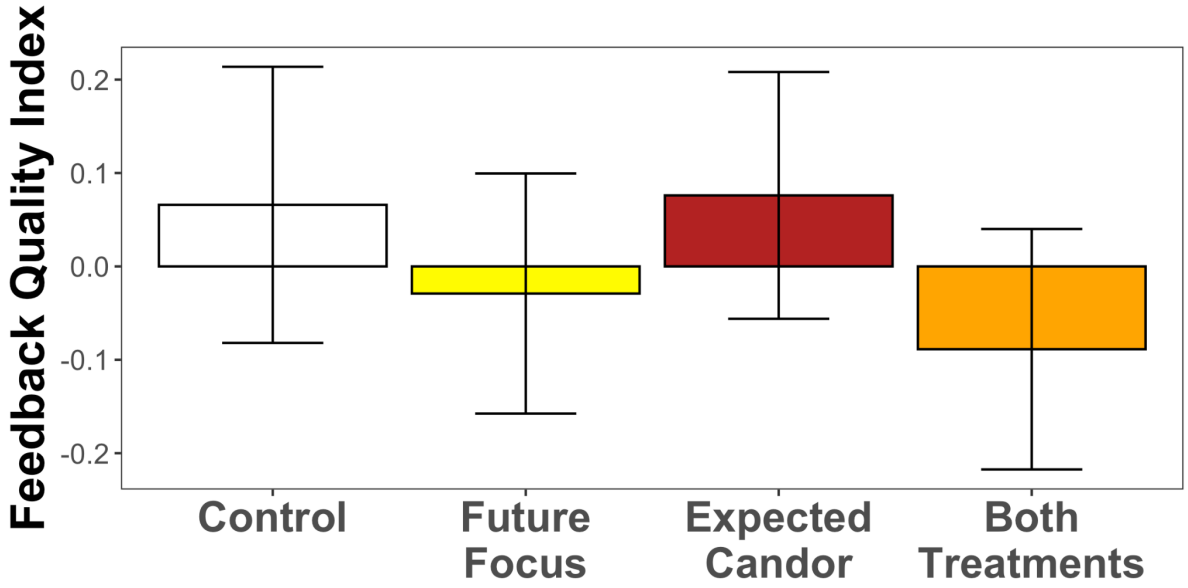
4.4.1.3. Motivation. There, we saw a slight negative effect of future focus on how motivating the feedback was (β = .190, SE = .090, t(856) = 2.1, p = .036) but no effect of expected candor (β = -.076, SE = .090, t(856) = 0.8, p = .405). And these results held in a model where we also control for gender, performance rating by the manager, gender and the average of how other people rated the participant (future focus: β = -.195, SE = .087, t(852) = 2.2, p = .025; expected candor: β = -.071, SE = .087, t(852) = 0.8, p = .416).

4.4.1.4. Combined Index. Although we intended to analyze the three questions separately, they ended up being highly correlated with one another (cronbach's α = .92) so we also combined them into a standardized index for clarity (the results on the individual questions mirror the aggregated results). There, we saw a slight negative effect of future focus (β =- .117,

SE = .063, t(856) = 1.9, p = .062) but still no effect of expected candor (β = -.025, SE = .063, t(856) = 0.4, p = .689). And these results held in a model where we also control for gender, performance rating by the manager, gender and the average of how other people rated the participant (future focus: β = -.120, SE = .060, t(852) = 2.0, p = .046; expected candor: β = -.021, SE = .060, t(852) = .35, p = .728). We plot the group means of this index across conditions in Figure 5.

Figure 5

Feedback Quality Results for Study 2B



Average standardized feedback quality index scores for all three text responses, by treatment assignment. Each point represents a group mean and 95% CI.

4.4.2. Other Analyses

4.4.2.1. Performance. The strongest predictor of feedback quality ratings was subjective performance - the better a reviewer thought the subject performed, the more the subject thought

the feedback was high-quality (β =.304, SE = .039, t(857) = 7.76, p < .001). These results held in a model where we also control for gender, performance rating by the manager, gender and the average of how other people rated the participant (β = .303, SE = .042, t(854) = 7.18, p < .001).

4.4.2.2. Concreteness. Concreteness of any question did not predict perceptions of feedback quality (Strengths: β = .083, SE = .077, t(857) = 1.08, p = .28; Development: β = - .041, SE = .078, t(857) = -.52, p = .601; Overall: β =- .120, SE = .094, t(857) = -1.28, p = .201). These results held in a model where we also control for gender, performance rating by the manager, gender and the average of how other people rated the participant (Strengths: β = .109, SE = .074, t(853) = 1.48, p = .140; Development: β = - .020, SE = .075, t(853) = -.27, p = .786; Overall: β =- .074, SE = .087, t(853) = -.848, p = .397).

3.4.2.3. Word count. When looking at the word count of each open-ended question on perceptions of feedback quality, while we find no impact of word count of the Strengths question (β = .001, SE = .001, t(857) = 1.29, p = .197), we find a significant negative impact of word count of the Development question on perceptions of feedback quality (β = - .002, SE = .001, t(857) = -2.42, p = .016) and a marginally negative impact of word count of the Overall question (β =- .002, SE = .001, t(857) = -1.663, p = .097). These results held in models where we also control for gender, performance rating by the manager, gender and the average of how other people rated the participant (Strengths: β = .000, SE = .001, t(853) =.542, p = .588; Development: β = - .001, SE = .057, t(853) = -2.241, p = .025; Overall: β =- .002, SE = .001, t(853) = -1.731, p = .084).

3.4.2.4. Subject Condition. Because our intervention was assigned at the respondent level, only some subjects were exposed (randomly) each of the interventions. As an exploratory analysis, we investigated whether subjects' condition affected their ratings of their manager. We

find that there is some evidence that the interventions changed subjects' expectations. While feedback evaluations were unchanged when the subject had seen a future focus message (β = .023, SE = .061, t(853) = 0.4, p = .701), subjects gave lower evaluations of their manager when they themselves had been in the expected candor condition (β = -.143, SE = .064, t(854) = 2.2, p = .027). And we found no significant interactions between subject and reviewer condition (all p > .20). This suggests that the expected candor intervention may have increased subjects' expectations of candid feedback, and thus been harder judges of their line managers' feedback.

4.5. Study 2 Discussion

In Study 2A, we tested two new ways to increase concreteness of developmental feedback. Unlike in Study 1, both of our interventions *increased* response rates on the Development question, while having no impact on the response rate on the Strengths question. However, whereas the future focus intervention increased response rates on the Overall question, the expected candor intervention decreased response rates on that final question. When respondents did write reviews, those in the future focus intervention wrote more than the control condition for all three questions. The expected candor intervention had no impact on the length of the responses to the Development question while significantly decreasing the length of the responses to the Strengths and Overall questions.

Focusing on the content of the feedback, we found that the expected candor intervention was the only one that significantly increased the concreteness of the Development responses. This intervention also significantly increased the concreteness of the Overall response. Meanwhile, the future focus intervention had no impact on the concreteness of the Development condition or Overall condition. Both interventions significantly reduced the concreteness of the Strengths review text. These results suggest that the relational concerns were posing a significant

barrier in the control condition, while the expected candor intervention reassured reviewers about the subjects' expectations, enabling reviewers to share more concrete development feedback without the fear of relational backlash.

In light of this result, it was essential for us to investigate whether concrete feedback does in fact produce a backlash effect. In Study 2B, we asked subjects to evaluate their managers' feedback, and we found the intervention did not affect reported helpfulness, accuracy, or motivation (or the index of all three). We interpret the null effect of expected candor in Study 2B as ruling out the possibility of large backlash effects. Although we had a large sample size, we were still concerned that we may have been underpowered to detect any backlash effect. However, we were well-powered enough to observe a slight backlash effect from the future focus intervention, giving us more comfort in interpreting the null effect of the expected candor intervention. These results imply that increasing the concreteness of developmental feedback does not demotivate recipients, nor do recipients consider more concrete feedback as less accurate. Recipients do not self-report more concrete feedback from their managers as more helpful, but future research should try to obtain observable downstream behavioral consequences such as future ratings, promotion, and pay increases.

In our exploratory analyses we also found two interesting results. First, that feedback recipients were more influenced by the numeric, versus qualitative, nature of their managers' feedback. In order to rule out that better quality employees are just more likely to find their feedback more helpful, accurate, and motivating, we are able to control for others' ratings of them. If it were the case that it has to do with the person being rated, not the manager's ratings, then others' ratings should also have predictive power; however, we do not find this to be the case. Instead, we conclude that participants weigh the numeric feedback given by their managers

more highly than the specific text feedback (just like students who flip to the back of a term paper to see their grade and do not engage with their teachers' feedback comments).

Finally, we also found that recipients evaluated their managers more harshly when they had themselves given feedback in an expected candor condition. Whether or not this effect is robust in future results, it speaks to a broader question about blinding during implementation. In our experiments, due to the constraints of the organizational context, we were unable to notify participants about their reviewer's condition (though some may have assumed their manager had the same prompt that they did). This also provided a cleaner test of the effect of the intervention on the language per se. However, in practice it may be impossible to blind recipients to the expected candor intervention (for example if the intervention is adopted firm-wide). We argue this makes the lack of a backlash effect in Study 2B all the more compelling. Had subjects known that their managers were told to be more candid, they may have avoided a negative reaction by attributing any undue candor to the prompt, rather than the manager, thereby preserving the relationship.[1] And yet, even though this rationalization was not available to most recipients, we still failed to detect a backlash in these data.

5. General Discussion

These studies were designed to test interventions that might improve developmental feedback based on three theoretical accounts of potential barriers to concrete feedback. Our null effects in Study 1, where we reminded people in a timely manner to provide more specific feedback (thereby activating the goal of giving concrete feedback), indicates that goal salience (Fishbach and Ferguson, 2007; Kruglanski et al., 2002) is not likely to improve concrete feedback provision. Not only did we fail to detect a significant effect of treatment on

[1] We do not think it was dishonest to construct the expected candor intervention without eliciting the recipients' expectations as most employees do want more candid feedback (Zenger & Folkman, 2014).

concreteness of developmental feedback, but we also detected a potential backlash, whereby people reduced the amount of feedback they gave, rather than expend the extra effort to provide feedback that met the expectations of the question.

In Study 2A we found positive evidence that the question prompt could be redesigned to elicit more concrete feedback. Our "expected candor" intervention (Hypothesis 2) framed candid feedback requests as something the target expects, licenses the feedback-giver to provide critical, constructive feedback without having to be worried about relational harm (Jampol & Zayas, 2021; Levine & Schweitzer, 2014; Levine, Roberts and Cohen, 2018; Lupoli, Jampol & Oveis, 2017; Lupoli, Levine & Greenberg, 2018; Yeomans, Schweitzer & Brooks, 2022). However, our "future focus" intervention (Hypothesis 3) prompted people to provide forward-looking advice will reduce the constraints of backward looking feedback (Linsey, Tseng, Fu, Cagan, Wood, & Schunn, 2010; Youmans & Arciszewski, 2014) and instead give more concrete advice focusing on future actions (Brooks, Gino, & Schweitzer, 2015; Levari, Gilbert & Wilson, 2022). We found that both treatments avoided the reduction in survey response rates that we observed in Study 1. Furthermore, both treatments increased the likelihood of writing text in the development box. However, only expected candor increased the concreteness of developmental feedback.

Finally, in Study 2B we ask feedback recipients about their impressions of the feedback. We observe no interpersonal backlash from the expected candor intervention. This result casts doubt on the idea that that concrete feedback will not be well-received by recipients.

5.1 Theoretical Implications

These studies help us understand a key barrier - and propose a viable solution - for prompting people to provide better developmental feedback. Our research contributes to the literature on the trade-off between benevolence and honesty (Levine & Cohen, 2018; Levine &

Gomez, 2019; Levine & Schweitzer, 2014; Levine, Roberts & Cohen, 2019; Yeomans, Schweitzer & Brooks, 2022). We provide new evidence showing how this trade-off is managed in the context of a common and important organizational task. While these feedback exercises are meant to improve the information flow within organizations, this goal can be hindered by the respondents' relational goals. The interventions aimed at addressing two potential non-social mechanisms - goal salience and backward looking evaluative mindsets - did not lead to more concrete developmental feedback. This adds to our understanding of the social restraints on effective feedback, and also shows how these restraints can be eased when feedback is elicited.

Moreover, we illustrate the role of expectations in feedback givers' minds that lead to fear of interpersonal conflict. Feedback givers may have genuine concerns that the information they provide might demotivate recipients, or else harm their relationship. These concerns may be an important contributor to the problems of "inflated" feedback (Waung & Highhouse, 1997). Our expected candor intervention directly targets the expectations of the feedback giver, and therefore encourages them to provide constructive, concrete feedback.

Finally, we demonstrate that some of these natural concerns may not be warranted. Recipients did not seem to be bothered by the increased concreteness in the expected candor intervention. Our failure to detect a backlash against managers who provided more concrete feedback or against those in the expected candor condition suggests that the interpersonal fears of concrete feedback provision may not be fully grounded in reality.

5.2 Practical Implications

In order to improve the quality of feedback in organizations, it is not enough to tell employees to provide more constructive feedback to their colleagues. As we have demonstrated, merely reminding people to give better feedback, or even "specific examples" is not enough

because it does not mitigate the relational concerns. Instead, employees can be prompted to provide feedback in a way that establishes the recipients' expectations of a candid response. In particular, our prompt explains that giving candid feedback is important for an employee's growth and development.

We do not believe this effect is unbounded - one can certainly imagine more strongly-worded feedback prompts that could elicit feedback which does incur a relational backlash. However, our simple re-writing of the prompt still nudges feedback givers, at the margin, to notice that there is room for them to be more concrete. The organizational constraints required us to design a light touch intervention - but these constraints are surely present in many other situations. By embedding this virtually costless intervention into pre-existing communications, our work demonstrates an actionable protocol for organisations to adopt our intervention in their own feedback procedures (DellaVigna, Kim & Linos, 2022).

5.3. Limitations

Our study design was a large field experiment conducted in collaboration with a large public sector organization. This has both benefits and costs. For one, we were limited in what we could add to their existing annual 360 review processes. Therefore, we were not able to directly ask participants about our proposed mechanism by which expected candor increases the quality of feedback - relational concerns. We were also unable to link our Study 1 or Study 2 outcomes to observed behaviors such as future ratings, salary increases or promotions. Finally, people could not be required to complete the survey; accordingly, response rates for all three studies were lower than both the organization and the researchers would have liked. While we analyzed non-response bias, we still wonder whether the interventions might encourage better feedback from the people who were unwilling to give it at all.

Our study was also limited to a single organization, and there are natural questions about the effects of our intervention in other contexts. We speculate that public-sector organization employees might be especially attuned to the needs of others, and thus these relational concerns might be especially focal for them. On the other hand, our study was conducted in a western industrialized nation, where people are perhaps more individually oriented and less collectively minded. Moreover, given that it was conducted in the United Kingdom, where there are specific cultural norms that dictate how directly people engage with one another, our findings may be a conservative test of this intervention or they may not generalize to different English-speaking contexts. It is worth mentioning that our measure of concreteness may also not generalize to other languages, or to other settings where the conversational norms are different.

Finally, we were limited in the extent of our intervention. We could only affect people's feedback processes at the time of recall, when they were writing within the 360 review survey. Other interventions might be more effective if they target feedback givers during encoding - that is, during the normal operations of an organization as the feedback giver is observing the recipients' performance. While annual 360 reviews are an important time for people to reflect and learn, feedback can be a year-round process, given throughout the course of their work together. In these settings, it may be especially difficult to support personal development while productivity goals are top-of-mind.

These reviews are also conducted via a third-party intermediary and the feedback recipient is not the one directly asking for feedback, so feedback givers may not pay attention to the prompt text, or may not believe that subjects had the expectation described in the expected candor intervention. Given the effect of medium on feedback specificity and communication (Waung & Highhouse, 1997; Schroderer, Kardas & Epley, 2017), it is possible that our

intervention may be limited in scope to when feedback-givers are asked to provide written communication via a third party. Therefore, there are many remaining questions about how development can be improved through these other communication channels and means of feedback solicitation.

5.4. Future Directions

In this organization, as in many, employees were asked to provide written feedback for their managers, peers, and direct reports. Future research needs to investigate how expected candor might affect feedback in different delivery modes - in a face-to-face meeting, for example. It is possible that expected candor may be more effective in person, where a feedback giver can express such intentions explicitly (e.g., "I am telling you this because you will benefit from the feedback and you deserve my honesty"). Research has shown that expressing expected intentions makes recipients more receptive to more critical feedback (Yeager et al., 2014), so this same mechanism may license feedback givers to be more direct, as well.

Additionally, in many organizations, including the one in this study, developmental feedback is coupled with evaluation, leading feedback givers to have to both provide appraisals as well as input that helps employees develop on the job (Li, Harris, Boswell, & Xie, 2011). Decoupling feedback from evaluation could provide a similar ability for employees to provide expected candid feedback without fearing for their relationship or fearing that their seemingly critical input would be mistaken for an assessment of poor performance.

Furthermore, future research needs to link feedback with more downstream consequences (both observationally, and when elicited by our expected candor intervention). Study 2B here measures some of the immediate aftereffects of increased concreteness, with respect to initial motivation and perceived accuracy and helpfulness. However, perceptions of accuracy and

helpfulness may differ substantially over time, after the recipients have a chance to act on the feedback they are given. In fact, these may be impossible to measure immediately, especially for complex jobs within a real organization. Future research could be conducted to study retrospective reports about helpfulness and accuracy. Further research could also be conducted to look at the behavioral consequences of more concrete feedback, especially in situations where researchers can objectively evaluate performance.

5.5 Conclusion

Interpersonal concerns seem to be unnecessarily holding people back from providing concrete feedback to their work colleagues. By making explicit that their colleagues expect direct and honest feedback, employees can be encouraged to give more concrete feedback without fearing or incurring relational penalties.

6. References

Austin, J., Sigurdsson, S. O., & Rubin, Y. S. (2006). An examination of the effects of delayed versus immediate prompts on safety belt use. *Environment and behavior,* 38(1), 140-149.

Ashford, S. J., & Cummings, L. L. (1983). Feedback as an individual resource: Personal strategies of creating information. *Organizational behavior and human performance,* 32(3), 370-398.

Baron, R. A. (1988). Negative effects of destructive criticism: Impact on conflict, self-efficacy, and task performance. *Journal of Applied Psychology,* 73(2), 199.

Bies, R. J. (2013). The delivery of bad news in organizations: A framework for analysis. *Journal of Management,* 39(1), 136-162.

Blunden, H., & Gino, F. (2018). How the other half thinks. In The Oxford handbook of advice.

Blunden, H., Yoon, J., Kristal, A.S., Whillans, A.V. (working paper). "Soliciting Advice Rather

Than Feedback Yields More Developmental, Critical, and Actionable Input." Harvard Business School Working Paper, No. 20-021, April 2021.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. Organizational behavior and human decision processes, 101(2), 127-151.

Bond, C. F., & Anderson, E. L. (1987). The reluctance to transmit bad news: Private discomfort or public display?. *Journal of Experimental Social Psychology,* 23(2), 176-187.

Brooks, A. W., Gino, F., & Schweitzer, M. E. (2015). Smart people ask for (my) advice: Seeking advice boosts perceptions of competence. Management Science, 61(6), 1421-1435.

Bruch, H., & Ghoshal, S. (2002). *Beware the busy manager.*

Cannon, M. D., & Witherspoon, R. (2005). Actionable feedback: Unlocking the power of learning and performance improvement. Academy of Management Perspectives, 19(2), 120-134.

Correll, S. J., Weisshaar, K. R., Wynn, A. T., & Wehner, J. D. (2020). Inside the black box of organizational life: The gendered language of performance assessment. *American Sociological Review,* 85(6), 1022-1050.

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: the aversion to harmful action. *Emotion,* 12(1), 2.

DellaVigna, S., Kim, W., & Linos, E. (2022). Bottlenecks for Evidence Adoption.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of personality and social psychology,* 70(5), 979.

Dibble, J. L., & Levine, T. R. (2010). Breaking good and bad news: Direction of the MUM effect and senders' cognitive representations of news valence. *Communication Research,* 37(5),

703-722.

Fedor, D. B. (1991). Recipient responses to performance feedback: A proposed model and its implications. *Research in personnel and human resources management*, 9(73), 120.

Fife-Schaw, C., Sheeran, P., & Norman, P. (2007). Simulating behaviour change interventions based on the theory of planned behaviour: Impacts on intention and action. *British journal of social psychology,* 46(1), 43-68.

Finkelstein, S. R., Fishbach, A., & Tu, Y. (2017). When friends exchange negative feedback. *Motivation and Emotion,* 41(1), 69-83.

Fishbach, A., & Ferguson, M. J. (2007). The goal construct in social psychology.

Fulham, N. M., Krueger, K. L., & Cohen, T. R. (2022). Honest feedback: Barriers to receptivity and discerning the truth in feedback. *Current Opinion in Psychology,* 101405.

Gino, F., & Grant, A. M. (2015). What's in it for me? The effects of learning and helping goals on advice giving. Unpublished manuscript.

Goffman, E. (1967). On face-work. Interaction ritual, 5-45.

Gong, Y., Wang, M., Huang, J. C., & Cheung, S. Y. (2017). Toward a goal orientation–based feedback-seeking typology: Implications for employee performance outcomes. Journal of Management, 43(4), 1234-1260.

Goodman, J. S., Wood, R. E., & Hendrickx, M. (2004). Feedback specificity, exploration, and learning. *Journal of Applied Psychology,* 89(2), 248.

Graham, N., Arai, M., & Hagstroemer, B. (2016). multiwayvcov: Multi-way standard error clustering. R package version, 1(3).

Higgins, E. T. (1996). Activation: Accessibility, and salience. Social psychology: Handbook of basic principles, 133-168.

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of applied psychology,* 64(4), 349.

Jampol, Lily, and Vivian Zayas. "Gendered white lies: Women are given inflated performance feedback compared with men." *Personality and Social Psychology Bulletin* 47, no. 1 (2021): 57-69.

Karlan, D., Ratan, A. L., & Zinman, J. (2014). Savings by and for the Poor: A Research Review and Agenda. *Review of Income and Wealth*, 60(1), 36-78.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin,* 119(2), 254.

Kopelman, R. E. (1986). Objective feedback. *Generalizing from laboratory to field settings,* 119, 145.

Kraft, M. A., & Rogers, T. (2015). The underutilized potential of teacher-to-parent communication: Evidence from a field experiment. *Economics of Education Review,* 47, 49-63.

Kristal, A. S., & Whillans, A. V. (2020). What we can learn from five naturalistic field experiments that failed to shift commuter behaviour. *Nature Human Behaviour,* 4(2), 169-176.

Kruglanski, A. W., Shah, J. Y., Fishbach, A., Friedman, R., Chun, W. Y., & Sleeth-Keppler, D. (2002). A theory of goal systems. Advances in experimental social psychology, 34(2), 331-378.

Lambert, B., Caza, B. B., Trinh, E., & Ashford, S. (2022). Individual-centered interventions:

identifying what, how, and why interventions work in organizational contexts. Academy of Management Annals, 16(2), 508-546.

Latham, G. P., & Locke, E. A. (1991). Self-regulation through goal setting. *Organizational behavior and human decision processes,* 50(2), 212-247.

Levari, D., Gilbert, D. T., & Wilson, T.D., (2022). Tips From the Top: Do the Best Performers Really Give the Best Advice? *Psychological Science*

Levine, E. E., & Cohen, T. R. (2018). You can handle the truth: Mispredicting the consequences of honest communication. *Journal of Experimental Psychology: General,* 147(9), 1400.

Levine, E., Hart, J., Moore, K., Rubin, E., Yadav, K., & Halpern, S. (2018). The surprising costs of silence: Asymmetric preferences for prosocial lies of commission and omission. *Journal of personality and social psychology,* 114(1), 29.

Levine, E., & Munguia Gomez, D. (2021). "I'm just being honest." When and why honesty enables help versus harm. *Journal of Personality and Social Psychology,* 120(1), 33.

Levine, E. E., Roberts, A. R., & Cohen, T. R. (2020). Difficult conversations: Navigating the tension between honesty and benevolence. *Current opinion in psychology,* 31, 38-43.

Levine, E. E., & Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. *Journal of Experimental Social Psychology,* 53, 107-117.

Li, N., Harris, T. B., Boswell, W. R., & Xie, Z. (2011). The role of organizational insiders' developmental feedback and proactive personality on newcomers' performance: an interactionist perspective. Journal of Applied Psychology, 96(6), 1317.

Linsey, J. S., Tseng, I., Fu, K., Cagan, J., Wood, K. L., & Schunn, C. (2010). A study of design fixation, its mitigation and perception in engineering design faculty.

Lix, K., Goldberg, A., Srivastava, S. B., & Valentine, M. A. (2022). Aligning differences:

Discursive diversity and team performance. *Management Science.*

Lupoli, M. J., Jampol, L., & Oveis, C. (2017). Lying because we care: Compassion increases prosocial lying. *Journal of Experimental Psychology: General,* 146(7), 1026.

Lupoli, M. J., Levine, E. E., & Greenberg, A. E. (2018). Paternalistic lies. *Organizational Behavior and Human Decision Processes,* 146, 31-50.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review,* 20(3), 709-734.

Mishra, G., & Farooqi, R. (2013). Exploring employee satisfaction with performance management and the challenges faced in context of IT industry. Compensation & Benefits Review, 45(6), 329-339.

Moore, A. K., Munguia Gomez, D. M., & Levine, E. E. (2019). Everyday dilemmas: New directions on the judgment and resolution of benevolence–integrity dilemmas. *Social and Personality Psychology Compass,* 13(7), e12472.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Sage.

Rhodes, R. E., & de Bruijn, G. J. (2013). How big is the physical activity intention–behaviour gap? A meta‐analysis using the action control framework. *British journal of health psychology,* 18(2), 296-309.

Rhodes, R. E., & Dickau, L. (2012). Experimental evidence for the intention–behavior relationship in the physical activity domain: A meta-analysis. *Health Psychology,* 31(6), 724.

Rogers, T., & Milkman, K. L. (2016). Reminders through association. *Psychological science,* 27(7), 973-986.

Rosen, S., & Tesser, A. (1970). On reluctance to communicate undesirable information: The MUM effect. *Sociometry,* 253-263.

Shea, S., DuMouchel, W., & Bahamonde, L. (1996). A meta-analysis of 16 randomized controlled trials to evaluate computer-based clinical reminder systems for preventive care in the ambulatory setting. *Journal of the American Medical Informatics Association,* 3(6), 399-409.

Schaerer, M., Kern, M., Berger, G., Medvec, V., & Swaab, R. I. (2018). The illusion of transparency in performance appraisals: When and why accuracy motivation explains unintentional feedback inflation. *Organizational Behavior and Human Decision Processes*, 144, 171-186.

Schooler, L. J., & Anderson, J. R. (1990). The disruptive potential of immediate feedback. In 12th Annual Conf. CSS Pod (pp. 702-708). Psychology Press.

Schroeder, J., Kardas, M., & Epley, N. (2017). The humanizing voice: Speech reveals, and text conceals, a more thoughtful mind in the midst of disagreement. Psychological science, 28(12), 1745-1762.

Sheeran, P., & Webb, T. L. (2016). The intention–behavior gap. *Social and personality psychology compass,* 10(9), 503-518.

Srivastava, S. B., Goldberg, A., Manian, V. G., & Potts, C. (2018). Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations. *Management Science,* 64(3), 1348-1364.

Sun, K. Q., & Slepian, M. L. (2020). The conversations we seek to avoid. *Organizational Behavior and Human Decision Processes,* 160, 87-105.

Taylor, M. S., Fisher, C. D., & Ilgen, D. R. (1984). Individual's reactions to performance

feedback in organizations: A control theory perspective. *Research in personnel and human resources management* (pp. 81-124). JAI Press.

Tesser, A., Rosen, S., & Tesser, M. (1971). On the reluctance to communicate undesirable messages (the MUM effect): A field study. *Psychological Reports,* 29(2), 651-654.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software, 45*, 1-67.

Webb, T. L., & Sheeran, P. (2006). Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological bulletin,* 132(2), 249.

Wiesenfeld, B. M., Reyt, J. N., Brockner, J., & Trope, Y. (2017). Construal level theory in organizational research. *Annual Review of Organizational Psychology and Organizational Behavior*, 4, 367-400.

Wigert, B., & Harter, J. (2017). Re-engineering performance management. Gallup. com. Viewed: June 21, 2022.

Waung, M., & Highhouse, S. (1997). Fear of conflict and empathic buffering: Two explanations for the inflation of performance feedback. *Organizational Behavior and Human Decision Processes,* 71(1), 37-54.

Yeager, D. S., Purdie-Vaughns, V., Garcia, J., Apfel, N., Brzustoski, P., Master, A., ... & Cohen, G. L. (2014). Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General,* 143(2), 804.

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

Yeomans, M., Schweitzer, M. E., & Brooks, A. W. (2022). The Conversational Circumplex: Identifying, prioritizing, and pursuing informational and relational motives in conversation. *Current Opinion in Psychology,* 44, 293-302.

Youmans, R. J., & Arciszewski, T. (2014). Design fixation: Classifications and modern methods of prevention. AI EDAM, 28(2), 129-137.

Zenger, J. (2016, May). How effective are yout 360-Degree Feedback Assessments. *Forbes*. Retrieved from: https://www.forbes.com/sites/jackzenger/2016/03/10/how-effective-are-your-360-degree-feedback-assessments/?sh=5418ae7da690

Zenger, J., & Folkman, J. (2014, January). Your employees want the negative feedback you hate to give. Harvard Business Review. Retrieved from: https://hbr.org/2014/01/youremployeeswant-the-negative-feedback-you-hate-to-give

APPENDIX A: Structured Measures from Studies 1-2

These are the structured measures asked during every review survey - this included responses on a five-item likert scale, and open text questions at the bottom of each page of likert scale questions. All likert scale questions were asked with the following prompt and labeled scale points: The unstructured text questions from Table 1 were asked after page 4, and before page D. The questions were identical in the two studies with two exceptions - the open-text questions on this list were only asked in Study 1, and in Study 2 only one version of question 3_2 was asked.

Please rate how effect the subject is at… [1 = Not effective, 2 = somewhat effective, 3 = effective, 4 = very effective, 5= extremely effective, NA= can't say]

1_1 Demonstrating their pride in and passion for the [ORGANISATION]

1_2 Communicating the purpose and direction of the department with clarity and enthusiasm

1_3 Valuing professional excellence and expertise in others

1_4 Encouraging innovation and rewarding initiative

1_5 Learning from what hasn't worked as well as what has

1_6 Please use this opportunity to provide further detail in relation to any of the ratings you have provided above **[OPEN TEXT]**

2_1 Being straightforward, truthful and candid, including to those in power

2_2 Surfacing tensions and resolving ambiguities

2_3 Giving clear and honest feedback to help staff succeed

2_4 Addressing performance concerns resolutely, fairly and promptly

2_5 Being a team player and working collaboratively

2_6 Encouraging others to take managed risks and learn from their mistakes

2_7 Building credibility and influence to strengthen relationships with ministers and external stakeholders

2_8 Developing and maintaining positive, productive relationships with junior staff

2_9 Please use this opportunity to provide further detail in relation to any of the ratings you have provided above **[OPEN TEXT]**

3_1 Giving others the space and authority to deliver

3_2 Being visible and approachable at all times[2]

3_3 Demonstrating receptiveness to being challenged, however uncomfortable

3_4 Championing difference, recognising the value it brings

[2] The treatment text in Study 1 had this as the prompt instead "Being accessible and approachable"

3_5 Demonstrating commercial awareness in decision making

3_6 Embedding flexible and responsive ways of working including digital where possible

3_7 Investing time in and showing commitment to their own development

3_8 Nurturing talent and investing in the development of others to be effective now and in the future

3_9 Please use this opportunity to provide further detail in relation to any of the ratings you have provided above **[OPEN TEXT]**

4_1 Taking responsibility for delivery of the [ORGANISATION]'s programme and [ORGANISATION'S LEADERS'] priorities

4_2 Developing strategies that focus on delivering results that have a long term impact

4_3 Showing resilience in making tough strategic decisions

4_4 Prioritising a consistently high quality customer outcome

4_5 Maintaining real focus on delivering efficiency and value for money

4_6 Please use this opportunity to provide further detail in relation to any of the ratings you have provided above **[OPEN TEXT]**

D_1 How long have you been with your organisation?

D_2 How old are you?

D_3 Are you:

D_4 Ethnicity:

D_5 Sexual Orientation:

D_6 What is your primary [ORGANISATION] profession?

D_7 How many departments, including your current one, have you worked for?

D_8 How many years of your working life have you been employed in Other Government sectors (local government, other public sector bodies)?

D-9 How many years of your working life have you been employed in Other non-Government sectors (e.g. not for profit, academia, voluntary organisations)?

D_10 How many years of your working life have you been employed in private sector (including as an employer or self-employed)?

APPENDIX B: Full regression tables for all results in Studies 1 & 2

Study 1

2.2.2. Word count of text responses in Study 1

	Strengths		Development		Overall	
Treatment Condition	-2.904*** (.828)	-2.598** (.813)	.821 (.735)	.962 (.719)	-1.756*** (.496)	-1.408** (.470)
Recipient Gender (0 = female)	—		—		—	
Male		-1.627* (.648)		-1.036 (.633)		-.775 (.435)
Other		.522 (1.618)		.954 (1.334)		.679 (1.038)
Subjective Performance	—	5.513*** (.331)	—	-2.441*** (.336)	—	1.755*** (.216)
Relationship (0 = manager)	—	—	—	—	—	—
Peer	—	-7.460*** (1.354)	—	-11.40*** (1.142)	—	-.536 (.695)
Report	—	10.75*** (1.488)	—	5.179*** (1.209)	—	12.67*** (.725)
Other	—	2.089 (1.500)	—	-2.993* (1.243)	—	6.497*** (.759)
Clustered SE	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**
Observations	23,103	23,103	23,053	23,053	22,967	22,967

*p < 0.05, **p <0.01, ***p<0.001

2.2.3. Concreteness of text responses in Study 1

	Strengths		Development		Overall	
Treatment Condition	-.035*** (.005)	-.035*** (.005)	-.010 (.006)	-.010 (.006)	-.040*** (.004)	.040*** (.004)
Recipient Gender (0 = female)	—		—		—	
Male		-.021*** (.005)		.010 (.006)		-.001 (.004)
Other		-.027** (.010)		.017 (.012)		.001 (.009)
Subjective Performance	—	-.004 (.002)	—	-.040*** (.003)	—	-.009*** (.002)
Relationship (0 = manager)	—	—	—	—	—	—
Peer	—	-.024** (.008)	—	-.090*** (.010)	—	-.007 (.006)
Report	—	.011 (.008)	—	-.035*** (.010)	—	-.010 (.006)
Other	—	-.003 (.009)	—	-.069*** (.010)	—	-.013 (.007)
Clustered SE	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**
Observations	23,103	23,103	23,053	23,053	22,967	22,967

*p < 0.05, **p <0.01, ***p<0.001

Study 2A

3.2.2. Word count of text responses in Study 2

	Strengths		Development		Overall	
Expected Candor	-4.960*** (1.059)	-4.913*** (1.036)	-.784 (1.021)	-.598 (1.014)	-2.313*** (.681)	-2.386*** (.679)
Future Focus	2.238* (1.043)	2.521* (1.022)	4.745*** (1.005)	4.739*** (1.000)	3.595*** (.688)	3.653*** (.686)
Recipient Gender (0 = female)	—		—		—	
Male		-3.662*** (.866)		-1.798* (.896)		-1.135* (.571)
Other		-3.743* (1.545)		-2.454 (1.674)		-.834 (1.089)
Subjective Performance	—	8.074*** (.408)	—	-6.837*** (.545)	—	-1.521*** (.314)
Relationship (0 = manager)	—	—	—	—	—	—
Peer	—	-12.96*** (1.579)	—	-15.31*** (1.449)	—	-10.37*** (1.092)
Report	—	10.92*** (1.652)	—	6.421*** (1.544)	—	2.141 (1.105)
Other	—	-2.088 (1.661)	—	-6.817*** (1.507)	—	-3.575** (1.101)
Clustered SE	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**
Observations	30,164	30,164	25,774	25,774	20,647	20,647

*p < 0.05, **p <0.01, ***p<0.001

3.2.3. Concreteness of text responses in Study 2

	Strengths		Development		Overall	
Expected Candor	-.018** (.006)	-.018** (.006)	.018** (.006)	.018** (.006)	.059*** (.005)	.058*** (.005)
Future Focus	-.018** (.006)	-.018** (.006)	-.001 (.006)	-.001 (.006)	.001 (.006)	-.000 (.005)
Recipient Gender (0 = female)	—		—		—	
Male		-.007 (.005)		.013* (.006)		.014** (.005)
Other		.012 (.009)		.003 (.012)		-.005 (.009)
Subjective Performance	—	-.004 (.002)	—	-.036*** (.003)	—	-.027*** (.003)
Relationship (0 = manager)	—	—	—	—	—	—
Peer	—	-.023* (.009)	—	-.102*** (.010)	—	-.011 (.009)
Report	—	.040*** (.009)	—	-.062*** (.010)	—	-.022* (.009)
Other	—	.000 (.009)	—	-.084*** (.010)	—	-.014 (.009)
Clustered SE	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**
Observations	30,164	30,164	25,774	25,774	20,647	20,647

*p < 0.05, **p <0.01, ***p<0.001

APPENDIX C: Analyses of Non-Response Rates

Study 1 Response Rates: The survey was not mandatory, and many respondents (23.3%) did not even begin to complete their assigned surveys. And this non-response was slightly different in treatment (23.9%) and control (22.8%; $t(45691)=1.87$, $p = .061$). Because each respondent was assigned to a single condition for all of their invitations, we think this may not be a failure of randomization. Instead, it could be a knock-on effect of previous surveys - and would be more likely to show up in later invitations. Unfortunately, at present we do not know the order in which invitations were sent.

Once a respondent entered the survey, the responses themselves were also optional. Among respondents who entered the survey, some respondents did not write a single word in the open responses at the end of the survey (Strengths: 12.7%; Development 22.1%; Overall: 17.4%). These rates were higher in the treatment condition (19.5%) than in the control condition (15.4%; $t(105103)= 7.56$, $p < .001$). We suspect the people who skipped these questions were also the ones who declined to open later invitations. This difference was similar across all three questions, though perhaps respondents in the treatment group were especially less likely to respond to the Development question than the Strengths question $t(105099)= 2.08$, $p = .038$).

We found different rates of non-response based on relationship. Line Managers were the least likely to ignore the survey invitation entirely (10.4%), followed by Direct Reports (22.3%), Peers (25.7%) and Others (28.0%), and there were no interactions with treatment. When all kinds of non-responses are pooled, Line Managers were the least likely to not write a word (Strengths: 16.8%; Development: 18.8%; Overall: 18.7%), followed by Direct Reports (Strengths: 32.4%; Development: 39.2%; Overall: 36.3%), then Peers (Strengths: 35.8%; Development: 44.7%; Overall: 39.6%) and finally Others (Strengths: 38.3%; Development: 46.4%; Overall: 41.6%).

When we condition on the responses that include at least one word, we find some interesting differences. Respondents wrote longer Strengths texts (m = 52.7 words; SD = 46.2) than Development texts (m = 48.2 words; SD = 43.2) or Overall texts (34.2 words; SD = 30.7).

We had planned to remove low-text responses, to focus on longer texts that indicated effort on the part of the respondent, and which would provide richer insight into the respondents' thoughts. We had hoped these short responses would be evenly distributed across treatment and control, and conducted balance checks for each of the three texts separately, at three different thresholds: <= 2 words, <= 5 words, or <= 10 words. All nine of these checks indicated significant imbalance across conditions (lowest $X^2(1) = 17.9$, $p < 2.2x10^{-5}$).

These checks indicate the treatment condition had a higher number of very short responses. However, this was driven entirely by non-responses (Strengths: 35.0 % vs 31.2%; Development 42.5% vs 38.1%; Overall: 38.7% vs. 34.7%). Among those who wrote at least one word, the treatment had no effect on word count for the Strengths texts ($\beta = -.236$, SE = .788, $t(30576) = 0.3$, $p = .765$) and the Overall texts ($\beta = -.331$, SE = .503, $t(28938) = 0.7$, $p = .510$), and increased the average word count among Development texts ($\beta = 3.22$, SE = .711, $t(27275) = 4.5$, $p < .001$).

These results were unexpected. Our preliminary interpretation of the effect on empty responses is that the treatment condition raised the expected effort of the writing task, so that respondents who were induced to skip the questions were low-effort writers who would have not written much anyways. However, our primary hypotheses concerned the content of the reviews, rather than the effort level per se. The differing pattern of results across questions also complicates our analytical strategy.

As a simplifying assumption, we decided to remove all responses that were five words or less from each condition. We have also confirmed that the results are similar if we use other thresholds (e.g. 2 words, 10 words). We find similar results when we instead remove a similar quantile of responses from each condition (e.g. bottom 50%, bottom 60%). This gives us additional confidence the results are genuinely due to treatment effects, rather than composition effects.

Study 2 Response Rates: The survey was not mandatory, and many respondents (20.4%) did not even begin to complete their assigned surveys. In Study 2, this survey-level non-response was not significantly different by condition ($X^2(3) = 2.50$, $p = .476$).

Once a respondent entered the survey, the responses themselves were also optional. Among respondents who entered the survey, some respondents did not write a single word in the open responses at the end of the survey (Strengths: 6.8%; Development 15.9%; Overall: 31.0%). Non-response rates did not vary much based on the treatment condition in the Strengths text (Expected Candor: $\beta = .006$, $SE = .004$, $t(38614) = 1.5$, $p = .137$; Future Focus: $\beta = -.007$, $SE = .004$, $t(38614) = 1.6$, $p = .101$), however both treatments increased response rates for Development (Expected Candor: $\beta = .019$, $SE = .006$, $t(38614) = 3.4$, $p = <.001$; Future Focus: $\beta = .017$, $SE = .006$, $t(38614) = 2.9$, $p = .003$), while for the Overall question Expected Candor seemed to decrease response rates (Expected Candor: $\beta = -.025$, $SE = .008$, $t(38614) = -3.3$, $p = .001$) while Future Focus somewhat increased them ($\beta = .015$, $SE = .008$, $t(38614) = 1.9$, $p = .053$).

We also found different rates of non-response based on relationship. Line Managers once again were the least likely to ignore the survey invitation entirely (9.1%), followed by Peers (20.9%), Direct Reports (21.0%) and Others (24.5%).

When we condition on the responses that include at least one word, we find some interesting differences. Respondents wrote longer Strengths texts (m = 72.0 words, SD = 66.2) than Development texts (m = 61.8 words, SD = 59.7) or Overall texts (43.3 words, SD = 41.5).

We had planned to remove low-text responses, to focus on longer texts that indicated effort on the part of the respondent, and which would provide richer insight into the respondents' thoughts. We had hoped these short responses would be evenly distributed across treatment and control, and conducted balance checks for each of the three texts separately, at three different thresholds: <= 2 words, <= 5 words, or <= 10 words. While there are no significant differences across conditions for Strengths text (lowest $X^2(3) = 5.01$, $p = 0.17$), there is significant imbalance in the six checks for Development and Overall texts (lowest $X^2(3) = 32.579$, $p < 3.95x10^{-7}$). These checks indicated differences in Development and Overall texts in the number of very short responses. For Development, the control had the highest non-response rate (34.3% compared with Expected Candor: 32.6% and Future Focus: 32.9%). For Overall, however, control (44.4%) fell in between Expected Candor (46.4%) and Future Focus (43.3%). There was no effect for Strengths (control: 25.9%; Future Focus: 25.8%; Expected Candor: 25.7%).

Among those who wrote at least one word for Strengths, the Expected Candor treatment decreased word count ($\beta = -5.16$, $SE = 1.04$, $t(36020) = 4.9$, $p < .001$), while Future Focus had a significant impact in the opposite direction ($\beta = 2.81$, $SE = 1.03$, $t(36020) = 2.7$, $p = .006$). For Development, Future Focus increased this word count ($\beta = 5.41$, $SE = 0.94$, $t(32499) = 5.7$, $p < .001$) while Expected Candor had no effect ($\beta = -0.02$, $SE = 0.95$, $t(36020) = 4.9$, $p < .001$) and a similar pattern was found for the overall question (Expected Candor: $\beta = -3.83$, $SE = .68$, $t(26690) = -5.6$, $p < 0.0001$; Future Focus: $\beta = 4.73$, $SE = .69$, $t(26690) = 6.9$, $p < 0.001$).

Once again, as a simplifying assumption, we decided to remove all responses that were five words or less from each condition. We have also confirmed that the results are similar if we use other thresholds (e.g. 2 words, 10 words). We find similar results when we instead remove a similar quantile of responses from each condition (e.g. bottom 50%, bottom 60%). This gives us additional confidence the results are genuinely due to treatment effects, rather than composition effects.

APPENDIX D: Analyses of Intervention Interactions in Study 2

C1. Exclusions

	Strengths		Development		Overall	
Expected Candor	0.00 (0.01)	0.00 (0.01)	0.03*** (0.01)	0.03*** (0.01)	-0.05*** (0.01)	-0.06*** (0.02)
Future Focus	0.01 (0.01)	0.00 (0.01)	0.04*** (0.01)	0.04*** (0.01)	0.03*** (0.01)	0.03*** (0.01)
Interaction	0.00 (0.01)	0.00 (0.01)	-0.03 (0.01)	-0.02 (0.01)	0.01 (0.01)	0.01 (0.01)
Controls	**NO**	**YES**	**NO**	**YES**	**NO**	**YES**
Clustered SE	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**
Sample	30,164	30,164	25,774	25,774	20,650	20,650

*p < 0.05, **p <0.01, ***p<0.001

C2. Word length

	Strengths		Development		Overall	
Expected Candor	-5.88*** (1.49)	-5.78*** (1.46)	-0.59 (1.36)	-0.34 (1.35)	-3.91*** (0.97)	-3.91*** (0.97)
Future Focus	1.31 (1.51)	1.66 (1.48)	4.94*** (1.45)	5.01*** (1.43)	2.12* (0.99)	2.24* (0.99)
Interaction	1.83 (2.13)	1.73 (2.09)	-0.39 (2.02)	-0.52 (2.01)	3.07* (1.36)	2.95* (1.36)
Controls	**NO**	**YES**	**NO**	**YES**	**NO**	**YES**
Clustered SE	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**
Sample	30,164	30,164	25,774	25,774	20,650	20,650

*p < 0.05, **p <0.01, ***p<0.001

C3. Concreteness

	Strengths		Development		Overall	
Expected Candor	-0.02** (0.01)	-0.02** (0.01)	0.02* (0.01)	0.02* (0.01)	0.06*** (0.01)	0.06*** (0.01)
Future Focus	-0.02** (0.01)	-0.02** (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)
Interaction	0.01 (0.01)	0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)	0.00 (0.01)	0.00 (0.01)
Controls	**NO**	**YES**	**NO**	**YES**	**NO**	**YES**
Clustered SE	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**
Sample	30,164	30,164	25,774	25,774	20,650	20,650

*p < 0.05, **p <0.01, ***p<0.001

C4. Study 2B Feedback Quality Index

	Strengths	
Expected Candor	.01 (0.07)	-0.01 (0.09)
Future Focus	-0.08 (0.09)	-0.10 (0.09)
Interaction	-0.06 (0.13)	-0.03 (0.12)
Controls	**NO**	**YES**
Clustered SE	**YES**	**YES**
Sample	859	859

*p < 0.05, **p <0.01, ***p<0.001