MANUSCRIPT ACCEPTED FOR PUBLICATION AT *JOURNAL OF EXPERIMENTAL PSYCHOLOGY: LEARNING, MEMORY, & COGNITION*, 22/12/17

A memory advantage for past-oriented over future-oriented performance feedback

Robert A. Nash[1], Naomi E. Winstone[2], Samantha E. A. Gregory[1], and Emily Papps[2]

[1] School of Life and Health Sciences, Aston University

[2] Department of Higher Education, University of Surrey

Acknowledgements

This research was funded in part by the Leverhulme Trust (Research Project Grant RPG-2016-189) and the Higher Education Academy (Grant GEN1024). The authors are grateful to Michael Parker for assistance with data collection, to King Edward VI College, Stourbridge for their support with running Experiment 4, and to Tillingbourne Junior School for their support with running Experiment 6. The authors also thank Sean Kang, Yana Weinstein, Bertram Opitz, and Paul Sowden for their insightful suggestions.

Corresponding author

Robert A. Nash
School of Life and Health Sciences
Aston University
Birmingham, B4 7ET
United Kingdom
Tel: +44 121 204 4522
Email: R.Nash1@aston.ac.uk

People frequently receive performance feedback that describes how well they achieved in the past, and how they could improve in future. In educational contexts, future-oriented (directive) feedback is often argued to be more valuable to learners than past-oriented (evaluative) feedback; critically, prior research led us to predict that it should also be better remembered. We tested this prediction in six experiments. Subjects read written feedback containing evaluative and directive comments, which supposedly related to essays they had previously written (Experiments 1-2), or to essays another person had written (Experiments 3-6). Subjects then tried to reproduce the feedback from memory after a short delay. In all six experiments, the data strongly revealed the opposite effect to the one we predicted: despite only small differences in wording, evaluative feedback was in fact recalled consistently better than directive feedback. Furthermore, even when adult subjects did recall directive feedback, they frequently misremembered it in an evaluative style. These findings appear at odds with the position that being oriented toward the future is advantageous to memory. They also raise important questions about the possible behavioral effects and generalizability of such biases, in terms of students' academic performance.

Keywords: feedback; education; future orientation; recall; assessment

A memory advantage for past-oriented over future-oriented performance feedback

In almost any profession or pastime—from education, to business, to sports and the performing arts—being able to improve our skills can hinge on receiving good quality feedback from others (Hattie & Timperley, 2007). Education researchers have accumulated substantial data concerning which kinds of feedback best enhance learning, which kinds people value, and how and when feedback is most effectively delivered (e.g., Gielen, Peeters, Dochy, Onghena, & Struyven, 2010; Winstone, Nash, Rowntree, & Menezes, 2016; Wollenschläger, Hattie, Machts, Möller, & Harms, 2016). But if any variety of feedback is to be truly effective, then the person who receives the feedback must be able draw upon it at a later time, when a need arises to develop an action plan or to directly implement the advice. In many cases this means it is highly advantageous to remember feedback; not least because many university students say they rarely read their written feedback more than once (Winstone, Nash, Rowntree, & Parker, 2017). For instance, imagine a student who receives critical feedback from her professor about her assignment. Ideally, the feedback should enable the student to improve her next assignment; however, if she never encodes the feedback in memory or is unable to recall it, then she may fail to reap those benefits.

Memory processes are therefore strongly implicated in determining whether feedback is effective. But what kinds of feedback stick in memory? In this paper we ask whether simple variations in the wording of written feedback—designed to orient people either toward past performance or toward future improvement—could influence the likelihood that people will remember it.

A cognitive perspective on receiving feedback

To begin asking how well people remember feedback, we can look to the research literature on memory as applied specifically to education. In that literature, cognitive psychologists have made sizeable contributions to our understanding of how students learn in general. For example, cumulative studies have neatly specified the mechanisms that underpin effective study practices, have examined how teaching environments and methods can be optimized to enhance learning, and have convincingly challenged educational myths (e.g., Agarwal, 2012; Chandler & Sweller, 1991; Pashler, McDaniel, Rohrer, & Bjork, 2008; Weinstein, McDermott, & Roediger, 2010). Eminent psychologists such as Roediger (2013; Roediger & Pyc, 2012) have called for greater translation of these research findings into teaching practices, arguing that better public awareness of educational science should lead to more effective, evidence-based teaching and learning practices.

Given the boom in memory research applied to education, and given the centrality of feedback to skill development, one might expect that cognitive psychologists would have amassed a wealth of research data on how effectively and under which circumstances people remember the feedback they receive on their performance. But in fact, very few such data currently exist. This is not to say that cognitive psychologists have ignored the topic of feedback; they certainly have not. However, most empirical studies on this topic to date have explored only very particular kinds of feedback. Specifically, many studies ask how receiving "correct/incorrect" feedback during or after a multiple-choice test can benefit students' performance in a subsequent test (e.g., Butler, Karpicke, & Roediger, 2008; Butler & Roediger, 2008; Hays, Kornell, & Bjork, 2010; Kang, McDermott, & Roediger, 2007; Smith & Kimball, 2010). One interesting finding from many of these

studies is that even when people are explicitly told "You answered this question incorrectly; the correct answer is [XXXX]," often those people still fail to answer the same question correctly when asked again, even just a few minutes later. Quite understandably, people do not reliably remember all of the feedback they receive.

In educational assessments, people typically receive feedback that goes far beyond being told whether they were correct or incorrect. Rather, people much more commonly receive detailed, descriptive feedback, with nuanced information about what was done well or less well. At present, the cognitive psychology literature tells us surprisingly little about how well and how accurately people remember these descriptive kinds of feedback. In fact, to our knowledge virtually no peer-reviewed research yet exists that examines people's memory for descriptive feedback about their performance (see Cutumisu & Schwartz, in press, for one exception with a sample of middle school students). Empirical research on this issue would clearly be valuable, especially when we consider that the educational literature on learners' engagement with feedback is notably lacking in experimental data (Winstone, Nash, Parker, & Rowntree, 2017). Might some kinds of feedback be better or more accurately remembered than others? Distinguishing two particular kinds of feedback offers a useful foundation for considering this question.

Evaluative vs. directive feedback

An issue often discussed in the education literature is whether learners gain most from receiving past-oriented *evaluative feedback*—focused on what the learner did well or badly—or from receiving future-oriented *directive feedback*, focused on how she or he could improve (sometimes called "feedforward"). Education experts often advise feedback-givers to focus on the future: the main purpose of feedback, after all, is to foster future improvement (Kluger & DeNisi, 1996). In line with this

reasoning, thinking about one's own future can encourage people to take decisions and actions that are of distal rather than only proximal benefit (see Prabhakar, Coughlin, & Ghetti, 2016). Not only do many education experts perceive greater value in directive feedback, but students, too, typically prefer receiving feedback about improvement, rather than feedback about what they did well or badly in the past (Winstone et al., 2016). But cognitive psychological research gives us reason to predict that these preferences for directive feedback would be supplemented by cognitive benefits. Specifically, the future-orientation of directive feedback could mean that learners are more likely to subsequently remember it, as compared with evaluative feedback.

Remembering for the future

It is now commonly accepted among researchers that memory evolved not only as a faculty for documenting the past, but also for enabling people to anticipate and plan for the future (Atance & O'Neill, 2001; Klein, 2013; Pillemer, 2003). Because remembering serves this evolutionarily adaptive, directive function, an implication is that memory systems and processes should in principle be especially well attuned to remembering information that concerns the future, as compared to information that concerns the past (Bluck, 2003; Klein, 2013; Pillemer, Picariello, Law, & Reichman, 1999).

Indeed, there is some evidence to support this view. In one study, Klein, Robertson, and Delton (2010) asked subjects to learn a list of object words. Whilst encoding the words, some subjects were asked to form a mental picture of a campsite, and to rate the likelihood that each object would appear at the campsite. Other subjects learned the same words, but were asked to remember a specific time in the past when they went camping, and to rate the likelihood that each object was at the

campsite during their trip. A third group, while learning the words, were asked to imagine planning a camping trip, and to rate the likelihood that they would take along each of the objects. Klein et al. found that the latter group, who planned for a future camping trip, subsequently recalled more of the words than did either of the other groups (see also Klein, Robertson, & Delton, 2011). These data point to memory benefits of being oriented toward the future, and on this basis we could predict that people would recall directive feedback better than evaluative feedback. Additional findings from the prospective memory literature might lead us to the same prediction. Research in that literature shows us that people develop more-accessible memory representations of instructions if they believe they will need to implement those instructions at a later time (Goschke & Kuhl, 1993; Koriat, Ben-Zur, & Nussbaum, 1990). Insofar that directive feedback—unlike evaluative feedback—explicitly guides people on what to do in future, these findings might lead us again to predict that directive feedback would be better remembered.

Overview of the present research

The main aim of the present research was to directly test these predictions of a memory advantage for directive feedback over evaluative feedback. To this end, subjects in Experiments 1 and 2 completed a short writing assignment, and afterwards they received detailed—and ostensibly personalized—written feedback. This feedback was in fact generic, and included some comments written in an evaluative style and others in a directive style. In Experiments 3-6 we excluded the initial assignment, and subjects simply read the feedback as though it were written for another person. In all experiments, shortly after subjects read the feedback we gave them a surprise recall test, and we assessed which feedback comments they were able to reproduce. To preview our findings, in all six experiments we discovered the exact

opposite effect to the one we predicted: subjects recalled evaluative feedback substantially better than directive feedback. Our data provide initial tests of some possible theoretical explanations of this finding.

Experiment 1

The procedures for all the experiments reported in this paper were reviewed and approved by an institutional research ethics committee.

Method

Subjects. A total of 61 psychology undergraduates (57 females and 4 males, M_{age} = 20.24, *SD* = 5.00, Range = 18-45) took part in exchange either for £10 or for course credit.

Materials.

Feedback scripts. We developed two versions of a script of standardized feedback to give to subjects. These scripts totaled 418 words (version A) and 411 words (version B) respectively, and can be found in the online supplemental materials. Both feedback scripts were divided into three subsections labeled "substance", "style", and "format," and each subsection contained several pieces of critique that were prefixed and suffixed by brief praise. The praise was not relevant to our experimental design, but was included merely to make the feedback as a whole seem less severe and more realistic.

Each feedback script contained 20 critique comments in total, and all subjects saw the same comments in the same order. The only difference between the two feedback scripts was the style in which each critique comment was written. Specifically, in both scripts half of the critique comments were written in an *evaluative style*; that is, they were presented as comments about the essays that the subject had produced, and thus they focused on past performance. The other half of

the critique comments were written in a *directive style*; that is, they were presented as comments about what the subject could improve next time, and thus they focused on future performance. We achieved this style manipulation using minimal re-wording of each critique comment, to cast the same general meaning in both an evaluative and a directive manner whilst keeping the comments' length and complexity approximately equal. For example, half of subjects were told "You didn't always demonstrate a sophisticated awareness of the issues you covered" (an evaluative comment), whereas the other half were told "You should aim to demonstrate a more sophisticated awareness of the issues you cover" (a directive comment). In both feedback scripts, critique comments were presented in pairs that alternated between the evaluative and directive style. We counterbalanced between scripts whether each individual critique comment appeared in the evaluative or directive style.

Achievement Goal Questionnaire—Revised (AGQ-R). All participants completed the AGQ-R, a widely-used and validated measure of trait achievement goals (Elliot & Murayama, 2008). The AGQ-R comprises 12 items that subjects rate on scales from 1 (strongly disagree) to 5 (strongly agree). The measure distinguishes mastery goals (developing competence relative to an absolute or intrapersonal standard) from performance goals (developing competence relative to a normative standard), and distinguishes approach goals (focusing on success) from avoidance goals (focusing on preventing failure). Four subscales of the AGQ-R, each calculated from 3 of the 12 scale-items, index each of the achievement goal-types in this 2 (mastery vs. performance) x 2 (approach vs. avoidance) framework. For example, one item from the mastery-approach subscale is "My goal is to learn as much as possible."

Procedure. Subjects signed up for a study purportedly investigating "personality and persuasive writing." Each subject individually attended two sessions

in the laboratory, separated by 1-2 days according to their availability. All instructions and information were presented to subjects on a computer screen.

Session 1. In the first session, subjects learned that they would be completing a persuasive writing task. To begin, we gave subjects a list of ten "contentious topics," and from these we asked them to choose four on which to write short essays. For example, two of the topics were "Should students have to pay for their university education?" and "Should Valentine's Day be abolished?". After selecting their topics, one of the chosen essay titles appeared at random on the computer screen, and we asked subjects to type a short persuasive essay on that topic, with a time limit of 5 min. A countdown timer at the top of the page indicated how much time was remaining. After 5 min, the page automatically changed, and the second essay title appeared. This process was repeated for all four essay titles, with a total duration of 20 min.

After finishing their fourth essay, subjects completed the AGQ-R. We then verbally informed subjects that a member of the teaching team would examine their persuasive essays prior to the second session, and would produce some detailed feedback on their performance. We told subjects that they would be given this feedback in Session 2, and we falsely informed subjects that after reading this feedback they would complete more persuasive writing.

Session 2. When subjects returned for the second session 1-2 days later, we presented to them, at random, one of our two feedback scripts on the computer screen. Despite all subjects receiving the same feedback comments, we told them that the feedback had been prepared specifically for themselves, based on their own persuasive writing in session 1. Subjects were allowed as much time as they needed to read their feedback script, and they clicked a button once they wished to proceed.

Next they completed a 5-min filler task, which involved solving reasoning puzzles similar to Raven's progressive matrices. When the time was up, subjects were automatically moved on.

On the next page, we gave subjects a surprise recall test. We asked them to think back to the feedback they received, and we gave them up to 10 min to type as much of the feedback as they could recall. We told them that although they probably would not remember the feedback verbatim, they should nevertheless try to be as accurate as possible with regard to the meaning of what had been written. Subjects were unable to move on to the next part of the experiment until they had spent at least 5 min on this recall task, but they were automatically moved on after 10 min.

Following this task, subjects used rating scales to judge the fairness (1 = Very unfair; 5 = Very fair) and helpfulness (1 = Very unhelpful; 5 = Very helpful) of the feedback. They also judged what percentage grade they would give themselves for their writing in session 1, and what grade they believed they could achieve next time in light of the feedback they received. We then asked them to write down any comments they had about the feedback.

To gain additional memory data, we next gave subjects a two-alternative forced choice (2AFC) recognition test, which included 10 questions. For each question, subjects saw one of the pairs of feedback comments written in an evaluative style, alongside the equivalent pair written in a directive style. We counterbalanced the presentation order. Subjects attempted to identify which of the pairs they had actually seen in their own feedback. After completing the recognition test, we finally asked subjects to write down what they believed the aim of the experiment was, and we debriefed and compensated them.

Data coding. A research assistant examined each subject's free recall response, blind both to our experimental hypotheses, and to which of the two feedback scripts each subject saw. Based on the gist of the responses, she then coded (a) which of the critique comments the subject had recalled, and (b) in which style (i.e., evaluative vs. directive) she or he had reproduced each comment. The coder ignored any praise that subjects recalled, and if a subject recalled a particular comment in both an evaluative and a directive style (for example, in two separate parts of their written response), this was coded twice. The first author, also blind to which version of the feedback each subject had received, independently coded 20% of responses in the same manner.

Our coding also permitted us to explore a secondary question: did subjects tend to recall the feedback in the same style as they had actually seen it, or did they systematically misremember comments in the incorrect style? This is an interesting question because systematic biases in *how* people reproduce the feedback from memory could provide insight into their spontaneous thought when reading the feedback (e.g., Brewer, 1977; Chan & McDermott, 2006; Garry, Strange, Bernstein, & Kinzett, 2007; Klepacz, Nash, Egan, Hodgkins, & Raats, 2016). For instance, if people systematically misremember evaluative comments in a directive style, this could suggest that they are spontaneously inferring what they would need to do differently next time. To answer this secondary question, after the initial coding we unblinded the data to reveal which of the feedback scripts the subject actually saw. Doing this enabled us to assess (separately for each coder's judgments) whether each of the recalled comments had been reproduced in the same style as it was actually presented (i.e., evaluative comments recalled as evaluative; directive comments

recalled as directive), or in the alternate, incorrect style (i.e., evaluative comments recalled as directive; directive comments recalled as evaluative).

The two coders' agreement was strong in terms of the total number of evaluative comments recalled in an evaluative style (r = .89); the number of directive comments recalled in a directive style (r = .79); the number of evaluative comments recalled in a directive style (r = .86); and the number of directive comments recalled in an evaluative style (r = .92). The analyses below are therefore based on the first coder's data.

Results

Subjects' appraisals of the feedback. Before addressing our main research questions, we first asked whether subjects seemed to believe our cover story that the feedback was personalized. Our data suggest that they did. Overall, when we asked subjects to write down any comments they had about the feedback, none indicated suspicion that the feedback was generic rather than personalized. Instead, many of the comments indicated that subjects were convinced by the feedback, and were even prepared to take it on board – a kind of "Barnum effect" (Johnson, Cain, Falke, Hayman, & Perillo, 1985). For example, one subject wrote "This feedback was useful and would be helpful for future work. It also makes me appreciate the process of reflection on my own work to improve." Another wrote "It was generally fair and I agreed upon most of what was said." In fact, subjects rated the feedback highly in terms of both fairness (M = 4.03 out of 5, SD = 0.77), and helpfulness (M = 4.11, SD = 0.49). They also believed that the feedback could help them to perform better next time. Specifically, although they believed they had performed poorly on the writing task, estimating their grade at just 49.93% (SD = 8.98%), they estimated that with the help of the feedback, they could achieve an average grade of 63.41% (SD = 8.14%)

next time. Finally, no subjects correctly guessed the aim of the study when asked, or guessed that the past vs. future orientation of the feedback was critical. Together, these data suggest that subjects truly believed they were receiving personalized feedback on their own writing.

Free recall. Our main analysis had two principal aims. The first aim was to assess the extent to which the evaluative vs. directive style of feedback comments would influence subjects' tendency to freely recall those comments. The second aim was to assess whether subjects systematically distorted this feedback style in their recollections. To answer these questions, we conducted a 2 (Feedback style: evaluative vs. directive) x 2 (Retrieval style accuracy: correct vs. incorrect) repeated-measures ANOVA on the number of critique comments that subjects recalled. Note that the first of these independent variables relates to the style in which the feedback comments were actually presented to subjects. The second variable relates to whether subjects recalled the comments in the same style as they actually saw them, or in the alternate, incorrect style.

As shown in Figure 1, our analysis revealed that contrary to our predictions, subjects recalled significantly more of the evaluative feedback than of the directive feedback, as indexed by a substantial main effect of feedback style, $F(1, 60) = 16.82$, $p < .001$, $\eta^2_p = .22$, $d = 0.76$, 95% CI on d [0.37, 1.14]. We also found that subjects had generally paid good attention to the wording and style of the feedback they received, as evidenced by a main effect of retrieval style accuracy whereby subjects recalled more feedback comments in the correct style than in the incorrect style, $F(1, 60) = 28.98$, $p < .001$, $\eta^2_p = .33$, $d = 1.02$ [0.60, 1.43].

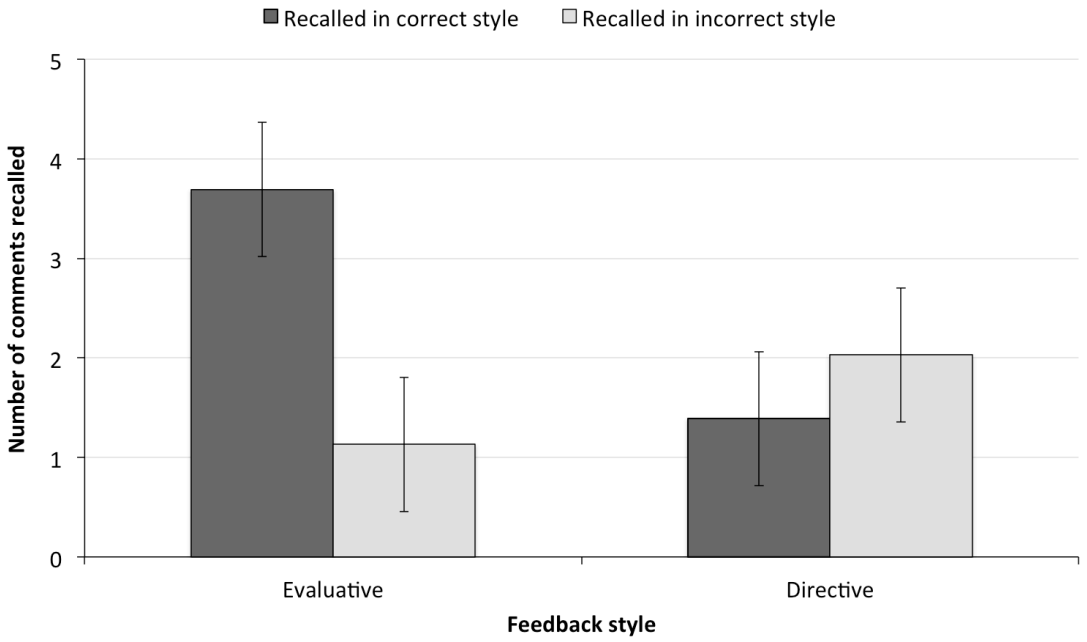

Figure 1. Recall of evaluative and directive feedback in Experiment 1, split according to retrieval style accuracy. Error bars are 95% within-subject confidence intervals (Loftus & Masson, 1994).

Finally, we found that regardless of the style in which the feedback comments were actually presented, subjects reproduced feedback in an evaluative style more frequently than they reproduced feedback in a directive style. This result is indexed by a significant interaction effect, again with a large effect size, $F(1, 60) = 22.51$, $p < .001$, $\eta^2_p = .27$, $d = 1.10$ [0.61, 1.59]. Follow up paired *t*-tests showed that people reproduced evaluative comments in the correct, evaluative style significantly more often than they reproduced evaluative comments in the incorrect, directive style, $t(60) = 6.32$, $p < .001$, $d = 1.40$ [0.90, 1.90]. Specifically, of all evaluative comments that were recalled, 77% were recalled as evaluative, and 23% as directive. In contrast, people reproduced directive comments in the correct, directive style no more often than they reproduced directive comments in the incorrect, evaluative style, $t(60) =$

1.80, p = .08, d = -0.38 [-0.81, 0.04]. Specifically, of all directive comments that were recalled, 41% were recalled as directive, and 59% as evaluative.

In short, our analysis points to what we might term an *evaluative recall bias*; that is, subjects were considerably more likely to recall feedback if it was presented in an evaluative rather than directive style, despite the substance of the feedback in both cases being virtually identical. Moreover, our data revealed that subjects tended to adopt an *evaluative retrieval style*; that is, they tended to reproduce the feedback comments in an evaluative style even when they had actually been directive.

Recognition performance. Subjects performed well overall in the 2AFC recognition test, identifying 70.98% (SD = 14.11%) of the comments correctly. In this test, subjects could make two possible types of recognition error: (1) recognizing evaluative feedback when in fact they had seen directive feedback ("evaluative errors"); or (2) recognizing directive feedback when in fact they had seen evaluative feedback ("directive errors"). A paired t-test revealed only a nonsignificant trend difference in how frequently each of these types of error was committed (Evaluative errors, M = 16.7%, SD = 13.0%; Directive errors, M = 12.3%, SD = 9.4%), $t(60)$ = 1.95, p = .06, d = 0.39, 95% CI on d [-0.01, 0.79].

Achievement goal orientation. We found no evidence that subjects' recall bias, retrieval style, or total number of comments recalled, were meaningfully related to their achievement goal orientations as measured with the AGQ-R (see Table S1 in the online supplemental materials for full details).

Experiment 2

The data from students in Experiment 1 suggest a strong memory advantage for evaluative feedback relative to directive feedback. This evaluative recall bias is rather surprising given that the prior evidence described above led us to predict the

exact opposite bias, and even the education literature suggests that students would be more interested in directive feedback than in evaluative feedback. In fact, it is also at odds with our students' own predictions. We defined evaluative and directive feedback to 36 psychology undergraduates who did not take part in these experiments, and asked them (a) which—all else being equal—they would prefer to receive (evaluative/directive), (b) which they would pay most attention to (evaluative/directive/both about the same), and (c) which they would be most likely to remember (evaluative/directive/both about the same). Overall, 72% said they would prefer to receive directive feedback (vs. 28% evaluative), 42% said they would pay more attention to directive feedback (vs. 17% evaluative), and 58% said they would be more likely to remember directive feedback (vs. 14% evaluative).

The fact that the unexpected recall bias was large by conventional standards makes it all the more intriguing. Nevertheless, it is important to replicate unexpected findings before interpreting them with any confidence. In Experiment 2 we aimed to do so. Assuming that our findings could be replicated, it would be valuable to gain a sense of the robustness of this evaluative recall bias. One straightforward interpretation is that the bias simply reflects the kinds of feedback that subjects consciously expect will be most important for them to encode within the specific task. If this were true, then it should be easy to amplify or attenuate the effect simply by instructing subjects about the kinds of information they should prioritize when reading the feedback. In Experiment 2, then, as well as aiming to replicate the evaluative recall bias, we also manipulated the instructions we gave to subjects. Whereas some were told simply to read the feedback, a second group were told to focus on finding out how they had performed, and a third group were told to focus on

finding out how to improve in future. The effectiveness of this simple manipulation should offer initial insights into the mechanisms underlying the evaluative recall bias.

Method

Subjects. In this experiment we added a further experimental manipulation—comprising three between-subjects conditions—to the within-subjects design used in Experiment 1. We conducted a power analysis to determine the sample size necessary to detect a medium-sized interaction effect (f = .25) in a 2 x 3 within-between subjects design, assuming power = .80, alpha = .05, and a correlation of zero between repeated-measures (approximated from the comparison of recalling evaluative vs. directive feedback in Experiment 1). This analysis suggested that 81 subjects would be required; we slightly oversampled and ultimately recruited 85 university students before coding or inspecting the data. Our final sample comprised 68 females and 17 males (M_{age} = 20.10, SD = 1.92, Range = 18-28) who took part in exchange for either £10 or course credit. Most (59%) were studying psychology; the remainder were from a variety of other study disciplines.

Materials and procedure. Subjects completed this experiment in two separate sessions, the first of which was the same as in Experiment 1 and involved producing short persuasive essays. The second session was also mostly identical, but at the start of this session we randomly assigned subjects to one of three conditions: Control, Past-orientation, or Future-orientation. Subjects in the control condition were simply told to read their feedback carefully, as per Experiment 1. However, subjects in the other two conditions received more specific instructions before reading their feedback. To those in the Past-orientation condition, we gave the following computerized instruction:

> When students receive feedback, it is very important for them to use the feedback to understand how they performed. Researchers have shown that when reading feedback, good students look for information that evaluates their work, and for information that explains why the marker judged the piece of work at a particular standard. Various evidence shows that students who engage with their feedback in this way – using it to understand how they performed – tend to get better value out of it. With this information in mind, please read the feedback on the next page carefully.

In contrast, we told subjects in the Future-orientation condition:

> When students receive feedback, it is very important for them to use the feedback to work out how to improve their performance. Researchers have shown that when reading feedback, good students look for information that directs them towards future improvement, and for information that explains where the marker thinks they should focus in order to improve the standard of their work. Various evidence shows that students who engage with their feedback in this way – using it to work out how to improve – tend to get better value out of it. With this information in mind, please read the feedback on the next page carefully.

With only two other exceptions, the procedure was identical to session 2 of Experiment 1. The first exception was that in this experiment we covertly recorded (via the experiment software) how long subjects spent on the feedback page before clicking to continue. The second was that at the end of session 2, we showed subjects

the three task instructions from the Control, Past-orientation, and Future-orientation conditions respectively, in a random order, and we asked subjects which instruction they had seen. This final question served to check whether subjects had read the instructions properly.

Data coding. We coded the data in the same way as in Experiment 1, with the second coder coding 20% of responses. The inter-rater agreement was high for the number of evaluative comments recalled in an evaluative style (r = .89); the number of directive comments recalled in a directive style (r = .79); the number of evaluative comments recalled in a directive style (r = .91); and the number of directive comments recalled in an evaluative style (r = .77). The analyses below are therefore based on the first coder's data.

Results

Subjects' appraisals of the feedback. Although five subjects noted in their written comments that the feedback seemed rather generic, many others wrote comments suggesting that they found the feedback useful, and none correctly guessed the aim of the study. As Table 1 shows, subjects again said that they found the feedback both fair and helpful, and believed it could help them to achieve a better grade next time. The fairness and helpfulness ratings did not differ significantly across instruction conditions (both $p > .14$, both $\eta^2_p < .05$). Likewise, our instruction manipulation did not significantly influence the grades subjects assigned themselves for their writing in session 1, $F(2, 82) = 1.25$, $p = .29$, $\eta^2_p = .03$. But it did affect their assessments of what grade they could achieve in the future, $F(2, 82) = 3.56$, $p = .03$, $\eta^2_p = .08$. Post-hoc Bonferroni tests showed that Future-orientation subjects projected higher future grades than did control subjects ($p = .03$); however, Past-orientation subjects' projections did not differ significantly from either those of control ($p = .28$)

or Future-orientation (p = .93) subjects. We should note that of the 85 subjects, 26 failed to correctly identify at the end of session 2 which instruction they had received at the start of that session. However, the pattern and statistical significance of all the findings described thus far were identical even when these 26 people were excluded from analyses.[1]

When we looked at the amount of time the full sample of subjects (i.e., N = 85) spent actually reading the feedback (first log-transforming these durations to successfully correct for positive skew), the means did not differ significantly between the groups, $F(2, 82) = 1.44$, $p = .24$, $\eta^2_p = .03$. However, when we excluded the 26 subjects who failed the attention check (i.e., restricting our analyses to those who remembered which instruction they received), a significant difference between conditions emerged, $F(2, 56) = 6.09$, $p < .01$, $\eta^2_p = .18$. Post-hoc Bonferroni tests showed that subjects in the control group spent significantly less time reading the feedback (M = 106 sec, SD = 36) than did those in both the Past-orientation (M = 132 sec, SD = 45, $p = .04$) and Future-orientation conditions (M = 135 sec, SD = 25, $p <$ .01). The latter conditions did not significantly differ (p = 1.00).

Table 1. *Experiment 2. Subjects' appraisals of the feedback according to condition (*SD*s in parentheses;* N = *85).*

	Condition			
	Control	**Past-orientation**	**Future-orientation**	**Total**
Fairness	4.23 (0.71)	4.00 (0.73)	4.14 (0.71)	4.12 (0.71)
Helpfulness	4.12 (0.52)	4.19 (0.65)	4.43 (0.63)	4.25 (0.62)
Session 1 self-assessment (%)	48.42 (13.40)	51.35 (11.76)	53.32 (8.71)	51.11 (11.45)
Projected grade (%)	60.00 (11.84)	64.52 (9.08)	67.18 (8.93)	64.01 (10.26)

[1] In the full sample (N = 85), $n_{control}$ = 26; n_{past} = 31; n_{future} = 28. Of those 26 who failed to identify the correct instruction, $n_{control}$ = 4; n_{past} = 14; n_{future} = 8.

Free recall. Our recall and recognition analyses reached identical conclusions whether or not we included the 26 subjects who failed the attention check, therefore we report the full-sample analyses here. To explore which parts of their feedback subjects recalled, we conducted a 3 (Instruction: Control vs. Past-orientation vs. Future-orientation) x 2 (Feedback style: evaluative vs. directive) x 2 (Retrieval style accuracy: correct vs. incorrect) mixed-measures ANOVA on the number of feedback comments recalled. As shown in Figure 2, we once again found an evaluative recall bias: subjects recalled significantly more of the evaluative feedback comments than of the directive feedback comments, $F(1, 82) = 40.21$, $p < .001$, $\eta^2_p = .33$, $d = 0.83$ [0.54, 1.11]. They also recalled more feedback in the correct style than in the incorrect style, $F(1, 82) = 31.34$, $p < .001$, $\eta^2_p = .28$, $d = 0.79$ [0.48, 1.09].

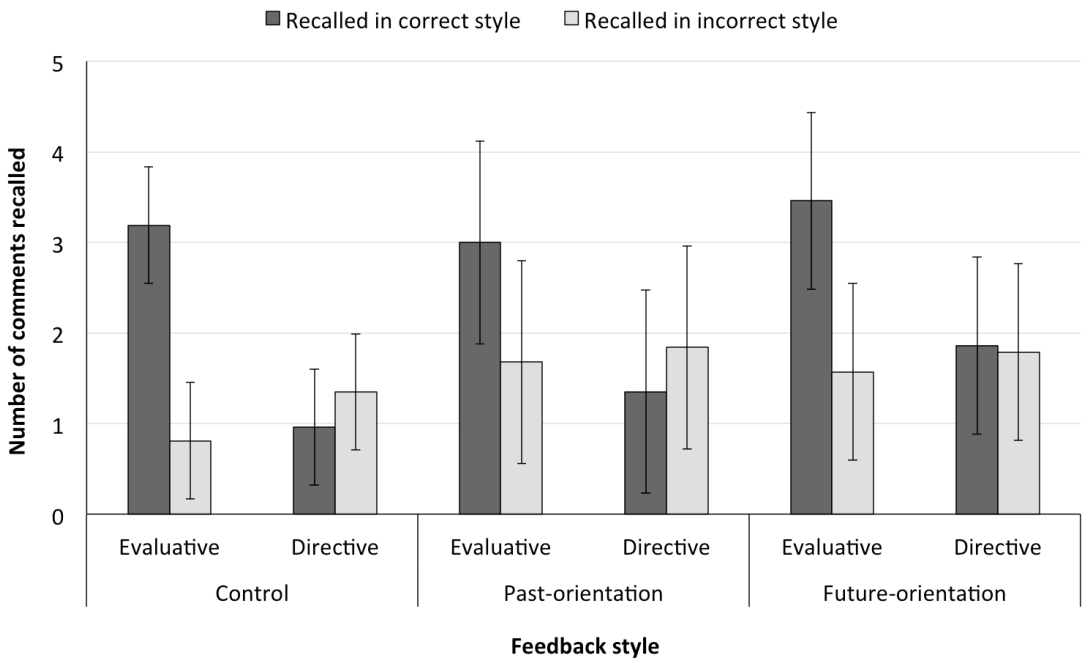

Figure 2. Recall of evaluative and directive feedback in Experiment 2, split according to retrieval style accuracy and instruction condition. Error bars are 95% within-subject confidence intervals, calculated separately for each instruction condition (Loftus & Masson, 1994).

A significant two-way interaction confirmed that subjects again adopted an evaluative retrieval style: that is, they reproduced feedback in an evaluative style—regardless of how comments had actually been framed—more frequently than they reproduced feedback in a directive style, $F(1, 82) = 15.31$, $p < .001$, $\eta^2_p = .16$, $d = 0.73$ [0.35, 1.11]. Follow up paired *t*-tests showed that people reproduced evaluative comments in the correct, evaluative style significantly more often than they reproduced evaluative comments in the incorrect, directive style, $t(84) = 5.57$, $p < .001$, $d = 1.02$ [0.63, 1.41]. Specifically, of all evaluative comments that were recalled, 70% were recalled as evaluative, and 30% as directive. However, people reproduced directive comments in the correct, directive style approximately as often as they reproduced directive comments in the incorrect, evaluative style, $t(84) = 0.97$, $p = .34$, $d = 0.17$ [-0.18, 0.53]. Specifically, of all directive comments that were recalled, 46% were recalled as directive, and 54% as evaluative.

Neither of the two-way interactions involving the instruction variable, nor the three-way interaction, was significant (all *p*s > .16, all $\eta^2_p < .05$). However, there was a significant overall main effect of instruction, $F(2, 82) = 4.87$, $p = .01$, $\eta^2_p = .11$. Post-hoc Bonferroni comparisons showed that subjects in the Future-orientation condition recalled more feedback overall compared with those in the Control condition ($p < .01$). However, overall recall in the Past-orientation condition did not differ significantly from either the Control condition ($p = .12$) or Future-orientation condition ($p = .83$). In other words, even though our instruction manipulation did not shift the evaluative recall bias, simply telling subjects to look for clues on how to improve in future nevertheless led them to subsequently recall more feedback overall.

Recognition performance. Subjects' recognition data are illustrated in Table 2. We conducted a 3 (Instruction) x 2 (Error type: Evaluative errors vs. Directive errors) mixed-factor ANOVA on the number of recognition errors made; this analysis revealed no significant main effect of instruction, $F(2, 82) = 0.65$, $p = .53$, $\eta^2_p = .02$, nor of error type, $F(1, 82) = 0.36$, $p = .55$, $\eta^2_p < .01$, $d = 0.11$ [-0.24, 0.47], nor a significant two-way interaction of these factors, $F(2, 82) = 0.05$, $p = .95$, $\eta^2_p < .01$.

Table 2. *Subjects' recognition test performance in Experiment 2, according to instruction condition (*SD*s in parentheses;* N = *85).*

	Instruction			
	Control	**Past-orientation**	**Future-orientation**	**Total**
Evaluative errors (%)	13.85 (11.34)	12.58 (9.99)	14.64 (12.32)	13.65 (11.11)
Directive errors (%)	14.23 (11.72)	14.52 (10.28)	16.07 (12.86)	14.94 (11.51)
Accurate recognition (%)	71.54 (15.67)	72.90 (11.31)	69.29 (11.52)	71.29 (12.80)

Achievement goal orientation. There was little evidence that subjects' recall bias, retrieval style, or total number of comments recalled, were meaningfully related to their achievement goal orientations (see Table S2 in the online supplemental materials for details).

Experiment 3

The results of Experiment 2 show that the evaluative recall bias can be replicated, and was relatively unaffected by an explicit instruction to prioritize directive (or evaluative) information. The latter finding gives us reason to believe that the bias is not a product of subjects assuming that evaluative information would be more important to encode.

So how might we explain this bias? One possibility is that subjects did not particularly care about improving their persuasive writing skills, and so were not motivated to care about the directive feedback. This explanation does not wholly fit with the spontaneous comments made by many subjects in Experiments 1 and 2; nevertheless it warrants some scrutiny. The explanation hinges on the assumption that whereas subjects cared little about improving (i.e., the directive feedback), they *did* care how they performed (i.e., the evaluative feedback). If this explanation were correct, then we should not observe the recall bias in a situation where the feedback is overtly irrelevant to the subjects themselves and to anything they have done. In this context, subjects should be no more motivated to read evaluative comments than to read directive comments.

Several other plausible accounts of the evaluative recall bias similarly assume that the effect would disappear when people read and recall feedback that is irrelevant to themselves. For example, a distinguishing feature of the evaluative feedback in Experiments 1 and 2 is that it was ostensibly related to essays that the subject already wrote, and that they could therefore visualize and remember producing, whereas the directive feedback only related to hypothetical future essays. Another account of the bias, then, is that evaluative feedback is more concrete, which could make those comments easier to encode and/or retrieve from memory than is directive feedback. A third account is that evaluative feedback is perceived as more self-relevant than is directive feedback, insofar as it relates to the subject's current self, rather than to a possible future self, and so is more easily or effectively encoded. Finally, a fourth account is that because directive feedback implies an obligation to act upon the advice, whereas evaluative feedback does not, the former might provoke an "information avoidance" response. This response, observed in many domains of

applied psychology, is characterized by an unwillingness to receive information that might require difficult actions to be taken (e.g., Howell & Shepperd, 2013; Sweeny, Melnyk, Miller, & Shepperd, 2010). If people selectively avoid feedback that obliges them to work hard to improve, then this would explain why directive feedback is poorly recalled.

Like the motivational account described above, if any of the concreteness, self-relevance, or information avoidance accounts is correct, then the evaluative recall bias should only occur when subjects receive feedback about *their own* prior performance, not when the feedback is irrelevant to themselves. In Experiment 3, we tested the plausibility of these four accounts by removing the persuasive writing task from our procedure entirely, and simply showing subjects—and asking them to recall—the feedback scripts that we used in our earlier experiments. In these circumstances, both evaluative and directive feedback should be equally motivating, concrete, and self-relevant, and neither should evoke information avoidance. If the evaluative recall bias disappeared in these circumstances, then the next step would be to determine which of the mechanisms had played a role.

Method

Subjects. A total of 40 volunteers (30 females and 10 males, M_{age} = 28.23, *SD* = 12.02, Range = 18-71) responded to an online advertisement, and took part without compensation.

Materials and procedure. We invited subjects to complete an online study on how people judge assessment feedback. To start, we informed subjects they would see some written feedback that another student had supposedly received after completing a set of short essays. We asked subjects to imagine the person receiving this feedback, and to read it carefully.

The remainder of the procedure was identical to the second session in Experiment 1, except that we shortened the filler task to 3 min, and subjects did not rate the fairness or helpfulness of the feedback, estimate grades, or complete a recognition task.

Data coding. We coded the data in the same way as in Experiments 1 and 2, but this time due to the smaller sample, the second coder coded 100% of responses. The inter-rater agreement was very high: number of evaluative comments recalled in an evaluative style (r = .95); the number of directive comments recalled in a directive style (r = .93); the number of evaluative comments recalled in a directive style (r = .93); and the number of directive comments recalled in an evaluative style (r = .96). The analyses below are therefore based on the first coder's data.

Results

The aim of this study was to find out whether subjects would still preferentially recall evaluative feedback over directive feedback, even when those feedback comments were not related to the subjects' concrete personal experiences. To answer this question, we conducted a 2 (Feedback style: evaluative vs. directive) x 2 (Retrieval style accuracy: correct vs. incorrect) repeated-measures ANOVA on the number of feedback comments recalled. Figure 3 shows that the results were highly similar to those of the previous experiments. Specifically, subjects once again recalled significantly more of the evaluative feedback than of the directive feedback, $F(1, 39) = 6.65$, $p = .01$, $\eta^2_p = .15$, $d = 0.54$, 95% CI on d [0.11, 0.97]. There was also a significant main effect of retrieval style accuracy, whereby subjects recalled more feedback comments in the correct style than in the incorrect style, $F(1, 39) = 18.61$, $p < .001$, $\eta^2_p = .32$, $d = 0.99$ [0.48, 1.48].

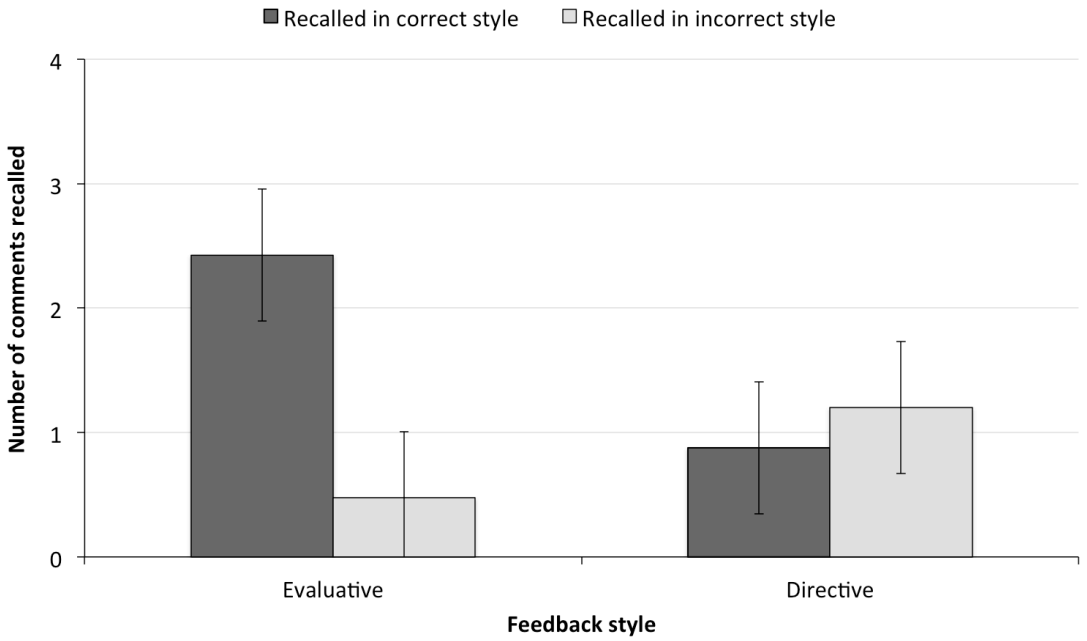

Figure 3. Recall of evaluative and directive feedback in Experiment 3, split according to retrieval style accuracy. Error bars are 95% within-subject confidence intervals (Loftus & Masson, 1994).

Finally, a significant interaction effect confirmed that subjects tended to reproduce feedback in an evaluative style—regardless of how comments were actually presented— more frequently than they reproduced feedback in a directive style, $F(1, 39) = 18.87$, $p < .001$, $\eta^2_p = .33$, $d = 1.13$ [0.56, 1.70]. Follow up paired *t*-tests showed that people reproduced evaluative comments in the correct, evaluative style significantly more often than they reproduced evaluative comments in the incorrect, directive style, $t(39) = 5.62$, $p < .001$, $d = 1.25$ [0.68, 1.80]. Specifically, of all evaluative comments that were recalled, 84% were recalled as evaluative, and 16% as directive. However, people reproduced directive comments in the correct, directive style no more often than they reproduced directive comments in the incorrect, evaluative style, $t(39) = 1.10$, $p = .28$, $d = -0.48$ [-1.00, 0.05]. Specifically, of all directive comments that were recalled, 42% were recalled as directive, and 58% as

evaluative. Overall then, we wholly replicated the main findings of Experiment 1 despite these subjects only reading feedback that apparently belonged to another person.

Experiment 4

Based on the results of Experiment 3, the evaluative recall bias cannot easily be attributed to subjects being primarily focused on how well they performed, and disinterested in future improvement. Nor are those results consistent with the interpretations that evaluative feedback is more concrete or self-relevant to subjects, or less likely to invoke information avoidance. If any of these four accounts were correct, then the evaluative recall bias should have disappeared in Experiment 3, and yet this was not the case. To validate this finding, in Experiment 4 we aimed to replicate the general method of Experiment 3.

We also set out to address a second question in Experiment 4: would the evaluative recall bias survive in a fully between-subjects design, where each subject sees either evaluative feedback only, or directive feedback only (as contrasted with the within-subjects design used in all experiments thus far, where all subjects saw both types of feedback)? Answering this question has practical relevance, given that feedback is undoubtedly delivered in many different formats in the real-world, rather than always with evaluative and directive advice interleaved. But this question is also theoretically relevant. For instance, another plausible explanation of the evaluative recall bias is that evaluative comments preferentially capture attention, perhaps because these comments can feel destructive where directive feedback feels constructive (Fong et al., 2016). Studies from many areas of psychological science show us that stimuli conveying threat can automatically attract attention, even when the observers themselves (i.e., the subjects) are not personally threatened (Öhman &

Mineka, 2001; Yiend & Mathews, 2001). If this attention capture account were correct, then we should predict the evaluative recall bias to disappear in a between-subjects design, where in principle no feedback comments should systematically draw attention away from others.

Method

Subjects. A total of 165 students from a large further education college in the West Midlands of England took part during class, without compensation. This sample size was based solely on the number of students available on one day of testing. In total, 13 subjects were removed from analyses because they failed to follow task instructions (e.g., reported that they had not read the feedback). All analyses are thus based on the remaining 152 subjects (112 females, 39 males, 1 other; M_{age} = 16.92, *SD* = 0.39, Range = 16-19). Each subject was randomly assigned to either the Mixed feedback condition (n = 50), the Evaluative-only feedback condition (n = 51), or the Directive-only feedback condition (n = 51).

Materials and procedure. Subjects followed the same general procedure as in Experiment 3, with two amendments. First, subjects in the Evaluative-only feedback condition saw a feedback script combining all the evaluative critique comments from both of the scripts used in Experiments 1-3, without any directive comments. Likewise, subjects in the Directive-only feedback condition saw a script containing all the directive critique comments, with no evaluative comments. Subjects in the Mixed feedback condition saw, at random, one of the two scripts used in Experiments 1-3, in which evaluative and directive comments were interleaved. In all conditions the same praise comments appeared at the start and end of each paragraph.

The second amendment from Experiment 3 was that after completing the recall task, all subjects made a series of additional judgments similar to those

collected in Experiments 1-2. Specifically, they rated (1) how helpful the feedback would be to the person who received it, (2) what percentage grade they think the person received, and (3) what grade the person might get on a subsequent task if they took the feedback on board. We also covertly measured the amount of time subjects spent on the feedback page before moving on. Mirroring the forced-choice questions used in our informal survey, mentioned in the introduction to Experiment 2, we asked subjects which kinds of comments they prefer to receive on their work (evaluative vs. directive), which they would pay most attention to (evaluative/directive/both about the same), and which they would be most likely to remember (evaluative/directive/both about the same).

Data coding. Subjects' recall data were coded following the same procedures as in the earlier experiments, and a second coder blind-coded 20% of responses. The inter-rater agreement was very high: number of evaluative comments recalled in an evaluative style (r = .95); the number of directive comments recalled in a directive style (r = .85); the number of evaluative comments recalled in a directive style (r = .90); and the number of directive comments recalled in an evaluative style (r = .88). Analyses are therefore based on the first coder's data.

Results

Subjects' appraisals of the feedback. As Table 3 shows, there were no meaningful differences between conditions in terms of the perceived helpfulness of the feedback, what grade the fictional student might achieve next time, or in terms of the time spent reading the feedback (all $p > .13$, all $\eta^2_p < .03$; distributions of reading times were already reasonably normal and so they were not log-transformed). However, there were significant differences in subjects' estimates of the grade the fictional student had achieved, $F(2, 149) = 4.08$, $p = .02$, $\eta^2_p = .05$. Post-hoc

Bonferroni tests showed that Directive-only subjects estimated higher grades than did Evaluative-only subjects (p = .01); however, neither group's estimates differed significantly from those of Mixed feedback subjects (both p > .46). In other words, when subjects only read directive critique, they believed the writer had performed better than when they read only evaluative critique. This finding may provide further insights into the underlying cause of the evaluative recall bias, and we return to this point shortly in an analysis of our study materials.

Table 3. *Experiment 4. Subjects' appraisals of the feedback according to condition (*SD*s in parentheses;* N = *152).*

	Condition			
	Mixed	**Evaluative-only**	**Directive-only**	**Total**
Helpfulness	3.64 (0.85)	3.69 (1.12)	3.80 (0.98)	3.71 (0.99)
Estimate of fictional student's grade (%)	62.92 (12.24)	59.63 (13.24)	66.25 (9.35)	62.93 (11.96)
Fictional student's projected grade (%)	75.06 (12.19)	75.73 (15.50)	79.88 (11.33)	76.90 (13.22)
Time spent reading feedback (sec)	114.9 (37.8)	116.3 (37.4)	114.7 (44.56)	115.3 (39.8)

Across the full sample, 76% of subjects said they would prefer to receive directive feedback (vs. 24% preferring evaluative). Whereas 41% believed they would pay more attention to directive feedback, only 13% believed they would pay more attention to evaluative feedback. And whereas 51% believed they would remember directive feedback better, only 30% believed they would remember evaluative feedback better. There were no significant differences across conditions (all χ^2 (4) < 4.4, all p > .36).

Free recall. We began our main analyses by examining the data from the Mixed feedback condition alone. As the leftmost part of Figure 4 shows, this analysis

wholly replicated our previous findings. Specifically, a 2 (Feedback style: evaluative vs. directive) x 2 (Retrieval style accuracy: correct vs. incorrect) repeated-measures ANOVA showed that subjects recalled significantly more of the evaluative feedback than of the directive feedback, $F(1, 49) = 13.80$, $p < .001$, $\eta^2_p = .22$, $d = 0.68$, 95% CI on d [0.29, 1.06]. There was also a significant main effect of retrieval style accuracy, whereby subjects recalled more feedback comments in the correct style than in the incorrect style, $F(1, 49) = 45.51$, $p < .001$, $\eta^2_p = .48$, $d = 1.27$ [0.82, 1.71].

A significant interaction effect showed that subjects reproduced feedback in an evaluative style more frequently than they reproduced feedback in a directive style, $F(1, 49) = 30.24$, $p < .001$, $\eta^2_p = .38$, $d = 1.11$ [0.65, 1.56]. Follow up paired *t*-tests showed that subjects reproduced evaluative comments in the correct, evaluative style significantly more often than they reproduced evaluative comments in the incorrect, directive style, $t(49) = 8.27$, $p < .001$, $d = 1.59$ [1.09, 2.07]. Of all evaluative comments that were recalled, 85% were recalled as evaluative, and 15% as directive. However, subjects reproduced directive comments in the correct, directive style and in the incorrect, evaluative style approximately equally, $t(49) = 0.31$, $p = .76$, $d = 0.07$ [-0.36, 0.49]. Of all directive comments that were recalled, 52% were recalled as directive, and 48% as evaluative. In short, the data from the Mixed feedback condition fully replicate the findings of Experiment 3.

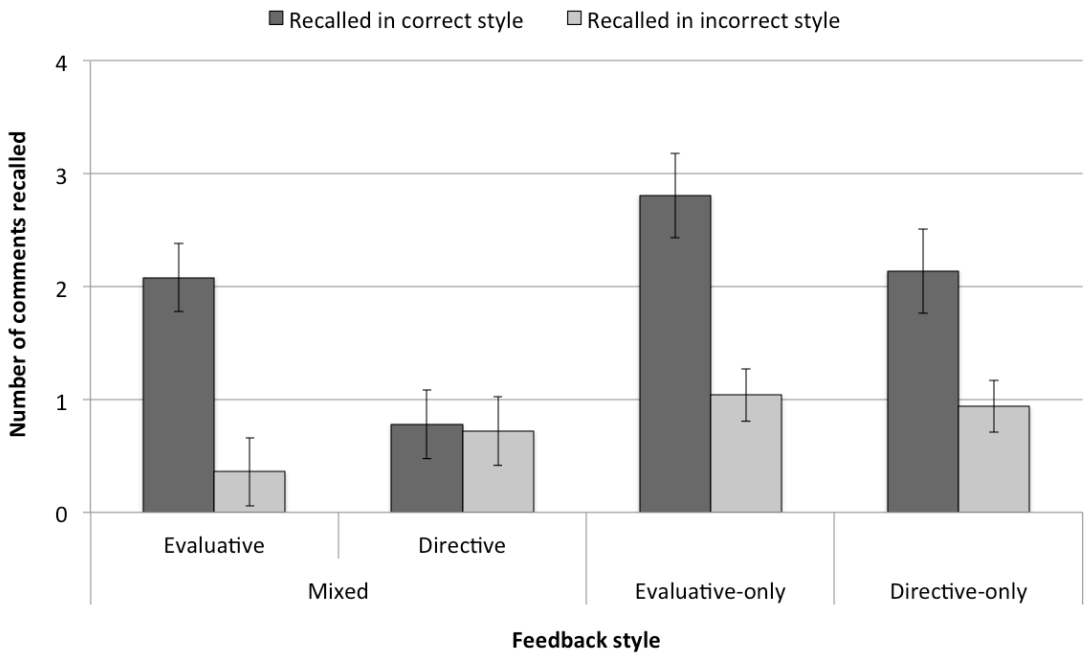

Figure 4. Recall of evaluative and directive feedback in Experiment 4, split according to retrieval style accuracy and feedback condition. Error bars for the Mixed condition are 95% within-subject confidence intervals (Loftus & Masson, 1994). Error bars for the Evaluative-only and Directive-only conditions are 95% between-subjects confidence intervals calculated separately for correct and for incorrect retrieval styles.

Looking next at the two between-subject conditions, we conducted a 2 (Feedback style: evaluative vs. directive) x 2 (Retrieval style accuracy: correct vs. incorrect) mixed-measures ANOVA on the number of feedback comments recalled, with the first factor manipulated between-subjects and the second manipulated within-subjects. As the rightmost parts of Figure 4 show, Evaluative-only subjects recalled significantly more feedback than did Directive-only subjects, $F(1, 100) = 4.26$, $p = .04$, $\eta^2_p = .04$, $d = 0.41$, 95% CI on d [0.02, 0.80]. In other words, we replicated the evaluative recall bias even in a between-subjects design; a finding that does not fit neatly with the notion that evaluative feedback captures attention relatively more than does directive feedback.

There was also a significant main effect of retrieval style accuracy, with subjects recalling more feedback comments in the correct style than in the incorrect style, $F(1, 100) = 34.54$, $p < .001$, $\eta^2_p = .26$, $d = 0.93$ [0.59, 1.27]. However, this time there was no significant interaction effect, $F(1, 100) = 1.27$, $p = .26$, $\eta^2_p = .01$, $d = 0.22$ [-0.17, 0.61], showing that Evaluative-only and Directive-only subjects were similarly accurate in their retrieval styles.

Experiments 1-4 materials analyses

The results across Experiments 1-4 are remarkably consistent. However, we used the same feedback materials in all four of these experiments, and by looking more closely at these materials, we can address some further accounts of our findings. Specifically, here we tackled three issues that might have contributed to the observed effects.

Item analysis of recall data

The first issue that needs to be addressed is that despite our efforts to ensure the evaluative and directive comments in our materials were closely matched, the evaluative recall bias might be driven by one or two particularly memorable evaluative comments within these scripts. We tested this explanation by conducting an item analysis, combining the data across all four experiments, and assessing how frequently each individual feedback comment was recalled when it was presented in an evaluative style versus a directive style. In this analysis we ignored the style in which subjects actually reproduced each comment.

Results. Our analysis revealed that of the 20 critical feedback comments presented to all subjects, 17 were recalled directionally more often when presented in the evaluative rather than the directive retrieval style (see Table S3 in the online supplemental materials). For example, of those subjects who were told "there was not

always a clear sense of where your points were leading" (an evaluative comment), 25% reproduced the gist of this comment in one style or the other. In contrast, of those subjects who were told "make sure there is always a clear sense of where your points are leading" (a directive comment), only 8% reproduced the gist of this comment. Our item analysis highlights concrete examples of how very subtle changes in wording led to sizeable effects on memory, and provides no evidence that our findings were driven by item-specific effects.

Ratings of individual feedback comments

The second issue to address, as illuminated in Experiment 4 (where subjects estimated higher percentage grades in the Directive-only condition than in the Evaluative-only condition), is that the evaluative and directive feedback may have differed not only in past vs. future orientation, but also in their perceived negativity or harshness. Confirming whether this is true may be important for identifying the cause of the apparent evaluative recall bias. The third issue is that people may infer that the intended meaning of feedback is evaluative, even when it is presented in a directive style. This inferred intention may be one possible explanation for the evaluative retrieval style that we have observed. To address both of these issues, we asked volunteers to individually appraise each of our feedback comments in terms of their harshness, their 'evaluativeness', and their 'directiveness'.

Subjects and procedure. A total of 40 volunteers (33 females, 7 males; M_{age} = 30.10, *SD* = 10.35, Range = 19-63) took part online in exchange for a £5 voucher. Each was shown all 40 of the feedback comments used in Experiments 1-4, in a random sequential order. For each comment, subjects were given the stem "If I received this feedback comment, I would think it is…", followed by three stem completions with 7-point response scales (1 = Not at all; 7 = Very much). These

involved judging whether each comment was (1) Worded negatively or harshly (hereafter, 'harshness'); (2) 'About' the quality of my work (hereafter, 'evaluativeness'); and (3) 'About' how I could improve next time (hereafter, 'directiveness'). The harshness scale served to test the prediction that evaluative feedback would be judged as more harsh than directive feedback. The evaluativeness and directiveness scales served to test the notion that people infer an evaluative intention to feedback comments, regardless of written style. We predicted that if this notion were true, evaluative comments would be seen to have a weak directive function, whereas directive comments would be seen to have a strong evaluative function.

Results. Analysis of subjects' harshness ratings revealed that evaluative feedback comments were indeed judged as significantly more harsh or negative (M = 2.29) than were directive comments (M = 1.76), $t(39) = 8.46$, $p < .001$, $d = 0.78$, 95% CI on d [0.53, 1.03]. We therefore return to explore this perceived harshness mechanism directly in Experiment 5.

Looking at evaluativeness and directiveness ratings, there was a significant interaction of feedback type and rating type, $F(1, 39) = 62.16$, $p < .001$. As we should expect, evaluative comments were judged to have more of an evaluative function (M = 5.46) than a directive function (M = 3.97), $t(39) = 5.98$, $p < .001$, $d = 1.22$ [0.73, 1.70] whereas directive comments were judged to have more of a directive function (M = 5.68) than an evaluative function (M = 4.85), $t(39) = 4.95$, $p < .001$, $d = 0.79$ [0.43, 1.15]. However, in line with our specific prediction, subjects judged the evaluative function of directive comments to be significantly greater than the directive function of evaluative comments, $t(39) = 4.08$, $p < .001$, $d = 0.74$ [0.34, 1.13]. In other words, people judged all comments to be evaluative regardless of style, and this

interpretation bias may partly or wholly explain the evaluative retrieval style in subjects' free recall responses.

Experiment 5

We now have two sets of converging evidence that people interpret evaluative feedback as more negative or harsh than directive feedback. This difference in harshness might explain why evaluative feedback is more effectively recalled than directive feedback. For example, Cutumisu and Schwartz (in press) recently showed that middle school students were significantly better at remembering negative feedback than they were at remembering positive feedback. In short, it is possible that subjects in our paradigm are simply best at remembering the most negative or harsh sounding feedback comments, regardless of their temporal orientation.

In Experiment 5 we tested this harshness mechanism directly, by trying to completely reverse the evaluative recall bias using new, adapted feedback scripts. Specifically, in some of our new scripts the evaluative feedback was intentionally written to seem harsher than the directive feedback (as was apparently the case in our prior experiments), but in other scripts, the directive feedback was written to seem harsher than the evaluative feedback. If perceived harshness, rather than temporal orientation per se, is the mechanism responsible for the evaluative recall bias, then we should expect to see the bias reverse entirely whenever directive comments are the harsher type. But if the effect does not reverse under these conditions, then we should conclude that differences in perceived harshness cannot explain the evaluative recall bias.

Method

Subjects. A total of 66 subjects completed the study online in exchange for a £5 voucher. This sample size was based on a power analysis, which showed that 66

subjects would permit the detection of a medium-sized interaction (f = .25) in a 2 x 2 mixed-measures design, assuming power = .80, alpha = .05, and a correlation of zero between repeated-measures. In total, 6 subjects were removed from analyses and replaced with new subjects, because they failed to follow task instructions (e.g., recalling the filler task instructions rather than the feedback). The final dataset comprised data from 49 females, 16 males, 1 other; M_{age} = 29.11, SD = 13.48, Range = 18-74). Each was randomly assigned to either the Evaluative-harsher condition (n = 34), or the Directive-harsher condition (n = 32).

Materials. We created a new set of feedback scripts for use in this experiment, adapted from those used in the earlier experiments. Specifically, we re-wrote each of the 40 original comments (20 evaluative, 20 directive) in two ways, one intended to seem 'supportive', and the other intended to seem 'stern', whilst otherwise carrying the same general critique. For example, one supportive-evaluative comment was "…you could have used evidence a bit more consistently to support your arguments", whereas the stern-evaluative variant of this comment was "…you failed to consistently use even a trace of evidence to support your arguments". The supportive-directive variant was "…you could try to make more consistent use of evidence to support your arguments", whereas the stern-directive variant was "you should consistently use at least some trace of evidence to support your arguments." To ensure that these new materials effectively manipulated perceived harshness, we piloted them with a separate group of volunteers.

Pilot study. Forty volunteers who were not involved in the main study (30 females, 10 males; M_{age} = 31.75, SD = 10.41, Range = 19-57) took part online in exchange for a £5 voucher. Each was shown all 80 feedback comments in a random sequential order (i.e., 20 comments presented in the Supportive-Evaluative,

Supportive-Directive, Stern-Evaluative, and Stern-Directive styles). For each comment, subjects rated their agreement with the statement "If I received this feedback comment, I would think it is negative or harsh" (1 = Not at all; 7 = Very much).

Mirroring our analysis of our original materials, reported above, evaluative feedback comments were judged as significantly more harsh (M = 3.01) than directive comments (M = 2.78), $F(1, 39) = 60.55$, $p < .001$, $\eta^2_p = .61$, $d = 0.55$, 95% CI on d [0.37, 0.74]. Confirming the effectiveness of our manipulation, "Stern" feedback comments were also considered significantly more harsh (M = 3.89) than were "Supportive" comments (M = 1.91), $F(1, 39) = 362.86$, $p < .001$, $\eta^2_p = .90$, $d = 3.80$ [2.87, 4.73]. Importantly, all four kinds of comments were judged to differ significantly from one another in terms of harshness. In particular, Stern-Evaluative comments were judged as more harsh (M = 4.03) than were Supportive-Directive comments (M = 1.82), $t(39) = 20.57$, $p < .001$, $d = 4.17$ [3.16, 5.17]. Stern-Directive comments were judged as more harsh (M = 3.74) than were Supportive-Evaluative comments (M = 1.99), $t(39) = 16.13$, $p < .001$, $d = 3.25$ [2.42, 4.06].

Having confirmed the effective manipulation, we created four new feedback scripts using these new materials (see the online supplemental materials). In the two Evaluative-harsher scripts (version A and version B as in earlier experiments, with the evaluative vs. directive style of each comment counterbalanced across versions), the evaluative comments were always stern and the directive comments were always supportive, thus replicating the scenario apparently seen in Experiments 1-4. In the two Directive-harsher scripts, the evaluative comments were always supportive and the directive comments were always stern.

Procedure. The procedure was identical to Experiment 3, with the exception that we used the new Evaluative-harsher and Directive-harsher feedback scripts in place of the original scripts. Subjects saw one of the four scripts at random.

Data coding. The data were coded in the same way as in previous experiments, and a second coder coded 20% of responses. The inter-rater agreement was very high: number of evaluative comments recalled in an evaluative style (r = .97); the number of directive comments recalled in a directive style (r = .93); the number of evaluative comments recalled in a directive style (r = .86); and the number of directive comments recalled in an evaluative style (r = .86). The analyses below are therefore based on the first coder's data.

Results

We conducted a 2 (Condition: Evaluative-harsher vs. Directive-harsher) x 2 (Feedback style: evaluative vs. directive) x 2 (Retrieval style accuracy: correct vs. incorrect) mixed-measures ANOVA on the number of feedback comments recalled, with the first factor manipulated between-subjects and the latter factors manipulated within-subjects.

Overall, as Figure 5 shows, we replicated the evaluative recall bias: subjects recalled significantly more of the evaluative feedback than of the directive feedback, $F(1, 64) = 5.30$, $p = .02$, $\eta^2_p = .08$, $d = 0.37$, 95% CI on d [0.05, 0.69]. They also recalled more feedback comments in the correct style than in the incorrect style, $F(1, 64) = 19.72$, $p < .001$, $\eta^2_p = .24$, $d = 0.81$ [0.42, 1.19], and a significant interaction effect showed that they reproduced feedback in an evaluative style more frequently than in a directive style, $F(1, 64) = 62.66$, $p < .001$, $\eta^2_p = .50$, $d = 1.39$ [0.97, 1.80]. Follow up paired t-tests showed that subjects reproduced evaluative comments in the correct, evaluative style significantly more often than they reproduced evaluative

comments in the incorrect, directive style, $t(65) = 8.78$, $p < .001$, $d = 1.56$ [1.12, 2.00]. Of all evaluative comments that were recalled, 89% were recalled as evaluative, and 11% as directive. However, subjects reproduced directive comments in the correct, directive style less often than in the incorrect, evaluative style, $t(65) = -2.05$, $p = .04$, $d = -0.39$ [-0.77, -0.01]. Of all directive comments that were recalled, 38% were recalled as directive, and 62% as evaluative.

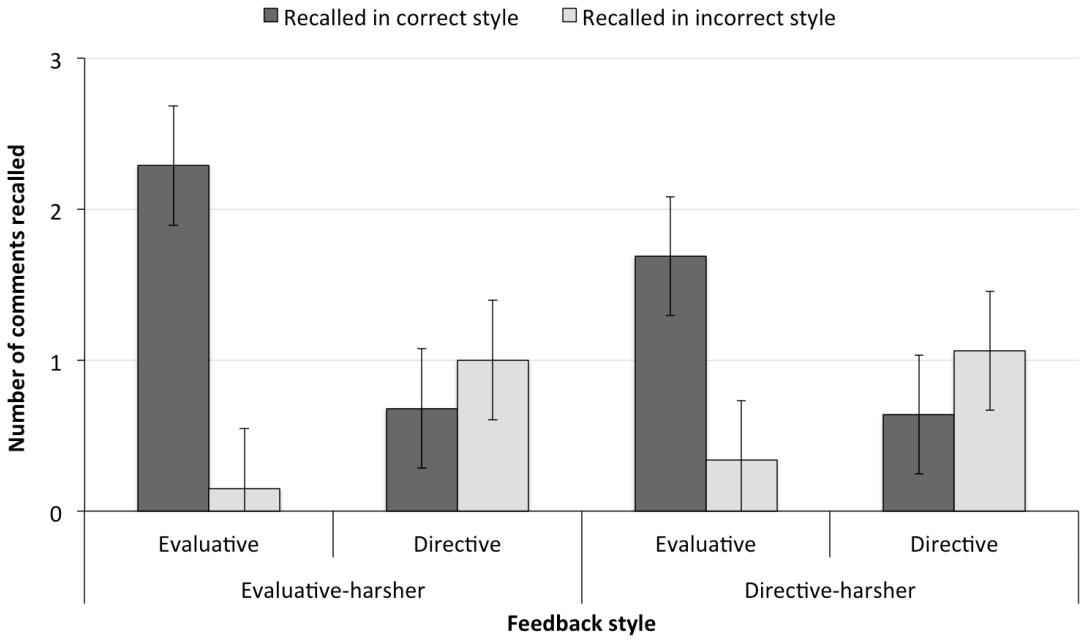

Figure 5. Recall of evaluative and directive feedback in Experiment 5, split according to retrieval style accuracy and script condition. Error bars are 95% within-subject confidence intervals, calculated separately for each script condition (Loftus & Masson, 1994).

Importantly, Figure 5 also shows that the results were highly comparable in the Evaluative-harsher condition and the Directive-harsher condition. As a reminder, if the evaluative recall bias were driven by differences in the perceived harshness of the comments, then we should see a switch to a directive recall bias in the Directive-

harsher condition. But this was not the case. In fact, the condition x feedback type interaction was very small and not statistically significant, $F(1, 64) = 0.93$, $p = .34$, $\eta^2_p = .01$, thus suggesting that the harshness of the comments cannot explain the evaluative recall bias. There was no other significant main effect or interaction involving condition (all $p > .09$, all $\eta^2_p < .05$).

Experiment 6

As well as testing the harshness mechanism directly, Experiment 5 also confirms that the evaluative recall bias can be replicated using different feedback scripts. Nevertheless, even in Experiment 5 the scripts we used were still similar to those used in the earlier experiments. We therefore carried out a final experiment using entirely different feedback scripts to extend the generalizability of the findings. Furthermore, this time we recruited a subject sample of 9- to 10-year old children: a group who just like adults, often have difficulties in remembering feedback that they know should help them to improve (Hargreaves, 2012). Extending our research with child subjects is potentially valuable because it could help us to begin tracing the roots of the evaluative recall bias. Developmental neuroscience research suggests that before around age 11, areas of the brain involved in cognitive control typically respond more to positive than to negative feedback during learning, whereas the opposite is true during adulthood (van Duijvenvoorde, Zanolie, Rombouts, Raijmakers, & Crone, 2008). Although our evaluative vs. directive manipulation does not map neatly onto a negative vs. positive distinction, nevertheless these findings may lead us to predict that children aged 9 to 10 would process critical feedback in a qualitatively different way to adults. Like in Experiments 3-5 then, in Experiment 6 we showed our child subjects a piece of generic written feedback and subsequently asked them to recall as much detail as they could.

Method

Subjects. A total of 46 children (25 females and 21 males, M_{age} = 9.80, *SD* = 0.40, Range = 9-10, five children did not give their age) from a school in the south-east of England took part during class, without compensation.

Materials and procedure. The procedure was largely identical to Experiment 3, but the materials were provided on paper rather than via a computer, and we developed new feedback scripts in collaboration with the subjects' class teachers to make the feedback more accessible and appropriate to the children. These feedback scripts were shorter than those used in Experiments 1-5 (both versions = 139 words), and they described developmentally relevant writing issues such as the appropriate use of capital letters. Both versions of the feedback can be found in the online supplemental materials. Subjects were told that the feedback had been received by another child, and were asked to read it carefully. Next they completed a similar filler task as in the previous experiments, for a total of 5 min. This task was presented as a "puzzle sheet" and subjects were asked to complete as many of the puzzles as they could within the allotted time. Following the filler task, subjects were given a further 5 min to write down as much of the feedback as they could remember. At the end of the study we asked subjects to report, by ticking the appropriate box on their sheet, whether they prefer to receive comments on their work that (a) tell them how they have done on that piece of work, or that (b) tell them how to improve for next time.

Data coding. Both the first and second coders coded 100% of the recall data. The inter-rater agreement was very high: number of evaluative comments recalled in an evaluative style (r = .96); the number of directive comments recalled in a directive style (r = .98); the number of evaluative comments recalled in a directive style (r =

.92); and the number of directive comments recalled in an evaluative style (r = .91). The analyses below are therefore based on the first coder's data.

Results

In total, 74% of the children told us that they prefer receiving directive feedback, whereas just 26% said they prefer receiving evaluative feedback. And yet, even though the consensus was clearly in favor of preferring directive feedback, Figure 6 shows that subjects nevertheless recalled significantly more of the evaluative feedback than of the directive feedback, $F(1, 45) = 16.90$, $p < .001$, $\eta^2_p = .27$, $d = 0.76$ [0.36, 1.16]. That is to say, we were able to replicate the evaluative recall bias both in a different subject population, and using completely different stimulus materials from the other experiments. Like in Experiments 3-5, this bias emerged despite the feedback apparently being destined for a stranger, rather than for the subjects themselves.

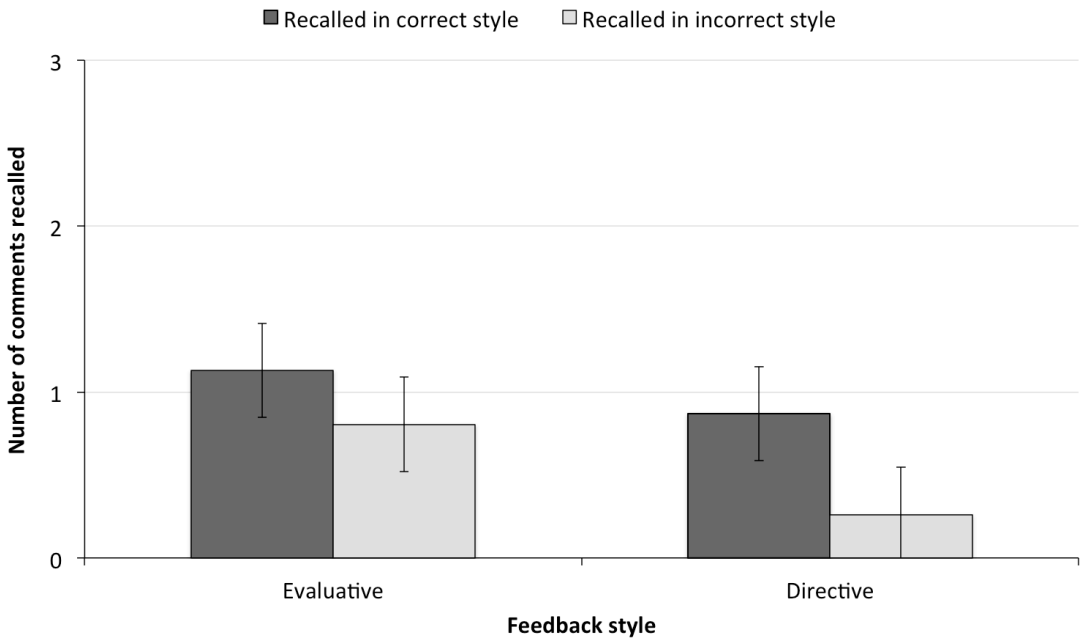

Figure 6. Recall of evaluative and directive feedback in Experiment 6, split according to retrieval style accuracy. Error bars are 95% within-subject confidence intervals (Loftus & Masson, 1994).

Subjects recalled more feedback comments in the correct style than in the incorrect style, $F(1, 45) = 13.22$, $p = .001$, $\eta^2_p = .23$, $d = 0.78$ [0.33, 1.23], but it is noteworthy that this time there was no significant interaction effect, $F(1, 45) = 0.79$, $p = .38$, $\eta^2_p = .02$, $d = -0.21$ [-0.67, 0.26]. In other words, unlike in our previous experiments with adult subjects (excepting the between-subjects conditions in Experiment 4), these children had no overall tendency to recall feedback in an evaluative rather than directive style. Of those evaluative comments that the children recalled, 58% were reproduced in the correct evaluative style, and 42% in the incorrect directive style. Of those directive comments that the children recalled, 77% were reproduced in the correct directive style, and 23% in the incorrect evaluative style. We comment on this difference in findings between experiments shortly.

Experiments 1-6 effect size analyses

In recent years, influential figures in psychological science have recommended a shift in how we report and interpret statistical findings, moving away from focusing solely on *p*-values, and instead paying greater attention to effect size estimates (e.g., Cumming, 2013). Taking a weighted average across all six of our experiments, the evaluative recall bias amounted to subjects recalling 46% more evaluative than directive feedback. To put this difference in standardized terms, we conducted a random effects mini meta-analysis of the data, the results of which are illustrated in Figure 7a. This meta-analysis gives a weighted effect size estimate for the evaluative recall bias of $d = 0.63$ [0.48, 0.77], $p < .001$.

Recall that in most of our experiments, subjects also tended to adopt an evaluative retrieval style, systematically reproducing the feedback comments in an evaluative style irrespective of how they were actually presented. Taking weighted averages across all six experiments, of all the directive feedback recalled, 50% was

reproduced in an incorrect, evaluative style; in contrast, of all the evaluative feedback recalled, only 23% was reproduced in an incorrect, directive style. A second random-effects mini meta-analysis, illustrated in Figure 7b, gives a weighted effect size estimate for the evaluative retrieval style of $d = 0.76$ [0.32, 1.20], $p < .001$.

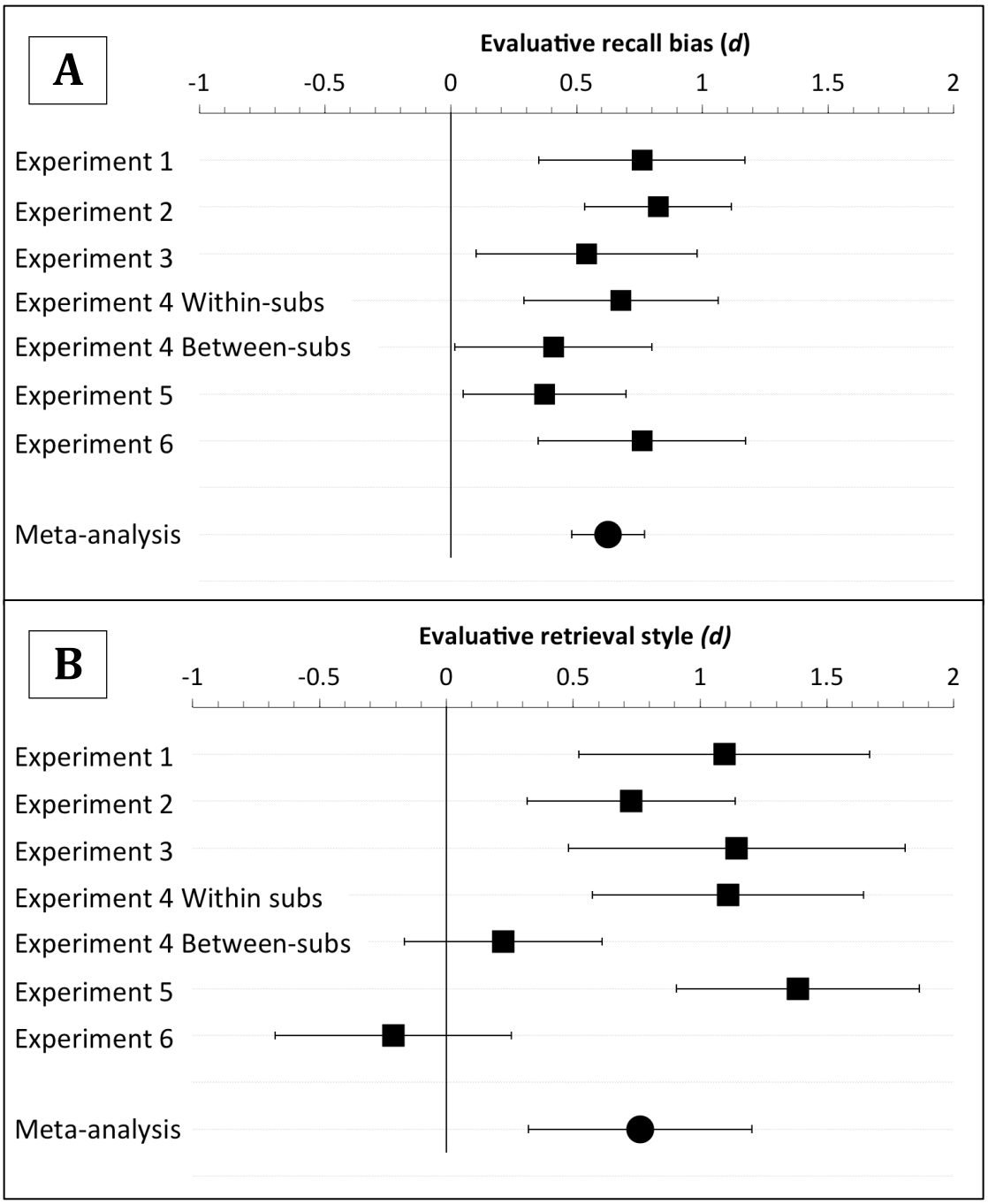

Figure 7. Forest plots illustrating standardized and meta-analyzed effect size data for (A) evaluative recall bias and (B) evaluative retrieval style across Experiments 1-6. Error bars are 95% confidence intervals around individual values of *d*.

General Discussion

Learners who can habitually remember the feedback they receive should in principle have a strong advantage as they strive to improve their skills. It has widely been argued that future-oriented directive feedback is more valuable to learners than is past-oriented evaluative feedback; however, our data indicate two crucial counterpoints. First, even though the wording of both forms of feedback was closely similar, directive feedback was less likely than evaluative feedback to be recalled by the adults and children in our studies. Our item analysis of Experiments 1-4 illustrates this point concretely, showing how small differences in wording often had sizeable effects on the likelihood of recall. Second, even when adults (but not children) did successfully remember directive feedback comments, those comments were very often misremembered as criticisms of prior performance (i.e., as evaluative) rather than guidance on future improvement. In short, we have good cause to believe that these memory biases are sizeable and robust in the kinds of contexts we have studied.

Notably, the preferential recall of evaluative over directive feedback is the exact opposite of the effect we predicted. Indeed, in a small informal survey and again in Experiment 4, we found that our participants largely predicted the opposite effect, too. The direction of the effect is puzzling then, not least because it seems at odds with the finding that people typically recall information more effectively when oriented toward the future (e.g., Klein et al., 2010, 2011). What is more, our data lend little support to several theoretical interpretations of why the bias might occur.

First, the bias occurred even when we told subjects to prioritize finding out how they could improve in future, and it was not exaggerated when we told them to prioritize finding out how they had performed (Experiment 2). This finding suggests that the bias is not driven by subjects' assumptions about which feedback comments

were more important to encode. Second, the bias occurred when subjects read another fictional person's feedback, rather than believing that the feedback pertained to their own persuasive writing (Experiments 3-6). This finding suggests that the bias is not a result of subjects being disproportionately interested in information on how they performed on the writing task, relative to information on how to improve. It also suggests that the bias is not driven by concreteness or self-reference effects on memory that bolster the encoding of evaluative information, because when reading another person's feedback, both evaluative and directive comments should have been equally concrete and equally self-relevant. Likewise, in Experiments 3-6 neither type of feedback should have invoked so-called information avoidance, and so our findings there suggest that this is not the responsible mechanism. In Experiments 1 and 2, there was little evidence that the bias was related to individual differences in (trait) achievement goal orientation, which provides some initial cause to doubt the role of stable motivational factors. In Experiment 4 we replicated the effect even in a between-subjects design, a finding that points away from an attentional capture mechanism, albeit more direct tests of this mechanism would be valuable. We found good evidence that evaluative feedback is perceived as harsher than directive feedback, yet the data from Experiment 5 suggest that these differences in harshness cannot explain the evaluative recall bias. And finally, observing the bias among children in Experiment 6 suggests that the bias is not learned through relatively formative experiences in education. We cannot rule out the role of even earlier experiences, of course, but this finding does suggest a mechanism rooted in more basic cognitive processes.

So in light of all these direct and indirect tests of theoretical accounts, how then might we otherwise explain this unexpected yet consistent effect? Future work

should examine the extent to which the evaluative recall bias is a retrieval effect (i.e., people are better able to retrieve evaluative feedback from memory), versus an encoding effect (i.e., people recall evaluative feedback better because they are more capable of committing it to memory). If evaluative feedback were encoded more effectively, then we might expect people to be more able to subsequently recognize this feedback, yet in Experiments 1 and 2, we found no reliable differences between evaluative and directive feedback in terms of recognition memory. This finding might therefore offer preliminary evidence that the effect in fact occurs at retrieval rather than encoding. Stronger and more direct tests of this question are needed before drawing firm conclusions though, especially given that subjects' recognition responses could plausibly have been contaminated by completing the free recall test. One as-yet untested retrieval-based account is that when attempting to retrieve feedback from memory, people selectively search for evaluative information, a strategy that could interfere with their ability to retrieve any directive feedback stored in memory. A direct test of this selective memory search mechanism might involve disrupting subjects' search-set, perhaps by instructing them to reproduce evaluative and directive comments separately. We are currently exploring this possibility.

Whatever the causal mechanism, these findings show that future-orientation does not always benefit memory; in this case it had quite the opposite effect. Gaining a better understanding of why this is the case may provide substantial contributions to theory on the directive and adaptive functions of episodic remembering. Moreover, it will be necessary to address the extent to which this memory bias generalizes across various real-world feedback scenarios: when there are higher-stakes involved in successfully implementing the advice, for example; when the initial task (e.g., persuasive writing) is lengthier and more meaningful; when the feedback is more

richly encoded; or when the delay-to-test is longer. To the extent that the memory bias could generalize across some different contexts, further studies may permit the design of interventions for supporting learners in remembering directive feedback more effectively, or in translating evaluative feedback into future actions. Whereas directive feedback is not always inherently more valuable than evaluative feedback, an ideal scenario should in principle be one wherein people can remember both kinds effectively.

Our secondary focus here was on whether people would tend to systematically misremember feedback in a directive or an evaluative style. Systematic patterns of misremembering of information can often tell us about the kinds of inferences people have spontaneously made when processing that information (Brewer, 1977; Chan & McDermott, 2006; Garry et al., 2007; Klepacz et al., 2016). One might hope that upon reading evaluative criticism, learners would spontaneously infer what they needed to do differently in future. If so, we might expect that in many instances they would later mistakenly believe they had read directive feedback. The systematic pattern we found in our studies with adult subjects was generally the opposite, pointing to an evaluative retrieval style. These findings suggest that in fact, when receiving directive feedback, adult subjects were often spontaneously inferring what had been done badly in the writing task. Our analysis of our study materials offers us a plausible account of this bias: people intuitively read the intent of feedback as evaluative, regardless of style. Specifically, we found that subjects judged the intent of directive feedback to be both evaluative and directive, whereas they perceived the intent of evaluative feedback to be purely evaluative.

Notably, as Figure 7 shows, the retrieval style bias was far more heterogeneous in magnitude than was the evaluative recall bias, and it did not appear

in the between-subjects design in Experiment 4, thus suggesting a relatively larger contribution of contextual factors. Indeed, we did not find the evaluative retrieval style bias among the children in Experiment 6, which may perhaps indicate that children typically read feedback in a more literal sense than do adults, and are less likely to read subtext into a feedback giver's directive language. It is possible that subjects' tendency to infer an evaluative intent to feedback mirrors the kinds of feedback they are most accustomed to receiving. We are unaware of existing data sources to validate this idea, but it is reasonable to speculate that children may more regularly receive developmental guidance compared to adult learners. More work with a developmental focus may elucidate the basis of systematic biases in people's feedback retrieval style.

It is clear that these findings raise many questions, but perhaps most important is this: Could these strong effects in people's recall of feedback ultimately translate into behavioral effects? That is, would a student be more likely to subsequently act upon evaluative feedback than upon directive feedback, all else being equal, and could these small changes in wording lead to genuine, measurable differences in students' subsequent performance? Answering these questions demands field studies, with data collected at multiple time-points and with generalizability assessed across different kinds of meaningful tasks. If performance consequences can be observed, then this unexpected cognitive bias could alter how much we stand to benefit from receiving feedback, shaping skill development far beyond the formal education context.

References

Agarwal, P. K. (2012). Advances in cognitive psychology relevant to education: Introduction to the special issue. *Educational Psychology Review*, *24*, 353-354.

Atance, C. M., & O'Neill, D. K. (2001). Episodic future thinking. *Trends in Cognitive Sciences, 5*, 533-539.

Bluck, S. (2003). Autobiographical memory: Exploring its functions in everyday life. *Memory, 11*, 113-123.

Brewer, W. F. (1977). Memory for the pragmatic implications of sentences. *Memory & Cognition, 5*, 673-678.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 918-928.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*, 604-616.

Chan, J. C. K. & McDermott, K. B. (2006). Remembering pragmatic inferences. *Applied Cognitive Psychology, 20*, 633-639.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*, 293-332.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Hove, UK: Routledge.

Cutumisu, M., & Schwartz, D. L. (in press). The impact of critical feedback choice on students' revision, performance, learning, and memory. *Computers in Human Behavior.*

Elliot, A. J., & Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology, 10*, 613-628.

Fong, C. J., Warner, J. R., Williams, K. M., Schallert, D. L., Chen, L. H., Williamson, Z. H., & Lin, S. (2016). Deconstructing constructive criticism: The nature of academic emotions associated with constructive, positive, and negative feedback. *Learning and Individual Differences*, *49*, 393-399.

Garry, M., Strange, D., Bernstein, D. M., & Kinzett, T. (2007). Photographs can distort memory for the news. *Applied Cognitive Psychology, 21*, 995-1004.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*, 304-315.

Goschke, T., & Kuhl, J. (1993). Representation of intentions: Persisting activation in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1211-1226.

Hargreaves, E. (2012). Teachers' classroom feedback: still trying to get it right. *Pedagogies: An International Journal*, *7*, 1-15.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Hays, M. J., Kornell, N., & Bjork, R. A. (2010). The costs and benefits of providing feedback during learning. *Psychonomic Bulletin & Review, 17*, 797-801.

Howell, J. L., & Shepperd, J. A. (2013). Behavioral obligation and information avoidance. *Annals of Behavioral Medicine, 45*, 258-263.

Johnson, J. T., Cain, L. M., Falke, T. L., Hayman, J., & Perillo, E. (1985). The "Barnum effect" revisited: Cognitive and motivational factors in the acceptance of personality descriptions. *Journal of Personality and Social Psychology, 49*, 1378-1391.

Kang, S. H., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*, 528-558.

Klein, S. B. (2013). The temporal orientation of memory: It's time for a change of direction. *Journal of Applied Research in Memory and Cognition*, *2*, 222-234.

Klein, S. B., Robertson, T. E., & Delton, A. W. (2010). Facing the future: Memory as an evolved system for planning future acts. *Memory & Cognition, 38*, 13-22.

Klein, S. B., Robertson, T. E., & Delton, A. W. (2011). The future-orientation of memory: Planning as a key component mediating the high levels of recall found with survival processing. *Memory, 19*, 121-139.

Klepacz, N. A., Nash, R., Egan, M. B., Hodgkins, C. E., & Raats, M. M. (2016). When is an image a health claim? A false-recollection method to detect implicit inferences about products' health benefits. *Health Psychology, 35*, 898-907.

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*, 254-284.

Koriat, A., Ben-Zur, H., & Nussbaum, A. (1990). Encoding information for future action: Memory for to-be-performed tasks versus memory for to-be-recalled tasks. *Memory & Cognition, 18*, 568-578.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*, 476-490.

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychological Review, 108*, 483-522.

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest, 9*, 105-119.

Pillemer, D. (2003). Directive functions of autobiographical memory: The guiding power of the specific episode. *Memory*, *11*, 193-202.

Pillemer, D. B., Picariello, M. L., Law, A. B., & Reichman, J. S. (1999). Memories of college: The importance of specific educational episodes. In D. C. Rubin (Ed.), *Remembering our past: Studies in autobiographical memory* (pp. 318-337). New York, NY: Cambridge University Press.

Prabhakar, J., Coughlin, C., & Ghetti, S. (2016). The neurocognitive development of episodic prospection and its implications for academic achievement. *Mind, Brain, and Education, 10*, 196-206.

Roediger, H. L. (2013). Applying cognitive psychology to education: Translational educational science. *Psychological Science in the Public Interest*, *14*, 1-3.

Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition*, *1*, 242-248.

Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay–retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 80-95.

Sweeny, K., Melnyk, D., Miller, W., & Shepperd, J. A. (2010). Information avoidance: Who, what, when, and why. *Review of General Psychology, 14*, 340-353.

Van Duijvenvoorde, A. C., Zanolie, K., Rombouts, S. A., Raijmakers, M. E., & Crone, E. A. (2008). Evaluating the negative or valuing the positive? Neural mechanisms supporting feedback-based learning across development. *Journal of Neuroscience, 28*, 9495-9503.

Weinstein, Y., McDermott, K. B., & Roediger, H. L. (2010). A comparison of study strategies for passages: rereading, answering questions, and generating questions. *Journal of Experimental Psychology: Applied, 16*, 308-316.

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52*, 17-37.

Winstone, N. E., Nash, R. A., Rowntree, J., & Menezes, R. (2016). What do students want most from written feedback information? Distinguishing necessities from luxuries using a budgeting methodology. *Assessment and Evaluation in Higher Education, 41*, 1237-1253

Winstone, N. E., Nash, R. A., Rowntree, J., & Parker, M. (2017). 'It'd be useful, but I wouldn't use it': Barriers to university students' feedback seeking and recipience. *Studies in Higher Education, 42*, 2026-2041.

Wollenschläger, M., Hattie, J., Machts, N., Möller, J., & Harms, U. (2016). What makes rubrics effective in teacher-feedback? Transparency of learning goals is not enough. *Contemporary Educational Psychology, 44*, 1-11.

Yiend, J., & Mathews, A. (2001). Anxiety and attention to threatening pictures. *Quarterly Journal of Experimental Psychology: Section A, 54*, 665-681.

Supplemental Materials (Online only)

Table S1. *Experiment 1. Correlation coefficients of relationship between subjects' achievement goal orientations, and outcome measures from the recall and recognition tasks (*N = *61).*

	Achievement Goal Orientation			
	Mastery-approach	Mastery-avoidance	Performance-approach	Performance-avoidance
Total feedback recalled	-.17 (p = .19)	.09 (p = .49)	-.09 (p = .50)	-.00 (p = .98)
Evaluative recall bias[a]	-.13 (p = .31)	-.10 (p = .43)	.04 (p = .74)	.02 (p = .86)
Evaluative retrieval style[b]	-.17 (p = .19)	-.17 (p = .19)	-.17 (p = .19)	-.17 (p = .19)
Evaluative recognition bias[c]	-.08 (p = .55)	-.05 (p = .71)	-.04 (p = .78)	.00 (p = .99)

[a] Total evaluative comments recalled, minus total directive comments recalled
[b] Total comments reproduced in an evaluative style, minus total comments reproduced in a directive style
[c] Total evaluative errors in recognition, minus total directive errors in recognition

Table S2. *Experiment 2. Correlation coefficients of relationship between subjects' achievement goal orientations, and outcome measures from the recall and recognition tasks (*N = *85).*

	Achievement Goal Orientation			
	Mastery-approach	Mastery-avoidance	Performance-approach	Performance-avoidance
Total feedback recalled	.16 (p = .15)	.20 (p = .06)	.05 (p = .66)	.06 (p = .58)
Evaluative recall bias[a]	.11 (p = .30)	-.05 (p = .64)	.12 (p = .29)	.15 (p = .18)
Evaluative retrieval style[b]	-.20 (p = .07)	-.13 (p = .24)	-.08 (p = .48)	-.10 (p = .39)
Evaluative recognition bias[c]	-.11 (p = .32)	-.13 (p = .25)	-.01 (p = .96)	.02 (p = .84)

[a] Total evaluative comments recalled, minus total directive comments recalled
[b] Total comments reproduced in an evaluative style, minus total comments reproduced in a directive style
[c] Total evaluative errors in recognition, minus total directive errors in recognition

Table S3. *Item analysis of subjects' free recall responses in Experiments 1-4.*

Evaluative feedback comment	**Recall rate**	**Directive feedback comment**	**Recall rate**
your responses tended to favour breadth at the expense of sufficient depth of detail	**50.0%** (86/172)	*you should aim to be more balanced, to avoid favouring breadth at the expense of sufficient depth of detail*	**39.2%** (65/166)
you sometimes neglected to follow arguments through and instead left them unexplained	**37.8%** (65/172)	*make sure you follow arguments through without leaving them unexplained*	**25.3%** (42/166)
you didn't always demonstrate a sophisticated awareness of the issues you covered	**10.8%** (18/166)	*you should aim to demonstrate a more sophisticated awareness of the issues you cover*	**3.5%** (6/172)
you were not often especially specific about the practical implications of the issues you discussed	**7.8%** (13/166)	*this can be improved by being more specific about the practical implications of the issues you discuss*	**4.1%** (7/172)
your arguments lacked some originality in places	**30.8%** (53/172)	*there is room for your arguments to demonstrate more originality in places*	**18.1%** (30/166)
You did not always try to provoke your reader's thinking, and focused instead on arguments that they would expect	**37.8%** (65/172)	*you could try to provoke your reader's thinking more, by focusing on arguments that they would find unexpected*	**20.5%** (34/166)
your responses were not always presented in a scientific style	**30.1%** (50/166)	*your responses should be presented in a more consistently scientific style*	**26.7%** (46/172)
on occasion your arguments sounded somewhat personal rather than objective	**54.2%** (90/166)	*wherever possible try to make sure that your arguments sound objective rather than personal*	**35.5%** (61/172)
you did not always use evidence to support your arguments	**41.3%** (71/172)	*you need to make stronger use of evidence to support your arguments*	**38.0%** (63/166)
Your work suggested that you did not look back over it to check that you had backed up all of your assertions	**16.3%** (28/172)	*You could do this by looking back over your work and checking that you have backed up all of your assertions*	**6.0%** (10/166)
I did not find all of your arguments wholly persuasive	**23.5%** (39/166)	*you might take additional efforts to ensure that all of your arguments are wholly persuasive.*	**5.2%** (9/172)
there was not always a lot of strength or impact in	**12.7%**	*you could particularly aim to strengthen the impact of*	**12.8%**

your concluding statements	(21/166)	*your concluding statements*	(22/172)
there was not enough critical evaluation of your ideas and arguments	**11.6%** (20/172)	*you should try to include more critical evaluation of your ideas and arguments*	**8.4%** (14/166)
you didn't always think about possible counterarguments to your position and defend against them	**36.0%** (62/172)	*you could try to think more about possible counterarguments to your position and defend against them*	**30.7%** (51/166)
the structuring of your arguments was a little unclear in places	**21.1%** (35/166)	*the structuring of your arguments is something that could be improved in places*	**15.1%** (26/172)
there was not always a clear sense of where your points were leading	**24.7%** (41/166)	*make sure there is always a clear sense of where your points are leading*	**8.1%** (14/172)
sometimes the way you made your points was not concise enough	**21.5%** (37/172)	*sometimes the way you make your points could be more concise*	**22.3%** (37/166)
you sometimes said in multiple sentences what you could potentially have said in just one	**48.8%** (84/172)	*avoid saying in multiple sentences what you could potentially say in just one*	**33.7%** (56/166)
there were a few examples of [grammar and punctuation] errors	**36.1%** (60/166)	*you should ensure to find and remove any [grammar and punctuation] errors*	**38.4%** (66/172)
you did not consistently use commas and semi-colons when appropriate	**62.7%** (104/166)	*you could make sure that you consistently use commas and semi-colons when appropriate*	**53.5%** (92/172)

Feedback scripts used in Experiments 1, 2, 3, and the Mixed feedback condition of Experiment 4. Evaluative comments are highlighted in yellow; directive comments are highlighted in blue (highlighting was not used in the actual study materials). In Experiment 4, the evaluative (yellow) comments from script Version 1 and script Version 2 were combined in the Evaluative-only condition. The directive (blue) comments from script Version 1 and script Version 2 were combined in the Directive-only condition.

Version 1.

Substance

I found your responses on these issues very interesting and thought-provoking, and they showed a good amount of thought and consideration. That said, your responses tended to favour breadth at the expense of sufficient depth of detail. That is, you sometimes neglected to follow arguments through and instead left them unexplained. Furthermore, you should aim to demonstrate a more sophisticated awareness of the issues you cover. For instance, this can be improved by being more specific about the practical implications of the issues you discuss. Finally, I felt that your arguments lacked some originality in places. You did not always try to provoke your reader's thinking, and focused instead on arguments that they would expect. Overall the substance of your responses was strong despite these specific issues.

Style

You demonstrated an engaging and mature writing style, and I had just a few suggestions to make in this regard. Specifically, your responses should be presented in a more consistently scientific style. In particular, wherever possible try to make sure that your arguments sound objective rather than personal. Another issue was that you did not always use evidence to support your arguments. Your work suggested that you did not look back over it to check that you had backed up all of your assertions. As a result, you might take additional efforts to ensure that all of your arguments are wholly persuasive. To this end, you could particularly aim to strengthen the impact of your concluding statements. Finally, in general there was not enough critical evaluation of your ideas and arguments. For example, you didn't always think about possible counterarguments to your position and defend against them. In general the presentation style of your responses was impressive, though, and shows a degree of flair.

Format

This was an area with which your responses were generally strong. I did find, though, that the structuring of your arguments is something that could be improved in places. So, for instance, make sure there is always a clear sense of where your points are leading. Additionally, sometimes the way you made your points was not concise enough. One example was that you sometimes said in multiple sentences what you could potentially have said in just one. Lastly, although your grammar and punctuation were generally very good, you should ensure to find and remove any errors. For example you could make sure that you consistently use commas and semi-colons when appropriate. Overall, you presented your ideas in a way that captured attention and interest.

Version 2.

Substance

I found your responses on these issues very interesting and thought-provoking, and they showed a good amount of thought and consideration. That said, you should aim to be more balanced, to avoid favouring breadth at the expense of sufficient depth of detail. That is, make sure you follow arguments through without leaving them unexplained. Furthermore, you didn't always demonstrate a sophisticated awareness of the issues you covered. For instance, you were not often especially specific about the practical implications of the issues you discussed. Finally, I felt that there is room for your arguments to demonstrate more originality in places. To do this, you could try to provoke your reader's thinking more, by focusing on arguments that they would find unexpected. Overall the substance of your responses was strong despite these specific issues.

Style

You demonstrated an engaging and mature writing style, and I had just a few suggestions to make in this regard. Specifically, your responses were not always presented in a scientific style. In particular, on occasion your arguments sounded somewhat personal rather than objective. Another issue is that you need to make stronger use of evidence to support your arguments. You could do this by looking back over your work and checking that you have backed up all of your assertions. As a result. I did not find all of your arguments wholly persuasive. In particular, there was not always a lot of strength or impact in your concluding statements. Finally, in general you should try to include more critical evaluation of your ideas and arguments. For example, you could try to think more about possible counterarguments to your position and defend against them. In general the presentation style of your responses was impressive, though, and shows a degree of flair.

Format

This was an area with which your responses were generally strong. I did find, though, that the structuring of your arguments was a little unclear in places. So, for instance, there was not always a clear sense of where your points were leading. Additionally, sometimes the way you make your points could be more concise. One suggestion would be to avoid saying in multiple sentences what you could potentially say in just one. Lastly, although your grammar and punctuation were generally very good, there were a few examples of errors. For example you did not consistently use commas and semi-colons when appropriate. Overall, you presented your ideas in a way that captured attention and interest.

Feedback scripts used in Experiment 5

Evaluative-harsher, Version 1.

Substance

I found your responses on these issues very interesting and thought-provoking, and they showed a good amount of thought and consideration. That said, your responses tended to favour breadth and were disappointingly shallow in terms of detail. That is, you neglected to follow arguments through and instead you left them inadequately explained. Furthermore, you should aim to more fully demonstrate how sophisticated your awareness is of the issues you cover. For instance, you should ideally be a little less vague about the practical implications of the issues you discuss. Finally, your arguments were lacking in any smallest sense of originality. You made virtually no effort to provoke your reader's thinking, and focused instead on entirely obvious arguments. Overall the substance of your responses was strong despite these specific issues.

Style

You demonstrated an engaging and mature writing style, and I had just a few suggestions to make in this regard. Specifically, your responses could be presented in a more scientific style. In particular, try to make sure that your arguments sound objective rather than personal. Another issue was that you failed to consistently use even a trace of evidence to support your arguments. Your work clearly showed that you had not even looked back over it, to check that you had backed up all of your weak assertions. As a result, you might try to ensure that all of your arguments are wholly persuasive. To this end, you could aim for a more distinct impact in your concluding statements. Finally, there was insufficient critical evaluation to compensate for the weaknesses in your ideas and arguments. For example, you didn't stop to consider possible counterarguments to your position, and actually defend against them. In general the presentation style of your responses was impressive, though, and shows a degree of flair.

Format

This was an area with which your responses were generally strong. I did find, though, that the structuring of your arguments could be slightly improved next time. So, for instance, ensure to always give a fully clear sense of where your points are leading. Additionally, the way you made your points was rambling, showing very little regard for conciseness. One example was that you sometimes said in multiple sentences what a good writer could have said in one sentence. Lastly, although your grammar and punctuation were generally good, you should ensure to find and remove any notable errors. For example you could make sure to consistently use commas and semi-colons when appropriate. Overall, you presented your ideas in a way that captured attention and interest.

Evaluative-harsher, Version 2.

Substance

I found your responses on these issues very interesting and thought-provoking, and they showed a good amount of thought and consideration. That said, you might try to avoid favouring breadth, and instead take opportunities for greater depth of detail. That is, you could think about making sure you follow arguments through, instead of leaving them unexplained. Furthermore, you didn't demonstrate even a minimally sophisticated awareness of the issues you covered. For instance, you were incredibly vague about the practical implications of the issues you discussed. Finally, your arguments could aim to demonstrate a clearer sense of originality. You could do more to provoke your reader's thinking, by focusing on arguments that they would not expect. Overall the substance of your responses was strong despite these specific issues.

Style

You demonstrated an engaging and mature writing style, and I had just a few suggestions to make in this regard. Specifically, you made little effort to present your responses in a scientific style. In particular, many of your arguments were dreadfully personal rather than objective. Another issue is that you could try to make more consistent use of evidence to support your arguments. You could do this by looking back over your work, and checking that you have tried to back up all of your good suggestions. As a result, I did not find most of your arguments even the slightest bit persuasive. In particular, there was a distinct lack of any impact in your concluding statements. Finally, you could try to include more critical evaluation to strengthen your ideas and arguments. For example, you could try to reflect more on possible counterarguments to your position, and defend against them. In general the presentation style of your responses was impressive, though, and shows a degree of flair.

Format

This was an area with which your responses were generally strong. I did find, though, that the structuring of your arguments was woefully unclear sometimes. So, for instance, you often gave no coherent sense of where your points were leading. Additionally, the way you make your points could be less wordy, placing greater emphasis on conciseness. One suggestion is to avoid saying in multiple sentences what you could potentially say in just one sentence. Lastly, although your grammar and punctuation were generally good, there were a few shameful examples of errors. For example you did not show the capability to use commas and semi-colons when appropriate. Overall, you presented your ideas in a way that captured attention and interest.

Directive-harsher, Version 1.

Substance

I found your responses on these issues very interesting and thought-provoking, and they showed a good amount of thought and consideration. That said, your responses tended to favour breadth and missed some opportunities for greater depth of detail. That is, you didn't always think to follow certain arguments through and instead left them unexplained. Furthermore, you should aim to demonstrate at least a minimally sophisticated awareness of the issues you cover. For instance, try not to be so incredibly vague about the practical implications of the issues you discuss. Finally, your arguments did not tend to demonstrate a clear sense of originality. You could have done more to provoke your reader's thinking, as you focused mainly on arguments that they might expect. Overall the substance of your responses was strong despite these specific issues.

Style

You demonstrated an engaging and mature writing style, and I had just a few suggestions to make in this regard. Specifically, you must make more effort to present your responses in a scientific style. In particular, make sure that your arguments are objective rather than dreadfully personal. Another issue was that you could have used evidence a bit more consistently to support your arguments. Your work suggested that you had not fully looked back over it, to check that you had backed up all of your good suggestions. As a result, you must work to ensure that your arguments are even the slightest bit persuasive. To this end, you should avoid having such a distinct lack of impact in your concluding statements. Finally, there could have been more critical evaluation to strengthen your ideas and arguments. For example, you didn't always reflect on possible counterarguments to your position, and defend against them. In general the presentation style of your responses was impressive, though, and shows a degree of flair.

Format

This was an area with which your responses were generally strong. I did find, though, that the structuring of your arguments could be less woefully unclear next time. So, for instance, do give at least some coherent sense of where your points are leading. Additionally, the way you made your points was sometimes wordy, not always placing enough emphasis on conciseness. One example was that you sometimes said in multiple sentences what you could potentially have said in just one sentence. Lastly, although your grammar and punctuation were generally good, you really must find and remove any shameful errors. For example you should show the capability to use commas and semi-colons when appropriate. Overall, you presented your ideas in a way that captured attention and interest.

Directive-harsher, Version 2.

Substance

I found your responses on these issues very interesting and thought-provoking, and they showed a good amount of thought and consideration. That said, you must avoid favouring breadth whilst being disappointingly shallow in terms of detail. That is, you really must learn to follow arguments through without leaving them inadequately explained. Furthermore, you didn't always fully demonstrate how sophisticated your awareness is of the issues you covered. For instance, you were occasionally a little vague about the practical implications of the issues you discussed. Finally, your arguments need to demonstrate at least some small sense of originality. You must make efforts to provoke your reader's thinking, by focusing on arguments that are not entirely obvious. Overall the substance of your responses was strong despite these specific issues.

Style

You demonstrated an engaging and mature writing style, and I had just a few suggestions to make in this regard. Specifically, your responses were not always presented in a scientific style. In particular, on occasion your arguments sounded somewhat personal rather than objective. Another issue is that you should consistently use at least some trace of evidence to support your arguments. You should do this by looking back over your work, and checking that you have at least backed up all of your weak assertions. As a result, I did not find that all of your arguments were wholly persuasive. In particular, there was sometimes room for more impact in your concluding statements. Finally, you must include sufficient critical evaluation to compensate for the weaknesses in your ideas and arguments. For example, you should stop to consider possible counterarguments to your position, and actually defend against them. In general the presentation style of your responses was impressive, though, and shows a degree of flair.

Format

This was an area with which your responses were generally strong. I did find, though, that the structuring of your arguments was slightly unclear sometimes. So, for instance, you didn't always give a fully clear sense of where your points were leading. Additionally, the way you make your points should be less rambling, showing greater regard for conciseness. One suggestion is to avoid saying in multiple sentences what a good writer could say in one sentence. Lastly, although your grammar and punctuation were generally good, there were a few notable examples of errors. For example you did not consistently use commas and semi-colons when appropriate. Overall, you presented your ideas in a way that captured attention and interest.

Feedback scripts used in Experiment 6

Version 1.

I thought that your work was interesting to read. However, you did not always remember to use capital letters. Next time think about where it might be helpful to add commas. Also, I felt you use the word 'said' too many times in your work. Remember to look back over your work for spelling mistakes. Your handwriting was not always the easiest to read in some places. Always remember to underline the date and title. I felt you could have used paragraphs in your work. I was impressed by the ideas you came up with, however next time make sure your ideas link together. I did not feel you fully understood the task. Next time please complete the task in full. I can tell that you tried really hard and you should be proud of the work you've done.

Version 2.

I thought that your work was interesting to read. However, you need to remember to use capital letters correctly. There were places where it would have been helpful to add commas. Also, you should try to think of different words to use other than 'said'. I noticed there were some spelling mistakes in your writing. You could improve your handwriting so your work is easy to read. You did not underline the date and title. It would be good to use paragraphs in your work. I was impressed by the ideas you came up with, however some of your ideas did not link together. In future, make sure you fully understand the task. You did not finish the task in full. I can tell that you tried really hard and you should be proud of the work you've done.