Weak memory for future-oriented feedback : investigating the roles of attention and improvement focus

Gregory, SEA, Winstone, NE, Ridout, N and Nash, RA

http://dx.doi.org/10.1080/09658211.2019.1709507

Running Head: Past- and future-oriented feedback

Weak memory for future-oriented feedback: Investigating the roles of attention and improvement focus

Samantha E. A. Gregory*, Aston University

Naomi E. Winstone, University of Surrey

Nathan Ridout, Aston University

Robert A. Nash, Aston University

* Corresponding author

Dr Samantha E. A. Gregory
Department of Psychology
Aston University
Birmingham B4 7ET
United Kingdom
s.gregory1@aston.ac.uk

Word count: 16560

Abstract

Recent research showed that people recall past-oriented, evaluative feedback more fully and accurately than future-oriented, directive feedback. Here we investigated whether these memory biases arise from preferential attention toward evaluative feedback during encoding. We also attempted to counter the biases via manipulations intended to focus participants on improvement. Participants received bogus evaluative and directive feedback on their writing. Before reading the feedback, some participants set goals for improvement (experiments 1 and 2), or they wrote about their past or future use of the writing skills, and/or were incentivised to improve (experiment 3); we objectively measured participants' attention during feedback reading using eyetracking. Finally, all participants completed a recall test. We successfully replicated the preferential remembering of evaluative feedback, but found little support for an attentional explanation. Goal-setting reduced participants' tendency to reproduce feedback in an evaluative style, but not their preferential remembering of evaluative feedback. Neither orienting participants toward their past or future use of the writing skills, nor incentivising them to improve, influenced their attention toward or memory for the feedback. These findings advance the search for a mechanism to explain people's weaker memory for future-oriented feedback, demonstrating that attentional and improvement-oriented accounts cannot adequately explain the effect.

Key words: eyetracking; education; feedback; goal-setting; recall

Receiving feedback is a crucial part of the learning process (Hattie & Timperley, 2007; Kluger & DeNisi, 1996), and feedback information can take many forms: in some cases it is *evaluative,* focusing on appraising a learner's past performance; in other cases it is *directive,* focusing on how the learner could improve in future. In a series of recent experiments we demonstrated—contrary to our initial predictions—that people were consistently worse at remembering directive feedback as compared with evaluative feedback (Nash, Winstone, Gregory, & Papps, 2018). In this paper we replicate and extend this discovery of an *evaluative recall bias* in two important ways. First, we use an eyetracking method to look for evidence that an attentional mechanism drives this robust bias. Second, we examine to what extent three interventions—designed to prompt participants to consider improving their performance—would counter the bias.

Memory for feedback

Regardless of the form that feedback takes, if learners are to be able to apply it in future contexts, then it is often important for them first to be able to remember it. This role for memory is especially crucial when we consider that students frequently say they read written feedback only briefly, and only once (Gibbs & Simpson, 2004; Robinson, Pope, & Holyoak, 2013; Winstone, Nash, Parker, & Rowntree, 2017; Winstone, Nash, Rowntree, & Parker, 2017). In assessment contexts where there are objective correct or incorrect answers—for example in a general knowledge or multiple-choice quiz— we know that receiving feedback prior to a second test of the same material can generally enable learners to improve their performance (e.g. Butler, Karpicke, & Roediger, 2008; Butler & Roediger, 2008; Kang, McDermott, & Roediger, 2007; Phye & Sanders, 1994; Shute, 2008). Yet we also know that people easily forget even these straightforward and unambiguous kinds of feedback, which can often lead to them making the same errors for a second time. For example, whilst completing a multiple-choice test, some of the participants in Butler and

Roediger's (2008) study received corrective feedback on each of their responses, either immediately or at the end of the test. One week later, they took a cued recall test of the same study material. Despite having been explicitly informed of the correct answer for every error they had made in the multiple-choice test, in the cued recall test participants nevertheless still gave incorrect answers around half of the time. Further, in many learning contexts, there is no such 'correct' answer, and therefore much of the feedback that learners receive is descriptive, guided by the feedback-giver's subjective appraisals of the merits of the work. The extent to which people can remember detailed, descriptive feedback comments like these, has received almost no empirical scrutiny.

Recently we investigated this issue by testing the extent to which people are capable of reproducing evaluative and directive feedback comments from memory (Nash et al., 2018). In our initial experiments, participants completed a persuasive writing task, and later they received detailed feedback about their performance. The feedback, although presented to participants with the suggestion that it was personalized, was in fact generic, and all participants received the same substantive comments. We presented each critical comment to different participants either in an evaluative style, written in a past-oriented manner (e.g. "you didn't always think about possible counterarguments to your position and defend against them"), or in a directive style, written in a future-oriented manner ("you could try to think more about possible counterarguments to your position and defend against them"). All participants saw a mixture of evaluative and directive comments, with each individual comment seen by half of participants in an evaluative style and by half of participants in a directive style. A short while after reading this feedback, participants completed a surprise recall test, in which they were asked to reproduce as much of the feedback as possible. We measured not only which statements the participants recalled, but also the style in which they

recalled them (i.e. whether they recalled a directive piece of feedback in a correct directive style or in an incorrect evaluative style).

Based on evidence from the feedback and memory literatures, for several reasons our prediction was that participants would recall more directive feedback than evaluative feedback. Firstly, students typically report a preference to receive feedback about future improvement rather than feedback about how they performed in the past (Carless, 2006; Dawson et al., 2019; Winstone, Nash, Rowntree, & Menezes, 2016), and we know that people are typically better able to remember information if they had been interested in finding it out (Fastrich, Kerr, Castel, & Murayama, 2018). Secondly, feedback and remembering are both fundamentally future-oriented processes: they both serve to guide future planning and execution of tasks (Klein, 2013; Klein, Robertson, & Delton, 2010; Kluger & DeNisi, 1996). Some studies, for example, have demonstrated enhanced memory for word lists when participants attempt to associate the to-be-learned words with future plans, rather than with memories of past events (Klein et al., 2010; Klein, Robertson, & Delton, 2011). Finally, key findings from the implementation intentions literature—wherein people are more likely to successfully memorize instructions if they believe they will be asked to implement them later—also led us to predict a memory advantage for feedback comments that provide future-oriented instructions (e.g. Chasteen, Park, & Schwarz, 2001; Goschke & Kuhl, 1993; McDaniel, Howard, & Butler, 2008).

However, in our studies we consistently found the opposite pattern to the one we predicted: participants recalled more evaluative feedback than directive feedback – a pattern that we termed the *evaluative recall bias* (Nash et al., 2018). This unexpected finding is all the more intriguing in light of the fact that when asked, most of Nash et al.'s (2018) participants claimed to prefer receiving directive feedback. This memory bias is compounded by a further effect whereby participants frequently tended to reproduce the feedback

comments from memory in an evaluative style, even when the comments had originally been directive. This latter bias—which we termed an *evaluative retrieval style*—is important because systematic patterns of misremembering can inform us about the kinds of inferences people make when processing information (e.g. Brewer, 1977; Chan & McDermott, 2006). The bias may reflect a tendency for people to infer criticisms about their past performance when receiving advice about how to improve in future. If so, then in practice this bias may result in students feeling that the feedback they received was negative rather than constructive, potentially leading to demotivation (Fong et al., 2016).

At present, despite extensive investigation of several candidate mechanisms (Nash et al., 2018) there is no clear explanation for why the evaluative recall bias occurs. In this paper, we develop two avenues of further investigation.

The role of attention

It is possible that the evaluative recall bias is driven by preferential attention toward evaluative feedback at the point of its encoding, rather than by memory processes per se, and that this preferential attention could be responsible for the effects subsequently seen in remembering. For example, although students often claim to have greater interest in directive feedback than in evaluative feedback, research from the economic psychology literature tells us that people tend to be strongly driven to obtain information about their prior performance, even when doing so is irrational in economic terms (Alós-Ferrer, García-Segarra, & Ritschel, 2018). If this finding applies to processing of feedback then it would lead us to predict that people should habitually pay preferential attention to evaluative feedback comments.

Two findings from Nash et al. (2018) seem to point against an attentional interpretation, but the evidence in both cases is far from conclusive. First, Nash et al. found no significant memory advantage for evaluative feedback over directive when participants were tested using a recognition test instead of a recall test (Experiments 1 and 2). This

finding might suggest that both kinds are encoded similarly in memory. However, Nash et al. only gathered recognition data in two relatively small experiments, and in one of these experiments there was still a moderate (but nonsignificant) recognition advantage in favour of evaluative feedback. Secondly, in a between-subjects design, participants who received only evaluative feedback spent an equivalent total amount of time reading to those participants who received only directive feedback (Experiment 4). This finding might suggest that participants paid equal attention to the feedback regardless of its written style. However, participants completed this experiment within a rather uncontrolled classroom environment that involved considerable distraction, and their reading times were measured only indirectly via their web browsers. This meant that there was likely considerable noise in the reading time data, which make these statistical comparisons difficult to interpret. Furthermore, a preferential attention mechanism could rely on relative rather than absolute attention. That is to say, people might pay less attention to directive feedback only when their attention is drawn by evaluative feedback. If so, then a between-subjects design cannot allow an attentional mechanism to be ruled out. In sum, it would be premature to draw theoretical conclusions about the possible role of attention in the evaluative recall bias based on the data published to date. To draw firmer conclusions, it is essential to use more reliable and direct measures of attention.

One objective and direct method of measuring attention at encoding is to track participants' eye gaze during reading. This approach would enable us to directly compare how long participants spend reading evaluative feedback comments relative to the interleaved directive comments. Whereas it is important to note that reading time does not necessarily equate to attention— for example, participants could be spending their time engaged in mind-wandering (Feng, D'Mello, & Graesser, 2013)—nevertheless eyetracking does provide an important step in testing an attentional account.

The role of improvement focus

Whether or not an attentional account of the evaluative recall bias holds, one might reason that this bias would not occur if greater emphasis were placed upon the importance of performing well. Put differently, the evaluative recall bias may rely on participants giving little consideration to the prospect of improving their skills, and so having little reason to attend to or mentally rehearse future-oriented, directive feedback comments. Indeed, in all of Nash et al.'s (2018) experiments, participants either knew that the feedback they received was not self-relevant (Experiments 3-6), or they were told it was self-relevant, yet given little reason to think that improving their performance would benefit them (Experiments 1-2).

If this lack of improvement focus is important, then one way to counter these memory biases might be to prompt participants to explicitly consider their skill improvement, which should in principle increase the goal-relevance of directive feedback, and could therefore promote their attention toward and memory for this information (e.g. Montagrin, Brosch, & Sander, 2013). When participants in Nash et al. (2018, Experiment 2) were explicitly advised to find out what they could improve, they appeared to spend longer reading the feedback overall, yet the evaluative recall bias was not reduced. However, simply telling people to heed advice about how to improve does not necessarily cause them to *want* to improve, or to perceive any value to this future-oriented information. We might expect that striving for the latter priorities would be a more effective way of undermining the evaluative recall bias.

In the present experiments, we therefore attempted to manipulate participants' focus on self-improvement, by using three different interventions immediately before they received their feedback: (1) requiring participants to set themselves goals for improving on the second writing task; (2) requiring them to reflect on how their writing skills had been valuable in the past, or would be valuable in the future; and (3) offering a performance-related monetary incentive.

Experiment 1

Here we investigated participants' memory for evaluative versus directive feedback, and for the first time we used eyetracking to provide a direct measure of attentional processes during participants' encoding of the feedback. If participants were found to spend relatively more time reading evaluative than directive feedback, then this would suggest that at least part of the evaluative recall bias is attributable to attentional mechanisms.

In addition to measuring visual attention, we also used a goal-setting intervention to test the prediction that the evaluative recall bias and/or retrieval style would be reduced or reversed when participants reflected (prior to receiving feedback) on how to improve their own performance. Goal-setting and action planning are considered especially vital in the context of learning and self-regulation, and have been conceptualized as central skills underpinning learners' ability to use feedback effectively (Latham & Locke, 1991; Locke & Latham, 2002; Winstone, Nash, Parker, et al., 2017). People's immediate goals have been found to direct both their allocation of attentional resources (Dijksterhuis & Aarts, 2010; Moskowitz, 2002; Vogt, De Houwer, Moors, Van Damme, & Crombez, 2010), and the subsequent memorability of goal-relevant material (Montagrin et al., 2013). Further, previous research demonstrates that manipulating participants' goals can influence qualitative aspects of remembering processes (e.g. Ikeda, Castel, & Murayama, 2015; Mangels, Rodriguez, Ochakovskaya, & Guerra-Carrillo, 2017; Murayama & Elliot, 2011), as well as affecting academic outcomes (Morisano, Hirsh, Peterson, Pihl, & Shore, 2010). It is therefore plausible that encouraging participants to set goals, and thus to actively reflect on their own self-improvement, would in turn reduce the evaluative recall bias and retrieval style. Further, we might also expect that goal-setting should lead participants to pay relatively more attention to directive feedback, as compared with participants who do not set goals, due to the increased goal-relevance of this directive information among the former group.

Method

Participants and design.

Based on *a priori* power analysis using G*Power, we estimated that 128 participants would be required to detect a medium-sized between-subjects effect in our study design, assuming power = .80, f = .25, and alpha = .05. A total of 128 students therefore took part in exchange for either £10 or for course credit. One participant was excluded from the final analysis because they failed to follow instructions on the memory task, and they were replaced with a new participant. The final sample comprised 99 females and 29 males, M_{age} = 22 years, SD = 4 years, range = 18–45. In the second session of the experiment, each participant was randomly assigned either to the Control condition (n = 65) or the Goal-setting condition (n = 63), as explained below.

Materials

Feedback scripts. During the experiment each participant received at random one of two versions of a script of standardized feedback, as used previously by Nash et al. (2018). These scripts totalled 418 words (version A) and 411 words (version B) respectively, and were divided into three subsections titled “substance”, “style”, and “format,” with each subsection containing critical feedback comments prefixed and suffixed by points of praise. The praise was not relevant to the present aims, but was intended solely to make the feedback seem less severe and more realistic.

Each feedback script contained 20 critique comments in total, which were presented in the same order to every participant. In both scripts half of the critique comments were written in an *evaluative style*: presented as comments about the essays written by the participant, and were thus focused on past performance. The other half of the critique comments were written in a *directive style*: presented as comments about what the participant could improve in a subsequent assignment, and were thus focused on future performance.

The written style of each individual comment was counterbalanced across the two scripts. We achieved this style manipulation by minimally re-wording each critique comment, thus maintaining the same general meaning using the evaluative and directive styles, while keeping the comments' length and complexity approximately equal. For example, half of participants were told, "You did not always try to provoke your reader's thinking, and focused instead on arguments that they would expect" (i.e., evaluative), whereas the other half were told "You could try to provoke your reader's thinking more, by focusing on arguments that they would find unexpected" (i.e., directive). We presented the critique comments in pairs that alternated between the evaluative and directive style throughout each feedback script.

Zimbardo Time Perspective Inventory (ZTPI: Zimbardo & Boyd, 1999). Directive feedback overtly relates to future learning and evaluative feedback to past performance; therefore memory biases may plausibly relate to individual differences in participants' general time orientation. For exploratory purposes, all participants completed the ZTPI, a frequently used measure of trait differences in people's relative orientation toward the past, present, and future. The ZTPI comprises 56 statements that participants rate on 5-point Likert scales (1 = very untrue; 5 = very true). These 56 items make up five subscales: Past-Negative, Past-Positive, Present-Hedonistic, Present-Fatalistic, and Future (an example item: "I try to live my life as fully as possible, one day at a time"). Published internal reliability estimates (Cronbach's alpha) for the five subscales range from 0.74 to 0.82 (Zimbardo & Boyd, 1999), which confirms that this measure is reliable.

Achievement Goal Questionnaire—Revised (AGQ-R). For exploratory purposes, all participants completed a task-adapted version of the AGQ-R, a measure of trait achievement goals (Elliot & Murayama, 2008). The AGQ-R comprises 12 items that participants rate on 5-point Likert scales (1 = strongly disagree; 5 = strongly agree). The measure contains four

subscales (3 items per scale) that distinguish mastery goal orientations (i.e., developing competence relative to an absolute or intrapersonal standard) from performance goal orientations (i.e., developing competence relative to other people or to a normative standard), and distinguish approach goal orientations (i.e., focusing on success) from avoidance goal orientations (i.e., focusing on preventing failure). Published internal reliability estimates (Cronbach's alpha) for the four subscales range from 0.84 to 0.94 (Elliot & Murayama, 2008), confirming a high degree of reliability. Small adaptations to the wording of the individual AGQ-R items were made, such that the items referred to the writing task used within this study (e.g. "My goal was to avoid performing poorly on the persuasive writing task").

Procedure.

Participants signed up for a study purportedly investigating "Personality and persuasive writing." Each participant individually attended two sessions in the laboratory spaced 2 days apart.

Session 1. We presented all instructions and information on a computer screen using an online tool (Qualtrics). To begin, participants were asked to complete a persuasive writing task consisting of four short essays. Participants chose four topics from a list of ten contentious titles (e.g. "Should students have to pay for their university education?"). Next, one of the four chosen titles was displayed on the computer screen at random, and the participant was asked to type a short persuasive essay on that topic within a time limit of 5 minutes. A countdown timer at the top of the page indicated how much time was remaining. After 5 minutes, the page automatically changed, and a new essay title appeared from those the participant initially chose. This process was repeated for all four essay titles, with a total duration of 20 minutes. Participants then completed the ZTPI, before being reminded that

their persuasive essays would be examined prior to the second session, and that during that session they would receive detailed feedback on their essays. Participants were told, falsely, that this feedback would be intended to help them to complete a second writing task in session 2, and they were asked to rate the extent to which they agreed that "I am looking forwards to receiving my feedback when I return for session 2" (1 = strongly disagree; 5 = strongly agree).

Session 2. On returning 2 days later, participants were informed that they would be given their feedback after completing a short task. At this point, the computer software pseudo-randomly assigned the participant to either the Goal-setting condition or the Control condition. Following the same principles as Avery and Smillie's, (2013) mastery-goal induction, Goal-setting participants were informed via text on the computer screen that the feedback they would shortly receive was designed to help them develop their skills, and that they should use the feedback to understand how to do so. Before seeing the feedback, they were asked to think of and type three goals that they could set themselves for improving in the upcoming second writing task. The computer screen presented three text boxes, and participants were required to type one goal in each box, without a time limit. Control participants, in contrast, were told nothing specific about the purpose of the feedback, except that they would receive it shortly. To equate approximately for the time taken by Goal-setting participants to complete the goals task, Control participants were asked to think of three interesting things that they might see on their journey back from the experiment, and to type short descriptions of these things into three text boxes on their computer screen. In short, all participants were asked to think about the near future, with only those in the Goal-setting condition asked to reflect specifically on their skills and goals in this writing context.

To receive the feedback, participants next moved to a separate computer, so that their eye movements and fixations could be measured. We used an Eyelink 1000 (SR Research,

Ontario, Canada), which is a video-based (infrared) eye-tracker, to record participants' eye position with accuracy of approximately 0.5° and a sampling rate of 2000 Hz. Participants placed their head on a chinrest in front of a LCD 22-inch monitor (resolution 1920 x 1080 pixels), at a viewing distance of 80 cm. A standard 9-dot calibration was first performed to ensure the correct recording of eye movements with maximum recording error of 1°; no further drift check was included. Once this calibration was complete, participants were then shown the feedback script that they had been randomly assigned, presented using Experiment Builder software (V1.10.1630). The feedback script was presented on three separate screen-pages, to ensure that the visual resolution of the text was sufficient to permit reliable eyetracking. Participants could move through the three pages at their own pace using arrows on the keyboard, and were permitted to go back and review previously viewed sections if they wished, before exiting the programme using the right arrow key after reading the final page.

After reading the feedback, participants completed the adapted AGQ—R questionnaire (Elliot & Murayama, 2008)before completing a 5-minute distracter task of logic puzzles. Participants were subsequently presented with a surprise memory test, in which they were given up to 10 mins to type as much of the feedback as they could remember; they were not permitted to move on until at least 5 mins had passed. The instructions explained that we did not require the feedback to be recalled verbatim, but that participants should try to convey the meaning of what was written.

Finally, participants were asked to rate the feedback overall in terms of its fairness (1 = very unfair; 5 = very fair) and helpfulness (1 = very unhelpful; 5 = very helpful), and then asked to estimate what percentage grade they might have received in the persuasive writing task, and what grade they might be able to achieve next time in light of the feedback. They were asked to provide any comments they had about the feedback, to guess the aim of the

study, and to choose which type of feedback they preferred to receive (feedback that tells me how I have performed; feedback that tells me what I should do next time; neither/ no preference).

Data processing.

Adherence to task instructions. We checked participants' responses to the intervening task (i.e., either setting improvement goals, or describing things they might see on their way home) to ensure they had followed instructions. Four Control participants misunderstood the instructions and instead wrote about what they might see during the remainder of the experiment; further, one wrote about what they had seen on their way to the experiment, and one wrote responses unrelated to the task. However, because this control task served primarily only as a time-filler, all of these participants were included in the final analysis. More importantly, all Goal-setting participants adhered to the task instructions by describing three goals for improvement; all participants were included in the final analysis.

Memory coding. One researcher coded the free recall data, blind to which script each participant had seen and to which experimental condition they were assigned. Specifically, the researcher coded which of the 20 feedback comments each participant had recalled, and whether each comment was recalled in an evaluative style or a directive style. Note that as per the instructions given to participants, these free recall data were recalled at the gist level such that, for example, a participant who wrote 'Be concise' would be coded as having successfully retrieved the feedback comment 'The way you make your points could be more concise'. Participants occasionally recalled certain pieces of feedback in both an evaluative and in a directive style; in these cases, the feedback comment was coded twice (i.e. once in

the directive style and once in the evaluative style)[1]. Any praise that participants recalled was ignored. After subsequently revealing which feedback script each participant actually saw, this coding approach therefore enabled us to establish four distinct measurements for each participant: (1) the number of evaluative feedback comments recalled in the correct, evaluative style; (2) the number of directive comments recalled in the correct, directive style; (3) the number of evaluative comments recalled in the incorrect, directive style; and (4) the number of directive comments recalled in the incorrect, evaluative style. To ensure the reliability of the coding, an independent coder also coded 20% of these data (also blind to the script and experimental condition). The agreement rates were high for all four of these interval-level measurements (all $r > .92$), and so all analyses of recall were based on the first coder's data.

Eyetracking. Interest areas were defined for each individual word in the critique comments of the feedback scripts, and each of these interest areas was categorized either as evaluative or directive, according to the written style of the full sentence within which it featured. For each interest area we recorded the total time (ms) spent looking at each word (hereafter, *dwell time*), and the number of times each word was fixated upon (hereafter, *run count)*. For run count a value of 1 implies that the word was read once and not returned to – note that this is different from 'fixation count' in that run count only measures how many times a word was looked at, not how many individual fixations were involved in reading

[1] Here and in both subsequent experiments, these double-coded statements were in the small minority (Experiment 1, 4.1% of the total number of statements recalled were double coded; in Experiment 2 this was again 4.1%, and in Experiment 3 it was 1.5%). None of the statistical findings reported in this paper were changed if rather than coding these statements twice, we instead excluded them from coding entirely.

individual words. For analysis, we averaged each of these variables separately for all evaluative and for all directive feedback comments, such that for each participant we calculated a single mean dwell time and a single mean run count for all evaluative comments combined, and the same for all directive comments combined. These mean values therefore equate for the slightly unequal number of words used for each feedback type, as well as the slightly unequal length of the feedback scripts. These mean dwell time and mean run count values were used in the analyses below. Dwell time and run count values are highly correlated, however we used both because they provide qualitatively different information about participants' reading patterns.

We checked the eyetracking data from each participant prior to inferential analysis, to ensure that the scan path corresponded with the predefined interest areas. This is often a requirement in eyetracking research due to the imperfect nature of the eyetracking software (Cohen, 2013). When reading English text, eye-gaze scan paths follow a clear pattern left to right and then a large leap down and across to the left to start the next sentence; therefore when the tracking is imperfect, it is possible to tell where the appropriate fixation location for the read words and sentence should be (Cohen, 2013). Fixations were therefore corrected manually in SR Research's Data Viewer software, using the drift correct function. Due to the short study duration and the large variation in the accurate recording of the scan path even within individual participants, this drift correction was conducted manually instead of using an automated process (Cohen, 2013). Further, due to the size of the lettering used and the narrow interest areas, some participants' fixations fell above or below the appropriate interest area for the word that they were fixating; in such cases we adjusted the interest areas in Data Viewer to properly accommodate their eye gaze, which led to a small, systematic, shift up or down of all interest areas for the participant. That is to say, the same correction was applied

to the entire feedback script for each individual participant, where necessary, rather than different corrections being applied to each individual feedback comment.

Results

Analytic approach. In all three experiments described here, main results are reported using standard null hypothesis significance testing, and also with Bayesian analysis conducted using default priors in JASP (JASP Team, 2018). Bayesian analysis allows us to test the relative likelihood of the alternative hypothesis being supported versus the null hypothesis, given the data obtained. With complex study designs such as ours, a large number of candidate models are tested (i.e., in JASP, the 2x2x2 design here results in 18 candidate models). Therefore, to consider all of the evidence in support of specific main effects and interaction effects, we take a Bayesian model averaging approach for all omnibus tests. whereby all of these candidate models that include a specific effect are averaged in JASP using the 'effects' output option (across matched models; see Wagenmakers et al., 2018 for further details of this process).

The Bayes Factors obtained from these averaged models can be interpreted as signifying the degree of evidence ($BF_{inclusion}$) in support of those models that include each specific main effect or interaction effect, relative to equivalent models that exclude that effect. When $BF_{inclusion}$ for an effect is equal to 1, the data lend equal support to the existence of that effect (H1) and to there being a null effect (H0). When $BF_{inclusion}$ is between 1 and 3, this is considered anecdotal support for H1; $BF_{inclusion}$ between 3 and 10 signify moderate evidence, $BF_{inclusion}$ between 10 and 30 signifies strong evidence, and $BF_{inclusion}$ above 30 signifies very strong evidence for H1. In contrast, anecdotal support for H0 is indicated when $BF_{inclusion}$ is between 0.33 and 1; $BF_{inclusion}$ between 0.1 and 0.33 signify moderate evidence, $BF_{inclusion}$ between 0.03 and 0.1 signifies strong evidence, and $BF_{inclusion}$ below 0.03 signifies

very strong evidence for H0. For follow-up *t*-tests, BF_{10} is the relevant outcome statistic instead of $BF_{inclusion}$; the same interpretations of these Bayes Factor values apply.

Participants' appraisals of the feedback. Most of our participants appeared unsuspicious about the suggestion that the feedback was personalised. In fact, only two spontaneously mentioned that they thought the feedback was generic. Participants rated the feedback highly in both fairness (M = 4.12 out of 5, SD = 0.81), and helpfulness (M = 4.15, SD = 0.68). They also believed that engaging with the feedback would allow them to do better in the future, so that whereas the average estimated grade for the initial writing was 52.51% (SD = 9.86) participants believed that this could increase to 66.12% (SD = 9.09) in the anticipated second writing task. None of the participants correctly guessed the experiment's aims, nor mentioned the differences in feedback style when prompted to guess. Interestingly, 76% of participants told us at the end of the experiment that in general they preferred to receive directive feedback, whereas only 21% said they preferred to receive evaluative feedback (the remaining 3% expressed no preference either way).

Analysis of recall data. To test our main predictions about memory for the feedback, we calculated a 2 (between subjects: condition: Goal-setting vs. Control) x 2 (within subjects: feedback type: evaluative vs. directive) x 2 (within subjects: retrieval style accuracy: correct vs. incorrect) mixed-factorial ANOVA on the number of feedback comments recalled. This analysis will confirm whether we replicated our previous findings of an *evaluative recall bias* and an *evaluative retrieval style* (Nash et al., 2018). Specifically, evidence of an evaluative recall bias would be revealed via a significant main effect of feedback type, whereby participants recalled more evaluative feedback comments than directive feedback comments. Furthermore, evidence of an evaluative retrieval style would be revealed via a significant interaction between feedback type and retrieval style accuracy, whereby evaluative comments

are reproduced in the correct (evaluative) style proportionately more often than directive comments are reproduced in the correct (directive) style.

Figure 1 about here.

Crucially, as Figure 1 shows, the analysis replicated the key patterns of findings demonstrated previously by Nash et al. (2018). That is to say, there was a significant evaluative recall bias, such that participants recalled more of the evaluative feedback (M = 4.59, SD = 1.67) than of the directive feedback (M = 3.27, SD = 1.55), $F(1, 126) = 50.00$, $p < .001$, $\eta^2_p = .28$ ($BF_{inclusion} > 30$). There was also a main effect of retrieval style accuracy, whereby participants recalled more feedback comments in the correct, original style (M = 4.98, SD = 1.94) than in the incorrect, opposite style (M = 2.88, SD = 1.55), $F(1, 126) = 90.39$, $p < .001$, $\eta^2_p = .42$ ($BF_{inclusion} > 30$). We also found a significant interaction between feedback type and retrieval style accuracy, thus signifying an evaluative retrieval style, $F(1, 126) = 41.69$, $p < .001$, $\eta^2_p = .25$ ($BF_{inclusion} > 30$). Follow-up paired t-tests showed that when participants recalled evaluative comments, they reproduced them in an evaluative style significantly more often (76% of the time) than in a directive style (24% of the time), $t(127) = 9.89$, $p < .001$, $d = 0.87$ ($BF_{10} > 30$). In contrast, when participants recalled directive comments, they reproduced these in the correct, directive style no more often (45% of the time) than in the incorrect, evaluative style (55% of the time), $t(127) = -1.29$, $p = .20$, $d = -0.11$ ($BF_{10} = 0.22$).

Looking next to the effects of goal-setting, we found that participants who completed the Goal-setting intervention did not significantly differ from those in the Control condition in terms of how much feedback they recalled overall, $F(1, 126) = 0.32$, $p = .57$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.11$). There was also no significant interaction between retrieval style accuracy and goal-setting condition, which tells us that both groups were similarly able to recall the

feedback comments in their original style, $F(1, 126) = 1.97$, $p = .16$, $\eta^2_p = .02$ ($BF_{inclusion} = 0.24$). However, of greater relevance to our main hypotheses, we also found that goal-setting condition did not interact significantly with feedback type, $F(1, 126) = 0.49$, $p = .49$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.30$), which suggests that the goal-setting manipulation had minimal effect on the evaluative recall bias.

There was, however, a significant three-way interaction between condition, feedback type, and retrieval style accuracy, signifying that the goal-setting intervention had a statistically significant effect upon participants' retrieval styles, $F(1, 126) = 6.13$, $p = .01$, $\eta^2_p = .05$ ($BF_{inclusion} > 30$). As Figure 1 shows, and in line with our predictions, participants in the Goal-setting condition demonstrated a smaller evaluative retrieval style, than did those in the Control condition. When analysing the data from the Control condition alone, we found a large significant interaction between feedback type and retrieval style recall accuracy, $F(1, 64) = 55.13$, $p < .001$, $\eta^2_p = .46$ ($BF_{inclusion} > 30$). Of the evaluative feedback recalled in this condition, participants reproduced 84% in an evaluative style, but of the directive feedback recalled, they reproduced just 40% in a directive style. When analysing data from the Goal-setting condition alone, we again found a statistically significant interaction between feedback type and retrieval style recall accuracy, $F(1, 62) = 6.13$, $p = .02$, $\eta^2_p = .09$ ($BF_{inclusion} > 30$), but the effect size for the latter was notably smaller than in the Control condition. This means that participants in the Goal-setting condition still demonstrated an overall tendency to reproduce feedback comments in an evaluative style, but that this tendency was less pronounced than in the Control condition. Of the evaluative feedback recalled in this Goal-setting condition, participants reproduced 68% in an evaluative style, but of the directive feedback recalled, they reproduced 51% in a directive style. In short, setting goals for improvement seemed to have little effect on which or how much feedback the participants

remembered, but it did influence how they remembered this feedback, with a significant but incomplete reduction in the evaluative retrieval style

Analysis of attention data. There were 115 participants with valid eyetracking data (60 Control; 55 Goal-setting). We analysed each of the two dependent variables (dwell time and run count) using 2 (between subjects: condition) x 2 (within subjects: feedback type) mixed-factor ANOVAs. We look first to the average dwell time data, which tell us how long on average participants spent reading each word of feedback in evaluative vs. directive comments. As Figure 2a shows, crucially we found no significant main effect of feedback type in these data, $F(1, 113) = 2.46$, $p = .12$, $\eta^2_p = .02$ ($BF_{inclusion} = 0.43$), such that participants spent approximately equal amounts of time reading directive feedback comments ($M = 244$ ms/word, $SD = 93$) and evaluative comments ($M = 239$ ms/word, $SD = 86$). There was also no significant main effect of condition, $F(1, 113) = 1.02$, $p = .32$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.64$), nor a significant interaction between feedback type and condition, $F(1, 113) = 0.49$, $p = .49$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.24$). In short, these results suggest that the goal-setting intervention did not notably affect participants' attention to the feedback.

We next looked to the run count data, which tell us how many times participants read each word of feedback on average. As shown in Figure 2b, we found no significant main effect of feedback type, $F(1, 113) = 0.48$, $p = .49$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.18$), which suggests that participants read the words of evaluative comments ($M = 0.98$, $SD = 0.32$) and of directive comments ($M = 0.99$, $SD = 0.37$) an approximately equal number of times. There was no significant main effect of condition, $F(1, 113) = 0.37$, $p = .54$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.55$), nor a significant interaction between feedback type and condition, $F(1, 113) = 0.00$, $p = .96$, $\eta^2_p < .001$ ($BF_{inclusion} = 0.21$). Again, these data show that the goal-setting intervention has no discernible effect on participants' attention toward the feedback.

Exploratory analyses. We examined group differences in participants' ratings of the feedback's fairness and helpfulness, their grading of their actual and expected performance, and their AGQ-R scores. We also analysed the extent to which individual participants' ZTPI and AGQ-R scores, their ratings of their interest in receiving the feedback, and their preference for evaluative vs. directive feedback, were associated with their (1) evaluative recall bias, (2) evaluative retrieval style, (3) selective attention toward evaluative feedback in terms of dwell time and run count. The results of these analyses are reported in supplemental materials, but there were few effects of note.

Figure 2 about here

Discussion

Our data replicate the evaluative recall bias for written feedback as reported in our previous work (Nash et al., 2018). They also suggest that this bias is not easily attributable to participants paying differential attention to evaluative versus directive feedback, nor was the bias reduced by a goal-setting intervention. The goal-setting intervention did, however, attenuate a secondary bias – the evaluative retrieval style. Specifically, after setting goals for improvement, participants were more likely to reproduce their feedback in a directive style than were those in the control group (whilst nevertheless still demonstrating a small and statistically significant evaluative retrieval style overall).

Reflecting on how goal-setting affected participants' retrieval styles, it is possible that the goal-setting task prompted participants to construe the feedback comments more as constructive advice (i.e., as directive), rather than as criticisms of their performance (i.e., as evaluative). It is also possible, though, that writing goals simply primed participants to write in a future-oriented style when subsequently completing the recall task. Therefore, in Experiment 2 we attempted to replicate the findings of Experiment 1, and to tease apart these

two competing explanations of the effect of goal-setting on participants' retrieval style. We did so by adding a third condition to our design in which prior to receiving their feedback, participants wrote advice for a hypothetical person who was going on vacation. This third task, in principle, could prime a directive writing style, but should not encourage participants to construe the feedback comments differently than would control participants.

Experiment 2

Method

To enhance the robustness and replicability of our findings, we pre-registered our method and analysis plan for Experiment 2, which can be found at http://aspredicted.org/blind.php?x=v27kz7

Participants and design.

Based on *a priori* power analysis using parameters estimated from Experiment 1, we estimated that 168 participants would be required to detect the 3 x 2 x 2 within-between interaction effect in our study design, assuming power = .80 and alpha = .05. A total of 168 students therefore took part in exchange for £10 or for course credit. Following our pre-registered protocol, one participant was excluded from analyses due to late completion of Session 2, and was replaced with a new participant. The final sample comprised 139 females and 29 males, M_{age} = 20 yrs, *SD* = 3 yrs, range = 18 – 45. In the second session of the experiment, each participant (*n* = 56 per condition) was randomly assigned either to the Control condition, the Advice-giving condition, or to the Goal-setting condition, as explained below.

Materials

Feedback scripts. As in Experiment 1, participants were presented at random with a version of our feedback scripts. Here, to further establish the replicability of the effects across

different materials, we developed a new feedback script (script 2), which we presented to half of the participants, while the script we had used in Experiment 1 (script 1) was presented to the other half. We prepared script 2 according to the same specifications of script 1 and so there were again two versions of script 2, which can be found in the online supplemental materials—both contained 20 critique comments and totalled 415 words (version A) and 407 words (version B).

Adapted Temporal Focus Scale (TFS). In this experiment, rather than using the lengthy ZTPI as in Experiment 1, we instead used a much shorter time perspective inventory that focuses on state, rather than trait time perspective; specifically, we adapted the TFS (Shipp, Edwards, & Lambert, 2009). The original TFS comprises 12 statements related to temporal focus, each of which is rated on a 7-point Likert scale (1 = Never; 7 = Constantly). The statements ask to what extent individuals attend towards the past, present or future. We adapted this scale to focus only on how they thought about the past or future (i.e., excluding the present) during the attention-based filler task (see below), and so we used only eight items: *I thought about things from my past (past); I thought about what my future has in store (future); I focused on my future (future); I replayed memories of the past in my mind (past); I imagined what tomorrow will bring for me (future); I reflected on what has happened in my life (past); I thought back to my earlier days (past); I thought about times to come (future).*

Achievement Goal Questionnaire—Revised (AGQ-R). Again, all participants completed our adapted version of the AGQ-R (Elliot & Murayama, 2008).

Filler task. Here, we used an alternative filler task to the one used in Experiment 1. This was a repetitive attention task, where participants had to respond using the left and right arrow keys to an asterisk that appeared to the left or the right of a central line. This task was chosen as it is an example of a simplistic, undemanding task, which are known to evoke

mind-wandering (e.g. Baird et al., 2012; Smallwood, Nind, & O'Connor, 2009). Using the TFS we could then measure to what extent any mind-wandering that occurred was directed more towards the past or the future.

Procedure.

The procedure matched that of Experiment 1, except for a few small details, as follows.

Session 1. We removed the ZTPI questionnaire from Session 1. We also removed the final question in which participants reported whether they were looking forwards to receiving the feedback. All other aspects were identical to Experiment 1.

Session 2. Participants returned 2 days later for the second session, in which they were presented with a short task prior to seeing their feedback. The computer software pseudo-randomly assigned the participants to one of three conditions: either the Goal-setting condition or the Control condition, as per Experiment 1, or a third, Advice-giving condition. In the Advice-giving condition, participants were asked to think of and type three pieces of advice that they could give to someone who is going on vacation, but, as with the Control condition, they were given no information about the purpose of the feedback they were to receive shortly.

As in Experiment 1, participants next saw their feedback, which was presented on a separate computer to allow the participants' eye movements and fixations to be recorded with an eyetracker. Immediately after they finished reading the feedback, participants completed the mind-wandering filler task, which lasted approximately 5 minutes, before completing the AGQ—R (Elliot & Murayama, 2008). Participants subsequently completed the surprise recall test, followed by the adapted TFS, which assessed how often during the filler task they had

thought about the past or the future. The study finished with participants answering the same feedback rating questions as were used in Experiment 1.

Data processing.

Adherence to task instructions. Prior to receiving the feedback, two Control participants mistakenly wrote what they might see during the experiment, further, one wrote about what they saw on their way to the experiment. However, all other Control participants completed the control task properly. Likewise, all participants in the Advice condition wrote appropriate advice, and all those in the Goal-setting condition set relevant goals. All participants were included in the analysis.

Memory coding. Both the memory data and the eye-tracking data were processed in the same way as in Experiment 1, with one researcher coding the free recall responses blind to condition. This time the coder was not blind to whether the participants had seen script 1 or script 2; however, they remained blind to which version of the script had been seen, and thus to the style in which each individual comment had been presented. Again, an independent coder also coded 20% of the data, and the agreement rates for all four measurements exceeded $r = .86$, therefore all analyses of recall were based on the first coder's data.

Results

Participants' appraisals of the feedback. Again, most participants appeared to believe that the feedback was personalised, with only seven of the 168 participants reporting that they suspected the feedback was generic. Participants rated the feedback highly in both fairness ($M = 4.05$ out of 5, $SD = 0.89$), and helpfulness ($M = 4.15$, $SD = 0.71$). They also believed that engaging with the feedback would allow them to do better in the future, as the estimated grade for the initial writing was 48.26% ($SD = 13.57$) whereas participants believed this could rise to 65.24% ($SD = 12.79$) if they completed the task again. No participant

correctly guessed the study's aims when prompted, nor noted the differences in feedback style. Consistent with Experiment 1, 73% of participants reported at the end of the experiment that, in general, they preferred to receive directive feedback, whereas only 21% said they preferred to receive evaluative feedback (5% expressed no preference either way).

Analysis of recall data. We conducted a 3 (between subjects: condition: Goal-setting vs. Advice-giving vs. Control) x 2 (within subjects: feedback type: evaluative vs. directive) x 2 (within subjects: retrieval style accuracy: correct vs. incorrect) mixed-factor ANOVA on the number of feedback comments recalled. As Figure 3 shows, this analysis replicated the key patterns of findings demonstrated in Experiment 1, and previously by Nash et al. (2018). Specifically, participants recalled significantly more of the evaluative comments (M = 4.15, SD= 1.70) than of the directive comments (M = 3.08, SD = 1.63), $F(1,165) = 28.84$, $p < .001$, $\eta^2_p = .15$ ($BF_{inclusion} > 30$). There was also a main effect of retrieval style accuracy, whereby participants recalled more feedback comments in the correct style (M = 4.51, SD = 1.67) than in the incorrect style (M = 2.73, SD = 1.55), $F(1, 165) = 89.70$, $p < .001$, $\eta^2_p = .35$ ($BF_{inclusion} > 30$). Further, we found a significant interaction between feedback style and retrieval style accuracy, signifying an overall evaluative retrieval style, $F(1, 165) = 41.14$, $p < .001$, $\eta^2_p = .20$ ($BF_{inclusion} > 30$). Planned follow-up paired t-tests showed that when participants recalled evaluative comments, they reproduced them in an evaluative style significantly more often (75% of the time) than in a directive style (25% of the time), $t(167) = 9.42$, $p < .001$, $d = 0.73$ ($BF_{10} > 30$). In contrast, when participants recalled directive comments, they reproduced these in the correct, directive style (45% of the time) no more often than in the incorrect, evaluative style (55% of the time), $t(167) = -1.59$, $p = .11$, $d = -0.12$ ($BF_{10} = 0.29$).

Figure 3 about here

Looking next to the effects of condition, we found that participants in all three groups recalled a similar number of feedback comments overall, $F(2,165) = 0.54$, $p = .59$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.03$). There was also no significant interaction between condition and feedback type, $F(2, 165) = 0.42$, $p = .66$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.04$), nor a significant interaction between condition and retrieval style accuracy, $F(2, 165) = 1.08$, $p = .34$, $\eta^2_p = .01$ ($BF_{inclusion} = 0.06$). The predicted three-way interaction between condition, feedback type, and retrieval style accuracy was also not statistically significant, $F(2, 165) = 2.84$, $p = .06$, $\eta^2_p = .03$, although the Bayesian analysis suggested strong evidence for this three-way effect ($BF_{inclusion} = 23.78$).

Analysis of attention data. There were 161 participants with valid eye-tracking data (52 Control; 53 Advice-giving; 56 Goal-setting). We separately analysed the two dependent variables (dwell time and run count) using two 3 (between subjects: condition) x 2 (within subjects: feedback type) mixed-factor ANOVAs. Looking first at dwell time, as illustrated in Figure 4a, we again found no significant main effect of feedback type, $F(1, 158) = 0.02$, $p = .90$, $\eta^2_p < .001$ ($BF_{inclusion} = 0.13$), meaning that participants spent approximately equal time attending to evaluative comments ($M = 239$ ms/word, $SD = 90$ ms) and to directive comments ($M = 239$ ms/word, $SD = 88$ ms). There was also no significant main effect of condition, $F(2, 158) = 0.27$, $p = .77$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.32$), nor a significant interaction between condition and feedback type, $F(2, 158) = 1.14$, $p = .32$, $\eta^2_p = .01$ ($BF_{inclusion} = 0.16$). These results suggest that the goal-setting and advice-giving interventions did not notably affect the attention that participants paid to the feedback.

We next examined the run count data, finding no significant main effect of feedback type, $F(1, 158) = 0.39$, $p = .54$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.15$). Specifically, as Figure 4b shows, participants read evaluative comments ($M = 0.98$, $SD = 0.33$) and directive comments ($M = 0.97$, $SD = 0.32$) an approximately equal number of times. There was also no significant main

effect of condition, $F(2, 158) = 0.71$, $p = .50$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.40$) nor a significant interaction between condition and feedback type, $F(2, 158) = 0.92$, $p = .40$, $\eta^2_p = .01$ ($BF_{inclusion} = 0.11$).

Exploratory analyses. When looking at the recall data, our findings were the same regardless of whether participants had read feedback script 1, or our new script 2 (adding script as an additional between-subject variable, all *p*-values for effects involving script > .27, all $BF_{inclusion} \leq 0.24$). This outcome demonstrates further evidence for the replicability of the evaluative recall bias and evaluative retrieval style across different materials.

Figure 4 about here

We examined group differences in participants' ratings of the feedback's fairness and helpfulness, their grading of their actual and future performance, their AGQ-R scores, and the tendency for their spontaneous thoughts after reading the feedback to be future-oriented (calculated as TFS Future score minus TFS Past score). We also examined the extent to which individual participants' TFS difference scores, AGQ-R scores, and their preference for evaluative vs. directive feedback were associated with their (1) evaluative recall bias, (2) evaluative retrieval style, and (3) selective attention toward evaluative feedback in terms of dwell time and run count. The results of these analyses are reported in supplemental materials, but again there were few effects of note.

Experiment 3

In Experiments 1 and 2 we found that the goal-setting manipulation had little influence on which feedback the participants went on to recall, though it did appear to influence the way in which the feedback was recalled (we return to look more closely at this effect later). However, setting explicit goals is only one way of leading people to explicitly consider their skill improvement. In Experiment 3 we used two different but related

manipulations that we believed could increase participants' focus on improvement. Firstly, we asked participants—before they received the feedback—to reflect on either (a) past occasions when they had to use persuasive writing skills, (b) future occasions when they will have to use these skills, or (c) neither. According to research stemming from Future Time Perspective Theory, students' mental orientation toward the future can reliably predict their goal motivation, goal-setting, and general academic engagement (Horstmanshof & Zimitat, 2007; Husman & Lens, 1999; Lasane & Jones, 1999; Nuttin & Lens, 1985). According to this line of 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). We therefore predicted that participants who focused on their future use of their writing skills would be less likely to exhibit an evaluative recall bias. Further, we also predicted that those who focused on their future use of writing skills would pay relatively greater attention to directive feedback, as measured by eyetracking.

Secondly, in many real learning scenarios students will be motivated to improve because they want to obtain a specific grade. In Nash et al.'s (2018) work and in our Experiments 1 and 2, there was no clear evidence as to whether participants were truly motivated to improve in the fictional second writing task. As such, it is important to ask whether the evaluative recall bias and evaluative retrieval style would replicate in a scenario where motivation is known to be high. We might predict that the memory effects seen in our previous studies would disappear if task improvement were explicitly incentivised. For instance, it may be that when participants have a clear reason to wish to improve in a subsequent task, they would pay relatively more attention to, and remember, directive feedback comments. Therefore, in Experiment 3 we also added an incentive manipulation whereby half of participants were told that they could receive a monetary bonus based on how much their persuasive writing improved in a (fictional) second writing task.

Method

We pre-registered our method and analysis plan for Experiment 3, which can be found at http://aspredicted.org/blind.php?x=c9z9au

Participants and design.

Based on *a priori* power analysis, we aimed to recruit 168 participants, but we slightly over-sampled and ultimately recruited 174 students, one of whom was excluded from analyses following our pre-registered protocol, due to not attempting to recall the feedback. The final sample of 173 participants took part in exchange for £10 or course credits, and comprised 133 females and 40 males, M_{age} = 20 yrs, *SD* = 3 yrs, range = 18 - 44. In the second session of the experiment, each participant was randomly assigned either to the Incentive condition or to the No Incentive condition. Orthogonally to this manipulation, each participant was randomly assigned to one of the three temporal focus conditions, thus resulting in six between subjects conditions: Control (*n* = 59; of which *n* = 30 in the Incentive condition, *n* = 29 in the No Incentive condition), Past (*n* = 57; of which *n* = 29 in the Incentive condition, *n* = 28 in the No Incentive condition) or Future (*n* = 57; of which *n* = 28 in the Incentive condition, *n* = 29 in the No Incentive condition).

Materials

All materials matched those used in Experiment 2. Participants were presented at random with one of the two versions of our two different feedback scripts, and all participants completed the adapted TFS and our adapted version of the AGQ-R (Elliot & Murayama, 2008). The filler task also matched that used in Experiment 2.

Procedure.

The procedure matched that of Experiment 2, except for a few small details, as follows.

Session 1. Participants signed up with the knowledge that they may receive a bonus of up to £5 plus the potential to win an additional £30 Amazon voucher. All other aspects were identical to Experiment 2.

Session 2. Unlike the previous experiments, we allowed participants to return for session 2 after a delay of anything between 1 day and 1 week.[2] Most participants returned within 3 days of completing session 1 (96%; $M = 1.69$). At the beginning of session 2, participants were presented with a short task prior to seeing their feedback. The computer software pseudo-randomly assigned the participants to one of the three temporal focus conditions. In the Past condition participants were instructed to '*take a few moments to think about the academic skills involved in the persuasive writing task. Then, using the boxes below, briefly describe three occasions in the past when you needed to use these skills'*. In the Future condition participants were instructed to: '*take a few moments to think about the academic skills involved in the persuasive writing task. Then, using the boxes below, briefly describe three occasions in the future when you will need to use these skills'*. In the control condition the participants were simply asked to: '*take a few moments to think about the academic skills involved in the persuasive writing task.'* They were given no information about the purpose of the feedback they were to receive shortly.

Participants were then pseudo-randomly assigned to one of two incentivisation conditions. In the Incentive condition participants were told: *In addition to your [payment/credits] for completing this survey, all participants who improve their writing in the second writing task will also receive a bonus of up to £5. The size of your bonus will depend on how fully you implement the feedback, as judged by our analysis of your writing.*

[2] The number of days between sessions was not meaningfully correlated (Spearman's rho) with either the evaluative recall bias ($r = -.07$, $p = .33$), or the evaluative retrieval style ($r = -.04$, $p = .63$).

Additionally, the six participants who implement their feedback most fully will receive a £30 Amazon voucher once we've completed the experiment in full. In contrast, participants in the No incentive condition were told: *In addition to your [payment/credits] for completing this survey, all participants who complete this second session in full will now receive an extra £5, and be entered into a prize draw to win one of six £30 Amazon vouchers.* In fact, all participants received an additional £5 at the end of the session and were entered into the prize draw. The participants were asked to rate using a sliding scale how motivated they were, on a scale of 0 (not motivated at all) to 10 (extremely motivated), to do better on the second writing task – note that this variable was not mentioned in our pre-registration.

As in Experiment 2, participants next saw their feedback, presented on a separate computer to allow the participants' eye movements and fixations to be recorded with an eyetracker. Immediately after they finished reading the feedback, participants completed the mind-wandering filler task, which lasted approximately 5 minutes. Participants then completed the AGQ—R (Elliot & Murayama, 2008) before completing the surprise recall test, followed by the adapted TFS, which assessed how often during the filler task they had thought about the past or the future. Participants then answered the feedback rating questions. Finally, as part of the debriefing participants were asked to recall which incentive condition they were in, so that it could be known whether or not they had properly read the instructions (this check was not mentioned in our pre-registration). Specifically, the experimenter asked them whether they had read the text explaining this, and to briefly reiterate what it had said, and prompted them with further questions if they were unsure what was being asked. We were particularly interested to ensure that Incentive participants did remember their instruction, whereas it was relatively unimportant whether or not No Incentive participants remembered their instruction.

Data processing.

Adherence to task instructions. The responses of participants in the Past and Future conditions were checked to ensure they had written about their past or future use of skills, respectively. All participants were judged to have completed the task properly.

Memory coding. Both the memory data and the eye-tracking data were processed in the same way as in Experiment 2, with one researcher coding the free recall responses blind to condition and script version, but not blind to script. Again, an independent coder also coded 20% of the data, and the agreement rates for all four measurements exceeded $r = .85$, therefore all analyses of recall were based on the first coder's data.

Results

Participants' appraisals of the feedback. Most participants appeared to believe that the feedback was personalised, with only one reporting that they suspected the feedback was generic. Participants rated the feedback highly in both fairness ($M = 4.09$ out of 5, $SD = 0.94$), and helpfulness ($M = 4.30$, $SD = 0.71$). They also believed that engaging with the feedback would allow them to do better in the future, as the estimated grade for the initial writing was 51.75% ($SD = 12.76$) whereas participants believed this could rise to 66.42% ($SD = 10.24$) if they completed the task again. Again, no participants correctly guessed the study's aims when prompted, nor noted the differences in feedback style. Consistent with the previous studies, 83% of participants reported at the end of the experiment that, in general, they preferred to receive directive feedback, whereas only 13% said they preferred to receive evaluative feedback (4% expressed no preference either way).

Incentive information and motivation ratings. When all participants were included in analyses, the incentive did not significantly influence participants' motivation ratings: Incentive condition ($M = 7.47$ out of 10, $SD = 2.03$) compared to the No Incentive condition ($M = 6.94$, $SD = 1.79$), $t(169) = 1.81$, $p = .07$, $d = 0.28$ ($BF_{10} = 0.75$). However, of those

participants in the Incentive condition, nine reported no recollection of the incentive. When these nine participants' ratings were removed from analysis we found that participants were significantly more motivated in the Incentive condition (M = 7.70, SD = 1.97) than those in the No Incentive condition, $t(160) = 2.57$, $p = .01$, $d = 0.40$ ($BF_{10} = 3.44$). All remaining analyses were nevertheless conducted with all participants included, as per our pre-registered plan; however we note that excluding these nine participants (which was not part of our plan) made no difference to any of the main findings in terms of statistical significance.

Analysis of recall data. We conducted a 2 (between subjects: incentivisation: Incentive vs No Incentive) x 3 (between subjects: temporal focus: control vs. past vs. future) x 2 (within subjects: feedback type: evaluative vs. directive) x 2 (within subjects: retrieval style accuracy: correct vs. incorrect) mixed-factor ANOVA on the number of feedback comments recalled. As Figure 5 shows, this analysis replicated the key patterns of findings demonstrated previously, whereby participants recalled significantly more of the evaluative comments (M = 3.84, SD = 1.77) than of the directive comments (M = 2.95, SD = 1.38), $F(1,167) = 24.06$, $p < .001$, $\eta^2_p = .13$ ($BF_{inclusion} > 30$). There was also a main effect of retrieval style accuracy, whereby participants recalled more feedback comments in the correct style (M = 4.40, SD = 1.86) than in the incorrect style (M = 2.49, SD = 1.50), $F(1, 167) = 94.49$, $p < .001$, $\eta^2_p = .36$ ($BF_{inclusion} > 30$). Further, we found a significant interaction between feedback style and retrieval style accuracy, signifying an overall evaluative retrieval style, $F(1, 167) = 31.03$, $p < .001$, $\eta^2_p = .16$ ($BF_{inclusion} > 30$). As seen previously, planned follow-up paired t-tests showed that when participants recalled evaluative comments, they reproduced them in an evaluative style significantly more often (75% of the time) than in a directive style (25% of the time), $t(172) = 9.18$, $p < .001$, $d = 0.70$ ($BF_{10} > 30$). In contrast, when participants recalled directive comments, they reproduced these in the correct, directive

style (50% of the time) no more often than in the incorrect, evaluative style (50% of the time), $t(172) = -0.12$, $p = .90$, $d = -0.01$ ($BF_{10} = 0.09$).

Looking next to the effects of incentivisation, we found that participants in both incentivisation groups recalled a similar number of feedback comments overall, $F(1,167) = 1.12$, $p = .29$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.11$). Importantly, there was also no significant interaction between incentivisation and feedback type, $F(1, 167) = 0.76$, $p = .38$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.22$), nor a significant interaction between incentivisation and retrieval style accuracy, $F(1, 167) = 1.64$, $p = .20$, $\eta^2_p = .01$ ($BF_{inclusion} = 0.19$), nor a three-way interaction between incentivisation, feedback type, and retrieval style accuracy, $F(1, 167) = 0.24$, $p = .63$, $\eta^2_p = .01$ $BF_{inclusion} = 0.22$). In other words, incentivisation had no meaningful effect on either the evaluative recall bias or the evaluative retrieval style.

Looking to the effects of temporal focus, we found that participants in the three conditions recalled a similar number of feedback comments overall, $F(2,167) = 0.60$, $p = .55$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.02$). There was also no significant interaction between temporal focus condition and feedback type, $F(2, 167) = 0.59$, $p = .56$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.04$), nor a significant interaction between temporal focus condition and retrieval style accuracy, $F(2, 167) = 0.29$, $p = .74$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.04$), nor a three-way interaction between temporal focus condition, feedback type, and retrieval style accuracy, $F(2, 167) = 0.55$, $p = .57$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.17$). In other words, asking participants to focus on their past or future use of writing skills had no meaningful effect on either the evaluative recall bias or the evaluative retrieval style.

Figure 5 about here

Finally, there was no significant interaction between incentivisation and temporal focus condition; $F(2, 167) = 0.59$, $p = .55$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.05$), nor any further three-

or four-way interactions of these two variables with feedback type and/or retrieval style accuracy, all $p > .32$ (all $BF_{inclusion} \leq 0.44$).

Analysis of attention data. There were 161 participants with valid eye-tracking data: (Control; Incentive $n = 28$, No Incentive $n = 26$; Past; Incentive, $n = 26$, No Incentive, $n = 27$; Future; Incentive, $n = 26$, No Incentive, $n = 28$). We separately analysed the two dependent variables (dwell time and run count) using two 2 (incentivisation) x 3 (temporal focus condition) x 2 (feedback type) mixed-factor ANOVAs. Looking first at the dwell time data, as depicted in Table 1, here we did find a significant main effect of feedback type, $F(1, 155) = 4.45$, $p = .04$, $\eta^2_p = .03$ ($BF_{inclusion} = 2.01$); interestingly, participants spent significantly more time attending to directive comments ($M = 268$ ms/word, $SD = 122$ ms) than to evaluative comments ($M = 260$ ms/word, $SD = 121$ ms), although in Bayesian terms the evidence for this effect is only anecdotal. There were no other statistically significant effects: incentivisation, $F(1, 155) = 1.02$, $p = .31$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.58$); temporal focus condition, $F(2, 155) = 0.14$, $p = .89$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.35$); incentivisation x feedback type, $F(1, 155) = 0.89$, $p = .35$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.29$); temporal focus condition x feedback type, $F(2, 155) = 0.74$, $p = .48$, $\eta^2_p < .01$($BF_{inclusion} = 0.12$); incentivisation x temporal focus condition x feedback type, $F(2, 155) = 0.65$, $p = .52$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.06$).

We next examined the run count data, also depicted in Table 1. Here we again found a significant main effect of feedback type, $F(1, 155) = 4.78$, $p = .03$, $\eta^2_p = .03$ ($BF_{inclusion} = 1.89$), whereby participants read directive comments significantly more times ($M = 1.08$/word, $SD = 0.39$) than evaluative comments ($M = 1.05$/word, $SD = 0.39$), again this is anecdotal evidence in Bayesian terms. There were no other significant effects: incentivisation, $F(1, 155) = 1.89$, $p = .17$, $\eta^2_p = .01$ ($BF_{inclusion} = 0.58$); temporal focus condition, $F(2, 155) = 0.29$, $p = .75$, $\eta^2_p < .01$($BF_{inclusion} = 0.39$); incentivisation x feedback

type, $F(1, 155) = 1.48$, $p = .23$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.36$); temporal focus condition x feedback type, $F(2, 155) = 2.06$, $p = .13$, $\eta^2_p = .03$ ($BF_{inclusion} = 0.39$); incentivisation x temporal focus condition x feedback type, $F(2, 155) = 0.47$, $p = .63$, $\eta^2_p < .01$ ($BF_{inclusion} = 0.13$). We return to look more closely at the data across studies shortly.

Table 1 about here

Exploratory analyses. As before, we examined group differences in participants' ratings of the feedback's fairness and helpfulness, their grading of their actual and future performance, their AGQ-R scores, and the tendency for their spontaneous thoughts after reading the feedback to be future-oriented (calculated as TFS Future score minus TFS Past score). We also examined the extent to which individual participants' motivation ratings, TFS difference scores, AGQ-R scores, and their preference for evaluative vs. directive feedback were associated with their (1) evaluative recall bias, (2) evaluative retrieval style, and (3) selective attention toward evaluative feedback in terms of dwell time and run count. The results of these analyses are reported in supplemental materials. Again there were few effects of note, and in particular it is important to note that participants' motivation ratings were not correlated significantly with the evaluative recall bias, r ($N = 169$) = .08, $p = .32$ ($BF_{10} = 0.16$). However, it is also interesting to note that these motivation ratings were correlated significantly and negatively with the evaluative retrieval style, r ($N = 169$) = -.18, $p = .02$ ($BF_{10} = 2.56$). This finding, though anecdotal and not predicted *a priori*, fits well with the earlier finding that our goal-setting manipulation reduced the evaluative retrieval style.

Effect size analysis

As a means to reach the most precise size estimates of the effects observed across Experiments 1-3, and thus to reach the most robust conclusions that these datasets can offer, we conducted a series of random effects mini-metaanalyses (Goh, Hall, & Rosenthal, 2016).

As the top half of Table 2 shows, looking first at the overall data for each experiment, the combined effect size estimate for the evaluative recall bias was medium in size (d = 0.45), somewhat smaller than the effect size of d = 0.63 [0.48, 0.77] estimated by Nash et al. (2018). The estimated size of the evaluative retrieval style was similar, at d = 0.48. Importantly though, and despite the significant differences observed in Experiment 3, there was little evidence across experiments of an 'evaluative attentional bias', with the combined effect size estimates for the two attentional measures being very small (d = -0.11 and -0.07 for dwell time and run count, respectively).

Because we used the same between subject manipulations in Experiments 1 and 2, we also conducted mini-metaanalyses of these two studies to assess the overall effects of goal-setting. Looking to the bottom half of Table 2, the effect of the goal-setting manipulation (relative to control participants) on the evaluative recall bias was very small, but its effect on the evaluative retrieval style differed statistically from zero, such that goal-setting significantly attenuated this retrieval style bias.[3] Goal-setting had no meaningful effect on participants' relative attention to evaluative over directive feedback comments.

Table 2 about here

General Discussion

[3] Note that in Table 2, the 95% CI for the effect of goal-setting on the evaluative retrieval style in Experiment 2 excludes zero, whereas in our main analyses of Experiment 2, the corresponding omnibus three-way interaction effect was not statistically significant at $\alpha = .05$. This is because the latter analysis included data from the Advice-giving condition, whereas the former only compares the Goal-setting and Control conditions.

For feedback to be used effectively and to therefore enhance our learning, we must often be able to retrieve it from memory at a later time. Our present findings replicate and extend those of Nash et al. (2018), demonstrating robustly that people recall evaluative feedback comments more readily than they recall directive feedback comments.

A key way in which these experiments extend the prior work is by obtaining direct, objective measures of attention during participants' engagement with the feedback. Looking at the combined data across the three experiments, we found that participants spent approximately equal amounts of time focusing on evaluative comments and directive comments, and read both types an equivalent number of times, even though they were subsequently able to recall the former type of feedback far more often. These findings therefore offer the first direct evidence that basic attentional differences are unlikely to account for the evaluative recall bias. Of course, reading times are not the only form of attentional encoding, and it remains possible that other differences may be implicated in the evaluative recall bias; for instance, participants may engage in deeper or more elaborative processing when reading evaluative feedback. Nevertheless, the lack of support here for an attentional account adds further evidence against the causal role of encoding processes in general, and therefore provides greater reason to look instead to retrieval processes for an explanation. For instance, one possibility is that when attempting to retrieve advice or feedback from memory, people tend to selectively search their memories for information that relates to prior performance. This might occur if people generally think of feedback as being evaluative information, and therefore engage in schema-driven retrieval processes that lead them to systematically neglect information that concerns future improvement. If people's mental schemas do indeed often represent feedback as being evaluative, then this might also neatly explain why participants tend to show an evaluative retrieval style when remembering feedback. That is to say, these schemas may provide the basis for Gricean implicatures,

whereby "do X better next time" is taken intuitively to mean, "I did X poorly this time". We are currently investigating these possibilities.

The strong evidence of no difference in the time taken to read evaluative vs. directive feedback, even despite a strong evaluative bias in recall, is particularly interesting in light of the boom in learning analytics research that relies on reading time data (e.g. Ada & Stansfield, 2017; Greller & Drachsler, 2012; Hatala et al., 2015; Kovanović et al., 2015; Nguyen, Tempelaar, Rienties, & Giesbers, 2016). In many such cases, measures of students' engagement with learning environments are inferred from covertly recorded, computerized data, such as the total time they have spent reading specific educational documents or instructions (Hatala et al., 2015; Zimbardi et al., 2017). What our present data clearly illustrate, though, is that even when two different kinds of information receive almost identical exposure in terms of students' reading time, there can still be sizeable differences in the students' cognitive engagement between these two kinds of information. Clearly, it is important not to conflate learners' reading time with their engagement.

The present experiments also extend the prior work by examining the effects of different manipulations, prior to receiving feedback, designed to lead participants to give greater thought to the prospect of improving their skills. In Experiments 1 and 2 we found that our goal-setting intervention had no statistically significant influence on the evaluative recall bias; nor did it influence the amount of feedback recalled overall. Therefore, despite goal-setting interventions being a successful method for improving performance in many tasks and for promoting a directive, future-oriented mindset (e.g. Latham & Locke, 1991; Locke & Latham, 2002; Morisano et al., 2010), there was no evidence that this simple goal-setting intervention improved the retention of feedback. These data therefore suggest that the evaluative recall bias is not easily attributed to participants merely being inadequately focused upon improvement.

The goal-setting intervention, however, did have a slightly different effect: our mini-metaanalysis shows that, as we had predicted, participants in the goal-setting groups showed a smaller evaluative retrieval style than did those in the control groups. Although they still recalled more evaluative feedback than directive feedback, the goal-setting participants were more likely than control participants to recall their feedback as suggestions for future improvement, rather than as critique of their past performance. Prior research has shown that goal-setting can improve students' academic performance, with those who engaged in a goal-setting intervention achieving better grades, better management of workload, and decreased negative affect (Morisano et al., 2010). It is possible, then that in setting goals, our own participants were more likely to process improvement-based feedback as having a directive, supportive intention, rather than as being a judgmental commentary on their performance.

Unfortunately, in Experiment 2 when we attempted to test two competing interpretations of this finding, the data were insufficiently clear to allow us to confidently rule out either interpretation. However, Experiment 3 offers suggestive evidence for a role of motivational factors, insofar that those participants who gave higher subjective motivation ratings tended to exhibit weaker evaluative retrieval styles. In broader terms, the fact that the goal-related factors had some apparent influence on the evaluative retrieval style, but not the evaluative recall bias, supports Nash et al.'s (2018) conclusions that the former is much more susceptible than the latter to contextual factors. It is an open question whether attenuating the evaluative retrieval style—whether via goal-setting or via other means—would be desirable in practice, insofar as doing so could have consequences for students' attitudes to learning, motivation, or acceptance of feedback. We do know, though, that people typically perceive evaluative feedback as more negative than directive feedback, and also that students typically prefer to receive directive rather than evaluative feedback (Nash et al., 2018). Therefore, it seems important that even when people successfully recall their directive feedback, they

typically recall it in an evaluative style: this bias could reasonably lead people to believe that they have not been given the feedback they want, and could perhaps even affect the feedback-receiver's self-esteem or their interpersonal relationship with the feedback-giver. These matters merit attention in future work.

To extend these findings using different techniques for influencing participants' improvement focus, in Experiment 3 we prompted participants to reflect either on how they would use their persuasive writing skills in the future, or how they have used these skills in the past (or neither). This manipulation had no meaningful influence on the evaluative recall bias, and unlike our goal-setting manipulation, it also had no apparent effect on participants' retrieval styles. In Experiment 3 we also tested the influence of an incentive to improve on these memory biases. This was an important question to address, as students are often motivated to do well in their real assignments in a way that is difficult to emulate in laboratory studies. The financial incentive we used in Experiment 3 had only a modest effect on participants' self-reported motivation to improve their writing, and there was no evidence that it had any effect on either of the memory biases of interest. Whereas stronger incentives and motivational manipulations would clearly be valuable for future research, it is noteworthy that even in the No Incentive condition, participants claimed to be quite motivated to improve, and so these results provide further confidence that the evaluative recall bias is not merely a result of participants being disinterested in improving. The fact that the evaluative recall bias was not correlated with participants' subjective motivation ratings provides further evidence to this end. This finding also fits with those from other research, which demonstrates that incentivising memory performance does not generally lead to meaningful improvements (Kang & Pashler, 2014; Ngaosuvan & Mäntylä, 2005; Wehe, Rhodes, & Seger, 2015). Overall, the Experiment 3 data therefore add further indications that the evaluative recall bias could generalize to more realistic contexts in which individuals are

motivated to learn from their feedback – this is a key issue that we are currently exploring further by taking our data collection out of the lab and into real classrooms.

Overall, the findings of these studies are highly relevant to educational practitioners. Whereas students typically prefer directive feedback, and whereas directive feedback is typically considered more valuable to learners than is evaluative feedback, our findings lend weight to a striking caveat: simply preparing feedback comments in a future-oriented style can make them less likely to be remembered. In real-world learning contexts, this bias in remembering could plausibly lead to students being less able to learn from and remedy their mistakes in future, at least when they receive future-oriented feedback without concrete, practical instructions on exactly what steps to take next, when, and how. Whereas the mechanism behind the evaluative recall bias still remains unidentified, these three experiments—by testing certain attentional and goal-based accounts of the bias, and by demonstrating conditions under which it can be observed—bring us closer to understanding this counterintuitive effect.

Acknowledgements

This research was conducted with the support of generous funding from the Leverhulme Trust, Research Project Grant RPG-2016-189. The authors are grateful to the funder, and also to Emily Papps and Danielle Robinson for their support with data coding.

Declaration of interest statement

Samantha Gregory, Naomi Winstone, Nathan Ridout and Robert Nash, declare that they have no conflict of interest.

References

Ada, M. B., & Stansfield, M. (2017). The potential of learning analytics in understanding students' engagement with their assessment feedback. In *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)* (pp. 227–229). IEEE. DOI:10.1109/ICALT.2017.40

Alós-Ferrer, C., García-Segarra, J., & Ritschel, A. (2018). Performance curiosity. *Journal of Economic Psychology, 64*, 1–17. DOI:10.1016/j.joep.2017.08.002

Avery, R. E., & Smillie, L. D. (2013). The impact of achievement goal states on working memory. *Motivation & Emotion, 37*, 39–49. DOI:10.1007/s11031-012-9287-4

Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W. Y., Franklin, M. S., & Schooler, J. W. (2012). Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological Science, 23*, 1117–1122. DOI:10.1177/0956797612446024

Brewer, W. F. (1977). Memory for the pragmatic implications of sentences. *Memory & Cognition, 5*, 673–678. DOI:10.3758/BF03197414

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 & Cognition, 34*, 918–928. DOI:10.1037/0278-7393.34.4.918

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. DOI:10.3758/MC.36.3.604

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education, 31*, 219–233. DOI:10.1080/03075070600572132

Chan, J. C. K., & McDermott, K. B. (2006). Remembering pragmatic inferences. *Applied Cognitive Psychology, 20*, 633–639. DOI:10.1002/acp.1215

Chasteen, A. L., Park, D. C., & Schwarz, N. (2001). Implementation intentions and facilitation of prospective memory. *Psychological Science, 12*, 457–461. DOI:10.1111/1467-9280.00385

Cohen, A. L. (2013). Software for the automatic correction of recorded eye fixation locations in reading experiments. *Behavior Research Methods, 45*, 679–683. DOI:10.3758/s13428-012-0280-3

Dawson, P., Henderson, M., Mahoney, P., Phillips, M., Ryan, T., Boud, D., & Molloy, E. (2019). What makes for effective feedback: staff and student perspectives. *Assessment & Evaluation in Higher Education, 44*, 25–36. DOI:10.1080/02602938.2018.1467877

Dijksterhuis, A., & Aarts, H. (2010). Goals, attention, and (un)consciousness. *Annual Review of Psychology, 61*, 467–490. DOI:10.1146/annurev.psych.093008.100445

Elliot, A. J., & Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology, 100,* 613–628. DOI:10.1037/0022-0663.100.3.613

Fastrich, G. M., Kerr, T., Castel, A. D., & Murayama, K. (2018). The role of interest in memory for trivia questions: An investigation with a large-scale database. *Motivation Science, 4*, 227–250. DOI:10.1037/mot0000087

Feng, S., D'Mello, S., & Graesser, A. C. (2013). Mind wandering while reading easy and difficult texts. *Psychonomic Bulletin and Review, 20*, 586–592. DOI:10.3758/s13423-012-0367-y

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 & Individual Differences*, *49*, 393–399. DOI:10.1016/j.lindif.2016.05.019

Gibbs, G., & Simpson, C. (2004). Does your assessment support your students' learning? *Learning and Teaching in Higher Education*, *1*, 3–31.

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*, 535–549. DOI:10.1111/spc3.12267

Goschke, T., & Kuhl, J. (1993). Representation of intentions: Persisting activation in memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *19*, 1211–1226. DOI:10.1037/0278-7393.19.5.1211

Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Journal of Educational Technology & Society*, *15*, 42–57.

Hatala, M., Joksimović, S., Gašević, D., Kovanović, V., Dawson, S., & Baker, R. S. (2015). Does time-on-task estimation matter? Implications for the validity of learning analytics findings. *Journal of Learning Analytics*, *2*, 81–110. DOI:10.18608/jla.2015.23.6

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 81–112. DOI:10.3102/003465430298487

Horstmanshof, L., & Zimitat, C. (2007). Future time orientation predicts academic engagement among first-year university students. *British Journal of Educational Psychology*, *77*, 703–718. DOI:10.1348/000709906X160778

Husman, J., & Lens, W. (1999). The role of the future in student motivation. *Educational Psychologist*, *34*, 113–125. DOI:10.1207/s15326985ep3402_4

Ikeda, K., Castel, A. D., & Murayama, K. (2015). Mastery-approach goals eliminate retrieval-induced forgetting: The role of achievement goals in memory inhibition. *Personality and Social Psychology Bulletin*, *41*, 687–695. DOI:10.1177/0146167215575730

JASP Team. (2018). JASP (Version 0.10.0)[Computer software].

Kang, S. H. K., 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. DOI:10.1080/09541440601056620

Kang, S. H. K., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation? *Journal of Applied Research in Memory and Cognition*, *3*, 183–188. DOI:10.1016/j.jarmac.2014.05.006

Klein, S. B. (2013). The complex act of projecting oneself into the future. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*, 63–79. DOI:10.1002/wcs.1210

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. DOI:10.3758/MC.38.1.13

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. DOI:10.1080/09658211.2010.537827

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. DOI:10.1037/0033-2909.119.2.254

Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R. S., & Hatala, M. (2015).

Penetrating the black box of time-on-task estimation. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 184–193). New York, New York, USA: ACM Press. DOI:10.1145/2723576.2723623

Lasane, T. P., & Jones, J. M. (1999). Temporal orientation and academic goal-setting: The mediating properties of a motivational self. *Journal of Social Behavior and Personality*, *14*, 31–44.

Latham, G. P., & Locke, E. A. (1991). Self-regulation through goal setting. *Organizational Behavior and Human Decision Processes*, *50*, 212–247. DOI:10.1016/0749-5978(91)90021-K

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*, 705–717. DOI:10.1037//0003-066X.57.9.705

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*, 476–490. DOI:10.3758/BF03210951

Mangels, J. A., Rodriguez, S., Ochakovskaya, Y., & Guerra-Carrillo, B. (2017). Achievement goal task framing and fit with personal goals modulate the neurocognitive response to corrective feedback. *AERA Open*, *3*. DOI:10.1177/2332858417720875

McDaniel, M. A., Howard, D. C., & Butler, K. M. (2008). Implementation intentions facilitate prospective memory under high attention demands. *Memory & Cognition*, *36*, 716–724. DOI:10.3758/MC.36.4.716

Montagrin, A., Brosch, T., & Sander, D. (2013). Goal conduciveness as a key determinant of memory facilitation. *Emotion*, *13*, 622–628. DOI:10.1037/a0033066

Morisano, D., Hirsh, J. B., Peterson, J. B., Pihl, R. O., & Shore, B. M. (2010). Setting,

elaborating, and reflecting on personal goals improves academic performance. *Journal of Applied Psychology*, *95*, 255–264. DOI:10.1037/a0018478

Moskowitz, G. B. (2002). Preconscious effects of temporary goals on attention. *Journal of Experimental Social Psychology*, *38*, 397–404.

Murayama, K., & Elliot, A. J. (2011). Achievement motivation and memory: Achievement goals differentially influence immediate and delayed remember-know recognition memory. *Personality and Social Psychology Bulletin*, *37*, 1339–1348. DOI:10.1177/0146167211410575

Nash, R. A., Winstone, N. E., Gregory, S. E. A., & Papps, E. (2018). A memory advantage for past-oriented over future-oriented performance feedback. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 1864–1879. DOI:10.1037/xlm0000549

Ngaosuvan, L., & Mäntylä, T. (2005). Rewarded remembering: Dissociations between self-rated motivation and memory performance. *Scandinavian Journal of Psychology*, *46*, 323–330. DOI:10.1111/j.1467-9450.2005.00462.x

Nguyen, Q., Tempelaar, D., Rienties, B., & Giesbers, B. (2016). What learning analytics based prediction models tell us about feedback preferences of students. *Quarterly Review of Distance Education*, *17*, 13–33.

Nuttin, J. R., & Lens, W. (1985). *Future time perspective and motivation: Theory and research method*. Hillsdale, NJ: Erlbaum.

Phye, G. D., & Sanders, C. E. (1994). Advice and feedback: Elements of practice for problem solving. *Contemporary Educational Psychology*, *19*, 286–301. DOI:10.1006/ceps.1994.1022

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. DOI:10.1111/mbe.12124

Robinson, S., Pope, D., & Holyoak, L. (2013). Can we meet their expectations? Experiences and perceptions of feedback in first year undergraduate students. *Assessment and Evaluation in Higher Education, 38*, 260–272. DOI:10.1080/02602938.2011.629291

Shipp, A. J., Edwards, J. R., & Lambert, L. S. (2009). Conceptualization and measurement of temporal focus: The subjective experience of the past, present, and future. *Organizational Behavior and Human Decision Processes, 110*, 1–22. DOI:10.1016/j.obhdp.2009.05.001

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153–189. DOI:10.3102/0034654307313795

Smallwood, J., Nind, L., & O'Connor, R. C. (2009). When is your head at? An exploration of the factors associated with the temporal focus of the wandering mind. *Consciousness & Cognition, 18*, 118–125. DOI:10.1016/j.concog.2008.11.004

Vogt, J., De Houwer, J., Moors, A., Van Damme, S., & Crombez, G. (2010). The automatic orienting of attention to goal-relevant stimuli. *Acta Psychologica, 134*, 61–69. DOI:10.1016/j.actpsy.2009.12.006

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin and Review, 25*, 58–76. DOI:10.3758/s13423-017-1323-7

Wehe, H. S., Rhodes, M. G., & Seger, C. A. (2015). Evidence for the negative impact of reward on self-regulated learning. *Quarterly Journal of Experimental Psychology, 68*,

2125–2130. DOI:10.1080/17470218.2015.1061566

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. DOI:10.1080/00461520.2016.1207538

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. DOI:10.1080/02602938.2015.1075956

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. DOI:10.1080/03075079.2015.1130032

Zimbardi, K., Colthorpe, K., Dekker, A., Engstrom, C., Bugarcic, A., Worthy, P., … Long, P. (2017). Are they using my feedback? The extent of students' feedback use has a large impact on subsequent academic performance. *Assessment and Evaluation in Higher Education, 42,* 625–644. DOI:10.1080/02602938.2016.1174187

Zimbardo, P. G., & Boyd, J. N. (1999). Putting time in perspective: A valid, reliable individual-differences metric. *Journal of Personality and Social Psychology, 77,* 1271–1288. DOI:10.1037/0022-3514.77.6.1271

List of figure captions

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

Figure 2. Attention toward evaluative and directive feedback in Experiment 1, split according to retrieval style accuracy and experimental condition. Panel A represents mean dwell time data; Panel B represents mean run count data. Error bars are 95% within-subject confidence intervals, calculated separately for each between-subject experimental condition (Loftus & Masson, 1994).

Figure 3. Recall of evaluative and directive feedback in Experiment 2, split according to retrieval style accuracy and experimental condition. Error bars are 95% within-subject confidence intervals, calculated separately for each between-subject experimental condition (Loftus & Masson, 1994).

Figure 4. Attention toward evaluative and directive feedback in Experiment 2, split according to retrieval style accuracy and experimental condition. Panel A represents mean dwell time data; Panel B represents mean run count data. Error bars are 95% within-subject confidence intervals, calculated separately for each between-subject experimental condition (Loftus & Masson, 1994).

Figure 5. Recall of evaluative and directive feedback by the No Incentive group (Panel A) and the Incentive group (Panel B) in Experiment 3, split according to retrieval style accuracy and experimental condition. Error bars are 95% within-subject confidence intervals, calculated separately for each between-subject experimental condition (Loftus & Masson, 1994).

Tables

Table 1. *Mean dwell time and run count for evaluative and directive feedback comments across each between-subject condition in Experiment 3 (standard deviations in parentheses).*

Incentivisation condition		No incentive			Incentive		
Temporal focus condition		Control	Past	Future	Control	Past	Future
Mean dwell time (ms/word)	Evaluative comments	256.46 (82.37)	260.54 (213.83)	238.97 (66.14)	257.94 (108.08)	276.64 (105.43)	268.18 (98.65)
	Directive comments	260.68 (84.21)	261.57 (199.29)	247.10 (67.27)	281.46 (105.39)	280.64 (138.40)	275.60 (99.56)
Mean run count (number of reads/word)	Evaluative comments	0.98 (0.32)	1.09 (0.72)	0.98 (0.26)	1.06 (0.31)	1.12 (0.31)	1.06 (0.22)
	Directive comments	1.01 (0.32)	1.05 (0.58)	1.03 (0.29)	1.16 (0.36)	1.13 (0.43)	1.11 (0.29)

Table 2. *Effect size estimates for the key outcome measures*

	Outcome measure	Experiment	Effect size *d* [95% CI]
Overall effect size estimate (all data)	Evaluative recall bias	Experiment 1	0.63 [0.44, 0.82]
		Experiment 2	0.42 [0.26, 0.57]
		Experiment 3	0.38 [0.22, 0.53]
		Metaanalytic effect	**0.45 [0.32, 0.61]**
	Evaluative retrieval style	Experiment 1	0.56 [0.38, 0.75]
		Experiment 2	0.49 [0.33, 0.65]
		Experiment 3	0.42 [0.27, 0.58]
		Metaanalytic effect	**0.48 [0.39, 0.58]**
	Evaluative attentional bias (dwell time)	Experiment 1	-0.14 [-0.33, 0.04]
		Experiment 2	0.01 [-0.14, 0.17]
		Experiment 3	-0.19 [-0.34, -0.04]
		Metaanalytic effect	**-0.11 [-0.23, 0.02]**
	Evaluative attentional bias (run count)	Experiment 1	-0.07 [-0.25, 0.12]
		Experiment 2	0.05 [-0.10, 0.21]
		Experiment 3	-0.19 [-0.34, -0.03]
		Metaanalytic effect	**-0.07 [-0.21, 0.08]**
Effect of goal-setting on outcome measure (Goal-setting condition vs. Control condition only)	Evaluative recall bias	Experiment 1	0.12 [-0.22, 0.47]
		Experiment 2	0.08 [-0.29, 0.45]
		Metaanalytic effect	**0.10 [-0.15, 0.36]**
	Evaluative retrieval style	Experiment 1	-0.44 [-0.79, -0.09]
		Experiment 2	-0.43 [-0.81, -0.06]
		Metaanalytic effect	**-0.44 [-0.69, -0.18]**
	Evaluative attentional bias (dwell time)	Experiment 1	-0.13 [-0.50, 0.24]
		Experiment 2	0.16 [-0.22, 0.53]

		Metaanalytic effect	**0.01 [-0.27, 0.29]**
	Evaluative attentional bias (run count)	Experiment 1	-0.01 [-0.37, 0.36]
		Experiment 2	0.20 [-0.18, 0.58]
		Metaanalytic effect	**0.09 [-0.17, 0.35]**

List of figure captions

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

Figure 2. Attention toward evaluative and directive feedback in Experiment 1, split according to retrieval style accuracy and experimental condition. Panel A represents mean dwell time data; Panel B represents mean run count data. Error bars are 95% within-subject confidence intervals, calculated separately for each between-subject experimental condition (Loftus & Masson, 1994).

Figure 3. Recall of evaluative and directive feedback in Experiment 2, split according to retrieval style accuracy and experimental condition. Error bars are 95% within-subject confidence intervals, calculated separately for each between-subject experimental condition (Loftus & Masson, 1994).

Figure 4. Attention toward evaluative and directive feedback in Experiment 2, split according to retrieval style accuracy and experimental condition. Panel A represents mean dwell time data; Panel B represents mean run count data. Error bars are 95% within-subject confidence intervals, calculated separately for each between-subject experimental condition (Loftus & Masson, 1994).

Figure 5. Recall of evaluative and directive feedback by the No Incentive group (Panel A) and the Incentive group (Panel B) in Experiment 3, split according to retrieval style accuracy and experimental condition. Error bars are 95% within-subject confidence intervals, calculated separately for each between-subject experimental condition (Loftus & Masson, 1994).