Author/s:
Ryan, A;McColl, GJ;O'Brien, R;Chiavaroli, N;Judd, T;Finch, S;Swanson, D

Title:
Tensions in post-examination feedback: information for learning versus potential for harm

Date:
2017-09-01

Citation:
Ryan, A., McColl, G. J., O'Brien, R., Chiavaroli, N., Judd, T., Finch, S. & Swanson, D. (2017). Tensions in post-examination feedback: information for learning versus potential for harm. Medical Education, 51 (9), pp.963-973. https://doi.org/10.1111/medu.13366.

Persistent Link:
https://hdl.handle.net/11343/293337

DR. ANNA RYAN (Orcid ID : 0000-0001-7458-027X)

PROF. GEOFFREY J MCCOLL (Orcid ID : 0000-0003-2741-5233)

Article type : Research Papers

Tensions in post exam feedback: Information for learning versus potential for harm

Abstract:

Objective:

Self-regulation is recognized as being a requisite skill for professional practice This study is part of a program of research designed to explore efficient methods of feedback that improve medical students' ability to self-regulate their learning. Our aim was to clarify how students respond to different forms and content of written feedback and to explore the impact on study behaviour and knowledge acquisition.

Methods:

Year two students in a 4-year graduate entry medical program completing four formative progress tests during the academic year were randomised into three groups receiving different feedback reports. All reports included proportion correct overall and by clinical rotation. One group received feedback reports including lists of clinical presentations relating to questions answered correctly and incorrectly, another group received reports containing this same information in combination with response certitude. The final group received reports involving normative comparisons. Baseline progress test performance quartile groupings (a proxy for academic ability) were determined by results on the first

progress test.

A mixed method approach with triangulation of research findings was used to interpret results. Outcomes of interest included progress test scores, summative examination results, and measures derived from study diaries, questionnaires and semi-structured interviews.

Results:

Of the three feedback types provided in this experiment, feedback containing normative comparisons resulted in inferior test performance for students in the lowest performance quartile group. This type of feedback appeared to stimulate general rather than exam-focused study.

Conclusions:

Medical students are often considered relatively homogenous and high achieving, yet results of this study suggest caution when providing them with normative feedback indicating poorer performance relative to their peers There is much need for further work to explore efficient methods of providing written feedback that improve medical students' ability to self-regulate their learning, particularly to those students who have the most room for improvement.

Introduction

The central aim of medical education is to help students acquire the knowledge, skills and attributes required for practice of the profession. Fundamental to successful practice as a medical doctor is the ability to self-regulate learning. This involves a continual cycle of reflection on performance against goals, recognition of strengths and weakness, identification and engagement in learning opportunities, application of new knowledge in everyday work, and subsequent re-evaluation of performance. [1] In order to improve knowledge and skill over the course of their career, medical trainees must learn how to regulate their own learning, starting with their initial medical training. The capacity for students to become self-regulating learners appears to be influenced by the quality of feedback they receive, [2] with feedback playing a key role in reflection on performance, recognition of strengths and weaknesses, and subsequent modification or maintenance of study strategies. [3]

While educators might not have a major influence on students' internal feedback (generated by the learners' own monitoring process such as their experience of the assessment, their reflection and consideration of factors contributing to their performance), external feedback *is* the domain of the educator. Feedback that strengthens learners' capacity to self-regulate requires information that facilitates reflection, motivates learners, builds self-esteem, encourages dialogue (between teachers and peers), and focuses on the achievement of explicit goals. [3]

In addition to high quality information, the relationship between the giver and the receiver of feedback is being increasingly recognized as a powerful influence on response to feedback. [4] Such relational factors are largely absent in automated feedback after written exams. Automated feedback can include grades, normative comparisons, or re-presentation of test questions and model answers but there are limitations with this approach. For example, grades alone don't allow learners to clarify unclear concepts or content [5] and high marks have been shown to drive learner focus towards outcomes and social comparison rather than task mastery. [6] However, students frequently want this information in order to gauge their performance relative to a standard. And, while normative feedback (representing students' performance relative to peers) provides an objective reference that can be important for self-monitoring and metacognitive components of self-regulation, [7] this can be demotivating for lower-scoring learners. [8] In addition, providing test questions and answers can improve long-term retention [9] (and is popular with students), but concerns have been raised that this improvement is due to memorisation rather than deeper understanding. [10]

Furthermore, giving face-to-face individual verbal feedback after written assessment is simply not feasible for the majority of medical schools in the current environment. Funding constraints, large class sizes and high academic workloads require large scale and efficient assessment methods and prevent faculty from engaging in individual verbal feedback. [11] While well-constructed multiple choice questions (MCQ) are scalable, efficient and can provide a valid and reliable assessment of clinical knowledge, [12,13] how and even whether to provide feedback to students after an MCQ exam remains unclear. We simply do not know what form of post-MCQ test feedback facilitates students' self-regulation and encourages learning, particularly for those students most needing to improve. The goal of this program

of research is to explore efficient methods of feedback that improve medical students' ability to self-regulate their learning. Our aim in this study was to clarify how students respond to different forms and content of automated written feedback following a MCQ exam. We asked: How do different forms and content of automated written feedback influence study behaviour and further acquisition of knowledge in medical students?

Methods

Study Overview

We employed a convergent parallel design [14] for the study involving concurrent collection of qualitative and quantitative data, separate methods of analysis and subsequent triangulation of all outcomes. We situated the study within year 2 (the first fulltime clinical year) of a four-year graduate-entry medical program. Students at nine sites (clinical schools) were recruited (N= 312 students out of a possible 332) to sit four MCQ format progress tests set at a graduation-level standard. Progress tests were administered after each rotation block, and students received automated written feedback within a week of completing each test. Of the 312 participants, 291 gave permission for the researchers to access the results of their formal end-of-year assessments, 250 completed study diaries during the course of the academic year, and 220 completed a questionnaire about their experience and behaviour (between the third and fourth progress tests). A smaller sample of these 220 (N=20) were interviewed to further explore the influence on their learning.

Participants were randomised into three feedback groups based on their rotation order and clinical school site resulting in 18 unique study combinations of clinical school, rotation order and feedback group. Participants in the three feedback groups received different feedback report content. All reports included an overall score and sub-scores by clinical rotation domain. As well as this information common to all feedback reports, each feedback group had additional (and different) feedback information.

Clinical presentation (CP) feedback reports included two lists of the clinical presentations featured in each of the test items – one list of clinical presentations answered correctly, and the other list of clinical presentations answered incorrectly. For example:

> Items answered correctly related to: chest pain; sudden loss of vision… Items answered incorrectly related to…

Clinical presentation plus (CP+) feedback reports contained two lists of the clinical presentations featured in each of the test items – one for clinical presentations answered correctly and another for those answered incorrectly (i.e. the same information as the CP reports) but also included the organ system of relevance to that item, and the level of certainty the student had recorded for each item answered (providing a measure metacognitive insight or reflection-in-action). [15]

For example:

> Items answered incorrectly with a high certainty of correctness related to: chest pain (respiratory system), sudden loss of vision (nervous system)...

Normative (N) feedback reports included information about how the student performed in relation to classmates (i.e. normative comparison). For example:

> The average of all participating students was x/80; medical items x%; surgical items x%...

Detailed examples of each type of report are included in Appendix S1 (available online as supporting information).

All MCQ items were originally sourced from the National Board of Medical Examiners (NBME); these were in the clinical vignette style used in the (USMLE) Step 2 Clinical Knowledge tests. Items were adapted for local use to reflect differences in spelling and terminology, units on diagnostic studies, drug formularies, etc. Four unique test forms were constructed using a two-dimensional test blueprint (physician task and organ system) matched to the school's year two curriculum content and balanced for item difficulty (using Step 2 item statistics) and patient characteristics (age, gender, setting). Participants in each of the 18 study group combinations were randomly assigned to one of four rotating test sequences (1234, 2341, 3412 or 4123) to ensure that changes in average test scores over time were unrelated to differences in test form difficulty. [16]

Participants were asked to indicate (a) the correct response and (b) how certain they were that this response was correct (low, medium or high) for each test item. Results were processed within a few days of each test, and bespoke software was used to generate personalised feedback reports emailed to individual participants.

Outcome Measures

Progress test results at test occasions 2, 3, and 4 were the primary outcomes of interest; summative end-of-year assessment results and study diary entries provided the secondary outcomes; and questionnaires and interviews were used to understand primary and secondary outcomes.

Knowledge acquisition

Progress tests, end-of-year summative MCQ and short answer question (SAQ) exams measured knowledge acquisition. Participants were allocated into four baseline quartile groups based on results on the first progress test, relative to the quartiles.

Study behaviour

Students used a diary instrument which captured and quantified their study activities (timetabled, informal and self-directed) in whole hours. We based the diary design on the approach of Wilkinson, Wells & Bushnell, [17] and adapted it for online delivery and reworded it to suit the local context. Each participant received 3 to 4 diary completion requests via automated email, with each diary covering a designated 24-hour period. Participants completed diaries retrospectively (requests were sent out the day after the specified diary date) with participants who failed to respond receiving a single reminder email a few days after the initial request.

Student perceptions

We asked students for their perceptions of the progress tests and their descriptions of how the test feedback changed their study behaviours through questionnaires and interviews.

We adapted the questionnaire from Wade & colleagues' validated progress testing instrument. [18] We reviewed items with reference to our focus on self-regulation, the study context and the relevant literature, [19-21] and added additional items in an iterative process involving four of the authors. The finalised questionnaire consisted of 18 Likert scale and 3 free text items and was pilot tested and subject to the content validity process described by Lynn. [22] All items had a content validity index of greater than 0.8 and were retained. The scale-level content validity index was 96.6%.[23]

We drew on feedback literature [9,24-26] to construct the interview schedule (Appendix S2 available online). We used open-ended questions in a one-to-one semi-structured interview

format We audio-recorded interviews and these were transcribed and anonymised by a professional data service.

Data Analysis

Quantitative Outcomes

We drew extensively on Linear Mixed Models (LMM) for the analysis of quantitative outcomes. Our selection of LMM was informed by the study's design and its accommodation of fixed and random factors. Ordinal outcomes were recoded to binary variables (0 or 1). In such cases we applied Generalised Linear Mixed Modelling (GLMM) rather than standard LMM.

Each of the fixed factors considered varied according to the outcome of interest. We included the key design factors - clinical school, rotation order (nested in clinical school) and feedback group (CP, CP+ or N) - in all models. Participant ID was included as a random factor for analyses involving multiple (repeated) measures of the same participant. Additional variables – time period, test occasion, test form, test order, baseline performance quartile group, baseline certainty, and gender – were included as appropriate.

The starting (most complex) model included an interaction between all primary explanatory factors. Small higher-order interactions were subsequently removed in a backwards stepwise manner. If all such interactions were small, only the main effects were modelled and reported. If the main effect of baseline quartile group was small, then it was also removed. We used pairwise comparisons to describe the main effects and planned all comparisons.

Consistent with recommendations, [27] we avoided relying solely on *P*-values in our interpretation, instead focusing on effect sizes (mean differences and odds ratios) and precision of estimates (confidence intervals). Within this framework, adjustments for multiple comparisons are not appropriate, [28] nor are they readily available for the complex models we report. We used triangulation of results from the various quantitative and qualitative outcome measures to avoid over interpretation of the results of multiple hypothesis tests. For detailed statistical methodology related to each quantitative outcome see Appendix S3 online.

Qualitative Outcomes

Three of the authors analysed free-text questionnaire responses with line-by-line open-coding thematic analysis. We identified common and group-specific themes through comparison among feedback groups.

Four authors independently coded four interview transcripts making notes of initial meanings and themes. We subsequently met to discuss the emerging themes and solidified these into a coding framework that was subsequently applied to the remaining sixteen interview transcripts. The group discussed any discrepancies and negotiated agreement.

Results

Seven of the 312 students who initially enrolled in the study withdrew, leaving 305 students with data for analysis. The presentation of results is focused on estimating effects for feedback group and baseline quartile group (as measured at the first progress test).

Knowledge acquisition

Progress tests

The number of students sitting each progress test varied from 296 on test occasion 1 to 257 on test occasion 4. The test forms had acceptable reliability with average coefficient alpha values for the 4 test forms ranging from 0.69 to 0.74. The estimated mean test result (adjusted for test occasion and test paper by averaging over tests 2, 3, and 4) for each feedback group by baseline quartile group reveals a clear effect for quartile group in all feedback groups (students in the fourth (highest scoring) quartile group continued to perform better) and a differential impact of feedback group on participants in the first (lowest scoring) quartile group (poorer performance when normative feedback given; Figure 1). Pairwise comparisons demonstrate that participants in the first quartile group at baseline who were in the normative feedback group scored an average of 4.42 marks out of 80 (5.5%) lower than those in the clinical presentation group (CP-N 95% CI 1.2, 7.7; $P = 0.007$) and 3.51 marks (4.4%) lower than those in the clinical presentation plus group (CP+-N 95% CI 0.7, 6.3; $P = 0.014$).

Summative exam results

Observed mean MCQ and SAQ scores by baseline quartile group and feedback group demonstrate a similar trend, but without statistically significant differences (Figure 2, *P* = 0.34; Figure 3, *P* = 0.07).

The overall effect of baseline quartile group was highly statistically significant. Participants who were in the first (lowest scoring) quartile group at baseline received an average of 1.98 marks less out of 17.5 (11.3%) on their MCQ score (95% CI 1.4, 2.6; *P* < 0.001) and 1.69 marks less out of 17.5 (9.7%) for their SAQ score (95% CI 1.3, 2.1; *P* < 0.001) than participants who were in the forth (highest scoring) quartile group. Both tests had acceptable psychometric properties with a coefficient alpha of 0.84 for the MCQ paper and 0.68 for the SAQ paper.

Certainty data

There was no statistically significant relationship between correctness and certainty within any feedback group. However, across groups, the odds of providing the correct response to items answered with a high degree of certainty were more than double at test occasion 4 compared with test occasion 1 (OR 2.09; 95% CI 1.9, 2.4; *P* < 0.001).

Study Behaviour

A total of 827 diaries were completed from 1056 requests sent (78% completion rate). After excluding incomplete and duplicate diaries, 804 diaries (contributed by 250 individual students) were retained for analysis. Participants studied progressively more in the lead up to their exams with an increase of greater than 6 hours/day over the final two months of the academic year (Mean difference (end November – end September) 6.20, 95% CI 4.4, 8.0; *P*< 0.001).

We were unable to detect any two-way interactions of feedback group by quartile group. However, participants in the lower two baseline performance quartiles studied for longer than those in higher two quartiles (Figure 4). The difference was greatest between students in the middle quartile groups, with participants in the second lowest quartile group studying nearly two and a half hours more per day than those in the second highest quartile group (mean difference (Quartile group 2 – Quartile group 3) 2.46; 95% CI 0.6, 4.3; *P* = 0.01)

At the time around the final test (the final two months of the year), the total of *study for an exam or test* combined with *general study in medicine* was similar, and not statistically

different, for all three feedback groups. The final model for *general study in medicine* demonstrated a two-way interaction of feedback group and test occasion. Feedback group differences were strongest on test occasion 4: participants in the normative feedback group spent an average of 69 minutes more than the clinical presentation group (mean difference (CP-N) -1.15 hours; 95% CI -2.3, -0.01; *P* = 0.04) and an average of 63 minutes more than the clinical presentation plus group (mean difference (CPplus -N) -1.05 hours; 95% CI 0.2, -2.3; *P* = 0.06) engaged in *general study in medicine* (Figure 4). Feedback group differences were weaker on the other test occasions (not shown). Consistent with the normative feedback group's higher mean in *general study in medicine* on test occasion 4 (but similar overall study time) was their lower mean score on *study for an exam or test* (Figure 4). Participants receiving clinical presentation feedback reported spending an average additional 54 minutes (mean difference (CP-N) 0.90 hours; 95%CI -1.6, 3.4; *P* = 0.44) and those receiving clinical presentation plus feedback reported an average additional 46 minutes on *study for an exam or test* (mean difference (CPplus -N) 0.77 hours; 95%CI -1.7, 3.2; *P* = 0.50). However, these differences do not reach statistical significance given the relative greater variability in the scores for *study for an exam or test*, compared with *general study in medicine*.

Student Perceptions

Interviews

The 20 participants who volunteered to be interviewed were drawn from a variety of clinical schools, rotation orders, feedback groups and test form sequences. Elements of self-regulation of learning, such as goal-setting, monitoring, reflection and adaption were evident in many interview data.

A number of students stressed the importance being an effective medical practitioner; "the goal is to be a good doctor" (Participant 2N), and felt this was a key purpose of feedback.

A desire to monitor their own performance also contributed to some participants' decision to participate:

> " I volunteered to be a part of this research project because I wanted a greater level of feedback about my progress throughout the year." (Participant 13CP+)

Others expressed a hope for receipt of information about their own individual strengths and weaknesses, and a guide to the effectiveness of current study. They outlined concern that without progress tests, they might be spending time on ineffective learning strategies but not receive feedback on this until after their end of year exams. Some expressed a sense of being overwhelmed with content in the medical program and needing help with navigating through their subject matter:

> "… you need a guide on where to spend your time." (Participant 13CP+)

Some students also felt taking the tests provided more useful feedback about their learning than the feedback reports themselves:

> "…what made me identify the areas I wasn't good at was the questions themselves. (Participant 14CP)

Many participants had an interest in knowing their academic rank within the cohort for both motivation and reassurance. They suggested that proportion-correct scores had limited utility and felt that peer comparison was more effective for this purpose.

While rotational information (percentage of medical, surgical and ambulatory care items answered correctly) was appreciated, students wanted more direction, particularly with respect to clinical presentations, in order to be able to improve their learning.

> "…I know a lot about abdominal pain so reading broadly on it is not going to do a lot for me but I may have missed some fundamental principle amongst all that reading." (Participant 19CP)

Not surprisingly, participants suggested that good results, particularly after a difficult test, improved their confidence and provided reassurance, while poor results could be a cause of stress and could undermine their confidence.

Questionnaire

A total of 210 participant questionnaires were suitable for analysis (68.9% effective response rate). Survey data generally reinforced the impressions generated through the interviews. As a result, the focus here is on items that provide further information.

Participants generally found the feedback was easy to interpret (61% agreement), many felt they could monitor their knowledge with the progress test (52% agreement), and half considered that the feedback provided useful information about their progress (49% agreement).

Generally the tests were more favourably viewed than the feedback. Approximately half the participants agreed that the progress tests helped them improve their knowledge (47% agreement) while only 22% agreed (most responding neutrally) that the feedback was useful for their learning.

Perceptions of feedback utility appeared to bear little relationship to feedback group. The two quotes below illustrate the typical dichotomy existing in the comments made by participants within the same feedback group – in this case those receiving normative comparisons.

> "It was good to see which areas of medicine I needed to work on more, and it was useful in indicating success in previous rotations, as well as a marker for how well others are doing as motivation" (N)

> "I didn't use it. I was given very little feedback and found it to be unhelpful in guiding my study" (N).

Most participants also reported that testing and feedback had not changed their study behaviours (15% agreement). The 32 participants who agreed that their study behaviour had changed (15% of respondents to that item) were spread across all feedback groups (10, 10 and 12 from the clinical presentation, clinical presentation plus and normative feedback groups respectively). In the free-text elaboration of this item, participants reported studying more, using "clinical scenarios", "practice questions" and "past [*exam*] papers". Others reported a change in general content of their study; "[*it*] highlighted the broad areas (Medicine/Surgery/Ambulatory) that needed revising" (N). Some described attempts to recall test items for later review and to focus on tested concepts during subsequent clinical learning.

> "I would look up things directly after the test that I could remember not knowing during the test" (CP)

"The tests helped me hone in on what's important and take note of those concepts when they came up in the ward rounds/clinics" (N)

Agree responses for each questionnaire item by feedback group and baseline quartile are presented in Appendix S4. A GLMM revealed three items with small but statistically significant effects with respect to baseline quartile group, and when the effect of baseline quartile group was removed, an effect of feedback group on agree responses was identified for two items (Appendix S5). The largest of these related to the ease of interpretation of the feedback reports ($P < 001$): pairwise comparisons revealed that the odds of agreeing that the feedback was "easy to interpret" was 3.6 times higher for normative feedback group participants compared with the clinical presentation group (95% CI 0.3, 41.8; $P = 0.016$) and were 4 times higher for normative feedback group participants compared with the clinical presentation plus group (95% CI 0.3, 50.0; $P = 0.012$).

Discussion

In this study, our aim was to clarify how medical students respond to different forms and content of written feedback. Our findings suggest that student responses to written feedback are quite variable and are influenced by the type of feedback provided and by students' relative position within their learning cohort.

Lower quartile group participants receiving normative feedback had an academic disadvantage of about 5% on subsequent tests compared to participants receiving other types of feedback. Participants who received this type of feedback, independent of their baseline test quartile group, also spent around one hour more than other participants undertaking *general study in medicine* prior to their summative examinations, which coincided with one hour less engaging in *study for an exam or test*. While the combined time spent in these two activities is roughly equivalent, the tendency to classify study differently may suggest those receiving normative feedback had a tendency to engage in more generalised (and less strategic and content specific) exam study.

Participants in the lowest two quartile groups, i.e. those performing below the median at baseline, spent more time studying for their exams than participants in the third quartile group. This is consistent with the view that the poorer performance of lower quartile group

students is not related to lack of (reported) effort, but the study they engage may be less effective.

Analysis of the questionnaire and interview responses indicates that many participants engaged in reflection on their learning - some as a result of internal feedback (through their interaction with the test itself) and others as a result of the external feedback (in the form of the feedback reports). Self-regulatory learning theory posits that feedback is used by a learner to guide self-reflection on performance relative to goals, and to assist in deciding whether to continue or modify learning strategies in the future. [29] Among the better performing participants, an accurate reflection on their test performance should lead them to see their progress as satisfactory with only minor (if any) changes in learning and study strategies required. However, given that tests were pitched at graduation-level, even these students should be able to identify some room for improvement. In reality, a very small number of participants (N=33) reported changing their study behaviour. The fact that these participants were from all feedback groups suggests the test experience served as a source of internal feedback for these participants, or that none of the specific features of the various feedback reports was more effective at stimulating behavioural change than the others.

The approximately 5% poorer performance on subsequent test scores for lowest quartile group participants who received normative feedback does, however, indicate a less effective, or even detrimental impact of this form of feedback (despite its ease of interpretation). Responses to feedback are complex and not always positive. [8] Reasons for a negative impact are more diverse than just the content of the feedback message itself, and can include the specific attitudes and beliefs of the learner, their relationships with the feedback provider and the wider educational and sociocultural environment of the learner. [30] Lower performing medical students may lack the capacity to accurately assess their ability, [31,32] in which case, feedback indicating low performance relative to their peers could have been unexpected.

Furthermore, emotions are an important mediator in perception of feedback information, [33] receptivity to feedback [34] and of misalignment between self-assessment and performance. [35,36] For lower performing students then, the inclusion of normative data in feedback might trigger emotional responses that lead to reduced confidence and create uncertainty in their

ability. Possibly the resultant decrease in their self-confidence had a subsequent impact on their performance rather than any decrease in their study effort.

Participants in the normative feedback group did not have access to the lists of correct and incorrect clinical presentations for each test item received by students in the other two groups. It is possible that this kind of direction is useful for lower-scoring participants, and the absence of this information in normative feedback reports produced a disadvantage (through lack of direction) for these students in terms of subsequent progress test performance.

Lower-performing medical students tend not to make learning goals aligned with their academic weaknesses and, therefore, may require more explicit direction than higher-achieving students. [7] Indeed, research involving medical students, consistent with other student groups, suggests that high-performing medical students are more likely to utilise self-regulated learning strategies such as strategic thinking and actions during learning. [29] The overall grade and percentage by rotating term information provided in the normative feedback reports may not have been explicit enough to direct lower performing students to areas of weakness – i.e. their reports may not have provided sufficient guidance for self-regulation. In contrast, the more detailed content (list of clinical presentations) in the reports of the other two feedback groups may have helped lower quartile group participants in these groups to identify misconceptions and hence offered more useful direction for learning. Again from a self-regulatory perspective, it appears that the feedback with more detailed content information (perhaps informing goal setting and strategic planning) was more helpful than that describing relative ability, which perhaps impacted affect and self-confidence to a greater extent than poor performance without normative feedback.

Limitations and Strengths

Although it was situated within a single medical curriculum, the study was conducted across nine different clinical schools in inner and outer metropolitan and rural settings. The inclusion of multiple sites, each with its own medical and learning and teaching context, substantially improves the generalizability of the study's findings relative to a single site.

However the *in situ* medical education setting of this study poses a number of challenges. In particular, our preferred study design would have included control groups that 1) did not

take progress tests and 2) took progress tests but did not receive feedback. However, neither was permitted by the responsible ethics committee because of a perceived potential for academic disadvantage.

The classification of participants into baseline test performance quartile groups relative to each other can only be considered a proxy for ability level, but this does provides a useful and effective lens through which to explore differential responses to feedback, and to consider groups of students who arguably need feedback more than others.

Implications

Our finding that student responses to written feedback are influenced by the type of feedback provided and by students' relative position within their learning cohort is important. This differential degree of performance improvement by lower baseline test performance participants who received normative information, perhaps mediated by differential focus of study activity, necessitates further investigation. In particular, given students' desire for normative information, we must explore whether the apparent negative impact of normative data might be reduced or overcome by providing more detailed item-level elaboration or through support of learners' self-regulation such as help to inform self-assessment [37] and facilitate reflection. [38] This work would help to identify whether normative feedback itself has a negative impact on these students, or if other factors (such as lack of goal setting and strategic planning) are involved, and if so to what extent they contribute to the effect.

We asked, how do different forms of automated written feedback influence study behaviour and further acquisition of knowledge in medical students? Our findings suggest that the act of sitting a test itself can be a stimulus for study activity and that a number of students will be motivated to study on the basis of this experience, independent of the type of feedback they receive. Our findings also suggest the need for care when providing students with automated post-test feedback which contains normative information confirming poorer performance relative to peers as this appears to have a negative impact on subsequent acquisition of knowledge when compared with feedback containing clinical presentations. These findings emphasise the importance of further work to explore efficient methods of written feedback that improve all medical students' ability to self-regulate their learning independent of their level of performance relative to their peers.

References

1. Zimmerman BJ. Self-regulated learning and academic achievement: An overview. *Educ Psychol*. 1990;25(1):3-17.

2. Murdoch-Eaton D, Whittle S. Generic skills in medical education: Developing the tools for successful lifelong learning. *Med Educ*. 2012;46(1):120-128.

3. Nicol DJ, Macfarlane-Dick D. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Stud High Educ*. 2006;31(2):199-218.

4. Eva KW, Regehr G. Effective feedback for maintenance of competence: From data delivery to trusting dialogues. *CMAJ*. 2013;185(6):463-464.

5. Urquhart LM, Rees CE, Ker JS. Making sense of feedback experiences: A multi-school study of medical students' narratives. *Med Educ*. 2014;48(2):189-203.

6. Butler R. Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *J Educ Psychol*. 1987;79(4):474-482.

7. Chang A, Chou CL, Teherani A, Hauer KE. Clinical skills-related learning goals of senior medical students after performance feedback. *Med Educ*. 2011;45(9):878-885.

8. Kluger AN, DeNisi A. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychol Bull*. 1996;119(2):254-284.

9. Butler AC, Karpicke JD, Roediger HL. The effect of type and timing of feedback on learning from multiple-choice tests. *J Exp Psychol App*. 2007;13(4):273-281.

10. Bangert-Drowns RL, Kulik CC, Kulik JA, Morgan M. The instructional effect of feedback in test-like events. *Rev Educ Res*. 1991;61(2):213-238.

11. Boud D, Molloy E. Rethinking models of feedback for learning: The challenge of design. *Assess Eval High Educ*. 2013;38(6):698-712.

12. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: From methods to programmes. *Med Educ*. 2005;39(3):309-317.

13. Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Med Educ*. 1985;19(3):238-247.

14. Creswell JW, Plano Clark VL, eds. *Designing and conducting mixed methods research.* 2nd edition ed. Thousand Oaks, California: Sage International; 2011.

15. Tweed M, Purdie G, Wilkinson T. Low performing students have insightfulness when they reflect-in-action. *Med Educ*. 2017;51(3):316-323.

16. Swanson DB, Holtzman KZ, Butler A, Case Western Reserve University School Of Medicine Cumulative, Achievement Testing Study Group. Cumulative achievement testing: Progress testing in reverse. *Med Teach*. 2010;32(6):516-20.

17. Wilkinson TJ, Wells JE, Bushnell JA. Using a diary to quantify learning activities. *Med Educ*. 2005;39(7):657-664.

18. Wade L, Harrison C, Hollands J, Mattick K, Ricketts C, Wass V. Student perceptions of the progress test in two settings and the implications for test deployment. *Adv Health Sci Educ*. 2012;17(4):573-583.

19. van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Med Teach*. 1996;18(2):103-109.

20. Arnold L, Willoughby TL. Ideas for medical education - the quarterly profile examination. *Acad Med*. 1990;65(8):515-516.

21. Wrigley W, Van Der Vleuten CPM, Freeman AC, Muijtjens A. A systemic framework for the progress test: Strengths, constraints and issues: AMEE guide no. 71. *Med Teach*. 2012;34(9):683-697.

22. Lynn MR. Determination and quantification of content validity. *Nurs Res*. 1986;35(6):382-385.

23. Polit DF, Beck CT. The content validity index: Are you sure you know what's being reported? critique and recommendations. *Res Nurs Health*. 2006;29(5):489-497.

24. Shute VJ. Focus on formative feedback. *Rev Educ Res*. 2008;78(1):153-189.

25. Hattie J, Timperley H. The power of feedback. *Rev Educ Res*. 2007;77(1):81-112.

26. Eva KW, Armson H, Holmboe E, et al. Factors influencing responsiveness to feedback: On the interplay between fear, confidence, and reasoning processes. *Adv Health Sci Educ*. 2012;17(1):15-26.

27. International Committee of Medical Journal Editors. Recommendations for the conduct, reporting, editing and publication of scholarly work in medical journals. http://www.icmje.org/recommendations/. Updated 2014. Accessed June/15, 2015.

28. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology*. 1990;1(1):43-46.

29. Cleary TJ, Sandars J. Assessing self-regulatory processes during clinical skill performance: A pilot study. *Med Teach*. 2011;33(7):e368-e374.

30. Sargeant JM, Bruce D, Campbell CM. Practicing physicians' needs for assessment and feedback as part of professional development. *J Contin Educ Health Prof*. 2013;33:S54-S62.

31. Langendyk V. Not knowing that they do not know: Self-assessment accuracy of third-year medical students. *Med Educ*. 2006;40(2):173-179.

32. Woolliscroft JO, TenHaken J, Smith J, Calhoun JG. Medical students' clinical self-assessments: Comparisons with external measures of performance and the students' self-assessments of overall performance and effort. *Acad Med*. 1993;68(4):285-294.

33. McConnell MM, Eva KW. The role of emotion in the learning and transfer of clinical skills and knowledge. *Acad Med*. 2012;87(10):1316-1322.

34. Fishbach A, Eyal T, Finkelstein SR. How positive and negative feedback motivate goal pursuit. *Soc Personal Psychol Compass*. 2010;4(8):517-530.

35. Kluger AN, Van Dijk D. Feedback, the various tasks of the doctor, and the feedforward alternative. *Med Educ*. 2010;44(12):1166-1174.

36. Eva KW, Munoz J, Hanson MD, Walsh A, Wakefield J. Which factors, personal or external, most influence students' generation of learning goals?[editorial]. *Acad Med*. 2010;85(10) (Supplement, RIME: Proceedings of the Forty-Ninth Annual Conference November 7-November 10, 2010):S102-S105.

37. Sargeant JM, Armson H, Chesluk B, et al. The processes and dimensions of informed self-assessment: A conceptual model. *Acad Med*. 2010;85(7):1212-1220.

38. Sargeant JM, Mann KV, van der Vleuten C, Metsemakers J. "Directed" self-assessment: Practice and feedback within a social context. *J Contin Educ Health Prof*. 2008;28(1):47-54.

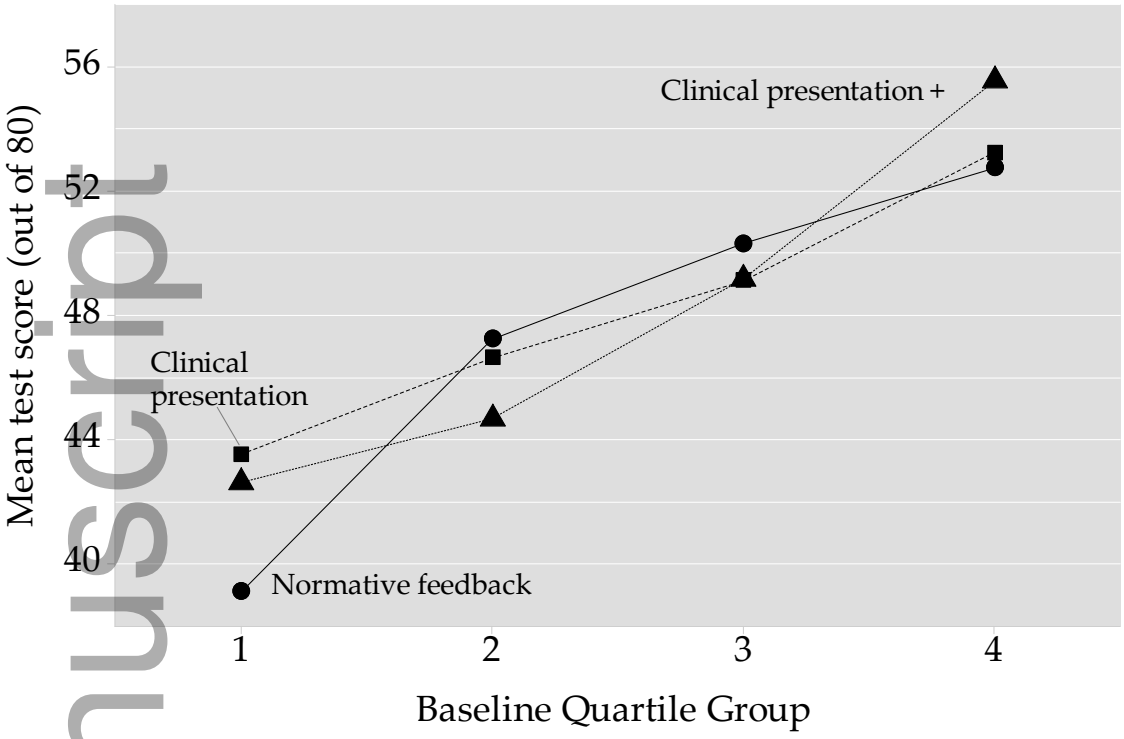

Figure 1: Progress test results (averaged over tests 2, 3 & 4) by baseline quartile group & feedback group

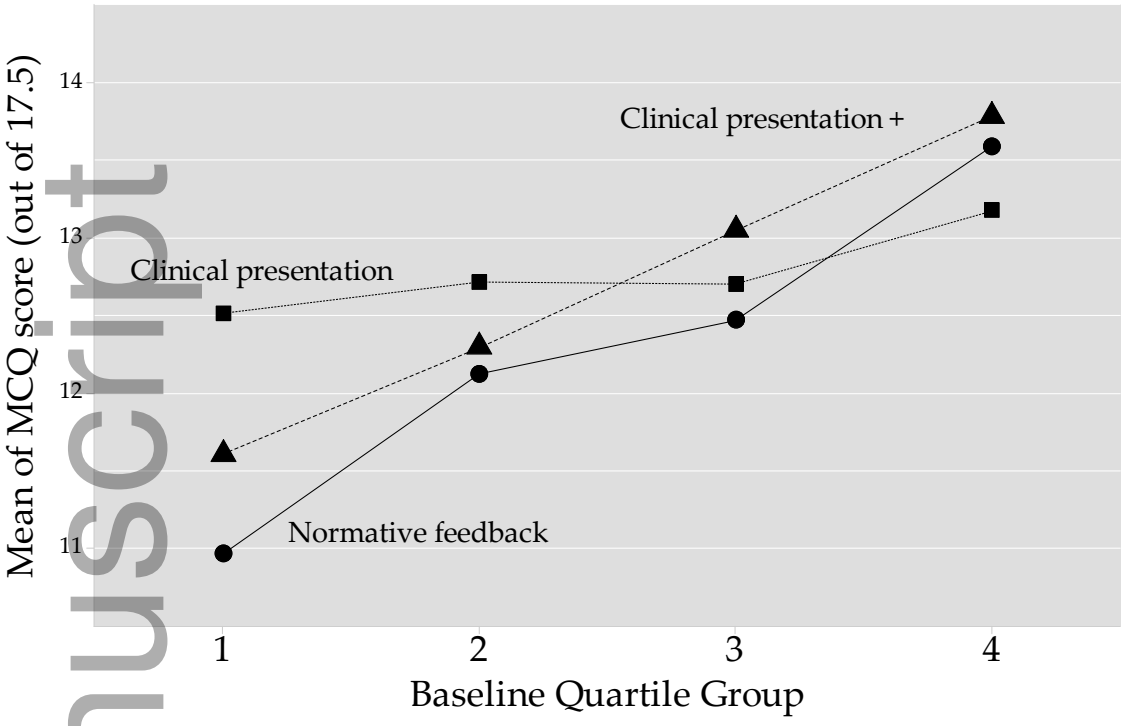

Figure 2: End of year MCQ score by baseline quartile group and feedback group

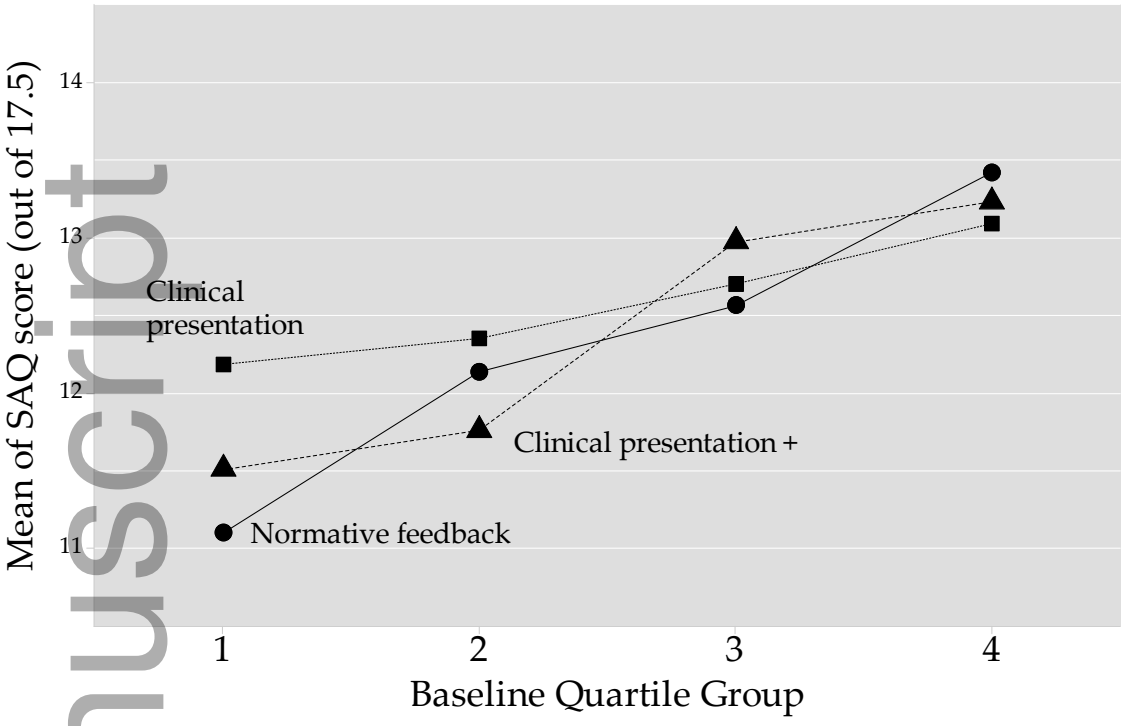

Figure 3 End of year SAQ score by baseline quartile group & feedback group

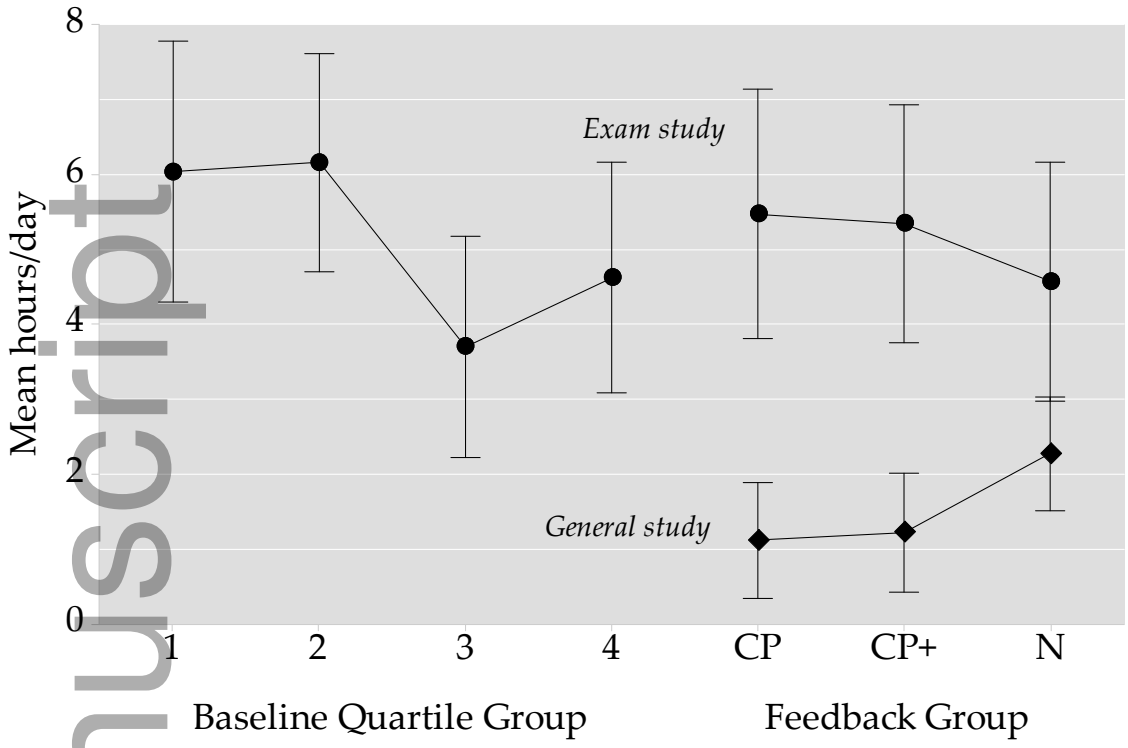

Figure 4: Estimated mean hours/day per study activities by baseline quartile group & feedback group (test occasion 4)