What Types of Feedback do Undergraduate Chemistry Students Give Each Other? A Case Study from Singapore

Norman T-Lon Lim[a], Yew-Jin Lee[a] and Peter Peng Foo Lee[a]

Corresponding author: Norman T-Lon Lim (norman.lim@nie.edu.sg)
[a]National Institute of Education, Nanyang Technological University, Singapore 637616, Singapore

Keywords: feedback, formative assessment, multiple-choice questions

Abstract

This study was part of a larger project to improve learning of undergraduate chemistry in Singapore through the use of self-authored three-tier multiple-choice questions (3TMCQs) and the giving/receiving of peer feedback. Specifically, we examined the quality of written feedback based on the classification by Hattie and Timperley (2007) that year 2 to 4 learners (N=31) gave each other on responses in their 3TMCQs (N=466 administered). It was found that the most common type of voluntary feedback given by test-makers was task (& self), followed by process (& self), self alone, and lastly regulation (& self) levels over seven chemistry courses. In addition, question type (based on revised Bloom's Taxonomy) had a marginal effect on the quality of feedback received; instead, items answered incorrectly garnered higher quality feedback and were four times more important than the cognitive level of questions. Feedback quality given by more experienced students was also no better than those given by less experienced ones. While there is growing evidence supporting the self-authoring of questions and giving/receiving peer feedback to enhance learning at undergraduate levels, further research is warranted into the types of peer feedback that learners may receive when attempting different question formats.

Introduction

Learning science at the undergraduate level is often described as difficult and challenging even for more talented students (Wieman, 2017). Over the years, universities have employed diverse teaching methods (e.g., team-based learning, problem-based learning, active learning, peer instruction, etc.) and resources such as educational technology (use of Massive Open Online Courses, gamification, response-based systems, learning analytics, Augmented Reality, etc.) to enhance the learning experience of undergraduates, albeit with varying results (e.g., Crouch & Mazur, 2001; NRC, 2011; Valverde-Berrocoso, et al., 2020). What perhaps has been an increasing focus of attention is improving learning through assessment, which has typically taken a back seat compared to the intensity of reform efforts in curriculum and pedagogy. Interest in the use of assessment for learning has risen since the late 1990s with the publication of "Inside the black box" by Black and Wiliam (1998) that made an argument for formative assessment (FA) and feedback as both precursors and outcomes of learning.

So instead of merely attempting to give a grade, FA shifts instructor attention to unpack/expose student thinking in order to adjust or realign one's teaching. Feedback is also widely acknowledged as one of the most influential FA factors that improves learning and can be given by an instructor, peers, self, or other authoritative sources such as reference texts (Hattie & Timperley, 2007). Feedback assists learning during one or all of the following inter-related aspects: knowing what is the current stage of one's learning; what are the attainment goals/learnings to strive towards; and what are the next steps of action or knowing how well one is proceeding (Black & Wiliam, 2010). In peer assessment that is another integral part of FA (Andrade, 2019), learners are encouraged to learn the disciplinary criteria for judging others'

work and thereby gain insights into their own thinking with increased responsibility for the learning of others (see Ashenafi, 2017).

Research on FA has likewise caused a rethink concerning the process of constructing and grading of tests of conceptual knowledge, which is a standard and unremarkable aspect of the work of university instructors everywhere. This process usually rests on the unquestioned expertise and role of the lecturer, but there are arguably more interesting outcomes when some of these responsibilities are turned over to the students. Students are meant to be graded during academic assessment; in the US, although likely true for other regions too, it is acknowledged that "[g]rades are a significant component within the American system of education. They are used to determine class placement, scholarships, and college admissions" (Randall & Engelhard, 2010, p. 1373).

Yet, studies have shown the many benefits that accrue when students are involved—partly or fully—in instructional processes such as test/quiz construction in science (e.g., Aflalo, 2021; Bottomley & Denny, 2011; Denny, Hamer, Luxton-Reilly, & Purchase, 2008; Hardy et al., 2014). Learners are now better able to grasp complex disciplinary content, identify one's gaps in knowledge and improve exam techniques (e.g., Guilding, Pye, Butler, Atkinson, & Field, 2021) when they are required to become more confident and think deeply about content matter through the interplay of their roles as test-maker and test-taker. These positive results may occur, for instance, in large classes (~1,200 students) with significantly favourable outcomes among more engaged learners (Hancock, Hare, Denny, & Denyer, 2018) or even within a context of minimal faculty involvement (Galloway & Burns, 2015). At no time does this imply that self-authoring strategies of questions are problem-free; undergraduates have expressed unhappiness over the allocation of topics in the course perceived as being "easier" and showed lack of understanding how their self-generated questions could be revised or improved among other challenges (Doyle, Buckley, & Whelan, 2019). Given the reported gains in learning from the use of feedback as well as self-authoring of test items, this formed the basis of our investigations in this paper. We now describe some past research on formative assessment (including peer assessment & the role of feedback) and three-tier multiple choice items in the literature review section that follows.

Literature Review

Formative assessment & feedback

Of all the influential educational practices that are currently available—and there are indeed many—nothing has quite so captured the imagination of educators everywhere as formative assessment (Morris, Perry, & Wadle, 2021). It is a widely implemented learning practice with both researchers and policymakers emphasizing its enactment in all classrooms and grade levels regardless of subject discipline (e.g., the Higher Education Academy https://www.advance-he.ac.uk/). For apparently relatively small investments in teacher training, large returns in terms of improved student learning and engagement have been among its compelling selling points, especially for the weakest learners in school contexts (Black & Wiliam, 1998). One of its major components, feedback, has been reported to be the number one factor open to school teachers (themselves a key factor for instruction) that leads to increased student achievement, an impressive claim by any account (Hattie, 2009). A study among masters students at one British university stood out in the review of FA and feedback by Morris et al. (2021). These authors described a study by Bandiera et al. (2015) that showed how instructor feedback on assessments given throughout the year improved subsequent test scores amounting to as much as 13% of a standard deviation. When feedback is given by peers,

there are similar benefits for learning although one should weigh "the complexities of implementing peer feedback effectively and the potential cost of substituting instructional time for peer feedback" (Morris et al., 2021, p. 18).

Nonetheless, characterizing what is FA with precision has been a harder task (Bennett, 2011). An accessible definition of FA explains it as what is "carried out during the instructional process for the purpose of adapting instruction to improve learning" (Penuel & Shepard, 2016, p. 788). It consists of a range of actions that instructors intentionally use to i) find out what students know (elicitation), ii) make decisions on the next courses of action (interpretation & judgment), and iii) to adjust their instruction to close the learning gap accordingly (taking action) (Wiliam & Black, 1996). For the student, FA generally results in higher achievement gains although its "benefits may vary widely in kind and size from one specific implementation of formative assessment to the next, and from one subpopulation of students to the next" according to Bennett (2011, p. 20). The theory of learning behind FA is indeed complex and not fully understood for learners may choose to consciously ignore helpful advice due to personal reasons or to protect their sense of self-esteem (Brown & Harris, 2013). Also, FA is distinguished from summative assessment, which is the process of testing or evaluating what learners know or can do after instruction has occurred. Summative assessment is commonly associated with grading or sorting functions for checking the extent and quality of learning although a hard distinction between this and FA is sometimes hard to make (Black & Wiliam, 2010). Keeping in mind the purposes, goals, timing, and how that information is used are probably the best criteria to distinguish formative and summative assessments (see Scriven, 1967).

While FA practices can occur over seconds, minutes, days or even weeks, the bulk of FA can be found during real-time teaching such as in the questioning process. Science educators have reported that among other features, FA is a highly responsive process characterised by "being ongoing; dynamic and progressive; informal; interactive; unplanned as well as planned; reactive as well as proactive; with the class, group, or individual; involving risk and uncertainty" as well as involving a number of dilemmas for practice; there are no "best" formulas for teachers to enact (Bell & Cowie, 2001, p. 547). Working in what has been called "medium cycle" FA (adjusting teaching based on feedback over days to a month; Wiliam, 2006) is very useful as this allows for intentional slipping in of more structured FA techniques into lesson plans. Popular strategies and techniques for FA (e.g., ranking exemplars, using mini-white boards, anchoring questions, using randomization devices, peer checklists, etc.) are listed by Wiliam (2009) while attempts to unpack what is meant by feedback are ongoing (Shute, 2008). Given such a multitude of options available in FA strategies, it has thus inspired us to explore the combined use of self-authored three-tier multiple choice items and the giving/receiving of peer feedback in our current study.

During peer assessment, which is an integral part of FA, learners evaluate the work of peers of similar status (in ability) in terms of the "amount, level, value, worth, quality, or success of the products or outcomes of learning" (Topping, 1998, p. 250). Peer assessment underscores how peers can act as valuable resources for each others' learning; they can effectively guide, instruct, and point out weakness in their fellow students that feels different from when an instructor is doing the very same actions. A study of university engineering undergraduates who participated in peer and self-assessment exercises showed that peer assessment compared favourably with expert assessors although their scoring tended to be overestimated and self-assessment scores showed lower reliability as compared to experts (Power & Tanner, in press).

Levels of feedback

It is to be remembered that not all feedback is the same: Hattie and Timperley (2007) reported that feedback can actually occur at four levels. These have been termed the i) task (how well tasks are understood, knowing the basic ideas, concepts, etc.), ii) process (processes/methods/formulae to understand/perform tasks), iii) self-regulation (metacognitive advice/actions), and iv) self (personal evaluations of learner) levels. These authors suggest that self-level feedback is generally not academically productive whereas process or self-regulation feedback have long-term benefits for learning because they help learners know what to do and how well they are actually doing it. It is not claimed that lower levels of feedback such as task feedback are not important or inferior to higher ones because knowing the basic facts, what is right/wrong, and how to acquire more information specific to complete a task builds the foundation of further knowledge. Conversely, regulation level feedback can play a critical role with fairly advanced or cognitively ready learners because they possess sufficient working knowledge of relevant content and processes. Only at these levels then can they meaningfully engage in sense-making across contexts and reflect on their own learning. This was indeed the case with a recent study on the effects of student-generated feedback on student-generated questions during a 7th Grade language unit. It was found that there was greater use of cognitive and metacognitive strategies associated with the former as well as enhanced perspective taking abilities among students (Yu & Wu, 2020). Of interest, students who gave higher-quality feedback as categorised by the instructor also demonstrated greater academic performance as determined by experts. This finding finds no conflict with other researchers who reported that academically successful students regard assessment as a means of enhancing their own learning (Brown, Peterson, & Irving, 2009).

Three-tier multiple choice questions

Treagust (1988) first modified the popular multiple-choice question (MCQ) into a two-tier format where test-takers were now requested to indicate their confidence levels on their answers. Since then, researchers have adapted it into three- as well as four-tier formats (see Caleon & Subramaniam, 2010), although this study focuses only on the three-tier MCQ (3TMCQ). This format consists of the typical question stem with answer key and distractors (first tier: reason), a confidence rating with variable Likert-style options (e.g., "unsure," "sure," "very sure") (second tier: confidence), and an open-ended (or selection type) question asking for reasons for choosing their answer (third tier: explanation). The advantages for using tiered formats are that they allow instructors to gauge how certain students are of their responses and assess if they hold onto any possible misconceptions or understand content at a deeper level rather than merely guessing. In addition, they allow learners to explain the reasons for their answers, which assist the instructor in knowing if a correct answer in the first tier was based on the application or recall of knowledge rather than happenstance. In this current journal, researchers have likewise employed tiered MCQ formats; Hill, Sharma, O'Bryne and Airey (2014) have used 3TMCQ to assess representational fluency while others have found the two-tier formats useful to uncover science misconceptions among university students (Kamcharean & Wattanakasiwich, 2016; Reinke, Kynn, & Parkinson, 2019). It is clear that 3TMCQ formats possess two major strengths: they are effective tests for conceptual knowledge and, as FA tools, they make thinking visible both to instructors as well as to the learners themselves. These strengths are enhanced when these items are self-authored and when test-makers are given opportunities to give feedback to assist their peers in learning, which were the potential learning outcomes that prompted our investigation.

Research Questions

Our research study was part of a larger project with undergraduate students in Singapore that sought to improve the learning of chemistry through: (1) cycles of self-authored test-construction in the subject, and (2) the giving and receiving of written feedback on (1) by peers. Our action research attempted the second focus, which is to ask what types of feedback chemistry undergraduates in Singapore give to their peers when they construct their own 3TMCQ. The quality of feedback students give each other will be classified according to the four levels by Hattie and Timperley (2007), which we related to the question type or cognitive levels based on the revised Bloom's Taxonomy (Anderson et al., 2001). For the three specific research questions (see Fig. 1), it is hypothesised that:

1. students are more likely to answer 3TMCQs correctly for easier types of questions.
2. higher-quality feedback will be provided for easier types of question and/or when answers are wrong.
3. students who had participated in this item-creation exercise more times will give higher-quality feedback over time.

Methodology

Participants

Each student (N=31) attending mandatory undergraduate chemistry courses (over Years 2-4) spanning academic years 2018 to 2020 was pre-assigned a specific chemistry topic to construct two original three-tier MCQs (3TMCQs) as well as craft their worked solutions. Of these 31 students from the same teacher education college within this research university in Singapore, four completed this exercise three times, two experienced it twice while the rest experienced it once during their undergraduate program. There was thus a total of seven chemistry courses that experienced this pedagogical intervention although some courses were repeated ones over the years. After university ethics approval was obtained, the nature of this study was explained at the beginning of each course to the students. Participation was entirely voluntary and it was explained there was no penalty for opting out—no student from any course declined to contribute in the research. All courses (class size 3-7 students) were also taught by the same instructor who was part of the research team.

Instruments

Over the duration of any single chemistry course, students administered their 3TMCQs in the class once to fellow classmates following teaching of the concepts by the instructor and then marked these items without assistance. While students as test-makers were strongly encouraged at the beginning of the course to give appropriate and relevant written feedback to their classmates, this was not mandated and left voluntary. After receiving the marked scripts, all test-takers could likewise give written feedback, if they wished, on any comments from the test-makers that they had received. Note that all 3TMCQs were checked for conceptual accuracy and technical quality by the instructor before they were administered to their peers in the class.

Data Analysis

We collected a total of 466 completed 3TMCQs with or without feedback and then classified all the 3TMCQs according to one of the six cognitive levels in revised Bloom's Taxonomy (i.e., termed as "question type"; in increasing cognitive difficulty or challenge: Remember, Understand, Appy, Analyse, Evaluate, and Create). Of these self-authored questions, 314 were answered correctly, of which 37 items did not receive any form of feedback. Of the 152 items that were answered incorrectly, 9 items did not receive any feedback. Samples of the coding

are attached in Supplementary Materials. Two of the authors also independently classified the levels of written feedback (if any; termed as "quality of feedback") given by the test-maker based on Hattie and Timperley (2007). If there were disputes in our coding, we discussed these discrepant codes until we reached consensus. In our analysis, self-level feedback that was present in conjunction with other types of feedback was recorded, but not deemed consequential except when it formed its own category of feedback. Thus, feedback that contained both process and self-levels were regarded as located at the process level. Also, if feedback contained both task and regulation feedback, for example, it was regarded as regulation feedback that is the "higher" level based on Hattie and Timperley (2007). Some examples of the coding for feedback by test-makers are shown in the Supplementary Materials, while Figure 1 summarises the study and the various data collected.

We then performed model selection (Burnham & Anderson, 2004) to investigate the three research questions. Model selection was conducted using information-theoretic approach and second-order Akaike's Information Criterion (AICc), and models with ΔAICc < 2 were selected as parsimonious models (Burnham & Anderson, 2010). Briefly, AICc provides a measure of model fit based on the Kullback-Leibler information loss. Firstly, to investigate if the correctness of answer was influenced by the question type, model fit of the model incorporating question type as a predictor was compared against a null model (i.e., without any predictor). Secondly, to investigate if the quality of feedback was influenced by correctness of answer and/or question type, a suite of candidate cumulative-link mixed models was produced to represent the various hypothesised relationships. Thirdly, model fit of the model incorporating experience of test-maker (i.e., number of times test-maker did the 3TMCQ exercise) as a predictor was compared against a null model to examine if test-makers give higher-quality feedback over time. All data were grouped by the test-maker and the course as random effects in the models to avoid pseudoreplication. The quality of feedback given was coded as an ordinal response (i.e., Self < Task < Process < Regulation). Relative variable importance (see Burnham & Anderson, 2010) was also calculated for predictor variables when multiple models were selected as parsimonious models. The analyses were conducted using packages ordinal (Christensen, 2019) and MuMIn (Barton, 2009) in software R (R Core Team, 2018).

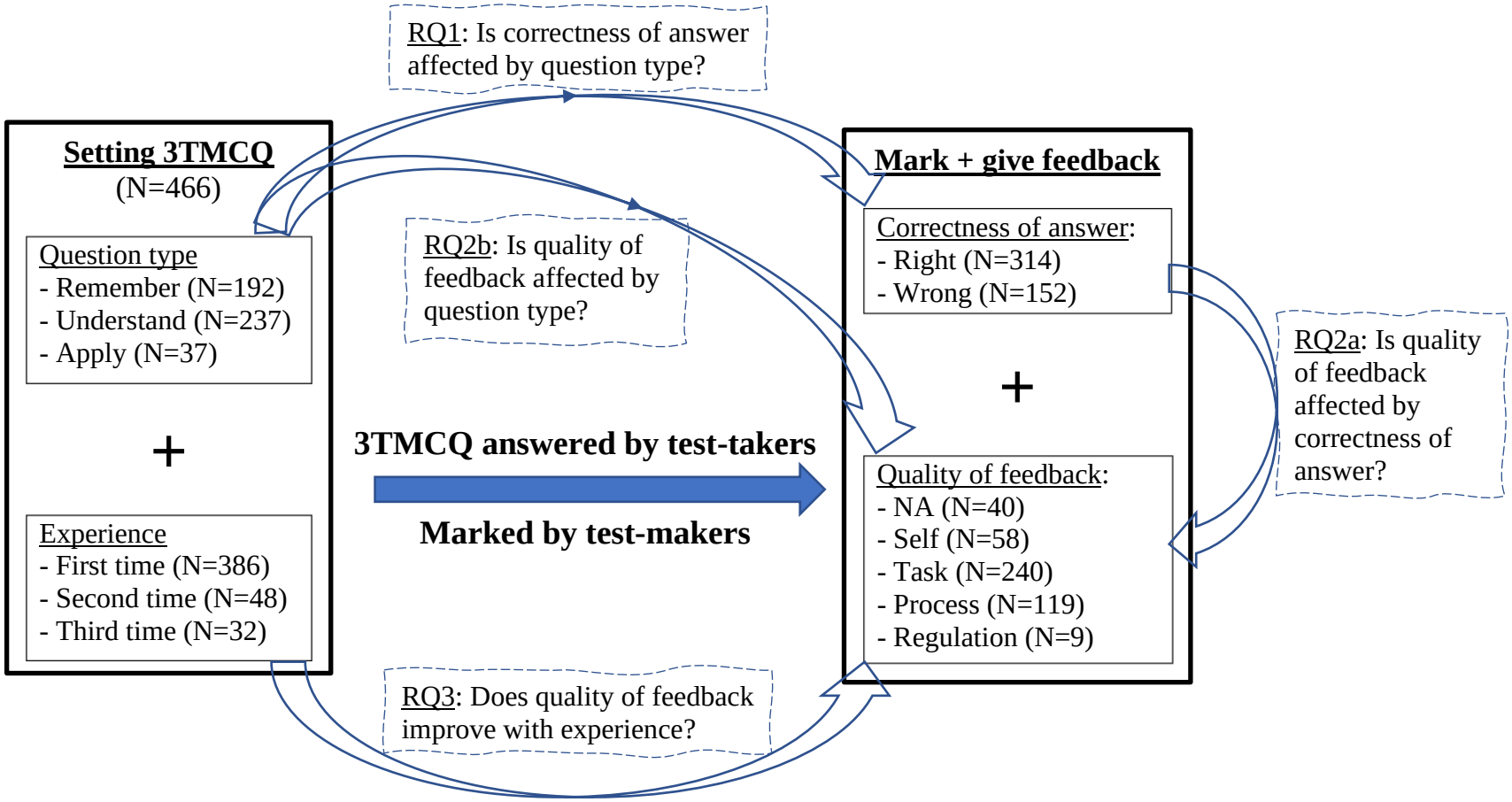

Figure 1. Schematic summarising the study on three-tier multiple-choice questions (3TMCQ) and its main research questions (RQ).

Results

Table 1. Frequency of feedback classified by type of question over correctly and incorrectly answered three-tier multiple-choice questions (3TMCQ). Items that did not receive any feedback were clustered together with those coded at the self-level with values indicated in brackets.

	Type of written feedback on 3TMCQ that were answered correctly				
Question type	**Task (& self)**	**Process (& self)**	**Regulation (& self)**	**Self (no feedback in brackets)**	**Totals (% in brackets)**
Remember	58	27	6	29 (11)	120 (38.2%)
Understand	78	36	0	61 (26)	175 (55.7%)
Apply	9	7	1	2 (0)	19 (6.1%)
Totals	145	70	7	92 (37)	N=314
	Type of written feedback on 3TMCQ that were answered incorrectly				
	Task (& self)	**Process (& self)**	**Regulation (& self)**	**Self (no feedback in brackets)**	**Totals (% in brackets)**
Remember	46	27	2	2 (0)	77 (50.7%)
Understand	39	14	0	4 (3)	57 (37.5%)
Apply	10	8	0	0 (0)	18 (11.8%)
Totals	95	49	2	6 (3)	N=152

Is correctness of answer affected by question type?

After coding, all students were found to have authored 3TMCQs that were exclusively at the first three cognitive levels of revised Bloom's Taxonomy, namely, Remember, Understand, and Apply (Table 1). For this research question, the question type had a strong influence over the correctness of answer given by test-taker as seen by the most parsimonious model (ΔAICc = 0; Table 2). A closer examination of the coefficients for the predictor question type (i.e., QnType) revealed that Understand questions had the highest proportion of correct answers (75.4%; N=232 questions in Table 1), followed by Remember questions (60.9%; N=197 questions) and then Apply questions (51.4%; N=37 questions). This result was unexpected as Remember-level questions should have been more manageable for getting an answer correct by test-takers compared to Understand and Apply levels.

Table 2. Model selection of question type (i.e., Remember, Understand, or Apply; QnType) influencing the correctness of answer by test-taker. *k*: number of predicted parameters, logLik: log-likelihood, AICc: second-order Akaike's Information Criterion, weight: Akaike's weights.

Model	*k*	logLik	AICc	ΔAICc	weight
QnType	4	−283.5	575.1	0.00	0.989
Null	2	−290.0	584.1	9.05	0.011

Is quality of feedback affected by correctness of answer and/or question type?

The quality of the feedback was only influenced by the correctness of the answer by test-taker (and not the question type) as evident from the most parsimonious model (ΔAICc = 0; Table 3). Additionally, as the predictor correctness of answer (i.e., AnsCor) had a negative coefficient in the most parsimonious model, this meant that test-makers gave lower-quality feedback whenever test-takers gave the correct answers. This finding stands to reason; test-makers may not feel compelled to give elaborate or detailed feedback when the test-taker appeared to have understood the concept tested and thus provided feedback that merely repeated the expected facts, concepts, or procedures.

Table 3. Model selection of predictors influencing the quality of feedback given by test-makers. AnsCor: correctness of answer, QnType: question type, *k*: number of predicted parameters, logLik: log-likelihood, AICc: second-order Akaike's Information Criterion, weight: Akaike's weights.

Model	*k*	logLik	AICc	ΔAICc	weight
AnsCor	6	−439.5	891.2	0.00	0.745
AnsCor + QnType	8	−438.5	893.3	2.15	0.255
Null	5	−447.7	909.3	18.15	0.000
QnType	7	−449.7	909.6	18.41	0.000

Conversely, the difficulty of question did not influence the type of feedback as models that incorporated the variable QnType had ΔAICc > 2. This can also be seen by the relative variable importance (RVI) of the two predictors where the correctness of answer is four times more important than the difficulty of question (RVI: AnsCor = 1.00; QnType = 0.255).

Does experience of the test-maker affect quality of feedback given?

Experience of the test-maker did not have a meaningful influence on the quality of feedback given as the null model was the most parsimonious model (i.e., ΔAICc = 0; Table 4). This

meant that the quality of feedback given by experienced students were not better than those given by inexperienced students.

Table 4. Model selection of experience (i.e., number of times test-maker did the exercise over the course of the programme) influencing the quality of feedback given by students who did this exercise in different courses of the program. *k*: number of predicted parameters, logLik: log-likelihood, AICc: second-order Akaike's Information Criterion, weight: Akaike's weights.

Model	*k*	logLik	AICc	ΔAICc	weight
Null	4	−449.9	907.9	0.00	0.544
Experience	5	−449.0	908.2	0.36	0.456

Discussion

Recent research has reported that there are many benefits when students are made to assume greater agency and ownership of their own learning, including attempting to craft test items and giving feedback that is part of FA to their peers as provisional expert instructors. Our study involving seven classes of undergraduate chemistry students in Singapore over a three-year span was part of a larger project investigating such possible benefits. Moreover, this action research has occurred in the learning of chemistry content at the undergraduate level that has been reported by many as difficult and challenging to master (e.g., Naiker, Wakeling, Johnson, & Brown, 2021).

In the current study that focused on the types of feedback that undergraduates give to their peers, we found that students' self-authored 3TMCQs were classified exclusively within the first three levels of revised Bloom's Taxonomy. This was not necessarily a negative or unexpected finding given that the chemistry content was almost always newly encountered in each week of these courses; test-makers were continuously attempting to make sense of the slew of science concepts as well as craft quality questions that could challenge their peers. Although we have predicted that the frequency of correct responses would decrease in the order of Remember > Understand > Apply, it was surprising that 3TMCQs that were located at the Understand cognitive levels elicited the highest frequency of correct responses (75.5%), followed by Remember (60.9%) and then Apply questions (51.4%) (Table 1). While question type definitely had a strong influence over the correctness of answer given from our statistical analysis, we are unable to suggest a clear explanation for this unexpected trend.

Feedback quality was only influenced by the correctness of the answer on the 3TMCQ as seen by the most parsimonious model (ΔAICc = 0; Table 3). Thus, test-makers gave lower-quality feedback whenever test-takers gave the correct answers, which is understandable given that test-takers can be assumed to have mastered the content already. Thus, there seems little more that can be added in terms of offering guidance or correction by test-makers. What is most interesting is that the correctness of the answer is four times more important in receiving higher-quality feedback than the difficulty of question based on the RVI values (RVI: AnsCor = 1.00; QnType = 0.255). What this means is that cognitively challenging or harder questions did not meaningfully influence the quality of feedback received. We speculate that one motive for addressing the presence of errors through feedback was that all fellow participants were soon-to-be science teachers who needed strong content mastery in the classroom, and such a consideration was likely more compelling than marking a difficult question.

Finally, we found that test-makers did not provide higher-quality feedback over time (null model: ΔAICc = 0; Table 4). This lack of improvement over time might perhaps be attributed to fatigue in giving feedback or that the novelty of this educational intervention has worn off. Alternatively, it may suggest that providing quality feedback is not a skill that requires honing over time, and this gives hope that providing feedback as a form of FA can be a fruitful exercise with little prior training.

Conclusion

What are some implications for teaching and learning undergraduate chemistry from this study? Anecdotally, our students had indicated that the process of being both test-makers as well as test-takers was useful for their learning without a clear preference for adopting either one mode, which suggests further research here. However, based on our findings, the implications are more nuanced and complex. The most direct outcome seems to be that setting more cognitively challenging 3TMCQ does not appear to influence the quality of feedback; learners do not receive feedback rated higher on the Hattie and Timperley (2007) classification when they attempt more difficult questions. This does not mean that there are no benefits to be obtained from attempting this; authoring challenging science items may assist test-makers (& test-takers) into thinking more deeply about content knowledge. At least in this educational intervention here, feedback quality is much more sensitive to wrong answers than the item type itself, where test-takers received higher-quality feedback when they answered questions incorrectly. Certainly, our findings do warrant further research into the quality of peer feedback that learners receive when attempting different question formats such as MCQ, short/structured items, and essay questions. If feedback quality differs across these kinds of questions, students might potentially receive unequal forms of feedback.

Acknowledgements

Ethical approval for this research was covered by Nanyang Technological University IRB-2018-07-032 protocol. We are grateful to the very constructive comments from the anonymous reviewers and our student participants over the years.

References

Aflalo, E. (2021). Students generating questions as a way of learning. *Active Learning in Higher Education, 22*(1), 63–75.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (Eds.). (2001). *A taxonomy for learning, teaching and assessing. A revision of Bloom's taxonomy of educational objectives*. New York: Addison-Wesley Longman.

Andrade, H. L. (2019). A critical review of research on student self-assessment. *Frontiers in Education, 4*, DOI=10.3389/feduc.2019.00087

Ashenafi, M. M. (2017). Peer-assessment in higher education – twenty-first century practices, challenges and the way forward. *Assessment & Evaluation in Higher Education, 42*(7), 226–251.

Bandiera, O., Larcinese, V., & Rasul, I. (2015). Blissful ignorance? A natural experiment on the effect of feedback on students' performance. *Labour Economics, 34*, 13–25.

Barton, K. (2009). *Mu-MIn: Multi-model inference*. R Package Version 0.12.2/r18. http://R-Forge.R-project.org/projects/mumin/

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education, 85*(5), 536–553.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*, 5–25.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74,

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*, 81–90.

Bottomley, S., & Denny, P. (2011). A participatory learning approach to biochemistry using student authored and evaluated multiple-choice questions. *Biochemistry and Molecular Biology Education*, *39*, 353–361.

Brown, G. T. L., & Harris, L. R. (2013). Student self-assessment. In J. McMillan (Ed.), *The SAGE handbook of research on classroom assessment* (pp. 367-393). London, UK: SAGE.

Brown, G. T. L., Peterson, E., & Irving. S. (2009). Beliefs that make a difference: Adaptive and maladaptive self-regulation in students' conceptions of assessment. In D. M. McInerney, G. T. L. Brown, & G. A. D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about Assessment for Learning* (pp. 159–186). Charlotte, NC: Information Age Publishing.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304.

Burnham, K. P., & Anderson, D. R. (2010). *Model selection and multi-model inference: A practical information-theoretic approach*. New York: Springer.

Caleon, I., & Subramaniam, R. (2010). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education*, *32*(7), 939–961.

Christensen, R. H. B. (2019). *Ordinal-regression models for ordinal data*. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics*, *69*(9), 970–977.

Denny, P., Hamer, J., Luxton-Reilly, A., & Purchase, H. (2008). PeerWise: Students sharing their multiple-choice questions. In *Proceedings of the Fourth International Workshop on Computing Education Research* (pp. 51–58). New York: Association for Computing Machinery.

Doyle, E., Buckley, P., & Whelan, J. (2019). Assessment co-creation: an exploratory analysis of opportunities and challenges based on student and instructor perspectives. *Teaching in Higher Education*, *24*(6), 739–754.

Galloway, K. W., & Burns, S. (2015). Doing it for themselves: students creating a high quality peer-learning environment. *Chemistry Education Research and Practice*, *16*, 82–92.

Guilding, C., Pye, R. E., Butler, S., Atkinson, M., & Field, E. (2021). Answering questions in a co‐created formative exam question bank improves summative exam performance, while students perceive benefits from answering, authoring, and peer discussion: A mixed methods analysis of PeerWise. *Pharmacology Research & Perspectives*, *9*(4), e00833.

Hancock, D., Hare, N., Denny, P., & Denyer, G. (2018). Improving large class performance and engagement through student-generated question banks, *Biochemistry and Molecular Biology Education*, *46*, 306–317.

Hardy, J., Bates, S. P., Casey M. M., Galloway, K. W., Galloway, R. K., Kay, A. E., Kirsop, P., & McQueen, H. A. (2014). Student-generated content: Enhancing learning through sharing multiple-choice questions. *International Journal of Science Education*, *36*, 2180–2194.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

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

Hill, M., Sharma, M., O'Bryne, J., & Airey, J. (2014). Developing and evaluating a survey for representational fluency in science. *International Journal of Innovation in Science and Mathematics Education*, *22*(5), 22–42.

Kamcharean, C., & Wattanakasiwich, P. (2016). Development and implication of a two-tier thermodynamic diagnostic test to survey students' understanding in thermal physics. *International Journal of Innovation in Science and Mathematics Education*, *24*(2), 14–36.

Morris, R., Perry, T., & Wadle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education*, *9*(3), e3292

Naiker, M., Wakeling, L., Johnson, J., & Brown, S. (2021). Attitudes and experiences among first-year regional Australian undergraduate students toward the study of chemistry. *Journal of University Teaching & Learning Practice*, *18*(4). https://doi.org/10.53761/1.18.4.15

National Research Council [NRC]. (2011). *Promising practices in undergraduate science, technology, engineering, and mathematics education: Summary of two workshops*. Washington, DC: National Academies Press.

Penuel, W. R., & Shepard, L. A. (2016). Assessment and teaching. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 787–850). Washington, DC: American Educational Research Association.

Power, J. R., & Tanner, D. (in press). Peer assessment, self-assessment, and resultant feedback: An examination of feasibility and reliability. *European Journal of Engineering Education*.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Austria. https://www.R-project.org/

Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching & Teacher Education, 26*, 1372–1380.

Reinke, N. B., Kynn, M., & Parkinson, A. L. (2019). Conceptual understanding of osmosis and diffusion by Australian first-year biology students. *International Journal of Innovation in Science and Mathematics Education, 27*(9), 17–33.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.) *Perspectives of curriculum evaluation* (*AERA Monograph series on curriculum evaluation, 1*) (pp. 39–83). Chicago, IL: Rand McNally.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153-189.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68(3)*, 249–276.

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education, 10*(2), 159–170.

Valverde-Berrocoso, J., Garrido-Arroyo, M. d. C., Burgos-Videla, C., & Morales-Cevallos, M. B. (2020). Trends in educational research about e-Learning: A systematic literature review (2009–2018). *Sustainability, 12*, 5153.

Wieman, C. (2017). *Improving how universities teach science: Lessons from the Science Education Initiative*. London, UK: Harvard University Press.

Wiliam, D. (2006). Formative assessment: Getting the focus right. *Educational Assessment, 11*(3–4), 283–289.

Wiliam, D. (2009). Content then process: Teacher learning communities in the service of formative assessment. In D. Reeves, (Ed.), *Ahead of the curve: The power of assessment to transform teaching and learning* (pp. 183-204). Bloomington, IN: Solution Tree Press.

Wiliam, D., & Black, P. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal, 22*(5), 537–548.

Yu, F. Y., & Wu, W.-S. (2020). Effects of student-generated feedback corresponding to answers to online student-generated questions on learning: What, why, and how? *Computers and Education, 145*, 103723.