10-2019

Formative assessments using text messages to develop students' ability to provide causal reasoning in general chemistry

Ryan D. Sweeder
Michigan State University

Deborah Herrington
Grand Valley State University, herringd@gvsu.edu

NRC Research Press

ARTICLE

Formative assessments using text messages to develop students' ability to provide causal reasoning in general chemistry

Ryan David Sweeder and Deborah G. Herrington

Abstract: Formative assessment is critical in providing students the opportunity to self-assess their content knowledge and providing data to inform instructional decisions. It also provides students with information about course expectations. If, as called for in numerous science instruction reform efforts, we expect students to be able to apply their chemistry knowledge to analyze data and construct coherent explanations, then not only must summative assessments include items that require this of students, but students must also be provided with frequent and ongoing opportunities to individually practice this difficult task and receive feedback. Although online homework systems can be quite effective at providing students with feedback regarding their mastery of basic skills, it is typically less useful in providing meaningful feedback on constructed student explanations. This study examined the impact of providing students with frequent out-of-class formative assessment activities initiated by text messages. Student responses were then used to facilitate in-class instruction. Increased student participation in these formative assessment tasks correlated positively with success on exams even after accounting for student prior knowledge. There was also evidence that students increased their ability to construct complete explanation over the course of the semester. All results were consistent across two different institutions and three instructors.

Key words: chemical education research, formative assessment, general chemistry.

Résumé : L'évaluation formative est un élément essentiel de l'enseignement, car elle permet aux étudiants d'autoévaluer leur connaissance de la matière et fournit des données qui serviront à prendre des décisions pédagogiques éclairées. Elle permet également aux étudiants de savoir ce qui est attendu d'eux dans le cadre du cours. Si, comme le prescrivent de nombreuses réformes de l'enseignement des sciences, on doit s'attendre à ce que les étudiants soient capables d'appliquer leurs connaissances en chimie pour analyser des données et formuler des explications cohérentes, alors, non seulement l'évaluation récapitulative doit-elle comporter des questions qui font appel à ces compétences, mais les étudiants doivent aussi avoir l'occasion à plusieurs reprises et sur une base continue de mettre personnellement en pratique cette tâche complexe et de recevoir une rétroaction. Même si les plateformes de devoirs en ligne peuvent permettre de fournir aux étudiants une rétroaction quant à leur maîtrise des notions de base, elles sont habituellement moins utiles pour fournir une rétroaction significative sur les explications qu'ils ont formulées. Cette étude a porté sur l'influence d'évaluations formatives fréquentes, administrées à la maison par voie de messages textes. Les réponses des étudiants ont ensuite été intégrées à l'enseignement en classe. La participation accrue des étudiants à ces activités d'évaluation formative a été corrélée à de meilleurs résultats aux examens, même après que les connaissances antérieures des étudiants eurent été prises en compte. L'étude a également permis d'observer au cours du trimestre une capacité accrue des étudiants à formuler des explications complètes. Tous les résultats concordaient entre les deux établissements d'enseignement et les trois enseignants participants. [Traduit par la Rédaction]

Mots-clés : recherche pédagogique en chimie, évaluation formative, chimie générale.

Introduction

Promoting and assessing deep, connected understanding of chemistry concepts requires students to meaningfully engage with the content to assess their knowledge. Yet, research indicates that the types of activities that students most frequently engage with outside of class such as rereading notes or the text or highlighting are not supportive of such self-assessment and have minimal effect on student outcomes.[1] In general chemistry, homework, often online, is frequently used as a mechanism to get students to engage with the material outside of class. Though homework can certainly provide a means of self-assessment, typical homework problems that require students to do a calculation or predict the product of a reaction are frequently completed by students following their notes or textbook or using a heuristic,[2] which may get them the correct answer without fully understanding the tested concept. Meaningfully assessing their own understanding of core chemistry concepts requires more of students. It requires them to analyze data, support claims with evidence, and provide particle level, causal–mechanistic reasoning for macroscopic phenomena.[3] To learn to think and reason in this way requires students to consistently engage in formative assessments that focus on these processes and provide students with constructive feedback from peers and instructors, not just before or on an exam, but throughout their course.

Received 14 August 2019. Accepted 12 October 2019.

R.D. Sweeder. Lyman Briggs College, Michigan State University, East Lansing, MI 48825, USA.
D.G. Herrington. Chemistry Department, Grand Valley State University, Allendale, MI 49401, USA.
Corresponding author: Deborah Herrington (email: herringd@gvsu.edu).

Published at www.nrcresearchpress.com/cjc on 18 October 2019.

Providing opportunities for students to engage with meaningful formative assessments is difficult, especially in larger enrollment classes. These types of assessments are typically challenging for students and effectively engaging with them requires a substantial amount of time, with some students requiring more time to grapple with them than others. This makes them less than ideal as in-class activities. However, most online homework systems are not structured to grade or provide meaningful feedback for these kinds of questions since they cannot be encapsulated fully in multiple choice or numerical answer questions. We have previously described a method for using a combination of a free text messaging system and Google Forms to engage students in these types of formative assessments out of class.[4] This paper describes a study that examines the abilities of students to answer these types of formative assessment questions, as well as the relationship between students' regular participation in engaging with such formative assessments and their course outcomes.

Student study strategies and self-assessment

Self-assessment is defined as a process by which learners monitor and evaluate the quality of behavior and thinking when learning and identify strategies that improve their understanding and skills.[5] Self-assessment is a critical component of metacognition[6] and self-regulated learning,[7] both of which are positively associated with greater student persistence and higher grades.[8] Unfortunately, many students do not know how to effectively assess their own understanding. This is evidenced by the fact that the study strategies most commonly employed by students (rereading notes or text and highlighting) are those that provide minimal opportunities for self-assessment and have been shown to have little or no impact on student outcomes.[1] This is perhaps not surprising given that for most of our students, these are the strategies that worked for them in high school[9] and how easy it is for us as learners to mislead ourselves regarding whether learning has been achieved.[10] Consider, for example, in studying for a test, a student reads something and because it looks familiar decides that they do not need to further study it, only to find out on a test when the information is not directly in front of them, they do not really know that concept. Further, though homework has been consistently shown to have positive correlations with course performance, particularly in the case of online homework that is frequently used in large general chemistry courses, multiple attempts at similar questions, low penalties for incorrect attempts, and immediate feedback have been shown to promote student focus on surface problem features rather than the concepts being addressed and can promote the use of guess and check approaches which reduces student use of metacognitive strategies.[11] Additionally, most online homework systems rely heavily on more traditional homework questions that focus on isolated skills and facts, which are important but do not encompass the entirety of what we want students to be able to do.

Further, what people tend to believe about activities that are and are not effective for learning are often at odds with the research data.[10] For example, many students believe that rereading is superior to testing, and students often view testing as a means to evaluate their learning rather than as opportunities to enhance it. Students also tend to view massed practice, where they practice a lot of the same thing at one time, as more beneficial than spaced practice, hence the prevalence of waiting to study until just before an exam. Moreover, though students often view errors and mistakes as something to be avoided, research shows that making errors is often essential for efficient learning. Students frequently believe that being presented with the correct method or answer prior to studying or performing is more beneficial;[12] however, research shows that trying to predict first, even unsuccessfully, can enhance learning.[13]

Finally, there are numerous factors that impact students' use of study strategies. In examining how and why STEM students choose to use certain study strategies, Hora and Oleson found that student use of high-impact study strategies was influenced by a variety of factors including student knowledge of these strategies, time to implement them, and resources required to study in this manner.[14] Further, they found that studying is not just using strategies, but also involves cues, largely instructor driven, about when to study, timing of the studying, and identification of which resources to use and what strategies to use. Most important to our work is that students look to instructor driven cues regarding when and what to study. Accordingly, if we want students to engage in systematic and frequent self-assessment that moves beyond just focusing on skills or isolated facts, we must provide opportunities and incentives within our course structure for them to do so, not just before an exam, but throughout the course. Such opportunities need to move beyond traditional homework questions found at the back of most textbooks or in online homework systems and require students to use and apply their knowledge to do more challenging things such as analyze data, support claims with evidence, and provide particle level, causal–mechanistic reasoning for macroscopic phenomena. There must also be opportunities for obtaining feedback to reflect on and identify strategies to improve their learning. One mechanism for this is effective formative assessment.

Formative assessment

Formative assessment involves the use of learning tasks that can elicit evidence about student learning and the use of such evidence by teachers and learners to make data-driven decisions about the next steps in instruction.[15] Effective formative assessment tasks are thought to improve student learning through five key strategies:

1. Clarifying learning expectations for students
2. Eliciting evidence of student understanding to guide instructor decision making
3. Providing feedback to move the learning forward
4. Activating students as peer resources
5. Activating students' ownership of learning

Given the focus on eliciting evidence of student understanding, using that evidence to provide feedback to the instructor and learners regarding what steps should be taken next, and empowering students to manage and regulate their own their learning, it is not surprising that the use of effective formative assessment improves learning for students.[16,17] However, another key factor influencing the impact of formative assessment is gaining student buy-in. Work by Brazeal and Couch showed that higher student buy-in on pre- and post-class forms of formative assessment predicted better performance on course exams and overall course grades.[18] This makes sense as reaping the benefits of formative assessment methods, students must meaningfully engage with them. More importantly, Brazeal and Couch identified factors that predict higher student buy-in. Specifically, greater student buy-in was associated with the use of relevant and challenging formative assessment questions and the frequent discussion of such questions. Further, the course section was also found to be a predictor of student buy-in towards formative assessment, suggesting that implementation affects how students perceive formative assessment.[18] Based on their findings, Brazeal and Couch suggest the following as means for fostering student buy-in towards formative assessment:

1. Make questions relevant and challenging
2. Encourage student discussion of formative assessment questions
3. Empower student ownership of learning through opportunities both in and outside of class for students to reflect on understanding, confusion, and study habits

Theoretical framework

The design of the formative assessment system and study described in this paper is underpinned by constructivism, a theory of learning in which the learner constructs understanding from their experiences, either individually or in groups.[19,20] Studies examining the use of student-centered instructional approaches in the classroom, where students engage in experiences through which they can construct knowledge, have overwhelmingly indicated that such instructional approaches can improve student learning during lecture courses.[21] However, it is equally important to consider the role of the learner and how they engage with the content outside of the classroom.

As described above, the psychology and education literature has clearly identified more and less effective study strategies employed by students.[1,22–24] From a constructivist perspective, it makes sense that study strategies that support self-assessment (such as practice testing and distributed practice) have the greatest impacts on student outcomes, as self-monitoring of learning and thinking is an important part of knowledge construction.[25] Knowledge is transferred to and stored in long-term memory through making connections to other knowledge. Retrieving knowledge in different situations such as though testing or distributed practice causes it to be reorganized and connected to more pieces of knowledge in long-term memory, making it easier for learners to retrieve and apply to different situations.[10] Thus, a formative assessment method that prompts students to consistently and regularly apply their chemistry knowledge to construct explanations of phenomena outside of class, followed by collaboratively reflecting on, correcting, and improving student responses in class, should support student self-assessment and knowledge construction.

Research also indicates that students do not always use the same study strategies, but rather, that the strategies they choose to use are influenced by a number of factors including the structure and expectations of the course.[26,27] This suggests that we should also consider learning from a situated cognition perspective that views the context in which the knowledge is constructed as an integral part of the learning.[28] Thus, if we want students to employ more effective study strategies outside of the classroom, then we need to structure our courses in such a way to integrate and value those types of experiences by dedicating time, attention, and priority to them. Accordingly, formative assessment will be most effective if it is integral to the course and aligned with summative assessments.

Research questions

The goal of our intervention was to use the technology that students most frequently interact with, their phones, to engage students in the types of meaningful formative assessment questions that the literature has linked to self-assessment, student buy-in, and greater conceptual understanding.[1,3,18,29] Ultimately the intent is to improve student understanding of core chemistry concepts and performance in our general chemistry courses. Thus, this research study sought to answer three key research questions:

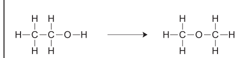
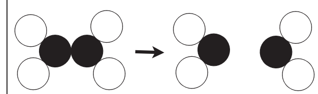
- How do students perform on these formative assessment questions?
- How does student participation with these types of questions relate to their exam performance?
- How does the quality of student explanations change over the course of a semester?

Methods

Question development

To develop the relevant and challenging questions that elicit evidence of student understanding and are foundational for student self-assessment and effective formative assessment, we used a framework of three-dimensional (3D) learning. This framework for learning and assessment aims to interweave content with science practices and cross cutting concepts and is described in detail in the *Framework for K–12 Education*[30] and the *Next Generation Science Standards.*[31] Several descriptions of how this framework has been applied to the development of college chemistry curricula[32,33] and the development of assessment items[34] have been published. Based on this framework, we developed a series of questions for a first semester general chemistry course that provided students with an initial phenomena or context to explain, explore, or evaluate. For calculation-based questions, students were asked to perform a calculation and then interpret or make a prediction based on their calculation. For non-calculation questions, students would make a claim or prediction and identify the evidence upon which they based their claim or prediction. Both styles of question would then require the students to provide an explanation or rationale. These explanatory response prompts frequently contained guidance for student answers (e.g., "Your answer should link energy required to break a bond, the type of radiation, and what sunscreen protects us from.") designed to promote explanations that included causal mechanistic reasoning.[35] See Fig. 1 for an example of each question type.

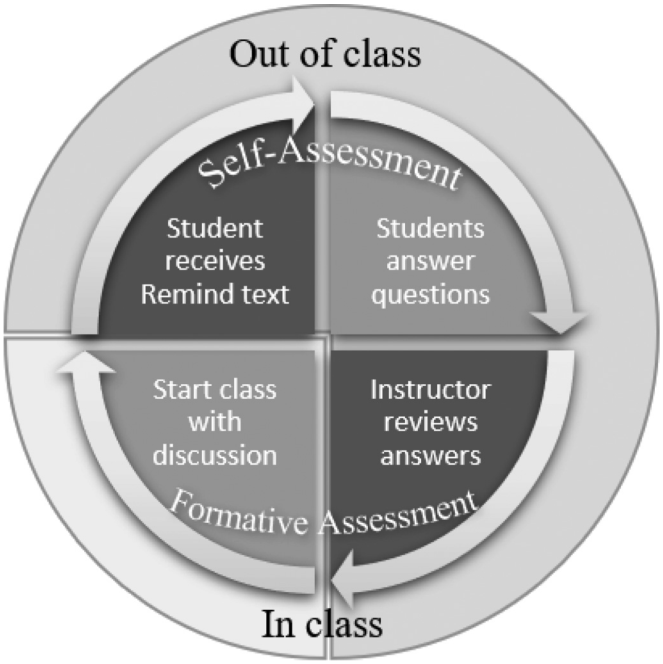
Fig. 1. Examples of calculation and non-calculation based questions administered using the Remind assessment system.

Calculation Based Question	Non-Calculation Based Question
Isomers are molecules that contain the same atoms, but the atoms are connected differently. The reaction shows an isomerization process, ethanol rearranges to form dimethyl ether. Using the bond energies provided, calculate ΔH for this process.  Ethanol Dimethyl ether Bond energies (kJ/mol) C-H 414 C-O 350 C-C 345 O-H 464 Which isomer is more stable? A. Ethanol B. Dimethyl Ether How did you decide on your answer to the question above?	Consider the following reaction of N_2O_4 going to $2NO_2$. What can you say about the energy change?  A. Exothermic B. Endothermic What about the reaction supports your decision? In other words, what reasoning did you use to determine your answer?

Intervention

To foster student engagement with these questions as a method of self and formative assessment, we used the "Remind Assessment System", the design of which is outlined in Fig. 2. Questions, like the ones shown in Fig. 1, were each entered into a Google form.[36] Students received one to three text messages per week sent using Remind,[37] each containing a link to a Google form question. Each question related to material that was covered in the previous class period and was due prior to the beginning of

Fig. 2. Remind assessment system design.



the next class period. Therefore, students generally had between 16 and 48 h to complete the question. As the purpose of these questions was to be formative, student answers were not graded for correctness, rather students were given credit for meaningfully attempting the question. Prior to class, instructors reviewed student responses, examining both the quantitative results and identifying patterns in the student written responses. This information would then be used to inform instruction for the beginning of the next class. Often this would involve students working in small groups comparing, evaluating, and critiquing sample student explanations to identify incorrect ideas and develop a high-quality explanation. A more detailed explanation of the implementation of this intervention has been published previously.[4] The alignment of this intervention with the key strategies of effective formative assessment[15] and suggested means for promoting student buy-in toward formative assessment[18] are illustrated in Fig. 3.

Data collection

This study was reviewed and approved as exempt by our Institutional Review Boards (GVSU Ref. No. 18-027-H; MSU x17-1192e). Participants were drawn from four different first semester general chemistry classes at two large, public institutions, taught by three different instructors (Table 1). Consent was requested from all students and responses for students who did not provide consent were not included in this study, although they may have been used for in-class instruction. Students who did not complete the Remind questions included in this study or did not complete the class were removed from the data set.

For one class at each institution, we also used two common exam questions as a summative assessment to measure student ability to make a claim and support it with appropriate evidence and reasoning.

Data analysis

The percentage of Remind questions completed by each student (participation rate) were calculated. A set of nine questions that were used at both institutions and required students to provide answers that fit a claim–evidence–reasoning format were identified for more in-depth analysis. Two questions varied between their implementation at the two institutions. For the most part, questions were administered to all four classes at about the same time in the semester. However, one question, focused on the energy associated both breaking bonds (Fig. 1), was given at institution two early in the semester and institution one late in the semester. A second question (bond energy) was given to only two classes (one at each institution).

To evaluate the quality of the student explanations to the nine common questions, a universal scoring rubric was developed. Students' claim and the evidence–reasoning parts of their answers were scored separately. Claims were scored based on their accuracy (0 or 1). In the case where students used the results of a calculation to formulate a claim, the accuracy of their claim was based on their calculated value. For the evidence and reasoning portion, a three-point scale was employed. Students were given one point each for correctly (based on their claim) identifying the key evidence used to make their claim and for providing correct reasoning to justify their claim. A third point could then be scored for accurately and explicitly linking their claim, evidence, and reasoning together. To ensure consistent coding for each question, using this general rubric, more specific expectations for evidence and reasoning were identified for each specific question. These criteria were then applied to a subset of responses (approximately 20% of total responses) by a single researcher. This coding was reviewed by at least one other researcher and responses that did not clearly fit into a single scoring category (about 10%–20% of responses depending on the question) were then discussed within the research team and coding criteria were refined as needed. The same initial researcher then coded all the remaining responses using the finalized coding scheme. Any responses that did not clearly fall into the previously defined categories (generally fewer than 5% of responses) were then discussed by the research team to identify a final score. The general rubric criteria along with an example of specific question criteria, and sample student responses for each score are shown in Table 2 below. The specific question criteria and sample student responses provided in Table 2 correspond to the calculation-based question shown in Fig. 1.

Student scores on midterm exams were obtained from the instructors. Students were then categorized into one of three roughly equal sized groups based on their exam 2 and 3 average; high scorers (≥80 out of 100), moderate scorers (70–80 out of 100), or low scorers (<70 out of 100). Rather than the final exam, an average of these two exams was deemed to be the best indicator of overall student course performance. At one institution, the course shared a common final exam with all sections meaning that the exam did not accurately reflect the instructor's expectation for students to provide complete explanations. At the other institution, some of the students were able to opt out of the standard final exam if their grade was sufficiently high at the end of the course, which would have only provided a partial data set. Because the early portion of all classes focused heavily on what is typically expected as "prior knowledge", exam 1 scores typically reflect a mix of college and high school experiences.

IBM SPSS Statistics 25[38] was used for statistical analysis of the data. Pearson correlations[39] were used to test for all correlational relationships. Means were compared using ANOVAs.[40] Linear regressions were used to predict the exam 2 and 3 average for students (course performance) based on a mix of scalar variables and dummy variables (to represent instructor or institution). Typically, variables were added stepwise until no further variance could be explained by the inclusion of more variables.

Results and discussion

Student performance on Remind questions

In looking at our first research question (How do students perform on these types of formative assessment questions?), we

Fig. 3. Alignment of texting formative assessment intervention characteristics with effective formative assessment practices and methods for fostering student buy-in.

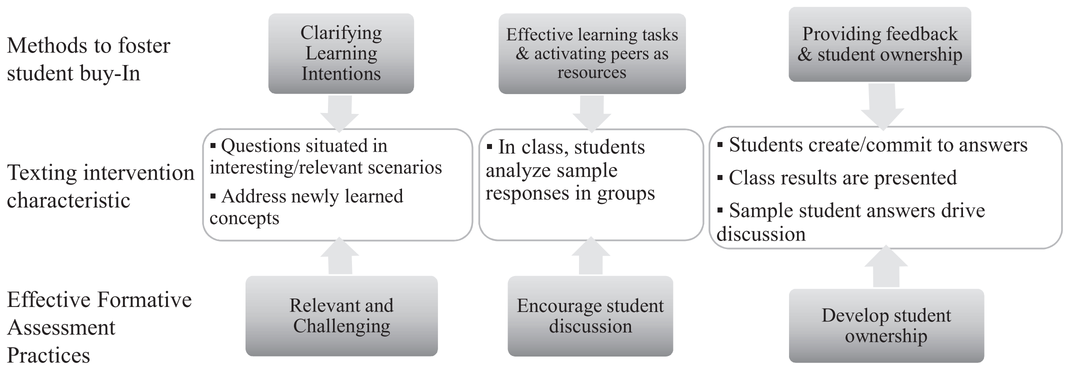


Table 1. Participants.

Instructor/no. of classes	No. of consenting students	No. of lectures per week
Institution one		
A/2 classes	179	3
B/1 class	69	2
Institution two		
C/1 class	65	3

found that student performance on individual questions varied considerably. This is perhaps not surprising as both the difficulty of content and the difficulty of questions vary throughout the semester. Examining just the claim portion of the questions, the percent of correct student answers ranged from 49% to 96% (Fig. 4). This suggests that the students are doing quite well at answering questions for having just encountered the material the previous day in class. In all but one question, over one-half of the submitted claims were correct. The combined evidence and reasoning scores were not quite as impressive, with between 18% (stability) and 80% (bond PE) of students providing both the evidence and reasoning consistent with their claim (score of 2 or 3) (Fig. 4). The average evidence and reasoning score ranged from 0.71 to 2.10 (out of a possible score of 3). As shown in Fig. 4, the percentage of students able to provide both evidence and reasoning consistent with their claim is generally substantially lower than the percentage of students able to provide a correct claim. This is not surprising given that providing both evidence and reasoning requires a more in-depth understanding of the concepts. These results reaffirm the importance of incorporating these questions as formative assessment and using the results and student responses to initiate instruction in the subsequent class session to help students recognize and reflect on the gaps in their understanding and underscore expectation for providing causal reasoning.

Though these patterns in student responses were evident regardless of instructor or institution, at the instructor or institution level, some statistical differences for individual questions were noted. At institution one, where both instructors worked from the same in-class materials and had equivalent rates of student participation, we see very few differences in student results. Only two questions show any statistically significant difference between the two instructors, the reasoning scores on the electron transition (1.86 vs. 1.46, $p = 0.037$, $\eta^2 = 0.025$) and electromagnetic radiation questions (1.64 vs. 1.24, $p = 0.025$, $\eta^2 = 0.028$). Looking across institutions more differences arise. Although only one question shows a difference in students' ability to correctly identifying a claim (bond PE), one-half of the questions show statistical differences (bond breaking, e-transition, ion size, bond energies, and stability) in the evidence and reasoning scores. However, this could be affected by the differences in timing of the questions relative to instruction or participation rates between the two institutions (see discussion in subsequent section). In general, we do not have enough information about specific student characteristics or differences in instruction prior to each question for the two different institutions to be able to meaningfully explain these differences.

Calculations in questions

In many courses, students are asked to complete calculations to demonstrate their understanding of content; however, it is well substantiated that many students are able to follow algorithmic processes to obtain correct answers without having a conceptual understanding of the concepts.[2] Examining how students perform on our formative assessment questions and requiring students to support their answer with an explanation (Research Question 1) not only supports these findings, but also highlights one additional potential concern, which is the students' ability to interpret their calculations. For example, on one question, students were asked, given the energy of an H–H bond, to calculate the frequency of light needed to break an H–H bond. They then had to use their calculated frequency and a figure of the electromagnetic (EM) spectrum to determine what type of EM radiation this corresponded to. Finally, students were asked to explain why it was important to use sunscreen, being told that their answers should link energy required to break a bond, the type of radiation, and what sunscreen protects us from. Of the submitted answers, 76% of the students provided a correct calculation; 70% of students could correctly identify the type of light that was consistent with their calculation (regardless of the accuracy of their calculation). However, only 44% (about one-half of those correctly completing the calculation) were able to provide correct evidence and reasoning to then support their claim (Fig. 5). For a second question involving a calculation (stability), we observe a much lower level of success on the calculation though a similar ability to correctly interpret their calculation to make a claim. We again see that roughly one-half of the students able to correctly complete the calculation were also able to provide valid evidence and reasoning (Fig. 5).

Participation

To address Research Question 2, "How does student participation with these types of questions relate to their exam performance?", we analyzed the participation of the students. Overall, student participation rates across the two institutions and were statistically different via an ANOVA (75% vs. 68%; $F_{1,312} = 5.268$, $p = 0.022$, $\eta^2 = 0.017$). Given the different class settings, this is not too surprising. The classes at the institution with the higher participation rate were a part of a residential college. Because the student received the message in the evening, students in the residential college were more likely to consistently be interacting

Table 2. Criteria for scoring student responses.

Score	General criteria	Specific criteria*	Sample responses*
0	No evidence or reasoning or evidence–reasoning that is not consistent with the claim	Evidence: Energy put in or energy released	More of a stable Lewis structure.
1	Evidence or reasoning that is consistent with the claim OR the information provided and (or) calculated	Reasoning: If energy needs to be put in then the product is higher in energy than the reactant and is less stable. OR if more energy is required to break bonds in ethanol than form dimethyl ether then ethanol must have stronger bonds and be more stable.	More energy was released in products than used in reactants. The higher delta H is more stable. (Evidence only)
2	Evidence and reasoning that are consistent with the claim OR the information provided and (or) calculated		The enthalpy value is positive, meaning you have to put energy in in order to change forms.
3	All pieces are explicitly tied together	Must explicitly connect overall potential energy of molecules or relative bond strengths to enthalpy and stability for a score of 3.	When changing from ethanol to dimethyl ether the overall enthalpy is a positive value. Since it is a positive value, we know that the bond energies in dimethyl ether must be weaker than those in ethanol, meaning that ethanol is more stable.

*Specific criteria and sample student responses for calculation-based question shown in Fig. 1.

Fig. 4. Percentage of students making a correct claim or providing both evidence and reasoning consistent with their claim (score of 2 or 3) by question. Questions appear in the order they were assigned in class.

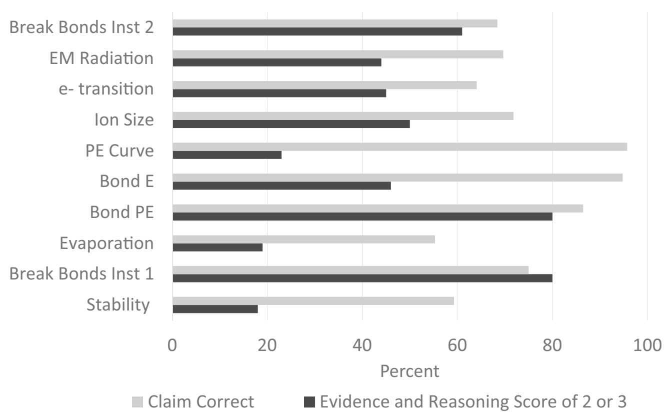
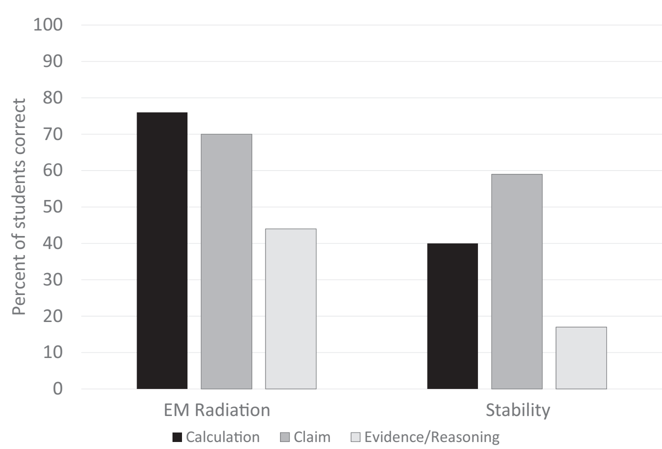


Fig. 5. Student performance on calculation questions.



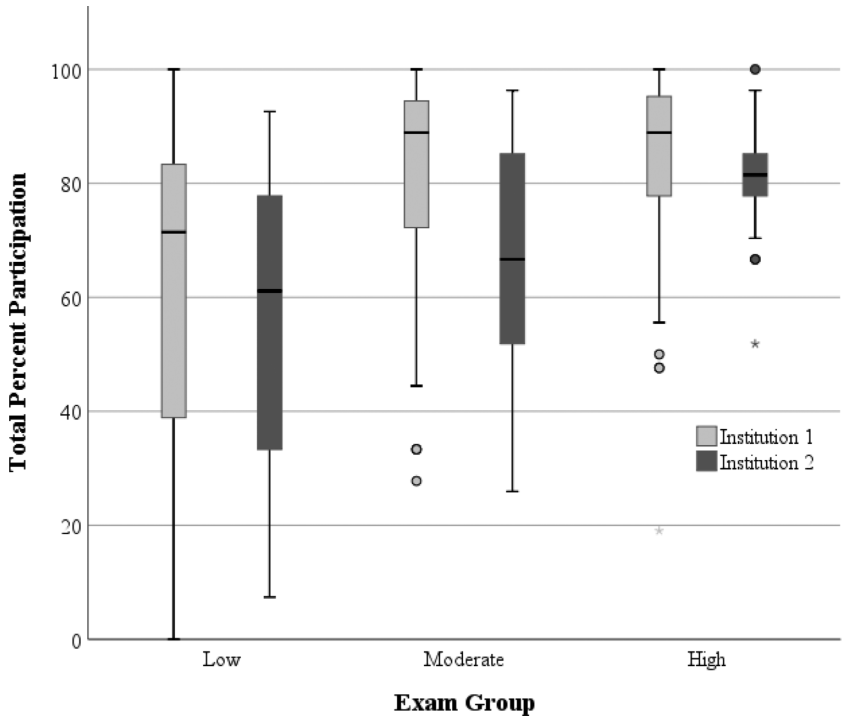
with other students enrolled in the same class who would have received the same message. This close proximity of students likely resulted in the increase in participation. Figure 6 shows a definite relationship between students' exam score grouping and participation rate across both institutions with higher scoring students completing more of the Remind questions. At institution one, the low scoring group participates at a statistically lower rate than the higher two groups, and at institution two, the high scoring group participates at a statistically higher rate than the other two groups. At institution one, there is no statistical difference between the two instructors. Across both institutions and all instructors, participation rate positively correlates with exam score ($r = 0.482$, $p < 0.001$). This, however, could be the result of better students being the ones who are more likely to complete the expected tasks rather than a causal relationship. Hence, to try to control for the academic ability of each students, we can employ a linear regression.

Linear regressions provide the opportunity to understand the influence of student participation in completing Remind questions on course performance after accounting for differences in students' academic ability. We opted to use the exam 2 and 3 average to represent the students' overall performance in the course and students' exam 1 score to account for students' prior knowledge and baseline academic abilities. A linear regression using just the exam 1 score predicts 44% of the variance in exam 2 and 3 averages. Adding in dummy variables to represent the instructors (which would account for differences in exam difficulty or averages) explains another 2% of the variance. If we then add in the total participation, it explains an addition 8% for a total of 54% of the exam 2 and 3 average explained. This suggests that the participation rate provides information that extends beyond simply the students' academic abilities and instructor (which should be captured in the first two variables). This likely represents a combination of enhanced learning through the engagement with this version of formative assessment, motivation and effort toward the class, or other personal attributes.

Quality of student explanations

Examining how the quality of student explanations change over the course of the semester (Research Question 3) is challenging given that the difficulty of the content generally increases as the course progresses. However, there are two ways in which we tried to address this question. First, there was one question that focused on the energy associated with breaking bonds (see Fig. 1 non-calculation based) that was administered at two different times during the course. At institution 2, this question was asked very early in the course; at institution 1, it was asked toward the end of the course. Given that for questions given at both institutions at approximately the same time during the course resulted in very similar types of student explanations, this question provides the ability to see how student responses change over time. Statistically there was a small improvement in the score of the

Fig. 6. Relationships between participation on Remind questions (as percent of total Remind questions asked) and exam score (based on low, moderate, or high grouping).



students completing the question later in the semester (1.90 vs. 1.61, $p = 0.020$, $\eta^2 = 0.022$). Though not statistically significant, we also see an increase (from 7% to 15%) in students earning a maximum score of 3, by noting both that a bond was broken and that no bonds were formed. This suggests that some students are recognizing the need to provide more complete answers to these formative assessment questions.

The second way to address Research Question 3 is to look at the quality of student responses on exam questions at the end of the course. The exams in each of the courses involved reinforced the need for students to develop their ability to provide explanations of content by including similar questions on the exams. This provided effective alignment between the formative and summative assessment tasks that encouraged student buy-in to the formative assessment activities. An independent rater evaluated each of the midterm course exams using the 3D-LAP[41] to determine the percentage of points allocated to these types of tasks that involve science practices (including explanation), core chemistry ideas, and cross-cutting concepts. Though the percentage of points allocated to 3D questions varied by exam and instructor, the average percentage of exam points allocated to 3D questions were similar (A: 49%; B: 34%; C: 54% - Classes A and B from Institution one and Class C from Institution two as noted in Table 1).

On the exams of two sections (one at each institution), we included questions that required students to provide a claim, evidence, and reasoning. One question (Fig. 7) focused on intermolecular forces (IMFs) and how they could be used to predict relative boiling points of compounds. This question was selected because students struggle with correctly identifying IMFs as occurring between molecules[42] and struggle with predicting relative boiling points.[43] Further, as students try to accomplish these tasks, they often rely upon heuristics rather than underlying conceptual understanding.[43,44] On exams at both institutions, 90% of students drew and labeled the IMFs as between two different molecules. Cooper et al. previously reported that a majority of second semester general chemistry students (60%) at a large research institute who gener-

Fig. 7. Exam question requiring explanation.

Methanal (CH_2O) and methanol (CH_3OH) are both common substances.

A. Draw the Lewis structure for methanal and methanol
B. Consider a liquid sample of each substance, identify **all** the types of **intermolecular forces** that would be present in each sample (given options)
C. Draw two molecules of CH_2O and then CH_3OH showing how the **strongest type** of intermolecular forces act between the molecules. Be sure to label the IMFs.
D. Identify which substance you would predict to have the higher boiling point.
E. Explain why this compound would have the higher boiling point. Be sure to use concepts of force and energy in your explanation.

ally score in the 75th percentile on the ACS general chemistry exam identified IMFs as existing within a molecule rather than between molecules.[42] Thus, our students appear to have a strong grasp of the concept of IMFs. More importantly, 85% of students (at both institutions) correctly predicted methanol to have the higher boiling point, identified the presence of H-bonding between the methanol molecules, and indicated that this stronger IMF was why it had a higher boiling point. Of those students, three-fourths (65% of total students) provided what we viewed as a fully complete answer, which also added that the stronger IMF required more energy to disrupt and thus a higher temperature was needed. A study by Kararo et al. using a very similar question with students at a large research institution using a transformed chemistry curriculum found that although 64% of students could correctly identify ethanol (compared with dimethyl ether) as having the higher boiling point, only 15% of those students gave complete answers that included H-bonding, relative strength of IMF, and energy.[45] The fact that such a large percentage of our students provided such complete answers we view as evidence that the repeated practice on the formative assessments had a positive impact on the students' approach to providing more complete explanatory reasoning. Further, at institution 2 (where participation in the Remind questions was more variable), we see a correlation of 0.63 between students' total participation on Remind questions and score on this exam question. This is supportive of the idea that engagement with these formative assessment questions helps practice these explanatory skills.

Although this study was conducted at two different institutions, both are geographically similar and draw heavily from the same high school student population with both schools having relatively challenging standards for admission.[46] Further, the formative assessment intervention used in this study was implemented by instructors who have experience and value the development and use of 3D assessments. This meant that the summative assessments very explicitly included questions with 3D components that helped to align student expectations and build buy-in. Additionally, though students responded to the formative assessment tasks outside of class, student answers were integral to the advancement of the classroom instruction as students regularly discussed and evaluated submitted answers. Thus, careful consideration must be taken in generalizing these results to more diverse student populations or classes where formative assessment is not an integral part of the course design or well aligned with summative assessment measures. It should also be recognized that these results focus on the class averages and the data do not allow us to make conclusions about how individual students or sub-populations of students may be impacted.

Conclusions

The results of this study indicate that students are quite successful at being able to make correct claims to questions that require them to apply their chemistry content knowledge but that they have lower levels of success when required to provide evidence and reasoning to support their claims. This is especially true when students must first complete a calculation. As we have shown, practice at creating explanations can be achieved through student participation in regular and frequent formative assessment tasks designed around providing such experiences, and such practice correlates positively with student exam scores even after accounting for students' prior knowledge and academic ability. As today's students are more apt to respond to text messages than email, using a system such as the one described here provides a means to easily engage students in such regular and frequent formative assessment tasks. Further, using such tasks consistently throughout the course shows evidence of helping students increase the quality of their explanations of phenomena using chemistry concepts. As a result, we see quite high levels of success on summative assessments, which include a mixture of questions that require students to provide these types of explanations in addition to more traditional skill or calculation focused questions. These results are consistently observed across three instructors at two different institutions.

Implications for research, practice, and learning

If our ultimate goal for instruction is for our students to be able to use their chemistry knowledge to construct coherent explanations, then not only must we assess this explicitly, but it is also incumbent upon instructors to provide students with the requisite practice needed to be successful in this difficult task. As we have shown, this practice can be achieved through student participation in regular and frequent formative assessment tasks designed around providing such experiences and that such practice correlates positively with student exam scores even after accounting for students' prior knowledge and academic ability. Although we assert that this formative assessment should be ongoing through the course, in this study, students completed only 1–3 questions per week through the entire semester. This suggests that such formative assessment does not need to be overly onerous for instructors or students. A further area of study could examine how the frequency and spacing of such tasks impacts student performance. Additionally, at this point, the results are only reported at the class level, so we have no ability to identify if the impact is even for all students or if this approach has differential impacts. For example, it may be that students' confidence in their answer plays a critical role, because errors, particularly those made with high confidence, appear to provide opportunities for learning, as feedback regarding errors is much more effective when made following errors made with high confidence vs. those made with low confidence.[47–50]

We also assert that the integration of these formative assessment tasks was integral to the course structure, and alignment with summative assessment expectations was key to their impact on student outcomes. The questions were sent to the students timed to align with the in-course instruction such that students were required to revisit and use newly acquired knowledge. Moving these questions outside of the classroom allowed students to spend as much time as needed to engage with the questions, as opposed to being limited by class time or the patience of the fastest working students. Ideally, this helps to build student ownership of their learning because each student has been able to construct their own answer. Further, to promote student buy-in and fully integrate these tasks into the course, student results and responses were used to drive discussion and instruction in the subsequent class meeting. This process was critical as it integrates the key criteria for effective formative assessment; providing feedback about students' written answers, clarifying learning expectations, engaging students with the content, and activating peers as learning resources. This was of course in addition to directly providing an opportunity to address any misunderstanding in student learning that may be evident in the student responses. Removing or changing any of these aspects may have a direct impact on the efficacy of this intervention.

Acknowledgements

The authors thank Justin Carmel for conducting the 3D-LAP evaluation of the course exams and Patrick Daubenmire and his undergraduate for their assistance in data organization, coding of some of the student responses, and helpful discussions.

References

(1) Dunlosky, J.; Rawson, K. A.; Marsh, E. J.; Nathan, M. J.; Willingham, D. T. *Psychol. Sci. Public Interest* **2013**, *14* (1), 4. doi:10.1177/1529100612453266.
(2) Talanquer, V. *J. Chem. Educ.* **2014**, *91* (8), 1091. doi:10.1021/ed4008765.
(3) Cooper, M. M. *J. Chem. Educ.* **2015**, *92*, 1273. doi:10.1021/acs.jchemed.5b00203.
(4) Herrington, D. G.; Sweeder, R. D. *J. Chem. Educ.* **2018**, *95* (12), 2148. doi:10.1021/acs.jchemed.8b00361.
(5) McMillan, J. H.; Hearn, J. *Educ. Horiz.* **2008**, *87* (1), 40.
(6) Siegesmund, A. *J. Microbiol. Biol. Educ.* **2016**, *17* (2), 204. doi:10.1128/jmbe.v17i2.954.
(7) Duckworth, K.; Akerman, R.; MacGregor, A.; Salter, E.; Vorhaus, J. *Self-regulated learning: a literature review*. Centre for Research on the Wider Benefits of Learning, London, 2009.
(8) Boekaerts, M.; Corno, L. *Appl. Psychol.* **2005**, *54* (2), 199. doi:10.1111/j.1464-0597.2005.00205.x.
(9) Cook, E.; Kennedy, E.; McGuire, S. Y. *J. Chem. Educ.* **2013**, *90* (8), 961. doi:10.1021/ed300686h.
(10) Bjork, R. A.; Dunlosky, J.; Kornell, N. *Annu. Rev. Psychol.* **2013**, *64* (1), 417. doi:10.1146/annurev-psych-113011-143823.
(11) Wilson, E. E.; Kennedy, S. A. Modern "homework" in general chemistry: an extensive review of the cognitive science principles, design, and impact of current online learning systems. In *ACS Symposium Series*; Sörensen, P. M., Canelas, D. A., Eds.; American Chemical Society: Washington, DC, 2017; Vol. 1261, pp. 101–130. doi:10.1021/bk-2017-1261.ch009.
(12) Huelser, B. J.; Metcalfe, J. *Mem. Cognit.* **2012**, *40*, 514. doi:10.3758/s13421-011-0167-z.
(13) Kornell, N. *Appl. Cogn. Psychol.* **2009**, *23* (9), 1297. doi:10.1002/acp.1537.
(14) Hora, M. T.; Oleson, A. K. *Int. J. STEM Educ.* **2017**, *4* (1). doi:10.1186/s40594-017-0055-6.
(15) Black, P.; Wiliam, D. *Educ. Assess. Eval. Account. Former. J. Pers. Eval. Educ.* **2009**, *21* (1), 5. doi:10.1007/s11092-008-9068-5.
(16) Eddy, S. L.; Hogan, K. A. *CBE Life Sci. Educ.* **2014**, *13* (3), 453. doi:10.1187/cbe.14-03-0050.
(17) Reimer, L. C.; Schenke, K.; Nguyen, T.; O'Dowd, D. K.; Domina, T.; Warschauer, M. *RSF* **2016**, *2* (1), 212. doi:10.7758/RSF.2016.2.1.10.

(18) Brazeal, K. R.; Couch, B. A. *J. Microbiol. Biol. Educ.* **2017**, *18* (1). doi:10.1128/jmbe.v18i1.1235.
(19) Piaget, J. *J. Res. Sci. Teach.* **1964**, *2* (3), 176. doi:10.1002/tea.3660020306.
(20) Vygotsky, L. S. *Mind in Society: The Development of Higher Psychological Processes*, revised ed.; Cole, M., John-Steiner, V., Scribner, S., Souberman, E., Eds.; Harvard University Press, Cambridge, Mass., 1978.
(21) National Research Council. *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*; National Academies Press, Washington, DC, 2012.
(22) Hartwig, M. K.; Dunlosky, J. *Psychon. Bull. Rev.* **2012**, *19* (1), 126. doi:10.3758/s13423-011-0181-y.
(23) Karpicke, J. D.; Butler, A. C.; Roediger, H. L. *Memory* **2009**, *17* (4), 471. doi:10.1080/09658210802647009.
(24) Karpicke, J. D.; Roediger, H. L. *Science* **2008**, *319* (5865), 966. doi:10.1126/science.1152408.
(25) Shepard, L. *The Role of Classroom Assessment in Teaching and Learning*; 517; National Education Policy Center, Boulder, CO, 2000; p. 99.
(26) Foerst, N. M.; Klug, J.; Jöstl, G.; Spiel, C.; Schober, B. *Front. Psychol.* **2017**, *8*. doi:10.3389/fpsyg.2017.01288.
(27) Scouller, K. *High. Educ.* **1998**, *35* (4), 453. doi:10.1023/A:1003196224280.
(28) Brown, J. S.; Collins, A.; Duguid, P. *Educ. Res.* **1989**, *18* (1), 32. doi:10.2307/1176008.
(29) Seery, M. Take It Easy on the Equations. Let Me Explain … . In *Education in Chemistry*. Royal Society of Chemistry, London, UK, 2018.
(30) National Research Council. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*, 2011. doi:10.17226/13165.
(31) NGSS Lead States. *Next Generation Science Standards: For States, By States*; National Academies Press: Washington, D.C., 2013.
(32) Cooper, M.; Klymkowsky, M. *J. Chem. Educ.* **2013**, *90* (9), 1116. doi:10.1021/ed300456y.
(33) Cooper, M. M.; Stowe, R. L.; Crandell, O. M.; Klymkowsky, M. W. *J. Chem. Educ.* **2019**, *96* (9), 1858. doi:10.1021/acs.jchemed.9b00401.
(34) Underwood, S. M.; Posey, L. A.; Herrington, D. G.; Carmel, J. H.; Cooper, M. M. *J. Chem. Educ.* **2018**, *95* (2), 207. doi:10.1021/acs.jchemed.7b00645.
(35) Cooper, M. M.; Kouyoumdjian, H.; Underwood, S. M. *J. Chem. Educ.* **2016**, *93* (10), 1703. doi:10.1021/acs.jchemed.6b00417.
(36) Google Forms — create and analyze surveys, for free [online]. Available from https://www.google.com/forms/about/. [Accessed 11 May 2018.]
(37) Remind [online]. Available form https://www.remind.com/. [Accessed 11 May 2018.]
(38) *IBM SPSS*; IBM, 2017.
(39) Taylor, R. *J. Diagn. Med. Sonogr.* **1990**, *6* (1), 35. doi:10.1177/875647939000600106.
(40) Field, A. *Discovering Statistics Using IBM SPSS Statistics, 4th Edition*; SAGE Publications Ltd, 2013.
(41) Laverty, J. T.; Underwood, S. M.; Matz, R. L.; Posey, L. A.; Carmel, J. H.; Caballero, M. D.; Fata-Hartley, C. L.; Ebert-May, D.; Jardeleza, S. E.; Cooper, M. M. Characterizing College Science Assessments: The Three-Dimensional Learning Assessment Protocol. *Plos One* **2016**, *11* (9), e0162333. doi:10.1371/journal.pone.0162333.
(42) Cooper, M. M.; Williams, L. C.; Underwood, S. M. *J. Chem. Educ.* **2015**, *92* (8), 1288. doi:10.1021/acs.jchemed.5b00169.
(43) Maeyer, J.; Talanquer, V. *Sci. Educ.* **2010**, *94*, 963. doi:10.1002/sce.20397.
(44) Cooper, M. M.; Corley, L. M.; Underwood, S. M. *J. Res. Sci. Teach.* **2013**, *50* (6), 699. doi:10.1002/tea.21093.
(45) Kararo, A. T.; Colvin, R. A.; Cooper, M. M.; Underwood, S. M. *Chem. Educ. Res. Pract.* **2019**, *20* (1), 316. doi:10.1039/C8RP00195B.
(46) CollegeSimply: college search, data, comparison and admission chances [online]. Available from https://www.collegesimply.com/. [Accessed 12 August 2019.]
(47) Butterfield, B.; Metcalfe, J. *J. Exp. Psychol. Learn. Mem. Cogn. Wash.* **2001**, *27* (6), 1491. doi:10.1037/0278-7393.27.6.1491.
(48) Butterfield, B.; Metcalfe, J. *Metacognition Learn.* **2006**, *1* (1), 69. doi:10.1007/s11409-006-6894-z.
(49) Butler, A. C.; Fazio, L. K.; Marsh, E. J. *Psychon. Bull. Rev.* **2011**, *18* (6), 1238. doi:10.3758/s13423-011-0173-y.
(50) Metcalfe, J.; Finn, B. *J. Exp. Psychol. Learn. Mem. Cogn.* **2011**, *37* (2), 437. doi:10.1037/a0021962.