JOURNAL OF CHEMICAL EDUCATION

pubs.acs.org/jchemeduc

Article

Formative and Summative Automated Assessment with Multiple-Choice Question Banks

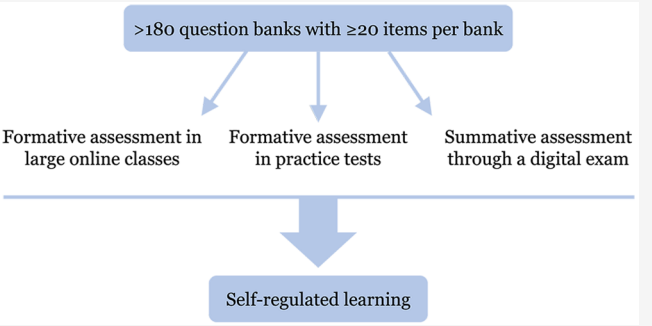
Maarten T. P. Beerepoot*

Cite This: *J. Chem. Educ.* 2023, 100, 2947−2955

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Digital automated assessment is a valuable and time-efficient tool for educators to provide immediate and objective feedback to learners. Automated assessment, however, puts high demands on the quality of the questions, alignment with the intended learning outcomes, and the quality of the feedback provided to the learners. We here describe the development and use of a large number ($N > 180$) of question banks with multiple items ($N \geq 20$) that are aligned to the intended learning outcomes of an undergraduate general chemistry course. Even though the development of a large number of high-quality question banks is a formidable task, they allow for flexible and easy-to-implement solutions for formative and summative assessment once developed. We here discuss three applications of the question banks: formative assessment in large online classes, practice tests that provide formative assessment outside classes, and summative assessment through a digital exam. We highlight the importance of aligning the question banks with intended learning outcomes, providing additional feedback to the learners and of quality assurance of the question banks, and show how the combined use of the question banks supports student self-regulated learning. We hope that the present work can inspire others to discover the various applications of question banks for formative and summative assessment.



KEYWORDS: *First-Year Undergraduate, General Chemistry, Testing/Assessment, Item Banks, E-Learning, Online Quizzes, Web-Based Learning, ChatGPT*

INTRODUCTION

Practice testing has been documented to be effective in learning, especially when practice is distributed over time and when learners are provided with feedback on their performance.[1] Spending a given amount of time on practice tests improves performance more than spending the same amount of time on restudying.[2] This so-called *testing effect* is present even when no feedback is provided, indicating that the act of testing itself improves retention of the material.[3] Testing concepts in an authentic classroom setting is beneficial for exam performance also when the concepts are assessed in a different way on an exam.[4,5] Thus, practice tests inside and outside the classroom can have a positive effect on student learning.

Automated assessment with multiple-choice (MC) items allows for immediate feedback to learners in a time-efficient manner, free from assessment bias. Indeed, learners can use digital practice tests whenever, wherever, and as often they want, without direct involvement of the educator. However, automated assessment also puts high demands on the quality of the items as well as on the reliability and validity of the test. Producing high-quality MC items is a challenging and time-consuming task that requires experience and training. Costello et al. investigated the quality of questions used in massive open online courses in different scientific fields and found common flaws in more than half of the questions, highlighting the importance of quality control and proper training of staff that write MC items.[6] Generating plausible and functioning distractors can be particularly challenging for educators.[7] A recent analysis of over 11,000 MC items across different undergraduate levels and disciplines revealed that 8% of all items were negatively discriminating (i.e., low-achieving students perform better than high-achieving students on that item), often because a distractor is erroneously set as the correct answer.[8]

Not only the quality of the questions, but also the educational context in which the digital tests are taken are decisive for success. Nicol[9] argues that MC items can be effectively implemented to develop students' self-regulated learning and provides explicit examples on how the educational context can be manipulated to satisfy the seven research-based principles of

Received: February 10, 2023
Revised: June 16, 2023
Published: July 18, 2023

ACS Publications

© 2023 The Author. Published by American Chemical Society and Division of Chemical Education, Inc.

https://doi.org/10.1021/acs.jchemed.3c00120
J. Chem. Educ. 2023, 100, 2947−2955

good feedback practice from Nicol and Macfarlane-Dick.[10] Hattie and Timperley reviewed the differential impact of different forms of feedback and found that computer-mediated feedback can be one of the most powerful forms of feedback.[11] In the absence of feedback, the danger of MC items is that students may endorse a lure, believing it to be the correct answer.[12] In the presence of feedback, however, this negative effect of MC testing is greatly reduced while a larger positive testing effect is observed,[13] highlighting the need for feedback when MC items are used. In addition to corrective feedback, frequent quizzing is another factor that is associated with high learning gains.[5]

MC items are widely used in undergraduate chemistry curricula and their use has been the object of various investigations in this journal.[14−27] Tellinghuisen and Sulikowski found that the performance on MC exams may depend on the order of the responses and to some degree also on the order of the questions in the exam.[14] Schroeder et al. found that these answer-ordering effects are more important for conceptual questions and that several consecutive difficult questions decrease performance compared to difficult questions that are spread over the entire test.[15] Hartman and Lin found that the percentage of students answering a particular question correctly decreases with five percent for each additional algorithmic step,[16] clearly showing that the difficulty of an item easily can be tuned by question design. Towns stresses the importance of well-formulated learning outcomes when making MC items and provides useful guidelines to help in the stem formulation, selection of distractors, and analysis of test results.[17] Domyancich shares strategies to (re)design MC items that assess higher-order cognitive skills.[18] Knaus et al. describe how a combination of performance data and mental effort measures on a practice exam with MC items can provide students and chemical educators with metacognitive information that can help shape self-study as well as teaching.[19] Several studies address academic integrity issues for online unproctored exams and discuss various strategies to minimize cheating including modification of the formulation of MC items.[20−22] These investigations as well as the concrete examples provided in the mentioned works are helpful in the development of MC items, question banks, and digital tests for chemistry education.

The aim of the current work is to show how MC *question banks* with similar items can contribute to various forms for formative and summative assessment. In the next section, we describe the development of the question banks in an undergraduate general chemistry course as well as the statistical item analysis used in this work. We then discuss three ways in which we use these question banks: for formative assessment in large online classes, for practice tests that provide formative assessment outside classes, and for summative assessment through a digital exam. Extensive analysis of student results and student perspectives from course evaluations is beyond the scope of this work. In the general discussion, we highlight the importance of quality assurance of the question banks and show how the three applications collectively support students' self-regulation of learning.

METHODOLOGY

Description of the Course

The question banks are used in a general chemistry course with approximately two hundred students from over ten different study programs including biology, biomedicine, pharmacy, biotechnology, chemistry, and geology. The contents of the course are divided in 14 topics with one topic per week and around five new intended learning outcomes per topic. The learning management system contains various resources for self-study, including prerecorded video lectures, short videos linked to the intended learning outcomes, references to relevant chapters in the textbook,[28] and practice tests. Students are expected to prepare for classes using these resources. Teaching activities in the course consists of on-campus seminar and laboratory classes in groups of approximately 20 students in combination with live online classes for all students together. A main objective of the seminar classes is to support the students in mastering the chemical concepts and chemical vocabulary for the present topic. Hence, focus in the seminar classes is on discussion of conceptual question in groups of around four students. Every group of approximately 20 students has one teaching assistant involved in both seminar and laboratory classes throughout the whole semester. A more detailed description of the large online classes and practice tests is included in the section on the application of the question banks.

Description of the Exam

Access to the exam is obtained by attending at least 10 out of 14 seminar classes and passing all five compulsory digital tests, two hand-in assignments and the laboratory course. The focus of the hand-in assignments is on drawing (structural formulas, structural isomerism, covalent vs ionic bonds, orbitals, Lewis structures, and three-dimensional molecular shapes) and explaining, which are learning outcomes that can only be indirectly assessed in MC items. In our experience, individual feedback on drawing skills often does not get enough attention in the classroom and certainly not in online classes. Formative assessment on the hand-in assignments is provided to the students in the form of rubrics and free-text comments. Compulsory digital tests, on the other hand, are automatically assessed. Ten out of 12 points are needed to pass all assignments and tests throughout the semester. This included preparatory tests for the laboratory course, but not the laboratory safety course where a full score is needed for the student to be allowed access to the laboratory. Students are given multiple attempts to pass all these course requirements. The high demands for passing the course requirements (10 out of 12 points corresponds to a passing threshold of 83%) typically result in students engaging with the course material from day one and using multiple attempts where needed, rather than dropping out when failing to meet the requirements. As is common in the Norwegian educational system, all assignments during the semester are formative and do not count toward the final grade. Thus, the final exam accounts for 100% of the students' grade.

Since 2020, the final exam is a 3 h digital exam which is automatically graded on an A−F scale where F represents fail. No points are subtracted for wrong answers. It is important for students to know the implications of the marking system.[29] Hence, students are urged to select an answer for each MC item. Alternative strategies exist and include discouraging guessing[24] and partial-credit scoring.[25,26]

The use of the question banks in the final exam is further discussed in the section on the application of the question banks, whereas aspects related to exam security are discussed here.

The exam was administered as a home exam in 2020 and 2021. The current version of the exam contains 50−55 questions, allowing students that have practiced adequately just enough time to finish the exam within the 3 h time frame. An

https://doi.org/10.1021/acs.jchemed.3c00120
J. Chem. Educ. 2023, 100, 2947−2955

Table 1. Example Questions Q1−Q7: Question Banks with a Large Number of Items Can Be Made by Exchanging the Set of Responses (Q1−Q4), a Part of the Stem (Q5−Q6), or Both

Q1	Which of the following is the longest bond? (A) H−Cl (B) H−F (C) H−I (D) H−Br
Q2	Which of the following is the strongest acid? (A) HCl (B) HF (C) HI (D) HBr
Q3	Which of the following ions has the largest radius? (A) O^{2-} (B) F^- (C) Na^+ (D) Mg^{2+}
Q4	Which of the following elements has the highest ionization energy? (A) He (B) Ne (C) Ar (D) Kr
Q5	How can HF be classified? (A) strong acid (B) weak acid (C) strong base (D) weak base
Q6	What kind of reaction is the following reaction? $6\ H^+\ (aq) + 2\ MnO_4^-\ (aq) + 5\ H_2O_2\ (aq) \rightarrow 2\ Mn^{2+}\ (aq) + 5\ O_2\ (g) + 8\ H_2O\ (l)$ (A) solution reaction (B) precipitation reaction (C) acid−base reaction (D) redox reaction
Q7	What is the oxidizing agent in the following reaction? $6\ H^+\ (aq) + 2\ MnO_4^-\ (aq) + 5\ H_2O_2\ (aq) \rightarrow 2\ Mn^{2+}\ (aq) + 5\ O_2\ (g) + 8\ H_2O\ (l)$ (A) Mn^{2+} (B) MnO_4^- (C) H_2O_2 (D) O_2

academic integrity pledge inspired by Nguyen et al.[21] and aligned with institutional and course-specific guidelines has been in place since the first administration in 2020. Randomization of the question order was introduced in 2021. Backtracking to earlier questions was permitted to facilitate the common exam-taking strategy of initially skipping difficult questions.[22] Even though the use of question banks,[20,23,30] integrity pledges,[21,30] time limits,[20,22,30] and randomization of the question order are all strategies that make collaboration more difficult, it is impossible to prevent collaboration in unproctored home exams entirely. Opportunities for collaboration give an unfair advantage to students who choose not to follow academic integrity regulations. Hence, the exam delivery was changed to an in-person proctored exam in 2022.

The use of a textbook, notes, and online resources during the exam was allowed between 2020 and 2022. The rationale behind was to focus on the *application* of knowledge and contribute to a more authentic assessment environment. Indeed, students can also use online resources when they apply general chemistry knowledge in work or studies later. It would be naive to *not* allow online resources in an unproctored home exam and assume that all students adhere to these guidelines.[22] In fact, Clark et al. have documented that exam questions where the answer can be searched online are answered correctly more often in an unproctored home exam than in an in-person proctored exam.[20] Our strategy has been—in line with others[20,22]—to formulate items such that they assess *application* of knowledge that is readily available by an online search. We predict, however, that the rapid advance of easily available artificial intelligence (AI) tools such as ChatGPT[31] will be a game-changer in chemistry assessment. Even though an early version of ChatGPT may struggle with nontext input and application questions[32] and provides answers that cannot be trusted,[27] it is likely only a matter of time before AI tools outperform most students on even the most advanced questions. Hence, we have decided to restrict the use of resources allowed during the exam to include only off-line resources (e.g., textbook, notes) from 2023.

Development of the Question Banks

We have so far developed over 180 question banks for the general chemistry course with at least 20 similar items per question bank. Originally, all items were MC items with one correct answer and usually three distractors. Other question types that allow for automated assessment have been added later and are being used for formative and summative assessment, but are not discussed specifically in the present work. The items are written in Norwegian and example questions (Q) in this work have been translated to English. For all purposes described in this work, it is essential that the question banks are aligned with the intended learning outcomes of the course. Indeed, each learning outcome is covered by at least one question bank to the extent to which this is possible. Conversely, one or more question banks together cover the essence of each learning outcome. As observed by Towns, well-defined learning outcomes facilitate item writing, and questions that are not directly related to an intended learning outcome should be avoided.[17]

Various strategies were used to construct 20 or more similar items in a question bank, as illustrated here using Q1−Q7 (Table 1). Some question banks contain items with an identical stem and different response sets containing for example bonds (Q1), molecules (Q2), ions (Q3), or elements (Q4). Other question banks contain items with identical response sets and similar stems, differing in for example a molecule (Q5) or a chemical equation (Q6) in the stem. In Q7, four different items with an identical response set were made per chemical equation by inverting the chemical equation and/or exchanging *oxidizing agent* with *reducing agent*. Since the typesetting of chemical equations is relatively time-consuming, chemical equations were to a large extent reused within the same (Q7) or different (Q6 and Q7) question banks. In response sets with numerical answers, distractors were mainly generated either from common computational errors—which is a recommended[17] but time-consuming strategy—or in a less time-consuming manner by creating a set of four MC items with an identical response set where each response is the correct answer to one of these items. The main guiding principle to construct distractors was to use plausible distractors only.[7,17]

AI text-generation software such as ChatGPT provides chemistry educators with a range of opportunities that were not easily available before, including generating assessment items and multiple versions and answers[33] and designing assignments where students assess ChatGPT responses to stimulate critical thinking.[27] ChatGPT and other resources could be helpful to generate multiple versions of an item and have the potential to assist in the creation of relevant and consistent distractors in a time-efficient manner. In addition to quality assurance—which is required for manual and AI-assisted

https://doi.org/10.1021/acs.jchemed.3c00120
J. Chem. Educ. 2023, 100, 2947−2955

item generation alike—this requires a workflow where the generated text output can be straightforwardly converted to actual MC items in the chosen digital environment.

The ordering of the responses in a MC item influences the performance, as has been demonstrated in the chemistry education literature.[14,15] In particular, performance is better when the correct answer appears earlier among the possible responses. We have not paid any particular attention to the ordering of responses in the items and instead randomized the response order for all questions in all tests where the question banks are used.

The question banks described here are used in online classes, in the compulsory digital tests that need to be passed to obtain access to the exam, in the digital exam, as well as in the practice tests available to the students throughout the entire semester.

Item Analysis

Item analysis[34] was used for quality assurance and analysis. Item statistics were downloaded from the learning management system for practice tests, compulsory tests and for the digital exam. These statistics include the number of students that have answered a particular item and the number of students that answered correctly. These item statistics are anonymous and contain data sorted per item rather than per student. The difficulty index p of an item is the ratio of correct answers (N_{correct}) to the total number of answers to that item (N_{tot}) in a particular test.[34]

$$p = \frac{N_{\text{correct}}}{N_{\text{tot}}}$$

The difficulty index ranges from 0 to 1 where 0 (1) represents the case where the question never (always) has been answered correctly. As a rule of thumb, an item with a difficulty index below 0.25 can be considered difficult whereas an item with a difficulty index above 0.75 can be considered easy.[17] Items with intermediate difficulty can in theory have the highest discriminatory power. In practice, however, adequate discrimination can be obtained for a wide range of difficulty indices and a target difficulty of $0.35 \geq p \geq 0.90$ has been established for MC items with four options.[8]

The discriminatory index D of an item indicates how well the item discriminates between students that have an overall high score on a test and students that have an overall low score on the test.[34] Thus, students are grouped together based on their overall test score. In this work, we have used the upper 27% and lower 27% groups to calculate the discriminatory index, which corresponds to the theoretically optimal choice of groups.[35] The discriminatory index is calculated by subtracting the difficulty index of the lower-performing group from the difficulty index from the upper-performing group. The discriminatory index thus ranges from −1 to 1 where an index of 1 represents the case where the upper 27% all answer correctly and the lower 27% all incorrectly. A discriminatory index above 0.35[8] or 0.40[17] is considered excellent, whereas an index below 0.15[8] or 0.20[17] indicates that the item is problematic. We note, however, that a more reliable measure of item discrimination can be obtained by excluding the score of the item in question from the total test score.[8]

We have similarly calculated the difficulty index and discriminatory index of a question bank by summing N_{correct} and N_{tot} over all items in a that question bank. In the same way, we can calculate the total difficulty of a test or of a series of MC items that is part of it. This approach yields statistically more robust indices. One should however be aware that the resulting difficulty and discriminatory power of the question bank are not necessarily representative for each of the individual items within that question bank.

We note that both the difficulty index and the discriminatory index are calculated from a specific test, in this work from practice tests, compulsory tests or the digital exam. Thus, the indices for the same item or question bank may differ between different tests. One might for example expect that the difficulty index of a particular question is higher in the final exam than in a practice test taken during the semester, reflecting among others the effect of practice.

APPLICATIONS OF THE QUESTION BANKS

The development of a large number of question banks is a formidable task. However, once developed, the question banks allow for flexible and easy-to-implement solutions for formative and summative assessment.[30] We here describe and discuss how we have used the question banks for formative assessment in large online classes, formative assessment in practice tests used outside classes and for summative assessment in a digital exam. In the general discussion that follows, we discuss quality assurance of the question banks and discuss how the various applications of the question banks in the course collectively support student self-regulated learning.

Formative Assessment in Large Online Classes

Teaching large classes may pose several challenges for student learning. Indeed, large classes often lead to low motivation, poor engagement, and poor student−teacher interaction. Students often assume a passive role in large classes and teachers may struggle to implement active learning with large student groups.[36] In large online classes, the importance of motivation, engagement and student-teacher interaction on student learning may be even more evident. The use of MC items in large classes has the potential to deal with some of the mentioned challenges. Their implementation in online classes is particularly straightforward since students are anyway connected to a computer or mobile device.

Discussion of items from the question banks constitute the central part of the online classes in the general chemistry course at UiT The Arctic University of Norway. The used teaching method can be classified as a *flipped-classroom* approach in the sense that information transmission is moved outside class, class time is used for active learning activities and students need to prepare for classes in order to fully benefit from in-class work.[37] A flipped-classroom approach to general chemistry has been shown to result in improved performance compared to a lecture-based approach for algorithmic and especially conceptual questions, as measured by a standardized exam.[38] With the exception of a short introduction on what is going on in the course in the particular week, the online classes consist of a series of polls with feedback (Figure 1). First, an item from the question banks is shown to the students, who then answer individually. Based on the distribution of the responses, the teacher then chooses to provide the students with feedback either through a hint or through discussion of the correct answer. This choice can be made on the basis of one of several criteria such as the difficulty index of the item or whether the most popular answer is correct or false. In either way, the teacher feedback is adapted based on (i) the percentage of students answering the item correctly (p); (ii) the relative frequency of each of the distractors chosen by the students; and (iii) the time

https://doi.org/10.1021/acs.jchemed.3c00120
J. Chem. Educ. 2023, 100, 2947−2955

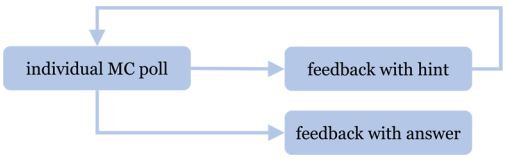


Figure 1. Flowchart illustrating the use of MC items and the role of feedback in large online classes. The choice whether to provide feedback through a hint or through an elaboration of the correct answer is made based on the distribution of the student responses to the individual poll.

it takes for the students to answer. After one to three—but usually one—round of polling and feedback, the same procedure is repeated for the next item, which can be selected from the same or from a different question bank. One round takes 6 min on average with a wide spread depending on the difficulty of the item and in particular on the time needed for the students to generate the answer.

Formative assessment is placed at the forefront in this poll-based teaching of large classes. The consecutive cycles of polling and acting on the students' responses provide continuous feedback in two directions. The teacher is provided with real-time data on the extent to which the students meet the various intended learning outcomes and can focus on identified areas of learning difficulty.[9] At the same time, the students are provided with clear examples of what they should be able to do (Feed Up), with information to which extent they master the intended learning outcomes (Feed Back), and with high-quality information on how to reduce the gap between current and desired mastery (Feed Forward), together providing them with the feedback needed to enhance their learning.[11] This teaching method thus allows for continuous diagnostic assessment to the teacher and continuous formative assessment to the students. One could in principle add a metacognitive dimension by having the students evaluate and report the mental effort used in answering each question, which provided the teacher with an additional layer of information to shape the feedback.[19]

In addition to well-documented learning gains from using testing in the classroom,[4,5] extensive use of MC items also allows for repetition and scaffolding in a straightforward manner. The first items in a class are usually taken from a previous topic and serve as repetition and to retrieve and activate relevant preknowledge. The relevance for the present topic can be stressed in the discussion of the correct answer. Indeed, the teacher can choose to provide the students with a mini-lecture introducing one of the main concepts of the present topic. An example of this scaffolding is provided in Table 2, where Q8 and Q9 test relevant concepts from a previous topic (molecular shape, bond polarity, and dipole moment), which are central to the topic of intermolecular forces (Q10 and Q11). A mini-lecture after the discussion of the correct answer to Q9 can consist of introducing hydrogen bonds, dipole−dipole forces, and dispersion forces as the three main types of intermolecular forces and explaining why only the latter act between two molecules of $SiCl_4$. Through this sequence of four questions, students not only receive feedback on the extent to which they master isolated learning outcomes from the previous and present topic, but also get exposed to the way in which the different topics are related to each other.

Taking the MC items from question banks has several advantages over constructing *ad hoc* items for teaching purposes only. First, students can practice with similar questions in practice tests before and after classes. Teacher feedback during classes can be a valuable supplement to the feedback (correct/false) that is provided in automatically assessed practice tests.[9] Second, the availability of multiple items in the same question bank easily allows for repetition of intended learning outcomes within and in between classes. Third, alignment of in-class MC items with items on compulsory tests and the final exam can be motivating,[9] resulting in high attendance and active participation. Finally, quality assurance of the question banks ensures the use of high-quality items, whereas items that are created *ad hoc* may suffer from various flaws.[6,8]

Table 2. Example Questions Q8−Q11: Sequence of MC Items Illustrating a Typical Scaffolding Procedure Used in Class

Q8	What is the predicted molecular shape of $SiCl_4$, based on VSEPR theory?
	(A) tetrahedral (B) square planar (C) seesaw (D) trigonal pyramidal
Q9	Which of the following molecules has a dipole moment?
	(A) $SiCl_4$ (B) SiO_2 (C) PCl_5 (D) PCl_3
Q10	Which are the strongest intermolecular forces between two molecules of PCl_3?
	(A) hydrogen bonds (B) dipole−dipole forces (C) ion−dipole forces (D) dispersion forces
Q11	Which are the strongest intermolecular forces between two molecules of PCl_5?
	(A) hydrogen bonds (B) dipole−dipole forces (C) ion−dipole forces (D) dispersion forces

Even though this use of MC items in large online classes may to some extent solve the challenges of poor motivation, poor engagement, and some aspects of poor student−teacher interaction, this is not the holy grail of teaching large (online) classes. Specifically, this teaching method is not particularly well-suited to test learning outcomes that require more time-consuming questions such as multiple-step calculations in a general chemistry course. Especially for these time-consuming questions, differences in response time between students lead to a situation in which some students are waiting after having given their response whereas others anyway do not get enough time to answer the question. As such, poll-based learning activities might not fully take advantage of the potential in flipped-classroom approaches to accommodate a mixed class of novices and experts.[37] Engaging large classes in problem-solving activities is, however, a general challenge of flipped-classroom approaches.[38]

Formative Assessment in Practice Tests

The question banks were used to construct one practice test for each of the 14 topics in the course to provide an opportunity for formative assessment outside classes. Each of the 14 practice tests is built up with 12 questions that are drawn from question banks. Where relevant, the first items are drawn from question banks from earlier topics to repeat relevant concepts that are related to the intended learning outcomes of the present topic. In this way, previous knowledge is repeated and relevant preknowledge is activated. We thus intentionally introduce priming effects in practice tests, but avoid those in compulsory tests and the final exam.[17] Even though related topics can be linked in this way (such as in Q9 and Q10 in Table 2), the same type of scaffolding as used in online classes cannot be used in practice tests since the specific items (specific molecules in Q9 and Q10) differ from one student to another and from one practice attempt to another. In some cases, easy and difficult versions of a question bank are used, such that the level of difficulty is deliberately increased for subsequent items in the

https://doi.org/10.1021/acs.jchemed.3c00120
J. Chem. Educ. 2023, 100, 2947−2955

practice test. This way of scaffolding questions in a practice test can also be used to provide systematic practice on the type of *essential skills* that a student needs to master before being able to solve more complex problems.[39]

The crucial aspect of the practice tests is that a student can take the test multiple times. For each attempt, the specific items are similar but not identical and cover the same intended learning outcomes. A table in the test instructions specifically links each item to the associated intended learning outcome and ideally also to a related learning resource. This resource is usually a short video on the learning outcome or a sample problem in the textbook.[28] The feedback on the practice tests consists of an overall score as well as feedback (correct/incorrect) on each question, such that students cannot believe an incorrect answer to be correct.[12] Even though providing the correct answer to a practice problem has been shown to result in increased *performance* on similar problems with an additional increase from providing a full solution, this effect is especially evident for problems with a high degree of similarity, suggesting a rather superficial effect on actual *learning*.[40] The practice tests described here are designed as a tool to support students in self-regulating their own learning. Students are advised to (i) use the practice tests iteratively and revisit (only) the topics related to the items that were answered incorrectly and (ii) revisit the practice tests after some weeks. This learning strategy is similar to *successive relearning*,[41] where a possible criterion could be to continue practicing until all 12 items have been answered correctly in the same attempt.

Even though the practice tests can be used throughout the entire semester and specifically before and after related teaching activities, most students use the practice tests after classes, in preparation to compulsory tests and in preparation for the final exam. Almost all students use the practice tests, although the percentage of students taking a given practice tests drops somewhat during the semester from roughly 95% for the first topic to roughly 80% for topics at the end of the course. The number of students that eventually obtain the maximum score for a given practice test drops more dramatically during the semester from roughly 70% to roughly 30% from the first to the last test, which could reflect among other things the difficulty of the practice tests and the number of opportunities to revisit earlier topics.

Summative Assessment

Items from the question banks were introduced in the final exam in two steps. In 2018 and before, the exam was a six-hour written exam on campus that only occasionally included one MC item as part of a larger assignment. In 2019, one assignment (out of six) with 15 items (the same items for all 149 students taking the exam) was introduced in the written on-campus exam, counting for 30% of the total number of points on the exam. Statistical analysis of the exam results in 2019 revealed that the difficulty index of this MC assignment ($p = 0.69$) was identical with that for the exam as a whole and that the assignment had an excellent discriminatory index of 0.43. The difficulty index and discriminatory index of the other five questions ranged from 0.60 to 0.78 and from 0.40 to 0.56, respectively.

In 2020, the exam was administered as a three-hour digital home exam consisting of 36 equally weighted MC items using the question banks described in this work. With 20 or more items per question bank, the total number of MC items used in the exam was between seven and eight hundred. Automated assessment of the exam was also introduced in the same year. Each student received a random item from the same 36 question banks, allowing for over 20^{36} different exam sets, by far exceeding the number of students that took the exam in that year ($N = 216$).

The resulting difficulty index of the total exam in 2020 was $p = 0.82$. The individual questions (here summed over all items in a question bank) had difficulty indices between $p = 0.48$ and $p = 0.96$ with five questions below $p = 0.72$ and five questions above $p = 0.88$. Even though almost all questions were within the target range for adequate discrimination,[8] most questions were "easy" according to Towns' rule of thumb ($p > 0.75$).[17] This probably reflects among other things the students' extensive practice in preparation for the exam.

Even though the use of the same question banks in practice tests and the final exam likely encourages students to practice extensively, one could also argue that MC items on the exam should be drawn from a different set of question banks. A different set of items on the exam may avoid students becoming proficient at the quiz items themselves rather than gaining in-depth understanding of the associated concepts.[30] In addition, using a novel set of exam questions in combination with a short time window for the exam might to some extent avoid exam questions appearing online on for example commercial tutoring Web sites.[22,23] We have therefore reduced the similarity of items between practice tests and the final exam in subsequent years by introducing different question banks and question types in the final exam such that currently only a minor part of the exam items can be encountered during the course.

Since the use of a textbook, notes and other resources was allowed during the exam in the period from 2020 to 2022, the use of items asking for reproduction of facts was limited. Instead, focus was on the application of knowledge such as examples Q12−Q15 (Table 3), which require a calculation. Q15 ($p = 0.86$) rewards understanding as an alternative to calculation if a student carefully examines the four alternatives and correctly applies qualitative knowledge on the titration of a weak acid with a strong base. For Q12, straightforward conversion of the given hydroxide concentration (1.0 mM) to pOH and subsequently to pH does not yield the correct answer since stoichiometry has to be considered. Indeed, this question was one of the most difficult ($p = 0.50$) and discriminatory ($D = 0.55$) questions on the exam. Q9 ($p = 0.90$) and Q13 ($p = 0.92$), on the other hand, were among the easiest ones. Items with a high difficulty index (easy items) cannot have a high discriminatory power. Hence, removing the easiest questions from the exam can be a strategy to develop a shorter exam with the same or better discriminatory

Table 3. Example Questions Q12−Q15: Test the Application of Knowledge Rather than the Recall of Facts

Q12	What is the pH of a 1.0 mM solution of barium hydroxide? (A) 11.0 (B) 11.3 (C) 14.0 (D) 13.7
Q13	A container contains only 2.00 g CO_2 (g) and 0.750 g He (g) at a pressure of 0.479 atm. What is the partial pressure of He? (A) 0.0915 atm (B) 0.128 atm (C) 0.341 atm (D) 0.377 atm
Q14	We prepare a buffer with pH = 1.50 from a weak acid HA ($K_a = 9.00 \times 10^{-2}$) and its natrium salt NaA. What is the ratio $[A^-]/[HA]$ in the buffer? (A) 1.4 (B) 1.8 (C) 2.3 (D) 2.8
Q15	Choose the right pH indicator for a titration of 20 mL of a 1.00 M solution of a weak acid ($K_a = 4.00 \times 10^{-6}$) with 1.00 M sodium hydroxide. (A) thymol blue (B) bromophenol blue (C) methyl red (D) phenolphtalein

https://doi.org/10.1021/acs.jchemed.3c00120
J. Chem. Educ. 2023, 100, 2947−2955

power. Other strategies to limit the advantage provided by using online resources are the using imaginary (but well-defined) units or constants[22] or irrelevant information or context.[21] With the advance of readily available AI tools such as ChatGPT,[31] however, these strategies are unlikely to prevent that MC exams where online resources are allowed will not be adequate to assess understanding in chemistry in the long run.

While question banks containing items with varying difficulty do not pose a problem in no-stake practice tests, using question banks in summative assessment requires the items within a question bank to have similar difficulty. In this way, all students are treated equally despite different students getting different items. Even though the effect of items with varying difficulty might average out to some extent, there is no guarantee that this happens in all cases. For example, the difficulty of drawing a Lewis structure depends to a large extent on the molecule. Indeed, items become more difficult when the selection of the central atom is nontrivial or when the valence shell of the central atom violates the octet rule, in particular in cases with an odd number of electrons.[42] The question banks selected for use in the final exam were therefore carefully examined and adapted where needed. For questions that require generating a correct Lewis structure, for example, we ensure that different version of the exam contain the same number of molecules where the central atom violates the octet rule. As another illustration, we consider Q12 in which students calculate the pH for a given solution of a given strong acid or base. Originally, the question bank consisted of items of varying difficulty according to item analysis from a compulsory test given during the course (Table 4). Straightforward conversion from the concentration of a monoprotic acid to pH ($p = 0.98$) is easier than conversion from alkali hydroxide concentration to pH ($p = 0.84$), which in turn is easier than conversion from alkali earth hydroxide concentration ($p = 0.53$), in which stoichiometry must be taken into account. Thus, the difficulty increases with the number of algorithmic steps in the solution.[16] For the final exam, two question banks were made from categories with a high discriminatory index: pH calculation of solutions of alkaline earth hydroxides and for strong bases with a concentration below 10^{-7} M. These two questions banks were the most difficult ones on the exam with difficulty indices (from the exam) of 0.50 and 0.48 and high discriminatory indices of 0.55 and 0.60, respectively.

Table 4. Difficulty Index p and Discriminatory Index D for Variants of the Q12[a,b,c,d]

acid or base	concentration	N_i	N_{tot}	p	D
HCl or HNO_3	≥1 mM	4	47	0.98	0.00
NaOH or KOH	≥1 mM	3	31	0.84	0.50
$Ca(OH)_2$, $Sr(OH)_2$, or $Ba(OH)_2$	≥1 mM	7	66	0.53	0.51
HCl or HNO_3	$\ll 10^{-7}$ M	2	18	0.28	0.20
NaOH or KOH	$\ll 10^{-7}$ M	1	12	0.17	0.50
$Ca(OH)_2$, $Sr(OH)_2$, or $Ba(OH)_2$	$\ll 10^{-7}$ M	3	28	0.04	0.25
all items in the question bank		20	202	0.57	0.47

[a]N_i is the number of MC items in the question bank for each category. [b]N_{tot} is the number of times an item in the specified category has been answered on the test. [c]The question bank contains 20 items differing in (i) the strong acid/base, which is either a monoprotic acid, an alkali hydroxide, or an alkaline earth hydroxide and (ii) the concentration of the acid/base, which is either at least 1 mM or much lower than 10^{-7} M. [d]The data are obtained from a compulsory test during the semester that was taken by 202 students. The last row is the sum row.

Advantages of using *question banks* for online exams include randomization of items among students, time-efficient generation of an exam set as well as straightforward generation of practice tests and practice exams that can be taken multiple times.[30] Additional advantages of using MC items is that a large number of learning outcomes can be assessed in a 3 h exam and that the possibility of automated assessment saves the instructor a lot of time. Indeed, one could argue that the time of the educator should rather be spent on assessment *for* learning than on assessment *of* learning. Disadvantages of using question banks include the increased time needed to generate items and technical challenges in the submission of student drawings.[30] Indeed, using the approach described in this work, learning outcomes based on drawing and explaining can at best be assessed indirectly. In the present general chemistry course, these learning outcomes are therefore assessed in compulsory hand-in assignments rather than on the final exam. A possible alternative is to use a combination of automatically assessed items and manually assessed assignment for summative assessment.

GENERAL DISCUSSION

Quality Assurance of the Question Banks

Extensive use of question banks with MC items requires thorough quality assurance. Making an occasional mistake is inherent in the process of generating a large number of items. Typical problems include ambiguous formulations,[6] non-functioning distractors,[7] and incidentally selecting a distractor as the key.[8] Such problems may cause frustration among students and are not acceptable for any of the discussed applications. In Table 5 we present four strategies that have been proven successful for quality assurance of the question banks presented in this work.

Table 5. Selected Strategies for Quality Assurance in the Development of Question Banks

Strategy 1	Discuss *one* item of a question bank thoroughly with a colleague before extending the question bank with multiple similar items.
Strategy 2	Encourage colleagues to go through all tests and discuss all items that may be unclear or problematic.
Strategy 3	Analyze test results systematically with a special focus on items with a low or even negative discriminatory index, which could indicate a wrong key.[17] Review the entire question bank when one of its items is found to be flawed.
Strategy 4	Implement routines for students to report questions that they suspect might be wrong. In these cases, either the item or the student's understanding of the material needs to be corrected.

Supporting Student Self-Regulation of Learning

A central argument of Nicol is that high-quality tests are just one of two ingredients for successful use of MC items to enhance student learning.[9] The other is carefully planning the educational context in which the tests are used. In Table 6 we show how the three applications of the question banks collectively support self-regulated student learning through discussion of the seven principles of good feedback practice to support self-regulated learning from Nicol and Macfarlane-Dick[10] as applied to MC tests by Nicol.[9]

https://doi.org/10.1021/acs.jchemed.3c00120
J. Chem. Educ. 2023, 100, 2947−2955

Table 6. Seven Principles of Good Feedback Practice That Support Self-Regulated Learning[9,10] in Relation to the Three Applications of the Question Banks Described in This Work

Clarifying goals, criteria, and standards	In online classes and practice tests, students are continuously exposed to the type of questions they need to be able to master on the exam.
Self-assessment and reflection	In practice tests, students find out which learning outcomes they master and receive feedback (correct/false) on their attempts. Items are drawn from question banks such that students can choose to continue practicing with similar questions, based on their self-assessment.
Delivering high-quality feedback	In online classes, the teacher provides the students with feedback based on the distribution of the students' responses. Drawing MC items from question banks ensures that these questions are similar to the items that students encounter in practice tests.
Encouraging dialogue around learning	In online classes, MC items encourage student−teacher dialogue. In practice tests, students get different items form the same question bank, which may encourage peer dialogue. This potential can be exploited further by encouraging students to take practice tests together.
Feedback and motivation	In practice tests, drawing MC items from question banks ensures that students have repeated opportunities to take practice tests with items that are aligned to the exam.
Closing the gap	The practice tests provide the students with opportunities to close the gap between current and desired performance. Drawing MC items from question banks ensures that students can practice with similar questions after checking their answers and revisiting the topics they did not master.
Feedback shaping teaching	Analysis of student results from practice tests provides the teacher with feedback to shape teaching. MC items in online classes can be chosen on the basis of this analysis. The question banks provide the teacher with a wide choice of high-quality items to select from.

CONCLUSION

We have discussed three applications of MC question banks that are aligned to the learning outcomes in a general chemistry course: formative assessment and student-active learning in large online classes, practice tests that provide formative assessment outside classes, and summative assessment through a digital exam. Even though the development of high-quality question banks is a formidable task, they allow for flexible and easy-to-implement solutions for formative and summative assessment. By carefully manipulating the educational context in which MC items are used, we can use automated assessment to provide students with repeated opportunities for formative assessment with the ultimate goal of improving student self-regulated learning. We hope that the present work can inspire others to discover the various applications of MC question banks for formative and summative assessment.

AUTHOR INFORMATION

Corresponding Author

Maarten T. P. Beerepoot − *Department of Chemistry, UiT The Arctic University of Norway, N-9037 Tromsø, Norway;* orcid.org/0000-0003-3976-9223; Email: maarten.beerepoot@uit.no

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jchemed.3c00120

Notes

The author declares no competing financial interest.

ACKNOWLEDGMENTS

The author thanks all colleagues who have contributed through discussion of the work as well as through additions to and quality assurance of the question banks.

REFERENCES

(1) Dunlosky, J.; Rawson, K. A.; Marsh, E. J.; Nathan, M. J.; Willingham, D. T. Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest* **2013**, *14*, 4−58.

(2) Rowland, C. A. The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin* **2014**, *140*, 1432−1463.

(3) Roediger, H. L., III; Karpicke, J. D. The power of testing memory: basic research and implications for educational practice. *Perspectives on Psychological Science* **2006**, *1*, 181−210.

(4) Nguyen, K.; McDaniel, M. A. Using quizzing to assist student learning in the classroom: the good, the bad, and the ugly. *Teaching of Psychology* **2015**, *42*, 87−92.

(5) Yang, C.; Luo, L.; Vadillo, M. A.; Yu, R.; Shanks, D. R. Testing (quizzing) boosts classroom learning: a systematic and meta-analytic review. *Psychological Bulletin* **2021**, *147*, 399−435.

(6) Costello, E.; Holland, J.; Kirwan, C. The future of online testing and assessment: question quality in MOOCs. *International Journal of Educational Technology in Higher Education* **2018**, *15*, 1−14.

(7) Tarrant, M.; Ware, J.; Mohammed, A. M. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education* **2009**, *9*, 1−8.

(8) Slepkov, A. D.; Van Bussel, M. L.; Fitze, K. M.; Burr, W. S. A baseline for multiple-choice testing in the university classroom. *SAGE Open* **2021**, *11*, 215824402110168.

(9) Nicol, D. E-assessment by design: using multiple-choice tests to good effect. *Journal of Further and Higher Education* **2007**, *31*, 53−64.

(10) Nicol, D. J.; Macfarlane-Dick, D. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education* **2006**, *31*, 199−218.

(11) Hattie, J.; Timperley, H. The power of feedback. *Rev. Educ. Res.* **2007**, *77*, 81−112.

(12) Roediger, H. L., III; Marsh, E. J. The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **2005**, *31*, 1155−1159.

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

(14) Tellinghuisen, J.; Sulikowski, M. M. Does the answer order matter on multiple-choice exams? *J. Chem. Educ.* **2008**, *85*, 572−575.

(15) Schroeder, J.; Murphy, K. L.; Holme, T. A. Investigating factors that influence item performance on ACS exams. *J. Chem. Educ.* **2012**, *89*, 346−350.

(16) Hartman, J. R.; Lin, S. Analysis of student performance on multiple-choice questions in general chemistry. *J. Chem. Educ.* **2011**, *88*, 1223−1230.

(17) Towns, M. H. Guide to developing high-quality, reliable, and valid multiple-choice assessments. *J. Chem. Educ.* **2014**, *91*, 1426−1431.

(18) Domyancich, J. M. The development of multiple-choice items consistent with the AP Chemistry curriculum framework to more accurately assess deeper understanding. *J. Chem. Educ.* **2014**, *91*, 1347−1351.

(19) Knaus, K. J.; Murphy, K. L.; Holme, T. A. Designing chemistry practice exams for enhanced benefits. An instrument for comparing performance and mental effort measures. *J. Chem. Educ.* **2009**, *86*, 827−832.

(20) Clark, T. M.; Turner, D. A.; Rostam, D. C. Evaluating and improving questions on an unproctored online general chemistry exam. *J. Chem. Educ.* **2022**, *99*, 3510−3521.
(21) Nguyen, J. G.; Keuseman, K. J.; Humston, J. J. Minimize online cheating for online assessments during COVID-19 pandemic. *J. Chem. Educ.* **2020**, *97*, 3429−3435.
(22) Raje, S.; Stitzel, S. Strategies for effective assessments while ensuring academic integrity in general chemistry courses during COVID-19. *J. Chem. Educ.* **2020**, *97*, 3436−3440.
(23) Clark, T. M.; Callam, C. S.; Paul, N. M.; Stoltzfus, M. W.; Turner, D.; Spinney, R. Testing in the time of COVID-19: A sudden transition to unproctored online exams. *J. Chem. Educ.* **2020**, *97*, 3413−3417.
(24) Campbell, M. L. Multiple-choice exams and guessing: results from a one-year study of general chemistry tests designed to discourage guessing. *J. Chem. Educ.* **2015**, *92*, 1194−1200.
(25) Carberry, T. P.; Lukeman, P. S.; Covell, D. J. Bringing nuance to automated exam and classroom response system grading: a tool for rapid, flexible, and scalable partial-credit scoring. *J. Chem. Educ.* **2019**, *96*, 1767−1772.
(26) Grunert, M. L.; Raker, J. R.; Murphy, K. L.; Holme, T. A. Polytomous versus dichotomous scoring on multiple-choice examinations: development of a rubric for rating partial credit. *J. Chem. Educ.* **2013**, *90*, 1310−1315.
(27) Clark, T. M. Investigating the use of an artificial intelligence chatbot with general chemistry exam questions. *J. Chem. Educ.* **2023**, *100*, 1905−1916.
(28) Silberberg, M. S.; Amateis, P. G. *Chemistry: The Molecular Nature of Matter and Change*, 9th ed.; McGraw-Hill Education: New York, 2021.
(29) Burton, R. F. Multiple-choice and true/false tests: myths and misapprehensions. *Assess. Eval. Higher Educ.* **2005**, *30*, 65−72.
(30) Krzic, M.; Brown, S. Question banks for effective online assessments in introductory science courses. *Natural Sciences Education* **2022**, *51*, No. e20091.
(31) Rudolph, J.; Tan, S.; Tan, S. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *J. Appl. Teach. Learn.* **2023**, *6* (1), 342−363, DOI: 10.37074/jalt.2023.6.1.9.
(32) Fergus, S.; Botha, M.; Ostovar, M. Evaluating academic answers generated using ChatGPT. *J. Chem. Educ.* **2023**, *100*, 1672−1675.
(33) Emenike, M. E.; Emenike, B. U. Was this title generated by ChatGPT? Considerations for artificial intelligence text-generation software programs for chemists and chemistry educators. *J. Chem. Educ.* **2023**, *100*, 1413−1418.
(34) Ding, L.; Beichner, R. Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics - Physics Education Research* **2009**, *5* (17), 020103.
(35) Kelley, T. L. The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology* **1939**, *30*, 17.
(36) Mulryan-Kyne, C. Teaching large classes at college and university level: challenges and opportunities. *Teaching in Higher Education* **2010**, *15*, 175−185.
(37) Abeysekera, L.; Dawson, P. Motivation and cognitive load in the flipped classroom: definition, rationale and a call for research. *Higher Education Research & Development* **2015**, *34*, 1−14.
(38) Weaver, G. C.; Sturtevant, H. G. Design, implementation, and evaluation of a flipped format general chemistry course. *J. Chem. Educ.* **2015**, *92*, 1437−1448.
(39) Mikula, B. D.; Heckler, A. F. Framework and implementation for improving physics essential skills via computer-based practice: Vector math. *Physical Review Physics Education Research* **2017**, *13* (23), 010122.
(40) Fakcharoenphol, W.; Potter, E.; Stelzer, T. What students learn when studying physics practice exam problems. *Physical Review Special Topics - Physics Education Research* **2011**, *7* (7), 010107.
(41) Rawson, K. A.; Dunlosky, J. Successive relearning: an underexplored but potent technique for obtaining and maintaining knowledge. *Current Directions in Psychological Science* **2022**, *31*, 362−368.
(42) Brady, J. A.; Milbury-Steen, J. N.; Burmeister, J. L. Lewis structure skills: Taxonomy and difficulty levels. *J. Chem. Educ.* **1990**, *67*, 491−493.

https://doi.org/10.1021/acs.jchemed.3c00120
J. Chem. Educ. 2023, 100, 2947−2955