Journal of Science Education and Technology (2023) 32:453–467
https://doi.org/10.1007/s10956-023-10043-2

Digitalization of Multistep Chemistry Exercises with Automated Formative Feedback

Carolin Eitemüller[1] · Florian Trauten[1] · Michael Striewe[2] · Maik Walpuski[1]

Accepted: 20 March 2023 / Published online: 31 March 2023

Abstract
For various reasons, students receive less formative feedback at post-secondary institutions compared to secondary school. Considering feedback as one of the most important influencing factors on learning processes, formative feedback is a promising approach to improving students' performances. In this context, new technologies, such as learning management systems (LMS) or intelligent tutoring systems (ITS), can make a valuable contribution to improving higher education teaching by providing automated and individualized error-specific just-in-time (JIT) feedback. However, the digitalization especially of paper-based open-ended tasks that can be used by LMS is currently still associated with a loss of quality. In this paper, we present an approach that allows us to transfer open-ended paper-based tasks in the field of chemistry into online tasks without losing quality and provide large university courses with automated and individualized error-specific JIT feedback. Results of a study of 238 first-year chemistry students reveal that the automated individualized error-specific JIT feedback had a significant positive influence on students' performance.

Keywords E-assessment · Automated formative feedback · Chemistry education · Higher education

Introduction

Formative Feedback in Higher Education

Professional assistance and academic support are of particular importance for students' study success since students receive feedback on their performances as well as hints on how to solve problems in personal contact with tutors or instructors (Heublein et al., 2017). In view of large numbers of participants in introductory courses, universities have to allocate vast personnel resources to provide appropriate support. Due to limited financial means, universities are often not in a position to satisfy students' needs for individual feedback sufficiently. As a result, students receive less feedback from tutors or professors in university than they do in school (York, 2003). In this context, the most common form of feedback is summative feedback given by the final exam grade at the end of the semester. However, formative feedback has the potential to support students in their exam preparation (Hattie, 2013) since students often overestimate their abilities, especially the low-performing students in general chemistry (Pazicni & Bauer, 2014), and do not prepare themselves sufficiently (de Bruin et al., 2017; Kruger & Dunning, 1999). This may lead to performance problems that, in turn, are a main reason for students' high dropout rates in chemistry in German higher education institutions (Heublein et al., 2020). Students who discontinue their studies are defined as persons who have taken up their first degree in chemistry at a German higher education institution through enrollment but leave the system without a (first) degree. Comparable results are reported by Chen (2013) for the USA.

In addition, students spend a lot of time learning outside lectures and tutorials, where they often do not receive any kind of professional feedback (Heublein et al., 2017). Kanuka (2001) found that students in distance learning programs identified a lack of timely and informative feedback as a problem. Thus, an important goal of academic teaching is to support students' learning process through

✉ Carolin Eitemüller
carolin.eitemueller@uni-due.de

[1] Department of Chemistry Education, University of Duisburg-Essen, Schützenbahn 70, 45127 Essen, Germany

[2] Paluno–The Ruhr Institute for Software Technology, University of Duisburg-Essen, Gerlingstraße 16, 45127 Essen, Germany

formative feedback during private study time. New technologies, such as learning management systems (LMS) or intelligent tutoring systems (ITS), can make a valuable contribution to improving higher education teaching by providing individualized JIT feedback (Ma et al., 2014). Immediate feedback is easy to implement there and has many advantages over manual correction regarding consistency and accuracy. In this context, a systematic literature review on automatic feedback generation in LMS shows that 82.5% of the included studies could not find evidence that manual feedback is more efficient than automatic feedback (Cavalcanti et al., 2021). In addition, the majority of the papers retrieved in the literature review (65%) concluded that the automatic feedback had a positive impact on students' performance. Immediate feedback in homework programs can also offer advantages in terms of timing of feedback to students compared to written homework where a large time gap between when the assignment is completed and when it is returned exists (Malik et al., 2014). Additionally, studies show that web-based homework with instant feedback has a positive influence on student achievement (Cole & Todd, 2003; Freasier et al., 2003).

Online Tasks—Opportunities and Limitations

Electronic tasks for digital learning systems are a promising approach to giving students formative feedback in their private learning time. These tasks have to be realistic in task procedure and have to automatically assess students' solutions and generate appropriate feedback. With the help of modern systems (e.g., the LMS Moodle), many paper-based tasks have already been successfully digitized without loss of quality (Trauten et al., 2019). However, traditional task formats like multiple-choice, drop-down, or fill-in that are provided by most of the LMS are not sufficient for the digitalization especially of paper-based open-ended tasks. This will be illustrated by a chemistry-specific exercise type below.

The concept of chemical reactions is of particular relevance for chemistry as it is internationally regarded as fundamental to the area (American Association for the Advancement of Science, 2001). Therefore, knowledge and skill acquisition in the field of chemical reactions are essential for the successful study in chemistry. However, students have major deficits in this field (DeBoer et al., 2009; Ferber, 2014; Walpuski et al., 2011), and studies confirm that some of the misconceptions still present at university level (Busker et al., 2010). Thus, online tasks that can provide formative feedback are required for the field of chemical reactions. A typical task is setting up a chemical reaction equation (cf. Fig. 1). However, the digitalization of corresponding

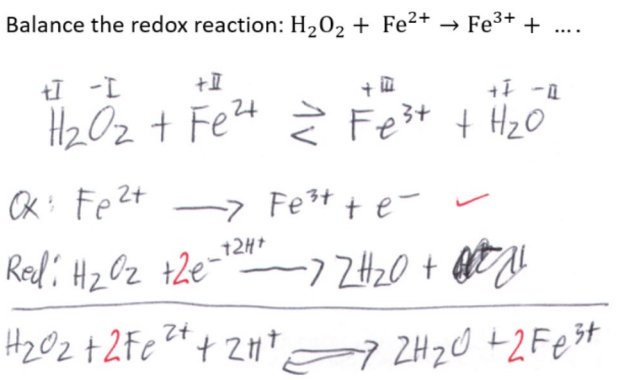

Fig. 1 Example of an open-ended exercise from the final exam in general chemistry at the end of the first semester

open-ended paper-based tasks using traditional item formats (e.g., multiple-choice, fill-in) currently leads to a loss of quality, either through a change of the learning objective or a limited assessment or reduced feedback (i.e., knowledge of correct response). We will discuss the loss of quality in more detail in the two subsections below. These difficulties can be summed up in two major challenges that have to be overcome when digitizing paper–pencil tasks with traditional item formats (closed-ended vs. open-ended). On the one hand, these difficulties relate to the aspect of user input and the automated assessment of solutions and on the other hand to the provision of formative individualized feedback.

Closed-Ended Tasks

User Input

When transferring open-ended paper-based tasks as seen in Fig. 1 into closed-ended online tasks, answers are inevitably predefined, even after thorough conception of distractors. A disadvantage of these tasks is that they require different skills. The recognition of a correct answer (e.g., a reaction equation) generally requires different, less complex cognitive processes than the independent reproduction of knowledge (Anderson & Bower, 1972). In the literature, it has been discussed for quite some time whether closed-ended task formats are able to represent higher learning goals at all in comparison to open-ended task formats (Lindner et al., 2015). In accordance with this assumption, multiple-choice tasks have been proven to be easier than open-ended tasks in some studies (Bonner, 2013; Hohensinn et al., 2011; Kastner & Stangl, 2011; Liu et al., 2011). Moreover, multiple-choice tasks seem to differentiate more poorly in the marginal areas of competence than open-ended tasks (Liu et al., 2011). Hence, questions in which students have to apply or use the

knowledge they have acquired (in the sense of the application level of Bloom's taxonomy) are difficult to create in a closed-ended item format.

Assessment and Feedback

Closed-ended tasks, like multiple-choice tasks, can be checked fully automatically and can, therefore, be analyzed time efficiently and objectively for a long time now. Various LMS, such as ILIAS or Moodle, already offer individual feedback components for closed-ended online tasks. However, the feedback is limited to information on the number of points achieved (knowledge of performance) and information on the correctness of an answer (knowledge of results) or the correct results (knowledge of correct response), which can be supplemented by multiple-try or answer-until-correct feedback depending on the setting. To support students' learning processes individually with appropriate feedback, distractors have to be created carefully (Haladyna & Rodriguez, 2013), which makes the transfer of paper-based tasks into online tasks a complex issue. Moreover, even with carefully designed distractors, students will not receive any appropriate feedback on their misconceptions if they reached an answer that is not available as distractor at all, and even feedback dealing with the various distractors remains highly speculative concerning the underlying misconception.

Open-Ended Tasks

Assessment and Feedback

In contrast, open-ended tasks allow individual feedback that can be tailored to the answers. Since there is no limit to the number of possible correct answers for open-ended tasks, generating meaningful feedback is much more complex than for closed-ended tasks. Another disadvantage of open-ended tasks is that solutions often have to be checked manually to provide feedback. Although better software solutions that enable free text input are available now, the assessment of these approaches does not go beyond checking keywords (Bridgeman et al., 2012; Shermis & Burstein, 2002), using complex rulesets with higher precision but low recall (Leacock & Chodorow, 2003) or employing machine learning techniques that require large sets of training data (Hussein et al., 2019; Shermis & Burstein, 2013). Assessment and feedback systems with open-ended tasks are, thus, mainly available for domains with comparatively structured content, since that makes it easier to check submitted solutions automatically. Examples can be found in programs for the fields of Boolean algebra (Herding et al., 2010), mathematical logic (Lodder & Heeren, 2011), and programming (Keuning et al., 2018). However, so far, we do not know of any digital tool that allows to realize both, the input and valid assessment of a chemical reaction equation in LMS like Moodle.

User Input

It is much easier to realize challenging learning goals such as setting up a reaction equation with open-ended tasks. For the input and processing of reaction equations by mouse or keyboard, there are various programs (e.g., ChemDraw,[1] ChemDoodle,[2] JSME[3]). However, these programs are unable to check the entered answers and provide feedback. A further disadvantage of these programs is that the degrees of freedom in drawing chemical compounds are often restricted. For example, these programs automatically flag violations of allowed valences. As these programs are used by experts, they usually know how to deal with the feedback. From a didactical point of view, however, the feedback is not sufficient.

Related Work

Individual Online Solutions

Surprisingly, there are only a few studies in literature that deal with the question of how paper-based chemistry-specific tasks can be digitized and provided with formative feedback. First of all, there are some studies, in which individual online solutions were developed to enable automated evaluation of free text responses in chemistry (Ashton et al., 2005; Chamala et al., 2006; Penn & Al-Shammari, 2008; Perry et al., 2007). For the PASS-IT Project (Project for Assessments in Scotland using Information Technology), traditional paper examinations in Higher Chemistry and Computing were transferred into an electronic format and compared to each other (Ashton et al., 2005). Although the authors report on modifications that had to be made when transferring the paper-based tasks into an electronic version, no differences between the two types of assessment could be determined. For example, as it was technically too difficult at that time to convert a paper-based task, in which a graph had to be drawn into an electronic version, this task was transferred into a multiple-choice question. Unfortunately, tasks that require students to set up reaction equations were not reported in the study. For the field of chemical engineering, Perry and colleagues developed a software solution that can automatically assess submitted solutions and provide formative feedback (Perry et al., 2007). However, the program, used by the University of Manchester, has no editor for entering chemical reaction equations. Although the program is an advanced tool for creating tasks, Perry and colleagues report difficulties in transferring graphics and formulas into the software (Perry et al., 2007).

[1] https://chemdrawdirect.perkinelmer.cloud/js/sample/index.html

[2] https://web.chemdoodle.com/demos/sketcher/

[3] https://peter-ertl.com/jsme/JSME_2017-02-26/JSME.html

These are only transferred as images to the online tasks. To digitally represent paper-based tasks with the learning objective of setting up chemical reaction equations, closed-ended tasks are used that have the disadvantages described in the section above. In addition, for organic chemistry, more recent software packages exist that include tools for drawing structures and organic reaction mechanisms. For example, the electronic program for organic chemistry homework (EPOCH) is a Java-based web application that offers graphical input and provides response-specific feedback (Chamala et al., 2006). EPOCH uses a series of conditions to analyze each response. Each condition is associated with feedback that EPOCH provides to a student whose response satisfies that condition. A further approach for drawing reaction mechanisms provides the curved arrow neglect (CAN) method from Penn and Al-Shammari (2008), which focuses on drawing and evaluation of reaction intermediates but ignores curved arrow notation itself. In addition, there are some online homework systems that allow for students to add the curved arrow notation for reaction equations (e.g., this is a feature in Achieve) (cf. section below).

Homework Systems

Beyond that, there are commercial online homework systems, which respond to individual mistakes. Responsive or responsive-adaptive online homework systems widely used in chemistry include Pearson's Mastering Chemistry,[4] ALEKS[5] (Assessment and Learning in Knowledge Spaces), and Achieve,[6] which all provide hints and feedback. For example, Pearson's Mastering Chemistry provides specific hints and tutorials to a student based on how the problem was missed. The system has graphical templates for inserting both mathematical and chemical formulas and symbols (Shepherd, 2009). However, homework systems are expensive and a German-language version does not exist yet. They are widely used in the USA and Canada. In contrast, a majority of German universities use LMS like Moodle, ILIAS, or Blackboard (Schmid et al., 2017), for which digital tools for *entering* and *assessing* chemical reaction equations or molecular formulas currently do not exist.

Learning Management Systems (LMS)

Hedtrich and Graulich (2018) pursued an initial approach to developing chemistry-specific tools for LMS that provide automated, individual feedback. They developed two digital tools with which it is possible to support students in their learning processes with formative feedback. One of the two developed software components reads out students' personal test results from the LMS, in which chemical-specific tasks were processed. The second software component automatically and individually generates and gives students feedback on their achieved performance level based on these data. Since students have to work on tasks in a LMS for this approach, only the classic task formats provided by the LMS are available.

Research Aim and Questions

Against this background, an important question is how tasks representing the learning goal of setting up chemical reaction equations can be digitized for the use of the LMS Moodle without loss of quality. After extensive research, we have not found an existing type of task that can fully meet the described requirements. There are only online homework systems with corresponding features available for a fee (cf. section below), but these are rarely used outside the USA and Canada and do not offer a German-language version, for instance. The aim of this study was to develop and evaluate a freely available digital tool for *entering* and *assessing* chemical reaction equations or molecular formulas as well as providing individualized error-specific JIT *feedback*. It is emphasized here that the assessment of students' solutions should be automated and should not require manual correction by a human tutor. As derived above, the domain-specific requirements posed two major challenges for the development of appropriate tasks. On the one hand, it must be possible to enter letters, numbers, indices, exponents, and special characters (e.g., arrows) in order to set up reaction equations. On the other hand, it is necessary to check students' responses based on subject-specific rules in order to provide individualized error-specific and fully automated JIT feedback. Besides the development of such a digital tool, the aim of the study was to investigate how students learn with tasks (i.e., whether they can correct their mistakes with the help of the automated and individualized error-specific JIT feedback). The underlying research question is as follows:

> RQ1: How helpful is the automated and individualized error-specific JIT feedback for error correction and solving the tasks?

Automated and individualized error-specific JIT feedback (IND-feedback) allows assisted multiple response tries for an exercise (a) by providing strategically useful information for error correction, but no immediate knowledge of correct response, and (b) by requiring the learner to apply the corrective information to a further attempt this exercise (cf. Narciss & Huth, 2006).

[4] https://www.pearsonmylabandmastering.com/northamerica/masteringchemistry/

[5] https://www.aleks.com/about_aleks

[6] https://www.macmillanlearning.com/college/us/digital/achieve

As an indicator for the usefulness of the feedback, students' performance in subsequent attempts to solve the task is considered. The percentage of students who were able to correct their mistake and solve the task after an incorrect solution attempt with the help of the feedback can be used to assess the usefulness of the feedback.

Considering that feedback is one of the most important influencing factors on learning processes (Hattie & Timperly, 2007), we wanted to further investigate whether the tasks providing automated and individualized error-specific JIT feedback (IND-feedback) improve students' performances compared to tasks providing automated corrective JIT feedback (COR-feedback), which are widely used in LMS like Moodle. Automated and corrective JIT feedback (COR-feedback) also allows assisted multiple response tries for an exercise. In contrast to the IND-feedback, it presents only knowledge of results, but no strategically useful information for error correction and no immediate knowledge of correct response. This type of feedback is usually implemented in computer based-trainings. The underlying research question is as follows:

RQ2: To what extent does the performance of students who learned with tasks providing individualized error-specific feedback (IND-feedback) differ from that of students who learned with tasks providing corrective feedback (COR-feedback)?

Research Design

In the winter terms 2019/2020 and 2020/2021, new learning tasks were provided to first-year B.Sc. Chemistry and B.Sc. Water Science students from authors' university in an online course on general chemistry via the LMS Moodle. In order to examine to what extent students' benefit in their learning success from the learning tasks with individualized and automated error-specific JIT feedback (IND-feedback), the newly developed tasks were compared with classic tasks that regularly provide only corrective JIT feedback (COR-feedback). Against this background, students were randomly assigned to one of two intervention groups. One intervention group learned with the tasks with individualized error-specific feedback (IND-feedback) while the other group worked on classic tasks with corrective feedback (COR-feedback). The processing of the tasks was voluntary and could be repeated at any time during the winter term. By solving the tasks correctly, students could gather bonus points at the end of the semester if they passed the exam. Since experience has shown that many first-year students initially need support in using Moodle and completing the tasks, they received a short introduction at the beginning of their studies.

Since the performance of the students in the tasks depends to a large extent on their prior knowledge, students' prior knowledge was also collected as a control variable via a standardized content knowledge test for the field of general chemistry (Averbeck, 2021; Freyer, 2013) at the beginning of the first semester. The content knowledge test consists of 35 multiple-choice single-select items and captures not only knowledge in the field of acid–base reactions or redox reactions but also knowledge in the field of atom models, stoichiometry, chemical equilibrium, and chemical bonds, for instance. As a result of the corona pandemic, the content knowledge test was administered online at the beginning of the winter term 2020/2021.

To answer the first research question, students' activities were recorded in an individual log file so that solved tasks, received feedback messages, time on task, and received credit points could be calculated, for instance. Depending on the number of solution attempts, students could receive up to a maximum of 3 credit points per task. One credit point was deducted for each feedback message received. Students who solved a task on the first solution attempt received 3 credit points, whereas students who solved a task correctly on the third attempt received 1 credit point. The deduction of credit points was not visible to the students and had no effect on their performance on the test items since the tasks were learning tasks. The point deduction served as an indicator for assessing the difficulty of the learning tasks and the quality of the feedback.

To answer the second research question, students' performance was operationalized using self-developed *test items*. For this purpose, the students had to complete a test item after each task set that was identical to the *learning tasks* providing feedback in terms of content and task setting. In contrast to the learning tasks with feedback (IND- or COR-feedback), test items are not used to acquire knowledge, but to determine performance. For this reason, they could only be completed once and were assessed by the program.

New Online Tasks for Reaction Equations

Based on the requirements presented in the previous sections, a new digital tool allowing the *input* and *assessment* of chemical reaction equations has been developed. The tool is based on the e-assessment system JACK® (Striewe, 2016), which is one of the standard e-assessment systems at the university level. It can be used as a standalone tool as well as in conjunction with an LMS like Moodle. The new online tasks can be integrated into LMS via JACK®, which implements the LTI (Learning Tool Interoperability) standard. This ensures a smooth exchange of data and learning content without having to create and manage additional accounts for the learners. A prerequisite for

using the tool is that the LMS used by the university also supports the LTI standard. This is the case with the LMS widely used at universities, such as Moodle or ILIAS. In addition to basic functions for the input and assessment of reaction equations, *individualized error-specific feedback* has been implemented in further development stages. Tasks were developed for two subtopics: acid–base reactions and redox reactions (cf. Figs. 2 and 3).

1st Stage of Development

A new input editor has been developed to enable the input and assessment of reaction equations in symbol notation. This allows the input of reactants and products as molecular formulas in two separate fill-in fields that are connected to each other via a predefined reaction arrow to form a reaction equation (cf. Fig. 2). A new editor was necessary since molecular formulas

Exercise 1

50 ml hydrocyanic acid solution ($c = 1\ ^{mol}/_{L}$) is dissolved in 300 ml water.

Questions

1. Formulate the reaction equation for the acid-base reaction. Neglect the states of matter of the substances (aq, s, l, g) and do not use dissociated forms.
2. Determine the Brønsted acid and Brønsted base for reactant and product side by entering molecular formulas.
3. Select an equation for calculating the pH value.
4. Calculate the pH value.

Answer:

1. ▯ ⇌ ▯

2.

reactants	products
Brønsted acid: ▯	conjugated Brønsted base: ▯
Brønsted base: ▯	conjugated Brønsted acid: ▯

3. Hydrocyanic acid is a ▯ acid. Hence, equation ▯ has to be used.
   - strong
   - medium strong
   - weak
   - 1
   - 2
   - 3
   - 4
   - 5
   - 6

4. $pH \approx$ ▯

1. $pH = -\lg(c_0)$
2. $pH = -\lg\left(-\frac{K_s}{2} + \sqrt{\frac{K_s^2}{4} + K_s \cdot c_0}\right)$
3. $pH = \frac{1}{2} \cdot (pK_s - \lg(c_0))$
4. $pH = 14 - (-(\lg(c_0))$
5. $pH = 14 - \left(-\lg\left(-\frac{K_B}{2} + \sqrt{\frac{K_B^2}{4} + K_B \cdot c_0}\right)\right)$
6. $pH = 14 - \left(\frac{1}{2} \cdot (pK_B - \lg(c_0))\right)$

hint submit

Fig. 2 Screenshot of an acid–base reaction task (lg, decadic logarithm; c_0, initial concentration)

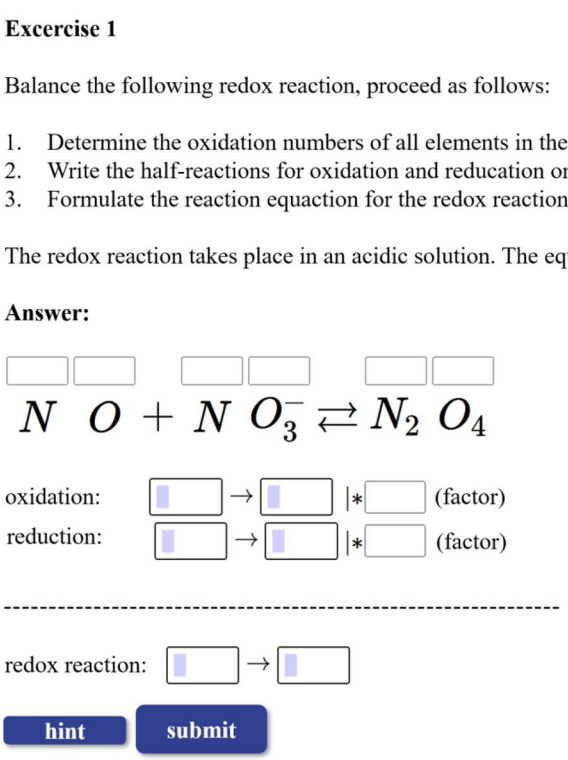

Fig. 3 Screenshot of a redox reaction task

can be entered with an already existing mathematical formula editor but assessment of solutions with this editor is not possible. More specifically, the existing editor stores input data in a format named Open Math,[7] which captures the semantics of mathematical formulas and equations. Since that format is not capable of capturing the semantics of chemical formulas, a similar data format named Open Chem was created and included in the formula editor (Pobel & Striewe, 2019). While mathematical formulas inherently provide appropriate functions to check equality or specific properties, similar functions are not automatically available for chemical formulas. Hence, some new functions were added to the assessment module of JACK®, so that item authors can use them to design feedback: (1) A new function named "contains" can be used to check whether a fill-in box contains the requested (element) symbol notation(s) (e.g., if correct reactants and products have been formed). The order, in which the reagents are given, can vary there, whereby several answers are recognized as correct. (2) With a second function named "compareNumberOfAtoms," it is possible to check whether the reaction equation is stoichiometrically balanced. This function ensures that several correct answers can be recognized. (3) With a third checker function "compareCharges," it is possible to compare the net sum of charges in one input field with another and if it is the same. This function ensures that in addition to the stoichiometrically balance, the charge balance is also considered. The only simplification is that the states of matter are currently excluded; an appropriate checker function is developed but has not been tested yet. Reaction conditions and catalysts also have to be determined in advance (e.g., by a drop-down menu). However, this should not change the necessary abilities for successful task processing in a fundamental way.

2nd Stage of Development

In order to identify typical solutions and provide automated feedback on standard errors using Intelligent Assessment (Müller et al., 2006), submitted exercises and exam solutions were analyzed in advance regarding typical errors. In this context, reaction equations had often not been stoichiometrically balanced and extension factors had been determined incorrectly. Against this background, feedback was developed that indicates the location of error and provides hints for solving the task. If there are errors in one of the partial equations, the entire equation is incorrect or extension factors were determined incorrectly (cf. Fig. 3).

3rd Stage of Development

Since the tasks were designed for self-assessment, they are linked to a three-stage algorithm that allows the same task

[7] https://www.openmath.org/

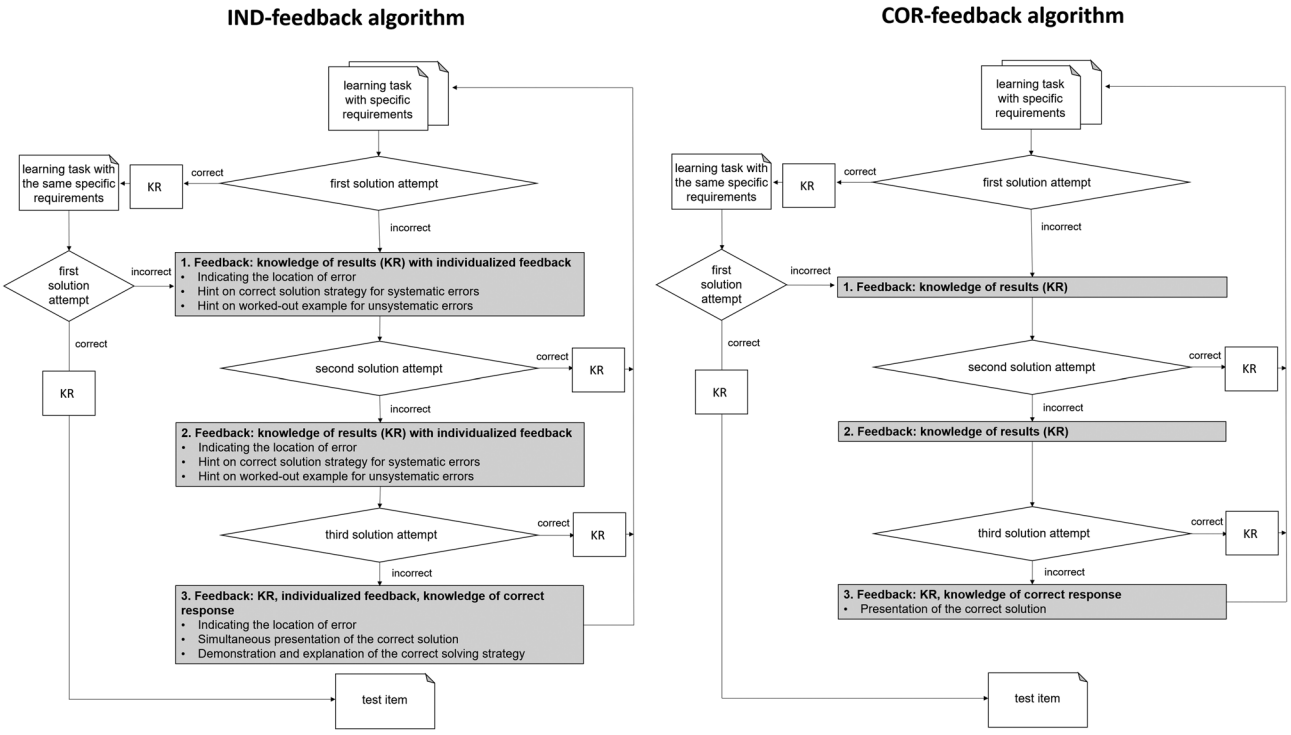

Fig. 4 IND- and COR-feedback algorithm for acid–base and redox reaction tasks

to be completed several times (cf. Fig. 4). If students do not find the correct answer in the first or second solution attempt, they are given the opportunity to correct their solution with the help of feedback. After the third wrong submission, the correct result is displayed. Considering these construction principles, two subtopics (acid–base (AB) reactions and redox (RD) reactions, cf. Figs. 2 and 3) were realized in two sets of six tasks each. To complete a task set, students had to correctly solve two tasks successively at the first attempt. Furthermore, a task set could be finished by working on all of the six tasks with various success.

The tasks with corrective feedback (COR-feedback) differ from the tasks with individualized error-specific feedback (IND-feedback) only in terms of the type of feedback (cf. Fig. 4). The corrective feedback is also given immediately and automatically by the program. In addition, students have three solution attempts per task for the correct answer. An example of IND-feedback for a typical systematic error in redox reaction tasks is shown in Fig. 5.

In order to practice entering reaction equations, students received some practical problems before the first exercise to familiarize themselves with the editor. This was necessary in order to understand the input editor, which can be used to write exponents and indices.

Data Collection

During the winter terms 2019/2020 and 2020/2021, the training was offered to 238 first-year B.Sc. Chemistry and B.Sc. Water Science students (60.7% male, average age 20 years). To 124 of them, individual error-specific JIT feedback (IND-feedback) was offered while to 114 corrective feedback (COR-feedback) was offered, although not all of them learned with the tasks. This group includes students who have received the learning tasks but have not entered anything. Overall, a maximum of 99 students worked on the tasks with IND-feedback (54 AB tasks, 45 RD tasks), and 89 students worked with the tasks with COR-feedback (53 AB task, 36 RD tasks). Data from students who repeated tasks at a later point in time (e.g., for exam preparation) were excluded from further analysis.

Results

The development of the new input editor allows digitizing tasks that require the input of chemical reaction equations or molecular formulas without using formats such as multiple-choice, which alter the learning objective (cf. "Closed-Ended

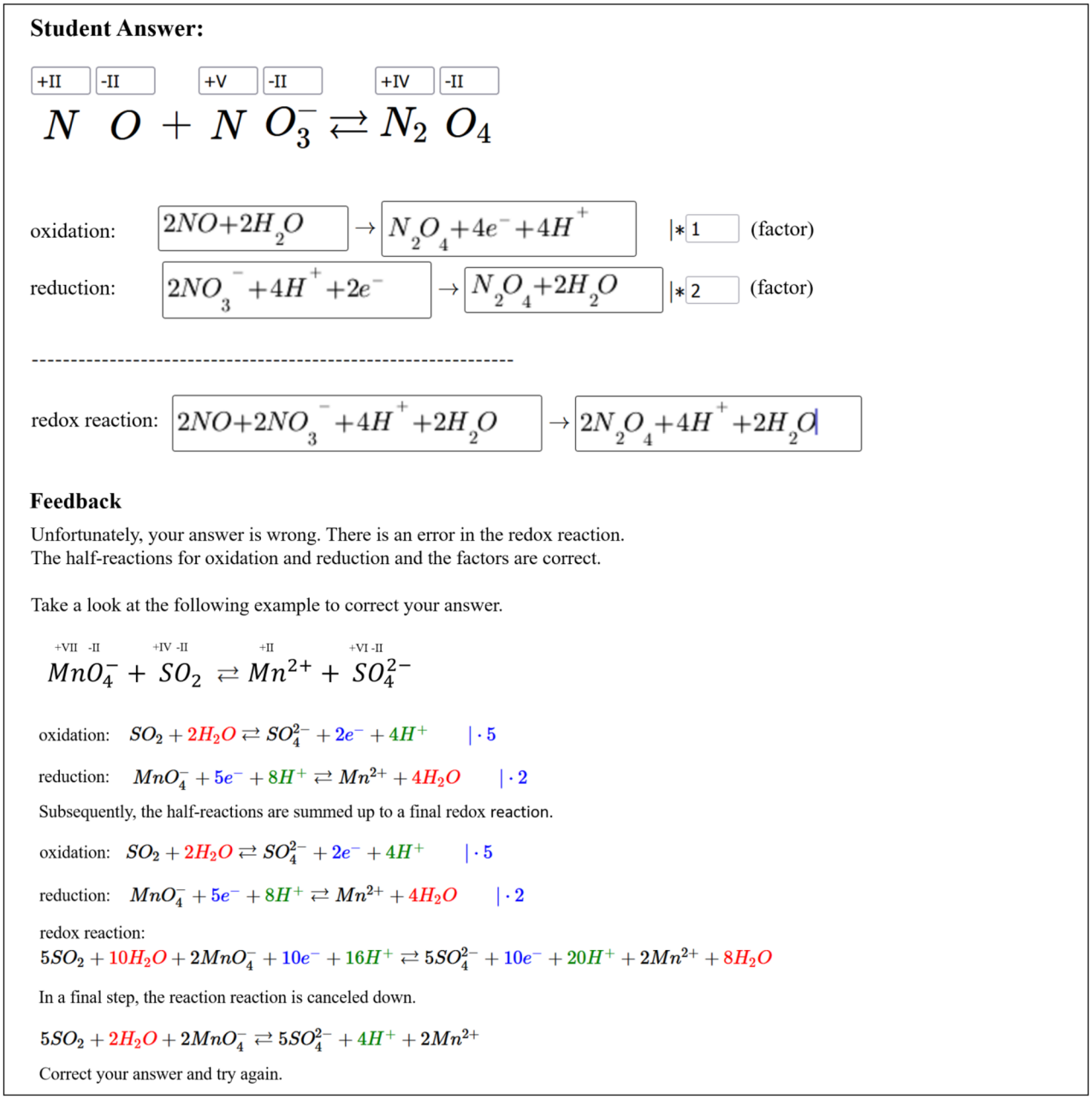

Fig. 5 Screenshot of an incorrect solution attempt and automatically generated IND-feedback for a systematic error

Tasks" section). A requirement for the digitalization of tasks was the automated assessment of students' solutions, which should be comparable to that of a human tutor. All of the three new functions of the editor ("contains," "compareNumberOfAtoms," "compareCharges") check students' solutions successfully to a large extent. This means that solutions cannot only be compared with previously deposited solutions, but alternative solutions can also be identified as such. In addition, various errors can be identified and tailored feedback can be provided. In case of the redox reaction tasks, 16 different errors could be distinguished and provided with feedback, while 6 errors were considered for the acid–base reaction tasks. Taking a closer look at how many feedback messages students from the IND-feedback group received, a log file analysis shows that, on average, 3 to 4 feedback messages were sent to the students during their training with both task sets, respectively (cf. Table 1). On average, 3–4 tasks are completed, with an average score of one point.

Table 1 Means and standard deviations for the solved tasks, received feedback messages, and credit points from students of the IND-feedback group

	Solved tasks		Feedback messages		Time on task (min)		Credit points	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Acid–base task set (AB) $N=54$	3.58	1.78	3.73	2.80	39.15	31.72	0.95	0.94
Redox reaction tasks (RD) $N=45$	3.77	1.87	3.65	3.10	46.36	31.64	0.96	0.95

Usefulness of the Automated Individualized Error-Specific Feedback (IND-Feedback) for Error Correction and Successful Task Completion

Based on the findings from the two cohorts, we can report that the AB tasks are, on average, of a reasonable difficulty. Log file analysis indicates that only 14.8% of the students ($N=54$) could solve the first AB task without any feedback at the first attempt. But in 33.3% of the cases, students manage to solve the first AB task with the help of the feedback (cf. Fig. 6). In the following AB tasks, the percentage of students who solve the tasks also increases, so that 90% of the students manage to solve the third task within two attempts. After three completed tasks, it strikes that the percentage of students increases who skipped the tasks (cf. Fig. 6). We will discuss the reasons for skipping the answer later on this paper.

For solving the RD tasks, the feedback seems to be helpful as well, even if the percentage of students who answer the tasks incorrectly or skipped them is higher compared to the AB tasks. Overall, 55.6% of the students ($N=45$) manage to solve the first RD task (cf. Fig. 6). In contrast, the percentage of students who answer the following tasks incorrectly or skipped them is greater than the percentage of students who answer the tasks correctly. Since there was no possibility to find out more about the reasons why these students skipped the tasks, we can only make assumptions at this point. One possible reason could be that the feedback is not targeted well enough to identify the mistakes and to correct the entered answer. Therefore, more precise and individualized error-specific feedback is needed for successful task completion, especially for more complex exercises like the RD tasks. Based on the individual incorrect answers of the students, we will develop more individualized error-specific feedback in further iteration steps. In addition, it could be helpful to offer students more than three solution attempts for more complex tasks, because students may need several attempts to correct all errors. Another reason could be that the students lost their motivation to complete the tasks because they had to enter all answers, even the correct ones, again with each attempt. This is particularly noticeable in the RD tasks, which consisted of 13 input fields per task. Therefore, tasks with many input fields, like the RD tasks, should save correct answers and only request a new input for incorrect answers.

Benefits of Automated Individualized Error-Specific JIT Feedback (IND-Feedback) over Corrective Feedback (COR-Feedback)

With regard to RQ2, an analysis of covariance (ANCOVA) was calculated, with prior knowledge serving as covariate. In

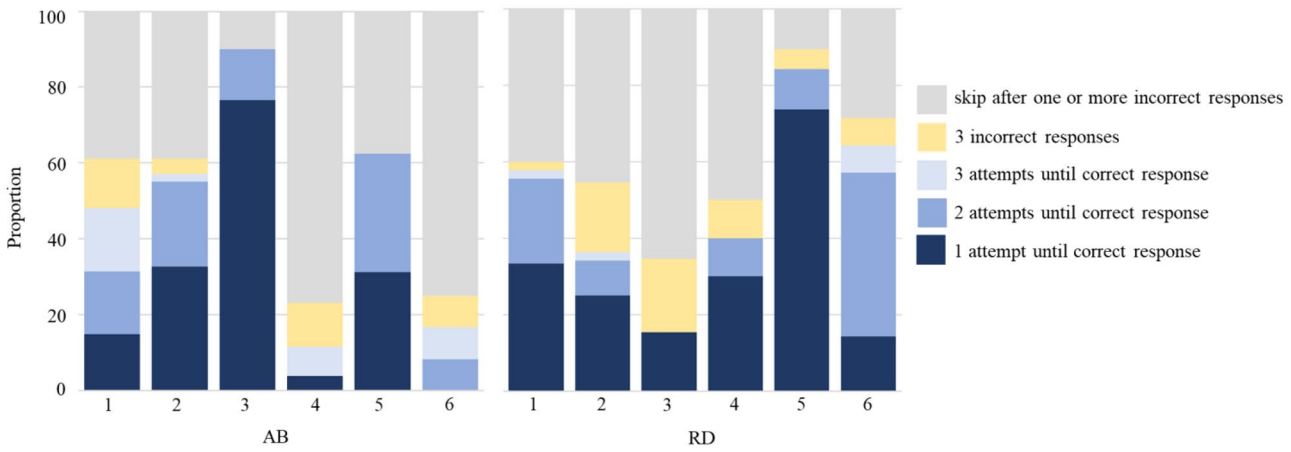

Fig. 6 Percentage of successful and unsuccessful solution attempts per task from students of the IND-feedback group (AB, acid–base reaction learning tasks; RD, redox reaction learning tasks)

Table 2 Results of a hierarchical linear regression analyses predicting students' performance in the test items of both task sets (AB, RD)

Model	M1			M2			M3		
Predictor	**B**	**s.e.**	**β**	**B**	**s.e.**	**β**	**B**	**s.e.**	**β**
Prior knowledge	.141	.081	.193	.165	.080	.226*	.167	.081	.229*
Feedback group	.334	.168	.221*	.361	.164	.239*	.384	.176	.255*
Mean score of credits				.187	.086	.239*	.215	.113	.275
Feedback group × mean score of credits							−.066	.174	−.380
R^2			.094			.150			.151

*$p \leq .05$

this context, the reliability of the content knowledge test can be valued as good in both winter terms with an EAP reliability of 0.765 and 0.847. After adjusting for prior knowledge, results show that students who received individual error-specific JIT feedback (IND-feedback) ($N = 40$, $M = 1.23$, $SD = 0.73$) perform significantly better in the *test items* than those who received only corrective feedback (COR-feedback) ($N = 38$, $M = 0.84$, $SD = 0.75$) ($F(1,75) = 4.78$, $p = 0.032$, $\eta^2 = 0.060$).

In addition, we examined whether students' performance on the *learning tasks* with feedback (AB and RD) had a further influence on their performance on the *test items* and whether this influence was moderated by the feedback. For this purpose, linear regression analyses were conducted, in which students' performance on the *test items* was modeled hierarchically with prior knowledge, feedback, and performance on the task with feedback as predictors. To better interpret the regression weights of the predictors, all variables were z-standardized before the regression analyses. Only the feedback group variable was left in its metric (COR-feedback: 0, IND-feedback: 1) because the weights can be easily interpreted.

In model 1 (M1), we first examined the extent to which the feedback group predicted performance on the test items while controlling for students' prior knowledge. The results indicate that feedback has a significant impact on performance on the test items, explaining 9.4% of variance (cf. Table 2). In a second model, the mean score of credits achieved while processing the learning tasks with feedback was included as another predictor. Results show that the mean score of credits also has a significant influence on performance on the test items, in addition to feedback and prior knowledge. Summing up, the predictors explain 15% of variance. In addition, the fit of the model M2 improves significantly compared to M1 ($\Delta R^2 = 0.056$, $p = 0.032$). In M3, we finally tested whether feedback had a moderating influence in predicting students' performance on the test items by including the interaction term of feedback group and mean score of credits as another predictor in the model. The results show that the regression weight of the interaction term does not become significant, nor does the inclusion of the predictor contribute to further explanation of variance. Likewise, this model has no significant better fit than M2 ($\Delta R^2 = 0.002$, $p = 0.705$). Thus, it can be assumed that prior knowledge, feedback, and the mean score of credits have an independent influence on performance on the test items. However, feedback as a moderator does not affect the relationship between mean score of credits and performance on the test items.

In contrast, further hierarchical regression analyses show a different interaction of the predictors if the performance on the test items was predicted separately for both task sets with feedback (acid–base reactions and redox reactions). Since passing the test item for both task sets is a dichotomous variable, two binary logistic regression analyses were calculated. Table 3 summarizes the results of the logistic regression analysis for the acid–base tasks while Table 4 summarizes the results for the redox reaction tasks.

In accordance with the hierarchical approach, the first model (M1) for the acid–base reaction test item considers only the feedback group and the students' prior knowledge, which was included in the model as a control variable. Results show that the feedback group has a significant

Table 3 Results of a hierarchical logistical regression analyses predicting students' performance in the test items of the task set AB

Model	M1			M2			M3		
Predictor	***b***	***p***	**exp[*b*]**	***b***	***p***	**exp[*b*]**	***b***	***p***	**exp[*b*]**
Prior knowledge	.179	n.s	1.196	.285	n.s	1.330	.343	n.s	1.409
Feedback group	1.523	0.008	4.586	1.606	.007	4.983	.2034	.002	7.648
Mean credits acid–base task set				.600	.050	1.822	1.298	.011	3.661
Feedback group × mean credits acid–base task set							−1.418	.030	.242
R^2_N			.167			.244			.332

n.s. not significant

Table 4 Results of a hierarchical logistical regression analyses predicting students' performance in the test item of the task set RD

Model	M1			M2			M3		
Predictor	***b***	***p***	**exp[*b*]**	***b***	***p***	**exp[*b*]**	***b***	***p***	**exp[*b*]**
Prior knowledge	.200	n.s	1.222	.194	n.s	1.214	.180	n.s	1.197
Feedback group	.650	n.s	1.916	.644	n.s	1.904	.605	n.s	.1832
Mean credits redox task set				−.048	n.s	.954	−.096	n.s	.908
Feedback group × mean credits redox task set							.121	n.s	1.128
R^2_N			**.048**			**.049**			**.050**

n.s. not significant

regression weight when controlling for prior knowledge (cf. Table 3). Thus, the chance of passing the test item increases by a factor of 4.58 if individual error-specific JIT feedback (IND-feedback) was received. This means that the chance of passing the test item increases approximately fourfold if students have received individual error-specific JIT feedback and prior knowledge remains constant.

In M2, the mean score of credits achieved in the exercise task providing feedback was included in the model as a further predictor, since it stands to reason that the probability of passing the test item increases if many tasks with feedback were solved correctly. The results indicate that the mean score of credits also has a significant regression weight. With respect to passing the test item, the coefficient $\exp[b] = 1.82$ states that the chance of passing the test item increases by a factor of 1.82 if the mean score improves by one unit (e.g., from 3 to 4 points). Likewise, the fit of the model improves significantly compared to M1 ($\chi^2(1) = 4.485$, $p = 0.034$).

In a third model, we tested again whether feedback had a moderating influence on the relationship between the mean score of credits and the performance on the test item. For this purpose, the interaction term of feedback group and mean score of credits was additionally included in M3. The results reveal that there is a moderating influence of the feedback group, as the interaction term of feedback group and mean score of credits becomes significant. This means that a high mean score in the tasks with feedback that could be achieved through individual error-specific JIT feedback has a positive effect on passing the test item. In summary, this model has a significantly better fit than M2 ($\chi^2(1) = 5.115$, $p = 0.024$) and explains an overall variance of 33.2%.

A comparable interaction of the predictors cannot be found for the redox reaction tasks. The results of the hierarchical logistic regression analyses indicate that none of the predictors have a significant influence on students' performance in the test item (cf. Table 4).

Discussion and Conclusion

Learning management systems are widely used in modern university courses from various fields. Since these digital learning systems can provide automated feedback, they are of particular importance for higher education teaching. Considering that feedback is one of the most important influencing factors on learning processes (Hattie & Timperly, 2007; Kluger & DeNisi, 1996; Wisniewski et al., 2020), it can make a valuable contribution to improving higher education. However, for the field of molecular formulas and chemical reaction equations, there is a lack of electronic tasks that automatically check students' input, generate appropriate individualized error-specific feedback, and can be implemented in LMS. Previous attempts to digitize tasks with the help of classical task formats (multiple-choice, etc.) were accompanied by loss of quality (cf. "Closed-Ended Tasks" section). Thus, the aim of the study was to develop and evaluate such a digital tool for entering and assessing chemical reaction equations or molecular formulas and providing formative individualized error-specific JIT feedback. In addition, a further aim was to analyze how helpful the individualized error-specific JIT feedback was in solving the tasks compared to corrective JIT feedback, which is regularly used in LMS.

The results of the study reveal that the development of a new editor allows the digitalization of paper-based tasks with the learning objective of setting up chemical reaction equations. The new exercise types automatically check students' solutions and offer detailed error-specific feedback, so manual correction by a tutor is no longer necessary. This could be of particular interest to universities that have not been able to provide individual error-specific JIT feedback through existing LMS. The results are relevant for many instructors in chemistry, as student understanding of acid/base and redox reactions are very important areas where a lack of understanding will affect student success in first year chemistry. In addition, the multistep exercises may also be of interest to other academic subjects where algorithmic problem solving is emphasized.

Based on the results so far, we assessed the usefulness of the individual error-specific feedback (IND-feedback) as satisfying, since a majority of students were able to correct their answers with the help of the feedback and solve the tasks within two or three attempts. Furthermore, the individual error-specific JIT feedback has a significant impact on students' performance in the test items. Given the same

prior knowledge, students who received individual error-specific JIT feedback outperform students who worked with traditional learning tasks providing only corrective feedback.

However, results show space for improvements especially in the quality of the IND-feedback of the redox reaction tasks. More individualized error-specific feedback as well as saving correct answers in the input fields may raise students' motivation and solution frequency of the redox reaction tasks. Further research is needed to increase the use and fit of the feedback, for example, through qualitative interviews. In the view of the low proportion of students who answer the tasks with IND-feedback correctly after the third solution attempt, it might be promising to expand the three-stage algorithm. For very complex tasks, such as the redox tasks with 13 input boxes, the students may need several attempts to find all errors. They should therefore be given several attempts to solve the task with the help of individual error-specific feedback.

In addition, one limitation compared to paper–pencil tasks is that states of matter and solution states can currently not be checked. An appropriate checker function is developed but has not been tested, yet. Moreover, the learning path of the digitalized chemistry tasks is more predefined compared to paper–pencil tasks. In order to be able to give individualized error-specific feedback, it was necessary to break down complex tasks into subtasks.

In summary, the present study indicates that the new digital tool for entering and assessing chemical reaction equations or molecular formulas is a promising approach to supporting students in large university courses by providing formative feedback via LMS. The new exercise types are a good opportunity to offer immediate individual error-specific feedback during private study time, which is usually not possible otherwise.

Acknowledgements We thank the German Research Foundation (DFG) for funding our research.

Funding Open Access funding enabled and organized by Projekt DEAL. Deutsche Forschungsgemeinschaft, 397641476, Carolin Eitemüller.

Data Availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethical Approval The research adhered to ethical standards and guidelines as the nature of study demanded.

Consent to Participate Consent was collected from the participants and the statistical analysis was performed using non-identifiable data.

Conflict of Interest The authors declare no competing interests.

References

American Association for the Advancement of Science (AAAS). (2001). *Atlas of science literacy* (Vol. 1). American Association for the Advancement of Science, National Science Teacher Association.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval process in free recall. *Psychological Review, 79*, 97–132.

Ashton, H., Beevers, C. E., Korabinski, A. A., & Youngson, M. A. (2005). Investigating the medium effect in computer-aided assessment of school chemistry and college computing national examinations. *British Journal of Educational Technology, 36*(5), 771–787.

Averbeck, D. (2021). Zum Studienerfolg in der Studieneingangsphase des Chemiestudiums: Der Einfluss kognitiver und affektiv-motivationaler Variablen [On academic success in the introductory phase of chemistry studies. *The influence of cognitive and affective-motivational variables*. Logos: Berlin.

Bonner, S. M. (2013). Mathematics strategy use in solving test items in varied formats. *The Journal of Experimental Education, 81*, 409–428.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*, 27–40.

Busker, M., Parchmann, I., & Wickleder, M. (2010). Eingangsvoraussetzungen von Studienanfängern im Fach Chemie: Welches Vorwissen und welche Interessen zeigen Studierende? [Entrance requirements for first-year students in chemistry: What prior knowledge and interests do students have?]. *Chemkon, 17*(4), 163–168. https://doi.org/10.1002/ckon.201010134

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 100027.

Chamala, R. R., Ciochina, R., Grossmann, R. B., Finkel, R. A., Kannan, S., & Ramachandran, P. (2006). EPOCH: An organic chemistry homework program that offers responsive-specific feedback to students. *Journal of Chemical Education, 83*(1), 164–169.

Chen, X. (2013). STEM attrition. College students' path into and out of STEM fields. *Statistical analysis report. Nces,* 2014–001. Washington, D.C.

Cole, R., & Todd, J. (2003). Effects of web-based multimedia homework with immediate rich feedback on students learning in general chemistry. *Journal of Chemical Education, 80*(11), 1338–1343.

DeBoer, G. E., Hermann-Abell, C. F., Wertheim, J., & Roseman, J. E. (2009). Assessment linked to middle school science learning goals: A report on field test results for four middle school science topics [Conference Paper]. *Annual Conference of the National Association of Research in Science Teaching*, Garden Grove, CA.

de Bruin, A. B. H., Kok, E. M., Lobbestael, J., & de Grip, A. (2017). The impact of an online tool for monitoring and regulating

learning at university: Overconfidence, learning strategy, and personality. *Metacognition Learning, 12*(1), 21–43.
Ferber, N. (2014). Entwicklung und Validierung eines Testinstruments zur Erfassung von Kompetenzentwicklung im Fach Chemie in der Sekundarstufe I. *Development and Validation of a Test Instrument to Measure the Development of Competencies in Chemistry at Lower Secondary Level,* Logos.
Freasier, B., Collins, G., & Newitt, P. (2003). A web-based interactive homework quiz and tutorial package to motivate undergraduate chemistry students and improve learning. *Journal of Chemical Education, 80*(11), 1344–1347.
Freyer, K. (2013). *Zum Einfluss von Studieneingangsvoraussetzungen auf den Studienerfolg Erstsemesterstudierender im Fach Chemie [On the influence of study entry requirements on the study success of first-semester students in chemistry]*. Logos: Berlin.
Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487
Hattie, J. (2013). Calibration and confidence: Where to next? *Learning and Instruction, 24*, 62–66.
Hedtrich, S., & Graulich, N. (2018). Using software tools to provide students in large classes with individualized formative feedback. *Journal of Chemical Education, 95*, 2263–2267.
Herding, D., Zimmermann, M., Bescherer, C., & Schroeder, U. (2010). Entwicklung eines Frameworks für semi-automatisches Feedback zur Unterstützung bei Lernprozessen. Development of a framework for semi-automated feedback to support learning processes [Conference paper]. *Proceedings of the DeLFI 2010*. Tagung der Fachgruppe E-Learning der Gesellschaft für Informatik eV., Bonn, Germany.
Heublein, U., Richter, J., & Schmelzer, R. (2020). Die Entwicklung der Studienabbruchquoten in Deutschland. *The Development of Dropout Rates in Germany*. (DZHW Brief 3|2020). Hannover: DZHW.
Heublein U., Ebert J., Hutzsch C., Isleib S., König R., Richter J., & Woisch A. (2017). Zwischen Studienerwartungen und Studienwirklichkeit: Ursachen des Studienabbruchs, beruflicher Verbleib der Studienabbrecherinnen und Studienabbrecher und Entwicklung der Studienabbruchquote an deutschen Hochschulen. *Between study expectations and study reality: Causes of termination of studies, occupational reasons for termination of studies and development of the rate of termination of studies at German universities.* DZHW: Hannover.
Hohensinn, C., & Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*, 732–746.
Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, 5, e208. https://doi.org/10.7717/peerj-cs.208
Kanuka, H. (2001). University student perceptions of the use of the web in distance-delivered programs. *The Canadian Journal of Higher Education, 31*, 49–71.
Kastner, M., & Stangl, B. (2011). Multiple choice and constructed response tests: Do test format and scoring matter? *Procedia – Social and Behavioral Sciences*, *12*, 263–273.
Keuning, H., Jeuring, J., & Heeren, B. A. (2018). Systematic literature review of automated feedback generation for programming exercises. *ACM Transactions on Computing Education (TOCE), 19*(1), 1–43.
Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284.
Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.
Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389–405.
Lindner, M. A., Strobel, B., & Köller, O. (2015). Multiple choice-Prüfungen an Hochschulen? Ein Literaturüberblick und Plädoyer für mehr praxisorientierte Forschung [Multiple choice exams at universities? A literature review and plea for more practice-oriented research]. *Zeitschrift Für Pädagogische Psychologie, 29*(3–4), 133–149.
Liu, O. L., Lee, H. S., & Linn, M. C. (2011). An investigation of explanation multiple choice items in science assessment. *Educational Assessment, 16*, 164–184.
Lodder, J. & Heeren, B. A (2011, June). *A teaching tool for proving equivalences between logical formulae* [Conference paper]. International Congress on Tools for Teaching Logic, Springer, Berlin, Heidelberg, Germany. https://doi.org/10.1007/978-3-642-21350-2_18
Ma, W., Adesope, O., Nesbit, J., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901. https://doi.org/10.1037/a0037123
Malik, K., Martinez, N., Romero, J., Schubel, S., & Janowicz, P. A. (2014). Mixed-methods study of online and written organic chemistry homework. *Journal of Chemical Education, 91*(11), 1804–1809.
Müller, W., Bescherer, C., Kortenkamp, U., & Spannagel, C. (2006). Intelligent computer-aided assessment in math classroom: State-of-the-art and perspectives [Conference paper]. *Proceedings of the Conference on Imaging the future for ICT and Education*, Alesund University College, Norway.
Narciss, S., & Huth, K. (2006). Fostering achievement and motivation with bug-related tutoring feedback in a computer-based training for written subtraction. *Learning and Instruction, 16*, 310–322.
Pazicni, S., & Bauer, C. F. (2014). Characterizing illusions of competence in introductory chemistry students. *Chemistry Education Research and Practice, 15*(1), 24–34.
Penn, J. H., & Al-Shammari, A. G. (2008). Teaching reaction mechanisms using the curved arrow neglect (CAN) method. *Journal of Chemical Education, 85*(9), 1291–1295.
Perry, S., Bulatov, I., & Roberts, E. (2007). The use of E-assessment in chemical engineering education. *Chemical Engineering Transaction, 12*, 555–560.
Pobel, S., & Striewe, M. (2019). Domain-specific extensions for an E-assessment system. In M. Herzog, Z. Kubincová, P. Han, M. Temperini (Eds.), *Advances in web-based learning – ICWL 2019. Lecture Notes in Computer Science*, vol 11841. Springer, Cham. https://doi.org/10.1007/978-3-030-35758-0_32
Schmid, U., Goertz, L., Radomski, S., Thom, S., & Behrens, J. (2017). Monitor Digitale Bildung: Die Hochschulen im digitalen Zeitalter. *Gütersloh: Bertelsmann Stiftung*.
Shepherd, T. D. (2009). Mastering chemistry. *Journal of Chemical Education, 86*(6), 694.
Shermis, M. D., & Burstein, J. C. (2002). *Automated essay scoring: A cross-disciplinary perspective*. Routledge: New York, NY.
Shermis, M. D., & Burstein, J. C. (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge: New York, NY.
Striewe, M. (2016). An architecture for modular grading and feedback generation for complex exercises. *Science of Computer Programming, 129*, 35–47. https://doi.org/10.1016/j.scico.2016.02.009
Trauten, F., Eitemüller, C., & Walpuski, M. (2019). Entwicklung und Evaluation von feedbackgestützten Online-Chemieaufgaben [Development and evaluation of feedback-supported online tasks in chemistry]. In C. Maurer (Ed.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Kiel 2018* (pp. 846–849). Kiel: IPN.
Walpuski, M., Ropohl, M., & Sumfleth, E. (2011). Students' knowledge about chemical reactions - development and analysis of standard-based test items. *Chemistry Education Research and Practice, 12*, 174–183.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*(3087). https://doi.org/10.3389/fpsyg.2019.03087

Yorke, M. (2003). Formative assessment in higher education. *Higher Education, 45*(4), 477–501.