Journal of Infrastructure, Policy and Development 2024, 8(11), 8839.
https://doi.org/10.24294/jipd.v8i11.8839

Article

A computer-based evaluation system: Its design, implementation and results obtained on a general chemistry course

Carmen Elena Stoenoiu[1], **Lorentz Jäntschi**[2,*]

[1] Faculty of Electrical Engineering, Technical University of Cluj-Napoca, 400114 Cluj-Napoca, Romania
[2] Department of Physics and Chemistry, Technical University of Cluj-Napoca, 400114 Cluj-Napoca, Romania
*** Corresponding author:** Lorentz Jäntschi, lorentz.Jantschi@gmail.com

CITATION

Stoenoiu CE, Jäntschi L. (2024). A computer-based evaluation system: Its design, implementation and results obtained on a general chemistry course. Journal of Infrastructure, Policy and Development. 8(11): 8839. https://doi.org/10.24294/jipd.v8i11.8839

ARTICLE INFO

Received: 27 August 2024
Accepted: 18 September 2024
Available online: 22 October 2024

Abstract: Using multiple evaluation methods and systems give a comprehensive assessment. A computer-based multiple-choice assessment system was designed, implemented, posted online, and used to assess students as part of their final evaluation marks for a discipline. The online system of evaluation was intended to be used multiple times for evaluating the assimilation degree of a specific course at the end of the course. The data recorded for the period 2017–2023 with about 1400 distinct users were used to analyze the performance of the evaluation system. The system worked fine and a slight modification of it served well on remote evaluation during COVID-19 period. However, the upturn of mobile phone applications requires the creation of a system adapted to the new virtual reality.

Keywords: online evaluation system; multiple choice questions; general chemistry; first-year undergraduate students

1. Introduction

The fourth industrial revolution (4IR) was determined by the emergence of new technologies (digital machines, artificial intelligence, robotics, big data, etc.) (Bloem et al., 2014) has created new challenges in the future demands of the labor market. Thus, new technologies have led to the emergence of needs for the creation of new skills, as a result of changes in tasks at work (Fallows and Steven, 2000), of new cognitive, non-cognitive and technical skills (Suleman, 2018), which have determined new challenges for engineering education.

Technological progress and the widespread use of ICT has affected all aspects of our lives, and education is no exception. Educational institutions have been forced to change their teaching methods (Baragash and Al-Samarraie, 2018; Ching-Ter and Hajiyev, 2017; Ramírez-Correa et al., 2017), and traditional classrooms are no longer constrained to conventional teaching and learning assessment methods (Alagarsamy and Vijay, 2019; Alsayyari et al., 2019; Dobre, 2015). Thus, classrooms have become smart learning environments (Tinmaz and Lee, 2020), which allow teachers to: (a) share teaching resources, expertise and advice, (b) teach flexibly, and (c) access remotely data (Munyengabe et al., 2017).

The importance of training and evaluating students through online environments was emphasized during the COVID-19 pandemic (Balkaya and Akkucuk, 2021; Camilleri and Camilleri, 2021) and continued later, developing in the form of learning management platforms (Barrot et al., 2021; Garcia et al., 2020; Joaquin et al., 2020).

In the specialized literature, there are studies that measure the evaluation of student and teacher satisfaction, which focus on the level of acceptance (Balkaya and

Akkucuk, 2021; Garcia et al., 2020; Raza et al., 2020), on the determining factors (Alzahrani and Seth, 2021; Cavus, 2021), but also on the effect generated among the academic community (Mehrolia et al., 2021).

If in the past assessment tools for teacher-student interaction (TSI) were characterized by: (1) a relatively single dimension and (2) the purpose paid less attention to the actual development of higher order thinking, today assessment tools are extremely complex and diverse (Mehrolia et al., 2021; Nguyen, 2021). This reflects the characteristics of higher-order interaction goals, multidimensional interaction content, diversified interaction methods, and rich and intelligent interaction environments (Nguyen et al., 2021).

Multiple-choice scoring systems are widely used for tests (Naşcu and Jäntschi, 2004a, 2004b; Nicol, 2007) and allow the use of statistics to obtain precise confidence intervals (Jäntschi, 2022). However, these multiple-choice systems are considered to be flawed by the possibility of guessing (Holmes, 2002). On the other hand, multiple-choice assessment systems are more difficult to construct (Jäntschi et al., 2007), maintain a high degree of inclusiveness (Loftis, 2019) and are more reliable, leading to less guesswork on the answer (Burton, 2001). Starting from the two considerations, there are authors who support hybrid assessment systems, most often used for broad subjects (Labrak et al., 2022). It should be noted that multiple choice assessment systems are not always the best choice (Chang and Akahorim, 1999) and online systems have some disadvantages compared to traditional paper tests (Bayazit and Askar, 2012).

In specialized literature, it is found that there are differences between the educational fields that cause the evaluation systems to encounter certain difficulties. Thus, in painting we need visual representations (Anglin et al., 2004), in music audio representations (Dannenberg and Hu, 2003), in mathematics equations and formulas (Matteson, 2006), in physics laws and principles (Hestenes, 1987) and in chemistry of equations (of chemical reactions), formulas (molecular, structural and geometric chemistry), pictorial representations (chemical processes and technologies, operating principles for methods), information structured in different ways (Jäntschi, 2013).

This study aims to create an evaluation system focused on making connections between information and measuring the degree to which it has been put into practice, rather than the degree of assimilation of information. Thus, for the multiple-choice assessment system, questions were designed to allow associations between information.

In addition, this research aims to discern the impacts and contributions of an online assessment system that requires the existence of an accessible database for storing information associated with the assessed content, a support system for database management and user interface, a computer system and classification based on recorded responses. The results of previous researches have shown that similar assessment systems are a valid and reliable solution in assessing students' knowledge (Jäntschi and Bolboacă, 2006; Omari, 2013 and Jiang et al., 2022), and the online version, because of its substructure dependent on web technologies, it can be accessed and applied from any place and offers advantages for improving the teaching-learning process (Hashemi Hosseinabad et al., 2021; Nguyen et al., 2021; Tasdemır et al., 2015).

This work aimed to research the online evaluation system of Multiple-choice evaluation systems (MCMA), starting from the design, implementation and evaluation of the obtained results. As the COVID-19 pandemic has imposed lockdown restrictions, they have led to significant changes in the way knowledge is transmitted. To observe the differences between the previous and the following period, two time periods were selected for comparison: from 2017 to 2019 and from 2022 to 2023. Considering the effects generated by the COVID-19 pandemic, the system was adapted for exclusive use online, being used as such for two years (2020 and 2021), and its modifications are provided in this study.

To support the argument that there is a visible difference in the presently developed software, we can say that it has a flexible and dynamic use, being able to face security problems but also some unexpected technical problems (regarding the power supply). Thus, when there is a power outage, the questions answered by the students are saved in the database and when the system is working again, they can resume the exam.

2. Materials and methods

The study was carried out by researching the results of the evaluation of engineering students, from the first year of the bachelor's degree, for the discipline of general chemistry. This discipline is not one of the basic areas of their specialization, and the summary is provided in Appendix A.2. This course, in general chemistry (Jäntschi, 2013), is intended for students who study in Romanian, English and German, and the evaluation system designed is a flexible one to be accessible to understanding. The content of the tests intended for evaluation allows the entire thematic area of chemistry to be covered, without using or accessing complex notions of chemistry, specific to specialized training.

The evaluation of the knowledge gained through the laboratory work was designed through two separate evaluations (Jäntschi, 2023).

2.1. Database structure

In order to obtain the online testing, a model was designed to include the common initial data, and then a database was created for each subject that takes the information in 3 files (a database, three tables for general chemistry).

The storage database topology for online assessment includes tables for keeping records on: User (identified by Id, Name, Pass, Date) where password Pass is stored encrypted with MD5 (32 characters)), Test (identified by Id, Qroenge, Rro, Ren, Rge, Aroenge) where Qroenge contains three texts separated by return; Rro, Ren, RGe and Aroenge contain a varying (but identical) number of lines; on each line, there is a possible answer (Rro, Ren, Rge) and the truth state of the answer 0/1 (in Aroenge) and Eval (with Id, subj and suid to manage secure connection, qlist, rlist, tlist and alist as ordered lists of space-separated values (qlist selected questions, rlist selected answers, tlist truth state for selected answers; alist truth for answered answers), tb, te and t storing times and p points earned).

The database includes a number of 54 possible questions and the user interface allows adding new subjects and then retrieving those registered separately for each of them. The information found in the database, according to the content intended for student evaluation, are:

- a number of 607 possible answers;
- a number of 296 true and 311 false answers;
- a possible number of answers of minimum 8, and maximum 26;
- a number of true answers of at least 3, and at most 13;
- a minimum number of false answers of 4, and a maximum of 13.

The evaluation, by completing a test, was designed to be carried out in a laboratory room equipped with computers that are connected in a network using the same class of IPs, this being carried out under the supervision of a teacher. The actual testing can be done by entering a password for the teacher and a password for the student, which are verified with the set of passwords stored on the server.

Each student can start the evaluation when he considers himself ready, by requesting a day and a time interval from the teaching staff, which is then established by mutual agreement. The database contains a number of 54 questions that are combined in a test pool, and each test will contain a number of 30 questions with 4 possible answers (at least one right and one wrong). The result is determined to be correct when only correct answers have been marked. For each correct solution, a number of 3 points is awarded. Each test has a time limit of 15 min, so when the test starts, it starts the timer, adding the start and end time inside the test (via tb and te inside the database). After completing the test, the related score obtained is established, as follows:

- the average correct response time to the current test (tmrc) is calculated;
- the total average time required for correct answers to all tests is calculated (tmrcnec);
- the coefficient c1 is calculated as a ratio between the average time per correct answer from the current test (tmrc) and the average time required for a correct answer from all tests in the database (tmrcnec);
- the coefficient c2 is calculated as a ratio between the number of correct answers from the current test (rcno) and the average number of correct answers from all the tests in the database (rcnom);
- the average value of the two coefficients (c1 and c2) is calculated, and the test score will be obtained as 10 times the calculated average;
- the test average is calculated for all tests given by a user, so we will have a list of test scores for each user; When there are at least two notes on the list, the smallest one will be eliminated, and the average will be obtained for the remaining ones.
- at the end, the grade is calculated, which will take values between 4–10 (where 4 is associated with the lowest grade, and 10 as the maximum grade).

According to the study by Coman et al. (2020) it was observed that higher education institutions were not prepared for exclusively online learning. The evaluation system presented in this study underwent a series of changes to adapt it to an exclusively online evaluation, among which we mention:

- the number of possible answers was reduced (3 instead of 4), and the recorded answer was simplified (out of 3 from {A, B, C} we have 1 or 2 correct, the possible answers being from the list: {A, B, C, AB, AC, BC});
- the assessment strategy has changed from individual counter-time, being adapted

to counter-time in series of students, using lists with more questions (up to 500). The first fastest response or the first two (if the second is different from the first) will be recorded;

- according to new strategies, students will receive positive points for each correct answer and negative points for each wrong answer. thus, the first two students (out of 6) who have 6 answers, of which more are correct than wrong, receive a grade of 1 + 1.5* (number of positive answers or number of negative answers) with a minimum of 1 or a maximum of 10. Then, for the next 2 for students the maximum decreases to 9, and for the last 2 students the maximum decreases to 8 and it is only conditioned by the positive answers being more than the negative ones;
- because during the online evaluation we have different locations for the evaluator and the evaluated student, in order to prevent fraud in the generated tests, the spaces between the words were replaced with double spaces (to prevent the textual question/answer search). To better understand the assessment method in Appendix A.3. 3) questions generated by MCMA are presented.

In order to comply with the GDPR (Regulation (EU), 2016), which imposed the pseudonymization of the information displayed on the statistics page, the evaluation system allowed a number to be associated with each student's name. They have thus been replaced by "Student < number >", where the number is generated uniquely for each student in the database when querying the system.

2.2. Programs and their topology

Testing involves accessing a test that is obtained by entering a web page, where the user is greeted with a welcome message (index, see **Table 1**), which is at the same time the root entry for the rest of the programs. The Universal Resource Locator (URL) is: http://l.academicdirect.org/Education/Evaluation/Chemistry/Chimie_Generala/

Access also requires the user to choose the desired language as follows:

- "?lang=ro" (default language) offers the welcome message and the menu in Romanian;
- "?lang=en" provides the welcome message and menu in English,
- "?lang=de" provides the welcome message and menu in German. The rating system allows adding a new language at any time.

The password and security modules are called inside programs whenever necessary to obtain credentials for a database connection (password) or to verify permission for an insert or update operation (security).

Table 1. Online evaluation software topology.

Program	**Actions**
Index	Welcome and menu
Insert	Add users (students)
Test	Generate a test and save an evaluation
Statistics	Calculate and display evaluation results
Password	Module containing credentials for database connection
Security	Module checking allowance of the testing (IP address based)

The software was implemented on a MySQL server (version 5.5.4) running on an intranet computer (otherwise). Storage database managed through a mysqli connection. The programs were implemented using PHP (version 7.4.10 is compiled and runs on Apache 2.4.46 HTTP server and FreeBSD 12.2 operating system).

Using the system for evaluation requires accessing the test program by going through three consecutive steps.

The first step to start testing is the verification of the following credentials:

- the IP address is from the designated address space (an intranet), and the last group of digits has a numeric value from a range;
- the teacher password exactly matches the correct value;
- the student encrypted password and the student's name match the value of the encrypted password in the User table; the student's name is selected from a drop-down combo box.

For the second step, a link is provided to register a new user (insert program). If the test credentials are passed and it is time for evaluation (a globally defined variable in the system with two states, TRUE and FALSE), a (new) test is generated. Note that you cannot take the test at night or during the semester, only during the day and during the exam session. A test contains (the second call to the test program):

- a number of m questions chosen at random (without replacement) from the list of n available questions (m is a predefined constant, set to 30, and n is queried from the database, it was 54);
- for each question (q1, ..., qm) a number of p (p is a predefined constant and were set to 4) possible answers (ri,1, ..., ri,p), each having associated a box check;
- Unix time for when the test was generated and sent to the client;
- username;
- a unique id (also recorded in the evaluation table, making it impossible to generate another test for the same user as long as the current test is not completed);
- a button to complete the test.

The third step for the test program involves the following checks:

- the user and the unique id to match an empty (not already completed) evaluation (a record in the Eval table);
- updates that record (from the Eval table) to contain numeric values for the fields: alist (list of $m \times p$ answers), te (Unix time for when the test was completed and sent back to the server), p (3 × the number of matches between the list of expected answers and the answered answers) and t (the time difference between the end and the beginning of the test);
- it displays a summary statistic for the evaluation to the user.

For practical reasons, the final evaluation (one evaluation from January of the current year) will be extracted from the database (which contains all the evaluations) for the students with the exam in the first semester. The reports with the grade from the exam will be completed with the information from this period (by default, for the statistics programs). Thus, for this, a number of 5 tables were generated:

- List all ratings, contains all ratings for all students that match the filter criteria of the time interval grouped by students and sorted ascending by date; when more than one assessment is recorded for a student, the one with the lowest results is null and is no longer considered a record of the student's average performance;

- Table with descriptive statistics of the database (numerical values from 2339 records in the database) average time to obtain a point (the value of this statistic is about 7.5 s); the average number of points (the value of this statistic is approximately 31.6 points);
- Table of test scores and table of means for all assessments included in the report (at the means above); the associated test scores are calculated; mean is given in bold;
- Database descriptive statistics table (numerical values from 2339 database records) containing the failed evaluation score (the one associated with a grade of 4 out of 10; the value of this variable was set to 3.5) and the best evaluation score (the value of this statistic is approximately 23);
- Scoreboard, containing the marks and points list (with hyperlinks to full assessment details) for each (pseudonymous) student.

3. Results and discussion

3.1. Statistics from the use of the system

Using the accessible data from the database, in **Figure 1** you can see the results regarding the statistics for the period 2017–2020, and in **Figure 2** we have the statistics related to the period 2021–2023. Analyzing the two figures we can see the gap between March 2020 and January 2022, where no new user was added and no assessment was made, this is the period when the COVID-19 restrictions prevented face-to-face meetings, and the system was used for doing remote online assessments following the procedure described in Section 2.3 (for 22 months).

In **Figure 3** you can see the average number of evaluations, (with variations between 1.0 and 2.0) and it shows us that *n* average students were satisfied with the first or the first two evaluations. Although there is no trend in the series described in **Figure 3**, we have statistical significance in the regression equation only for an intercept (y(time) = $1.43_{\pm 0.26} + 0.0003_{\pm 0.006}$·time, $r^2 = 0.0006$). For this series we see the probability associated with the intercept not belonging to the model as P (value = 1.43; t-value = 11.5, n = 19) = 5×10^{-8}% and the probability associated with the slope not belonging to the model as P (value = 0.0003; t-value = 0.11, n = 19) = 91%.

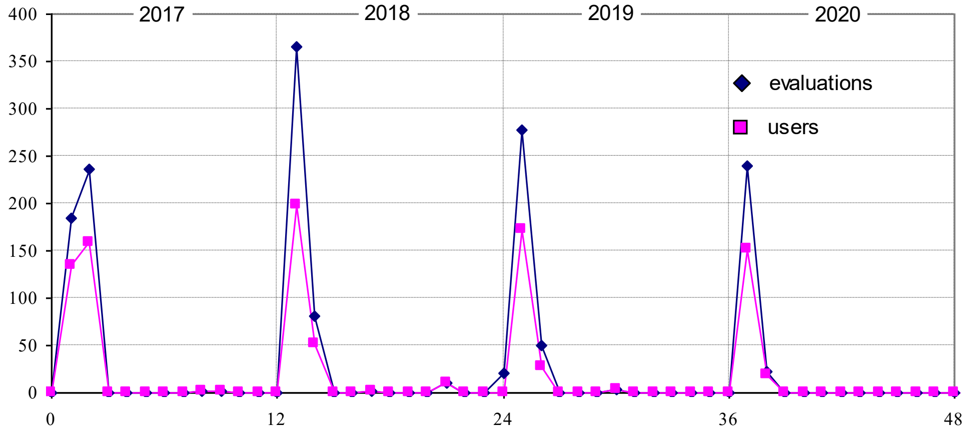

Figure 1. Monthly distinct evaluations and users for 2017–2020.

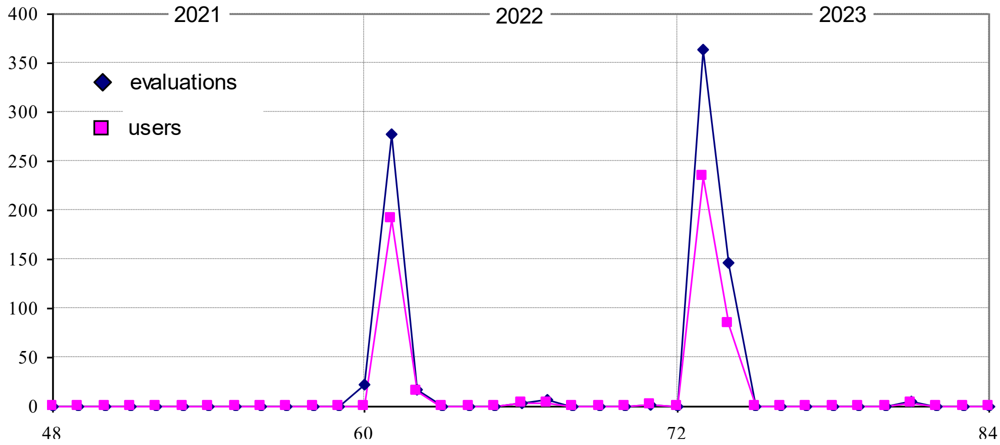

Figure 2. Monthly distinct evaluations and users for 2021–2023.

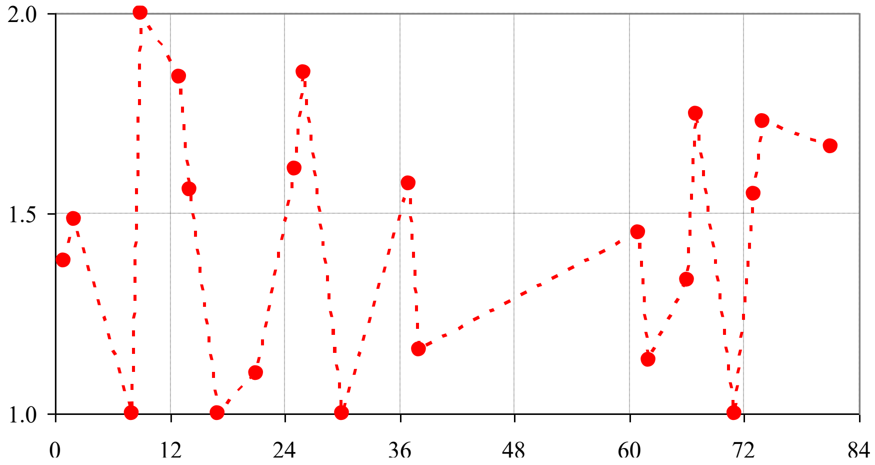

Figure 3. Per user average number of evaluations.

According to the results obtained, the normal distribution hypothesis for the average number of evaluations cannot be rejected (the probability of a better random extraction from the normal distribution N ($\mu = 1.44$, $\sigma^2 = 0.10$). Thus, we observe with the Kolmogorov-Smirnov statistic that we have 26.5%, with the Anderson-Darling statistic we have 22.4%, and with the Chi-Square statistic 35.1% and the conventional limit from which a random draw from the normal distribution must be rejected is 5%. Results obtained show us that all the probability values are far above this limit.

3.2. Evaluated content degree of assimilation

After obtaining the answers for each question, the report was made as a result of all the evaluations in the database. **Figure 4** shows the proportion of questions with correct answers.

Thus, we observe that the average proportion of correct answers is 22.41% (from an average of 1299.4 choices for each question and an average of 290.5 correct answers for each question). From the multitude of questions, it emerged that question 3 is the question with the lowest proportion of correct answers of 11.13% (143 correct answers out of 1284 in total) which suggests that the associated content and its presentation should be improved. Similar reasoning can still be applied to the rest of the questions with a low proportion of correct answers.

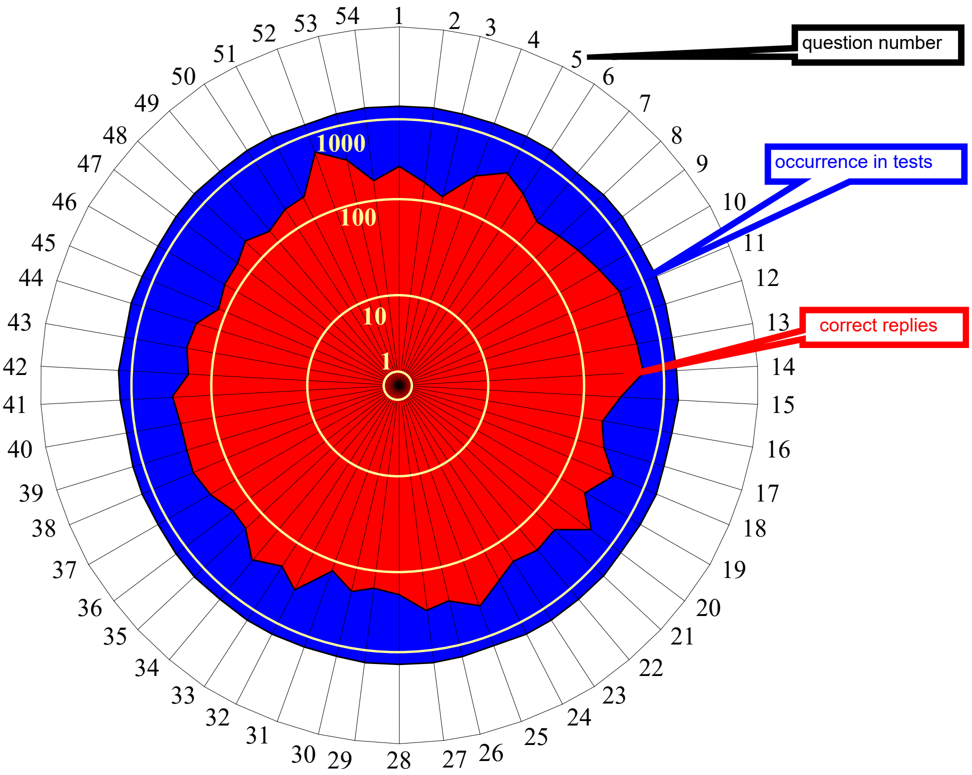

Figure 4. Knowledge coverage: correct replies by question (logarithmic scale).

At the opposite pole is question number 52, where the content seems to be understood by a high percentage of respondents (45.35% correct answers, i.e., 576 out of 1270).

Another way of analysis is that at the level of answers. Previously I mentioned that there are 607 possible answers in the database (296 true, 311 false, with an average of 462.4 occurrences in each test). From the evaluation results we can see that the distribution of answers between the medians of the correct classification is unbalanced, the first 303 wrong classifications contain 231 true answers (and 72 false answers), while the last 303 wrong classifications contain only 65 (and 238 false answers).

For example, we present two cases of each, as follows:

- the statement “Mg is present in chlorophyll” (true) collected 65.15% incorrect answers (129 out of 198), referring to other elements: “In relation to transient elements”); a change to the question is requested;
- “$4AlBr_3 + 3O_2 \rightarrow 6Br_2 + 2Al_2O_3$” (true) collected 64.61% incorrect answers for no obvious reason (answers 294 out of 455); relating to other elements “Relating to the production and use of oxygen”);
- “Irrational formulas” (for “Chemical formulas are:”) was correctly identified as the wrong answer in 90.67% (answers 583 out of 643);
- “Silicon (78%), Oxygen (21%), Others (1%)” as a possible answer to “Atmospheric planetary boundary layer has:” was correctly identified as the wrong answer in 89.34% (444 responses from 497).

From this analysis it follows that the statistical information provided by the evaluation and included in the database is useful information that will allow the improvement of both the course and the evaluation.

For a more complex analysis, **Table 2** was built where the answers are collected by a contingency. Thus, the variables in **Table 2** (CT, CF, WT, WF) are considered as a series whose answers are those answers that are the result of each possible answer, without there being a determined order.

Table 2. Replies on statements (2 × 2) contingency.

Replies			**On**
Correct	**Wrong**		
CT	WT	True	Statements
CF	WF	False	

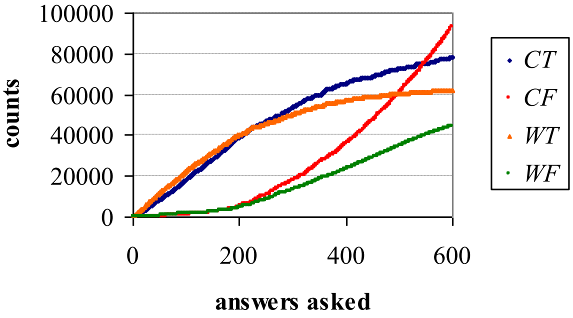

Figure 5. Cumulative gained knowledge.

However, it is possible to order according to the increasing size of the proportion of correct answers identified. With this ordering, the value of the variables can be replaced with the cumulative frequencies (according to **Figure 5**).

The existence of 2339 assessment records with 30 questions each and 4 possible answers for each question, brings us to a total of 280,680 counts. Although the prevalence of correct answers is visible (CT and CF) compared to the wrong ones (WT and WF), identifying the statements becomes more difficult for True (CT and WT) versus False (CF and WF). Thus, we see that for 36.6% (i.e., 222 out of 607) wrong answers were more frequent than correct ones for true statements (WT (i) > CT (i) for $1 \leq i \leq 222$ in **Figure 5**). Much smaller we find the proportion corresponding to false statements (161 out of 607, 26.5%, WF (i) > CF (i) for $1 \leq i \leq 161$ in **Figure 5**). Finally, if we make the difference between CT and WT (CT (607) − WT (607) = 16049) and between CF and WF (CF (607) − WF (607) = 50,979), we notice that the last difference determines a risk in excess of 12.5% (Jäntschi, 2021) for excess risk general considerations and algorithms for exact calculation). Since all variables and sample sizes are quite large, confidence intervals were calculated from the normal approximation of the binomial distribution (using Equation (2) of Bolboacă and Achimaş (2004b)). Following the calculations, for this particular case an excess risk of 12.5 ± 0.3% was obtained, being the excess risk that the correct answer is identified by the respondents to be false and not true. The odds ratio (Bolboacă and Achimaş, 2004a) for normal approximations of its confidence intervals and (Jäntschi, 2021) for general considerations and exact calculation) may be used to express the chance to correctly identify a false answer than a true one. Following the calculations, odds ratio gives almost a doubled chance to correctly identify a false answer than a true one: 1.70 ± 0.02.

3.3. Students' progress

According to the results obtained, it was concluded that the database offers a multitude of information, from which many descriptive and inferential statistics can be extracted, such as subjects with difficulty in understanding (those exemplified in section 3.2). Student progression (first, second, third assessment and more) can also be measured, which is another relevant statistic (for educational purposes). Some of the students, despite numerous attempts, do not make significant progress from one assessment to another. However, what is important is what is obtained on average and as a trend. In order to obtain a relevant statistic, a procedure was implemented here that allows the exploitation of its performance. If a student made a good assessment and stopped there, then that assessment can still count as the next and final assessment. By doing this, all student records that have one, two, three ratings (and so on) have the same number (**Table 3**). The statistics in this series can be used to observe the progress of respondents between assessments, as well as to reveal any trends, if any. Interestingly, the information listed in **Table 3** were made from consecutive assessments, but without taking into account the time interval between them.

Table 3. Learning curves from consecutive evaluations.

Eval	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Avg	32.55	39.40	41.03	41.47	41.60	41.65	41.67	41.67	41.70
StD	17.19	17.64	17.23	17.01	16.94	16.91	16.90	16.90	16.89
Cnt	1246	1246	1246	1246	1246	1246	1246	1246	1246

Eval: Evaluation; Avg: Average; Std: Standard deviation; Cnt: count of.

From the analysis of the results, it can be observed that a different result is obtained when the date and time of the evaluation is taken into account. From the information provided by **Table 3**, a student's t test reveals that statistically there is no difference when more than one assessment occurs (e.g., the 9th assessment provides more points than the first for a random chance probability of 70.42%, and to be considered significant, it had to be less than 5%). At the same time, an increasing trend can be observed in the Avg data, while the Std data has a decreasing trend. A linear regression can be considered significant when the slope of the mean according to Eval is 0.748 and the probability of not being null is 4.3%. Thus, it can be understood that learning is a limiting non-linear curve that we will learn and learn, and complete knowledge is a limit target. From this it follows that a distinct statistically significant Dose-Response pattern can be found in the Avg data (Equation (1)), which indicates that the greatest gain is in between the first and the second evaluation (1.276 coefficient in the model).

$$\text{Avg (Eval)} = 28.58_{\pm 0.29} + \frac{13.13_{\pm 0.80}}{1 + (\text{Eval}/1.276_{\pm 0.057})^{-3.43_{\pm 0.17}}} \tag{1}$$

The free quotient value of 28.58 (from Equation (1)) suggests that 28 points can be obtained with 0 ratings, so it should not be considered a passing score.

If we go further, we can also have a significant exponential model

$$(\text{Avg (Eval)} = 41.66_{\pm 0.04} + 36.20_{\pm 0.50} \cdot \text{Exp}(-\text{Eval}/0.724_{\pm 0.021})) \tag{2}$$

where, the free coefficient (41.66) has a different interpretation, it is the average score obtained after an infinite number of evaluations. This score can also be assigned a grade (e.g., 42 points to a 7 for a rating from 4 to 10, or to a 6 for a rating from 1 to 10).

The results are consistent with the existing literature and show us that the different qualities of the online system have a direct impact on users when they start to learn and use the system (Al-Hunaiyyan et al., 2021; Tilak and Kumar, 2022; Zhu and Liu, 2020).

4. Conclusion and perspectives

The COVID-19 pandemic has triggered unprecedented challenges in most sectors of activity, and in the field of education, the need for adaptation and the imposed changes have created a new approach. This study examines an online assessment system that allows the generation of multiple-choice tests to meet a certain level of difficulty, monitored total test time, and many other constraints designed to identify knowledge level.

From the analysis of the database, you can see the results regarding the statistics for the period 2017–2020, and those related to the period 2021–2023, which show us a gap due to the lack of data for the second post-COVID period (see **Figures 1** and **2**). Also, from the analysis of the average number of evaluations, we can see that there are students who were satisfied with the first or first tests, and in the regression equation statistical significance was obtained for the intersection (see **Figure 3**). Thus, the assumption of normal distribution for the average number of ratings cannot be rejected.

From the analysis of the degree of assimilation of the evaluated content, we observe from the report of all evaluations (see **Figure 4**) that the average proportion of correct answers is 22.41%. At the same time, there are questions that have a small proportion of correct answers (11.13%, see question 3), but there are also questions where the weight is high (45.35%, see question 52). This statistical information obtained from the study are useful information that will allow the improvement of both the course and the evaluation. Using a more complex analysis, by collecting the answers through a contingency (see **Table 2** and **Figure 5**), it was observed that the chance of correctly identifying a false answer than a true one is doubled (1.70 ± 0.02), which always leads us to the need for improvement.

Analyzing the progress registered by students, we can say that there is a number of students who, despite numerous attempts, do not make significant progress from one assessment to another. From the analysis of the results, it can be seen that a different result is obtained when the date and time of the assessment is taken into account. Although, Student's *t*-test reveals that statistically there is no difference when more than one assessment takes place, from the results obtained (using Equation (1)) it can be said that the biggest gain is between the first and second assessment. Thus, we can understand that learning is a non-linear limiting curve that we will learn and learn and complete knowledge is a limit target.

Although the adaptation, through the use of an implemented online assessment system (MCMA) was achieved (as a viable and reliable solution for the assessment of students' knowledge), the results show us that it requires improvement. If, at first glance, it may seem complicated and burdensome, it is necessary in today's educational environment, creating the possibility of new approaches and improvements. Thus, student progress between assessments, as well as subjects with difficulty in understanding, can be leveraged through the assessment system, and obtaining the average and any overall trend is particularly useful.

It therefore aims to inform policy makers, guide educational practices, and inspire future studies by identifying adaptation techniques for solving similar problems. We believe that the adaptation process is not over, that once started it will continue, but the results of this study can be a useful tool in designing a more robust, egalitarian and innovative educational system.

5. Policies suggestion

Below is a collection of stakeholder policies that could be considered, given the experience of online learning and assessment during the COVID-19 pandemic:

The improvement of digital infrastructure and widespread access to the Internet, as well as the availability of digital devices, are intended to reduce the digital gap between students in urban and rural areas and to support education.

Accessible technological solutions for distance learning and the implementation of extensive training programs can lead to improved skills.

A flexible curriculum that transitions between in-person and online learning modes, as well as accessible resources can help and support overcoming the psychological impact of a pandemic.

Encouraging the development and adoption of innovative learning platforms and educational technologies can contribute to the permanent improvement of the quality of online education.

Fostering partnerships between educational institutions and technology companies to create engaging and interactive digital learning resources.

These policies, mentioned above, can help address a number of obstacles inherent in crisis situations and establish a more solid and equitable educational framework. This framework will enable effective navigation in situations similar to the COVID-19 pandemic and will meet the educational requirements of all students through unity and coherence.

Author contributions: Conceptualization, CES and LJ; methodology, CES; software, LJ; validation, LJ; formal analysis, CES; investigation, LJ; resources, LJ; data curation, LJ; writing—original draft preparation, CES; writing—review and editing, CES; visualization, LJ; supervision, LJ; project administration, LJ. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The content of this paper is a development of what was presented at the '10th Annual Conference on Computational Science & Computational Intelligence' (CSCI 2023). We would like to thank the participants of the conference, who asked questions and comments, which subsequently added value. We also thank the blind reviewers from the journal, who also contributed significantly to increase the value of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

Alagarsamy, S., Vijay, M. (2019) Construction and Validation of the Learning Management System Success Scale in the Higher Education Setting. Available online: http://gatrenterprise.com/GATRJournals/GJBSSR/pdf_files/GJBSSRVol7(2)2019/5.Subburaj.pdf (accessed on 9 October 2023).

Al-Hunaiyyan, A., Alhajri, R., Al-Sharhan, S., et al. (2021). Factors Influencing the Acceptance and Adoption of Online Learning in Response to the COVID-19 Pandemic. International Journal of Web-Based Learning and Teaching Technologies, 16(6), 1–16. https://doi.org/10.4018/ijwltt.20211101.oa5

Alsayyari, A., Alblawi, A., Elhajji M. (2019). Engineering students' acceptance and experience of learning management systems: a case study at Shaqra university. Available online: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8632633 (accessed on 5 October 2023).

Alzahrani, L., & Seth, K. P. (2021). Factors influencing students' satisfaction with continuous use of learning management systems during the COVID-19 pandemic: An empirical study. Education and Information Technologies, 26(6), 6787–6805. https://doi.org/10.1007/s10639-021-10492-5

Anglin, G. J., Vaez, H., & Cunningham, K. L. (2004). Visual Representations and Learning: The Role of Static and Animated Graphics. Available online: https://www.taylorfrancis.com/books/edit/10.4324/9781410609519/handbook-research-educational-communications-technology-david-jonassen-marcy-driscoll?refId=ad1dd4f7-2f5c-4cee-b954-852075fee1fa&context=ubx (accessed on 15 October 2023).

Balkaya, S., & Akkucuk, U. (2021). Adoption and Use of Learning Management Systems in Education: The Role of Playfulness and Self-Management. Sustainability, 13(3), 1127. https://doi.org/10.3390/su13031127

Baragash, R. S., & Al-Samarraie, H. (2018). Blended learning: Investigating the influence of engagement in multiple learning delivery modes on students' performance. Telematics and Informatics, 35(7), 2082–2098. https://doi.org/10.1016/j.tele.2018.07.010

Barrot, J. S., Llenares, I. I., & del Rosario, L. S. (2021). Students' online learning challenges during the pandemic and how they cope with them: The case of the Philippines. Education and Information Technologies, 26(6), 7321–7338. https://doi.org/10.1007/s10639-021-10589-x

Bayazit, A., & Aşkar, P. (2012). Performance and duration differences between online and paper-pencil tests. Asia Pacific Education Review, 13(2), 219–226. https://doi.org/10.1007/s12564-011-9190-9

Bloem, J., Doorn, M. V., Duivestein, S., et al. (2014). The Fourth Industrial Revolution. Available online: https://docslib.org/doc/12519523/the-fourth-industrial-revolution-things-to-tighten-the-link-between-it-and-ot (accessed on 9 October 2023).

Bolboacă, S. D., Achimaş-Cadariu, B. A. (2004a). Binomial Distribution Sample Confidence Intervals Estimation 6. Available online: http://lejpt.academicdirect.org/A04/01_21.pdf (accessed on 9 October 2023).

Bolboacă, S. D., Achimaş-Cadariu, B. A. (2004b). Binomial Distribution Sample Confidence Intervals Estimation 5. Available on: http://ljs.academicdirect.org/A04/26_43.pdf (accessed on 9 October 2023).

Burton, R. F. (2001). Quantifying the Effects of Chance in Multiple Choice and True/False Tests: Question selection and guessing of answers. Assessment & Evaluation in Higher Education, 26(1), 41–50. https://doi.org/10.1080/02602930020022273

Camilleri, M. A., & Camilleri, A. C. (2021). The Acceptance of Learning Management Systems and Video Conferencing Technologies: Lessons Learned from COVID-19. Technology, Knowledge and Learning, 27(4), 1311–1333. https://doi.org/10.1007/s10758-021-09561-y

Cavus, N., Mohammed, Y. B., & Yakubu, M. N. (2021). Determinants of Learning Management Systems during COVID-19 Pandemic for Sustainable Education. Sustainability, 13(9), 5189. https://doi.org/10.3390/su13095189

Chang, J. C., & Akahori, K. (1999). An Evaluation of Japanese CALL Systems on the WWW Comparing a Freely Input Approach with Multiple Selection. Computer Assisted Language Learning, 12(1), 59–79.

https://doi.org/10.1076/call.12.1.59.5717

Chang, C. T., Hajiyev, J., & Su, C. R. (2017). Examining the students' behavioral intention to use e-learning in Azerbaijan? The General Extended Technology Acceptance Model for E-learning approach. Computers & Education, 111, 128–143. https://doi.org/10.1016/j.compedu.2017.04.010

Dannenberg, R. B., & Hu, N. (2003). Pattern Discovery Techniques for Music Audio. Journal of New Music Research, 32(2), 153–163. https://doi.org/10.1076/jnmr.32.2.153.16738

Dobre, I. (2015). Learning Management Systems for Higher Education—An Overview of Available Options for Higher Education Organizations. Procedia—Social and Behavioral Sciences, 180, 313–320. https://doi.org/10.1016/j.sbspro.2015.02.122

Fallows, S., & Steven, C. (2000). Building employability skills into the higher education curriculum: a university—wide initiative. Education + Training, 42(2), 75–83. https://doi.org/10.1108/00400910010331620

Garcia, J. G., Gangan, M. G., Tolentino, M. N., et al. (2020). Canvas adoption: Assessment and acceptance of the learning management system on a web-based platform. Available online: https://www.researchgate.net/publication/347404623_Canvas_Adoption_Assessment_and_Acceptance_of_the_Learning_Management_System_on_a_Web-based_Platform#fullTextFileContent (accessed on 8 April 2024).

Hashemi Hosseinabad, S., Safayani, M., & Mirzaei, A. (2021). Multiple answers to a question: a new approach for visual question answering. The Visual Computer, 37(1), 119–131. https://doi.org/10.1007/s00371-019-01786-4

Hestenes, D. (1987). Toward a modeling theory of physics instruction. American Journal of Physics, 55(5), 440–454. https://doi.org/10.1119/1.15129

Holmes, P. (2002). Multiple evaluation versus multiple choice as testing paradigm feasibility, reliability and validity in practice [PhD thesis]. University of Twente.

Jäntschi, L., & Bolboacă, S. D. (2006). Auto-calibrated online evaluation: database design and implementation. Leonardo Electron. J. Pract. Technol., 5(9), 179–192.

Jäntschi, L., Stoenoiu, C. E., & Bolboaca, S. D. (2007). Linking Assessment to e-Learning in Microbiology and Toxicology for Undergraduate Students. In: Proceedings of the EUROCON 2007—The International Conference on "Computer as a Tool". pp. 2447–2452.

Jäntschi, L. (2013). General Chemistry. Available online: http://lori.academicdirect.org/books/pdf/2013_gcc.pdf (accessed on 4 November 2023).

Jäntschi, L. (2021). Formulas, Algorithms and Examples for Binomial Distributed Data Confidence Interval Calculation: Excess Risk, Relative Risk and Odds Ratio. Mathematics, 9(19), 2506. https://doi.org/10.3390/math9192506

Jäntschi, L. (2022). Binomial Distributed Data Confidence Interval Calculation: Formulas, Algorithms and Examples. Symmetry, 14(6), 1104. https://doi.org/10.3390/sym14061104

Jäntschi, L. (2023). General chemistry laboratory work: a practical guide (Romanian). Available online: http://ph.academicdirect.org/llcggp.pdf (accessed on 5 November 2023).

Jiang, P., Yan, K., Chen, H., et al. (2022). Building of online evaluation system based on socket protocol. Computer Science and Information Systems, 19(1), 185–204. https://doi.org/10.2298/csis210201047j

Joaquin, J. J. B., Biana, H. T., & Dacela, M. A. (2020). The Philippine Higher Education Sector in the Time of COVID-19. Frontiers in Education, 5. https://doi.org/10.3389/feduc.2020.576371

Labrak, Y., Bazoge, A., Dufour, R., et al. (2022). FrenchMedMCQA: A French multiple-choice question answering dataset for medical domain. In: Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI); Abu Dhabi, United Arab Emirates. pp. 41–46.

Loftis, J. R. (2019). Beyond Information Recall. American Association of Philosophy Teachers Studies in Pedagogy, 5, 89–122. https://doi.org/10.5840/aaptstudies2019121144

Matteson, S. M. (2006). Mathematical Literacy and Standardized Mathematical Assessments. Reading Psychology, 27(2–3), 205–233. https://doi.org/10.1080/02702710600642491

Mehrolia, S., Alagarsamy, S., & Indhu Sabari, M. (2021). Moderating effects of academic involvement in web-based learning management system success: A multigroup analysis. Heliyon, 7(5), e07000. https://doi.org/10.1016/j.heliyon.2021.e07000

Munyengabe, S., Yiyi, Z., Haiyan, H., & Hitimana, S. (2017). Primary teachers' perceptions on ICT integration for enhancing teaching and learning through the implementation of One Laptop Per Child Program in primary schools of Rwanda. Eurasia Journal of Mathematics, Science and Technology Education, 13(11), 7193e7204. https://doi.org/10.12973/ejmste/79044

Nașcu, H. I., & Jäntschi, L. (2004a). Multiple choice examination system 1. Database Design and Implementation for General

Chemistry. Available online: http://ljs.academicdirect.org/A05/18_33.htm (accessed on 4 November 2023).

Naşcu, H. I., & Jäntschi, L. (2004b). Multiple Choice Examination System 2. Online Quizzes for General Chemistry. Available online: http://ljs.academicdirect.org/A05/18_33.htm (accessed on 4 November 2023).

Nguyen, N. T. (2021). A study on satisfaction of users towards learning management system at International University—Vietnam National University HCMC. Asia Pacific Management Review, 26(4), 186–196. https://doi.org/10.1016/j.apmrv.2021.02.001

Nguyen, T., Bui, T., Fujita, H., et al. (2021). Multiple-objective optimization applied in extracting multiple-choice tests. Engineering Applications of Artificial Intelligence, 105, 104439. https://doi.org/10.1016/j.engappai.2021.104439

Nicol, D. (2007). E-assessment by design: using multiple-choice tests to good effect. Journal of Further and Higher Education, 31(1), 53–64. https://doi.org/10.1080/03098770601167922

Omari, A. (2013). An Evaluation and Assessment System for Online MCQ’s Exams. International Journal of Electronics and Electrical Engineering, 1(3), 219–222. https://doi.org/10.12720/ijeee.1.3.219-222

Ramírez-Correa, P. E., Rondan-Cataluña, F. J., Arenas-Gaitán, J., et al. (2017). Moderating effect of learning styles on a learning management system’s success. Telematics and Informatics, 34(1), 272–286. https://doi.org/10.1016/j.tele.2016.04.006

Raza, S. A., Qazi, W., Khan, K. A., et al. (2020). Social Isolation and Acceptance of the Learning Management System (LMS) in the time of COVID-19 Pandemic: An Expansion of the UTAUT Model. Journal of Educational Computing Research, 59(2), 183–208. https://doi.org/10.1177/0735633120960421

Regulation (EU). (2016). 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 (accessed on 8 October 2023).

Suleman, F. (2018). The employability skills of higher education graduates: insights into conceptual frameworks and methodological options. Higher Education, 76(2), 263–278. https://doi.org/10.1007/s10734-017-0207-0

Tasdemir, S., Balcı, M., Cabi, A., et al. (2015). The Design and Application of Online Exam System Supported by Database. International Journal of Applied Mathematics, Electronics and Computers, 3(3), 204–207. https://doi.org/10.18100/ijamec.43348

Tinmaz, H., & Lee, J. H. (2020). An analysis of users’ preferences on learning management systems: a case on German versus Spanish students. Smart Learning Environments, 7(1). https://doi.org/10.1186/s40561-020-00141-8

Tilak, J. B. G., & Kumar, A. G. (2022). Policy Changes in Global Higher Education: What Lessons Do We Learn from the COVID-19 Pandemic? Higher Education Policy, 35(3), 610–628. https://doi.org/10.1057/s41307-022-00266-0

Zhu, X., & Liu, J. (2020). Education in and After COVID-19: Immediate Responses and Long-Term Visions. Postdigital Science and Education, 2(3), 695–699. https://doi.org/10.1007/s42438-020-00126-3

Appendix

A.1. Abbreviations

- MCMA: multiple choice multiple answers
- HTTP: hypertext transfer protocol
- GDPR: general data protection regulation
- URL: universal resource locator
- COVID-19: Corona virus disease 2019
- EU: European Union
- PHP: software (pre and post processed hypertext)
- MySQL: software (relational database management system)
- Apache: software (cross-platform web server)
- FreeBSD: software (Unix-like operating system)
- IP (address): Internet protocol address (usually referring its v4 version)

A.2. General chemistry subjects covered in the evaluation

- Periodic system; periodic properties; electronic structure
- The abundance of elements; chemical formulas; stoichiometry
- Minerals; physical and chemical properties; chemical reactions
- Hydrogen; oxygen; water
- Alkali and alkaline earth metals
- p3–p6 block of elements (groups 15–18)
- d1–d5 block of elements (groups 3–7)
- d6–d10 block of elements (groups 8–12)
- f1–f14 elements block (lanthanides and actinides)
- Boron group; carbon group
- Organic chemistry; hardness and hard materials
- Ceramics; semiconductors; superconducting
- Advanced materials; polymers and plastics; biomolecules and reaction mechanisms
- Methods and models; structure activity/property relationships

A.3. Example of generating remote-based evaluation files

1) By the solid-state density, the chemical elements can be ordered as follows:
   a) B < Be < Li < He < H
   b) B < C < N
   c) N < O < F
   (A and C are correct, B is wrong—the order is opposite)
2) In connection with isotopes of hydrogen:
   a) 3M(T) = M(p)
   b) T = 3015H is tritium
   c) D = 21H is deuterium
   (C is correct, A and B are wrong)
3) For the reaction $H_2 + O_2 \rightarrow H_2O$:
   a) H_2 and O_2 are products of reaction
   b) The correct coefficients are 1 (H_2), 2 (O_2), 2 (H_2O)

c) H_2 and O_2 are reactants

(C is correct, A and B are wrong)

4) (450 questions in a file; many generated files)