2023 International Conference on Computational Science and Computational Intelligence (CSCI) | 979-8-3503-6151-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CSCI62032.2023.00282

# A computer based evaluation system: design, implementation and results on general chemistry

Carmen Elena Stoenoiu
*Department of Electric Machines and Drives*
*Technical University of Cluj-Napoca*
Cluj-Napoca, Romania
http://orcid.org/0000-0003-2306-5942

Lorentz Jäntschi
*Department of Physics and Chemistry*
*Technical University of Cluj-Napoca*
Cluj-Napoca, Romania
http://orcid.org/0000-0001-8524-743X

***Abstract*—A computer based multiple choice multiple answers evaluation system has been designed, implemented, posted online and used to evaluate students on general chemistry for the last 10 years. In this paper, the design, implementation, and some results regarding the students' performances are presented and discussed.**

***Index Terms*—online evaluation system, multiple choice questions, general chemistry, first-year undergraduate students**

## I. Introduction

The mission of a knowledge assessment system is to provide an objective answer regarding the degree of knowledge assimilation on a particular subject. A number of studies addresses this issue [1]–[5].

Multiple choice evaluation systems with single answer questions (pick one answer, implemented with radio buttons) are easy to be built [6], [7], widely used for tests [8], and statistics associated with them provide exact confidence intervals [9]. Unfortunately, these systems are flawed by the possibility of guessing [10].

Multiple choice evaluation systems (MCMA; pick one or more answers, implemented with checkboxes) are more difficult to be built [11], keep the inclusiveness high [12], and are more reliable. This leads to significantly diminishing the possibility of providing an answer by guessing [13].

It is worth noting that, for large topics, some authors [14] communicate the use of hybrid (single answer and multiple answers) evaluation systems. One should note that the MCMA system is not always the best choice, as some authors report [15], and online systems have some disadvantages when compared with traditional paper based tests as well [16].

The contents can be very varied and complex. For example, in painting there is a great emphasis on visual representations [17], in music - on audio representations [18], in mathematics - on equations and formulas [19], in physics - on laws and principles [20].

In chemistry, the constituent elements of content construction are equations (of chemical reactions), formulas (chemical molecular, structural, and geometric), pictorial representations (chemical processes and technologies, operating principles for methods), information structured in different manners (tables - eg: Periodic Table; strings - eg: electronic configuration of atoms and ions; scales - eg: pH scale) [21].

An assessment system must emphasize the formation of connections between information and measure the degree to which these have been made, rather than measuring the degree of information assimilation. Thus, in a multiple-choice assessment system, the questions must be designed so that, wherever possible, they refer to the associations that occur between information.

It is important that, among the possible answers, there be some that induce cognitive conflict by presenting anomalous and contradictory data [22]. At the same time, other authors note that people who hold strong opinions on complex social issues are likely to examine relevant empirical evidence in a biased manner, being apt of accepting "confirming" evidence at face value while subjecting "disconfirming" evidence to critical evaluation, As a result, undue support is drawn for the initial positions, from mixed or random empirical findings. [23].

The construction of an online assessment system frequently requires the use of a database to store the information associated with the assessed content: the use of a support system for managing the database and the user interface, a computer system, and classification based on recorded responses [24]–[26]. Similar systems were previously reported [27]–[29].

This paper puts forward the design and implementation of an online MCMA assessment system, and the results obtained with it in the assessment of first-year undergraduate students in general chemistry over several years.

Considering the Covid isolation has led to significant changes in the way knowledge is transmitted, to illustrate the differences between the period before and after this, two time periods have been selected for comparison: from 2017 to 2019 and from 2022 to 2023.

The system was adapted for online exclusive use, it was used as such for two years (2020 and 2021), and its changes are provided for convenience.

## II. Material and Method

A one-semester general chemistry content has been developed for first-year undergraduate students, which is aimed

Submited as **Full/Regular Research Paper**. To be deliverd in **RTED** at *The 2023 International Conference on Computational Science and Computational Intelligence (CSCI'23). December 13-15, 2023, Las Vegas, USA.*

2769-5654/23/$31.00 ©2023 IEEE
DOI 10.1109/CSCI62032.2023.00282

especially at students for whom chemistry is not one of the core areas from which their training derives. Its summary is provided in the Appendix B.

The general chemistry course [21] that deploys the content in the appendix is addressed to students studying in Romanian, English and German, so the evaluation system was designed in a flexible way to suit all these languages. The elaborated content was intended to cover as much of the thematic area of chemistry as possible without addressing complex notions of chemistry, specific to specialized training. Laboratory skills and associated knowledge assessment is elaborated in two separate evaluations [30].

### A. Database structure

A template was designed (see Tab. I), and a database containing 3 files, according to the template, is to be created for each topic (one database, three tables for general chemistry).

TABLE I
ONLINE EVALUATION STORAGE DATABASE TOPOLOGY

| Table | Fields | Notes |
|---|---|---|
| User | \|Id\|Name\|Pass\|Date\| | n1 |
| Test | \|id\|qroenge\|rro\|ren\|rge\|aroenge\| | n2 |
| Eval | \|id\|subj\|lang\|name\|suid\|qlist\|rlist\|alist\|tb\|te\|p\|t\| | n3 |

n1: Password *Pass* is stored encrypted with MD5 (32 characters)
n2: *qroenge* contains three texts separated by carriage return; *rro*, *ren*, *rge* and *aroenge* contains a varied (but identical) number of lines; on each line, there is one possible answer (*rro*, *ren*, *rge*) and the state of the truth of the answer 0/1 (*aroenge*)
n3: *subj* and *suid* are to manage secure connection; *qlist*, *rlist*, *tlist* and *alist* are ordered lists of values separated by space (*qlist* - selected questions, *rlist* - selected answers, *tlist* - truth state for the selected answers; *alist* - truth for the replied answers); *tb*, *te*, and *t* stores times and *p* earned points

The database was populated with 54 questions. The user interface allows adding users and saving records of each evaluation. Descriptive statistics for the evaluation content of the database are as follows:

- There are 607 possible answers (an average of 11.24 answers per question);
- There are 296 true answers and 311 are false (averages of 5.48 and 5.76 and a ratio true:false of $\approx$ 19:20);
- The lowest number of possible answers is 8, the highest number is 26;
- The lowest number of true answers is 3, the highest number is 13;
- The lowest number of false answers is 4, the highest number is 13.

### B. Principles of the evaluation system

The testing is designed to take place in a certain room on computers from the same class of IP addresses under the professor's supervision (a test begins when both the professor's password and students' password match the stored passwords).

A student can apply the test whenever wanted (in a reasonable timeframe of the day). The database contains 54 questions. Each test will extract 30 questions, with 4 possible answers. For convenience, among the 4 possible answers are included at least one true (correct) and one false (wrong). A problem is considered to be solved correctly when only the correct answers have been marked. Each correct solution brings 3 points.

The testing is timed, with a time limit of 15 minutes. The time when the test is generated and the time when the solution to the test is transmitted are recorded (*tb* and *te* in Tab. I).

The test score is calculated as follows:

- the average time per correct answer from the current test is calculated ($tm\_rc$);
- the average time required for a correct answer from all the tests in the database ($tm\_rc\_nec$) is calculated;
- the two times are divided to calculate the coefficient $c1 = tm\_rc/tm\_rc\_nec$;
- it is proceeded in the same way for the number of correct answers, and the coefficient $c2 = nr\_rc/nr\_rc\_nec$ is calculated;
- a mean value of the 2 coefficients ($c1$ and $c2$) is calculated, and the test score is 10 times the calculated average;
- The test average is calculated for all tests applied by a user: the list of test scores is built; if there are at least 2 grades on the list, the smallest one is eliminated; the remaining ones are averaged;
- Grade calculation: the lowest test average (redefined as a constant and fixed at 3.5) is associated with grade 4 (four); the highest test average is associated with a grade of 10 (ten); the grade is given by placing the test average between the lowest test average and the highest test average.

### C. Covid-19 era - generating and applying tests competitively (first to response)

A study reveals that higher education institutions were unprepared for exclusively online learning [31]. In the case of the system presented here, a number of changes have been made to adapt it to an exclusively online assessment, namely:

- the number of possible answers has been reduced, from 4 to 3, so that the recorded answer is greatly simplified - as at least one answer is correct, out of 3 answers exactly 1 or two are true, so the possible answers are then included in the list: A, B, C, AB, AC, BC;
- the competitive assessment strategy has changed: from individual counter-time (in which each student is assessed with an individual test) to counter-time in series of 6 students, using lists with a large number of questions (e.g. 500) from which one question is chosen at the time of evaluation, and recorded is the first (quickest) answer or the first two (if the second is different from the first);
- in this new evaluation strategy, positive points are received for each good answer, and negative points for each wrong answer; the first two students (out of the 6) who reach a total of 6 answers with more positive than negative answers receive the grade 1 + 1.5*(number of positive answers - number of negative answers) - with a minimum of 1 and a maximum of 10. For the next 2, the maximum goes down to 9, and for the last 2, the maximum goes down again to 8, and it is no longer

conditioned that the positive answers are more than the negative ones;
- since the evaluator and the evaluated are not in the same location, cheating is always possible; in order to prevent the textual searching (for the questions and for the answer), in the generated tests the spaces between words were replaced with double spaces.

Three MCMA genererated questions are listed as example in Appendix C.

### D. The general data protection regulation

GDPR is an EU regulation on information privacy, effective from 25 May 2018 [32], which enforced pseudonymization of the information displayed on the statistics page. The names of the students were replaced with "Student $< number >$", where $number$ is uniquely generated for each student in the database at the query of the system.

### E. Programs and their topology

A welcome page ($index$, see Tab. II) has been created, which is at the same time the root entry for the rest of the programs. Its universal resource locator (URL) is: http://l.academicdirect.org/Education/Evaluation/Chemistry/Chimie_Generala/. Accessing it with "?lang=ro" (default language) provides the welcome message and the menu in Romanian, while "?lang=en" provides it in English, and "?lang=de" in German. The system is flexible, so that any language can be added at any time.

TABLE II
ONLINE EVALUATION SOFTWARE TOPOLOGY

| Program | Actions |
|---|---|
| $index$ | welcome & menu |
| $insert$ | add users (students) |
| $test$ | generate a test & save an evaluation |
| $statistics$ | calculate & display evaluation results |
| $password$ | module containing credentials for database connection |
| $security$ | module checking allowance of the testing (IP address based) |

The $password$ and $security$ modules are called inside of the programs whenever necessary to get credentials for a database connection ($password$), or to check the allowance of an insert or update operation ($security$).

## III. SOFTWARE IMPLEMENTATION

A MySQL server (version 5.5.4) is running on a (different) intranet computer. The storage database managed through a mysqli connection. The programs were implemented using PHP (version 7.4.10 is compiled and running on Apache 2.4.46 HTTP server and FreeBSD 12.2 operating system).

### A. Using of the system for evaluation

$test$ program interacts with the user in three contexts. By precedence, the first is before beginning a test. At this point, the following credentials need to be verified:

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

A link for registering a new user ($insert$ program) is provided. 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. To be noted that it is not possible to give the test during the night, or during the semester, only during the day and during the examination session. A test contains (second call for the $test$ program):

- a number of $m$ questions picked at random (without replacement) from the list of $n$ available questions ($m$ is a predefined constant, set to 30, and $n$ is queried from database, was 54);
- for each question ($q_1$, ..., $q_m$) a number of $p$ ($p$ is a predefined constant and were set to 4) possible answers ($r_{i,1}$, ..., $r_{i,p}$), each having associated a check box;
- Unix time for the moment when the test was generated and sent to the client;
- user name;
- an unique id (recorded in the $eval$ table as well, making the generation of another test for the same user impossible as long as the current test is not finalized);
- a button to finalize the test.

Finally, the third call for the $test$ program:

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

### B. Querying the system for getting marks

The database contains records for all evaluations, but only records for students having at least one evaluation since January of the current year are listed for practical reasons (the students have their examination of the first semester exams in January). Furthermore, an examination session is no longer than one month, and the students exam marks list reports must be filled with the information from this time period - the default (for $statistics$ program) is to display the records for the students having evaluations in the last month. A number of five tables are generated:

- All evaluations list - contains all evaluations of all students that match the filtering criteria of timeframe grouped by the students and sorted ascending by date; when more than one evaluation is recorded for a student, the one with worse results is null, and not further considered a record for the student's average performance;

- Database descriptive statistics table (numeric values from 2339 records in the database) - average time for getting a point (the value of this statistic is about 7.5 s); the average of the number of the points (the value of this statistic is about 31.6 points);
- Testing marks and means table - for all the included evaluations relative (to the above averages) ratios are listed; associated testing scores are calculated; the average is given in boldface;
- Database descriptive statistics table (numeric values from 2339 records in the database) containing the failing evaluation score (the one associated with a mark of 4 out of 10; the value of this variable has been fixed at 3.5) and the best evaluation score (the value of this statistic is about 23);
- Results table, containing the marks and list of points (with hyperlinks to full details of the evaluations) for each (pseudonymized) student.

## IV. Results

### A. Statistics from the use of the system

Basic descriptive and inferential statistics can be extracted from the database, including its usage statistics (monthly evaluations, Fig. 1; monthly distinct users, Fig. 2) One can

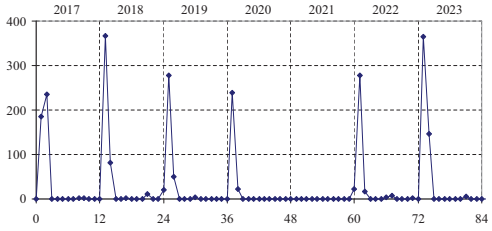


Fig. 1. Monthly (distinct) evaluations

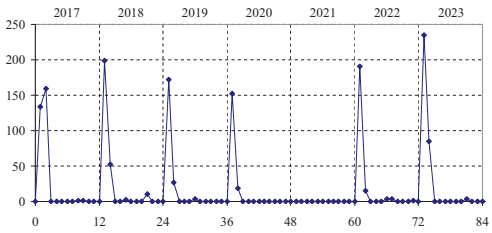


Fig. 2. Monthly (distinct) users

notice the gap between March 2020 and January 2022, where no new user has been added and no evaluation has been made - it is the timeframe when Covid-19 restrictions prevented face to face meetings, and the system was used to do online remote evaluations, following the procedure described in Section II-C (for 22 months). Another statistic is about the average number of evaluations (Fig. 3), which ranged from 1.0 to 2.0, showing that n average students were satisfied with their first or first two evaluations. There is no trend in the time series from Fig.

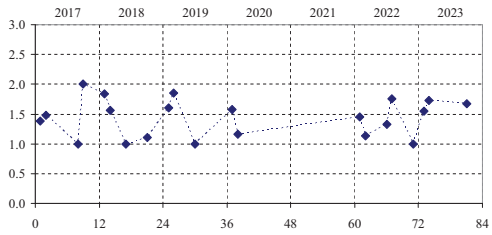


Fig. 3. Per user average number of evaluations

3, the regression equation has statistical significance only for the intercept ($y(t) = 1.43_{\pm 0.26} + 0.0003_{\pm 0.006}, r^2 = 0.0006$; probability associated with the intercept not belonging to the model is $P(1.43; t = 11.5, n = 19) = 5 \cdot 10^{-8}\%$; probability associated with the slope not belonging to the model is $P(0.0003; t = 0.11, n = 19) = 91\%$). Additionally, the hypothesis of normal distribution for the average number of evaluations cannot be rejected (probability of a better draw at random from normal distribution $\mathcal{N}(\mu = 1.44, \sigma^2 = 0.10)$ assessed by Kolmogorov-Smirnov statistic is 26.5%, assessed by Anderson-Darling statistic - 22.4%, and assessed by Chi-Squared statistic - 35.1%; the conventional limit from which one must reject a hypothesis of a random draw from normal distribution is 5% and all the probability values are well above this limit).

### B. Evaluated content degree of assimilation

When inspecting the answers provided to each question, one must construct the report by querying all evaluations in the database. For each evaluation, the questions in the list add one unit to the frequency of the question, and add one unit to the frequency of the correct answered question, if all provided answers for it are correct. Proportion of the correct answered questions can be further calculated (Fig. 4).

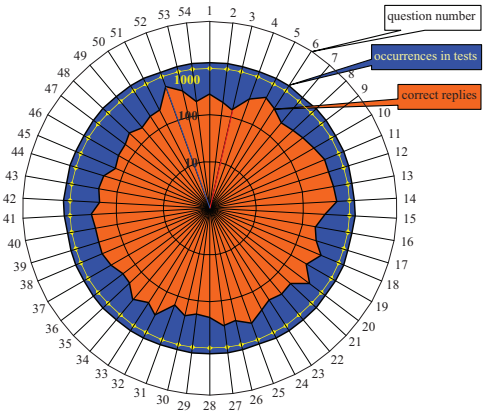


Fig. 4. Knowledge coverage: correct replies by question (logarithmic scale)

The average proportion of correct answers was 22.41% (from an average of 1299.4 picks for each question and an average of 290.5 correct answers for each question). The question with the lowest proportion of correct answers was question

number 3 (”By the first ionization potential, the chemical elements can be ordered as follows:”) with a proportion of 11.13% (143 correct answers from 1284 in total) and this suggests that the associated content and its presentation to the students must be improved. Similar reasoning can be applied further for the rest of questions with low proportion of correct replied answers.

On the other side of the statistic is question number 52 (”In connection with polymers:”) which appears to be the best understood topic of the course (45.35% correct replied answers, 576 out of 1270).

A similar way of extracting information is when the analysis is taken at answers level. As mentioned before, there are 607 possible answers in the database (296 true, 311 false, with an average of 462.4 appearances on tests each). The most difficult to be identified were the true answers. The distribution of the answers between the medians of the correct classification of the answers is very unbalanced: first 303 misclassifications contain 231 true answers (and 72 false answers), while the last 303 misclassifications contain only 65 (and 238 false answers). Picking the first two cases from each extreme:

- ”Mg is present in chlorophyll” (true) collected 65.15% (129 out of 198) incorrect answers (most of them probably because it was attached to a question relating to other elements, ”In connection with the transitional elements:”); a change of this assignment will probably improve the correct classification;
- ”$4AlBr_3 + 3O_2 \longrightarrow 6Br_2 + 2Al_2O_3$” (true) collected 64.61% (294 out of 455) incorrect answers with no obvious reason why (is in the proper connection In connection with the production and use of the oxygen:”);
- ”Irrational formulas” (for ”Chemical formulas are:”) was correctly identified as wrong answer in 90.67% replies (583 out of 643);
- ”Silicium (78%), Oxygen (21%), Others (1%)” as a possible reply to ”The atmospheric planetary boundary layer it has:” was correctly identified as wrong answer in 89.34% replies (444 out of 497).

Once again, inspecting the statistics of the evaluation database, useful information is retrieved, helping the improvement of the course and of the evaluation.

A more complex analysis is conducted if the replies are collected by a contingency (Tab. III). Each of the variables

TABLE III
REPLIES ON STATEMENTS (2 × 2) CONTINGENCY

| **Replies** | | | **on** |
|---|---|---|---|
| Correct | Wrong | | |
| $CT$ | $WT$ | True | **Statements** |
| $CF$ | $WF$ | False | |

from Tab. III ($CT$, $CF$, $WT$, $WF$) can be seen as a series of which answers are the replies to each possible answer with no particular order specified. But actually a particular order makes sense: the ascending order of the proportion of correct identified answers. For this particular order, the values of the variables can be replaced with their cumulative frequencies (Fig. 5). In the database, there are 2339 records

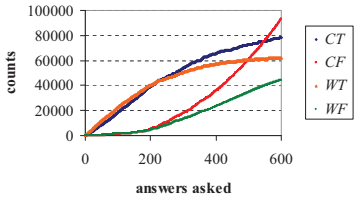


Fig. 5. Cumulative gained knowledge

of evaluations with 30 questions each and 4 possible answers for each question, which gives a total of 280,680 counts. $CT(607) + CF(607) + WT(607) + WF(607)$ from Fig. 5 is 280,680. In Fig. 5, the prevalence of correct answers is visible ($CT$, $CF$) over the wrong ones ($WT$, $WF$). More interesting is the issue of the identification of the True ($CT$, $WT$) vs. False ($CF$, $WF$) statements. For about half of the answers (for 222 out of 607 exactly, 36.6%) wrong replies were more frequent than correct ones for True statements ($WT(i) > CT(i)$ for $1 \leq i \leq 222$ in Fig. 5). Significantly smaller is the proportion corresponding to the False statements (161 out of 607, 26.5%, $WF(i) > CF(i)$ for $1 \leq i \leq 161$ in Fig. 5). In the end, from all answers, the fissure is even greater: $CT(607) - WT(607) = 16049$, while $CF(607) - WF(607) = 50979$, with an excess risk (see [33] for excess risk general considerations and algorithms for exact calculation) of 12.5%. Since all variables and samples sizes are quite large it is safe to calculate the confidence intervals from normal approximation of the binomial distribution (using Eq. 2 from [34]). In this particular instance, the excess risk is $12.5 \pm 0.3\%$ and it is the excess risk that the correct answer identified by the students is a false one and not a true one. The odds ratio (see [33] for odds ratio general considerations and algorithms for exact calculation and see [35] for normal approximations) gives almost a doubled chance to correctly identify a false answer than a true one: $1.70 \pm 0.02$.

### C. Students progress

Numerous other descriptive and inferential statistics can be extracted from the database, such as topics with difficulty in understanding (as the ones exemplified in §IV-B). Students' progress trend (first, second, third evaluation and more), is another relevant (for education purposes) statistic. Some of the students, despite numerous trials, do not achieve a significant progress from one evaluation to another. However, important is what is obtained in average, and as a trend. In order to obtain a relevant statistic, a procedure has been deployed here. If a student made a good evaluation and has stopped there, then that evaluation can be considered further as its next and as its last evaluation. By doing this, all records, of students having one, two, three (and so on) evaluations have same count (Tab. IV). Statistics from this series can be used to reveal students progress between evaluations as well

as to reveal trends, if any exist. One should note that the

TABLE IV
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

information listed in Tab. IV has been made from consecutive evaluations but with disregard of the timeframe between them. A different outcome is retrieved when the date and time moment of the evaluation is considered. Upon checking the information from Tab. IV, a Student t test reveals that, statistically speaking, there is no difference (no progress) from making more than one evaluation (for instance, the probability that the 9th evaluation provides more points than the 1st one gets a probability by random chance of 70.42%, and in order to be considered significant, it was supposed to be no greater than 5%). At the same time, one may observe a clear increased tendency in the *Avg* data, and a clear decreasing tendency in *Std* data. A linear regression may be considered significant (the slope of the Avg as function of Eval is 0.748 and has a probability of not being null of 4.3%) but it can be guessed that learning is a limitative nonlinear curve - one will learn and learn, but finally the complete knowledge is a limit target. Indeed, one may find a Dose-Response distinct statistically significant model in the *Avg* data, $Avg(Eval) = 28.58_{\pm 0.29} + \frac{13.13_{\pm 0.80}}{1+(Eval/1.276_{\pm 0.057})^{-3.43_{\pm 0.17}}}$, which indicates that the greatest gain is in between the first and the second evaluation (1.276 coefficient in the model). The 28.58 value of the free coefficient suggests that 28 points may be gotten with 0 evaluations, so it should not be considered a passing score. One may find an exponential model significant as well ($Avg(Eval) = 41.66_{\pm 0.04} + 36.20_{\pm 0.50} \cdot \exp(-Eval/0.724_{\pm 0.021})$). Here, the free coefficient (41.66) has a different interpretation, it is the average score gained after an infinite number of evaluations. Also, this score can be assigned to a mark (for instance 42 points to a 7 for an evaluation from 4 to 10, or to a 6 for an evaluation from 1 to 10).

## V. CONCLUSIONS AND PERSPECTIVES

Students' performances from the use of the implemented system allows extracting valuable educational information. Improvement of the evaluation system and of the course will follow after implementing the lessons learned from the use of it.

## APPENDIX

### A. Abbreviations

- MCMA: multiple choice multiple answers
- HTTP: hyper text 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)
- CSCI: Computational Science and Computational Intelligence (international conference on)

### B. 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
- $p^3$-$p^6$ block of elements (groups 15-18)
- $d^1$-$d^5$ block of elements (groups 3-7)
- $d^6$-$d^{10}$ block of elements (groups 8-12)
- $f^1$-$f^{14}$ elements block (lanthanides and actinides)
- Boron group; carbon group
- Organic chemistry; hardness and hard materials
- Ceramics; semiconductors; superconducting
- Advanced materials; polymers & plastics; biomolecules & reaction mechanisms
- Methods & models; structure activity / property relationships

### C. Example of generating remote-based evaluation files

1 By the first ionization potential, the chemical elements can be ordered as follows:
   A. K < Na < Li < H
   B. Rn < Xe < Kr < Ar < Ne < He
   C. He < Ne < Ar < Kr < Xe < Rn
   (A and B are correct, C is wrong - the order is opposite)
2 In connection with rare gases (He, Ne, Ar, Kr, Xe, Rn):
   A. Xe behave similar to $H_2$
   B. Xe(g) + $PtF_6$(g) $\longrightarrow$ Xe[$PtF_6$](s)
   C. $O_2$(g) + $PtH_6$(g) $\longrightarrow$ Xe[$PtH_6$](s)
   (B is correct, A and C are wrong)
3 In connection with halogens (X: F, Cl, Br, I, At):
   A. F is the most electropositive element
   B. $X_2$ + $X'_2$ $\longrightarrow$ $XX'_7$ (X=I, X'=F)
   C. $X_2$ + $X'_2$ $\longrightarrow$ $XX'_5$ (X=Br, X'=F)
   (B and C are correct, A is wrong)
4 ... (450 questions in a file; many generated files)

## REFERENCES

[1] J. J. Ferreira, L. Maguta, A. B. Chissaca, I. F. Jussa, and S. S. Abudo, "Cohort study to evaluate the assimilation and retention of knowledge after theoretical test in undergraduate health science," *Porto Biomed. J.*, vol. 1, no. 5, pp. 181–185, 2016. [Online]. Available: http://sciencedirect.com/science/article/pii/S2444866416300599

[2] J.-N. Lee and B. Choi, "Determinants of knowledge management assimilation: An empirical investigation," *IEEE Trans. Eng. Manag.*, vol. 57, no. 3, pp. 430–449, 2010. [Online]. Available: http://ieeexplore.ieee.org/document/5357438

[3] K. Kulasegaram and P. K. Rangachari, "Beyond "formative": assessments to enrich student learning," *Advances in Physiology Education*, vol. 42, no. 1, pp. 5–14, 2018. [Online]. Available: http://doi.org/10.1152/advan.00122.2017

[4] Y. Zhen, S. Shi, W. Wang, and G. Wang, *Construction of Teacher-Student Interaction Evaluation Index System for High School Mathematics Concept Assimilation Learning Based on Artificial Intelligence*. Singapore: Springer Singapore, 2021, pp. 125–148. [Online]. Available: http://doi.org/10.1007/978-981-16-6502-8_13

[5] A. Bizy, C. Calvo, L. Such-Miquel, M. Bernabé Villodre, and M. Zarzoso-Muñoz, "The utility of kahoot! to evaluate conceptual knowledge assimilation in the subject "cell biology" in the degree of dentistry," in *INTED2023 Proceedings*, ser. 17th International Technology, Education and Development Conference. IATED, 6-8 March, 2023 2023, pp. 7110–7113. [Online]. Available: https://doi.org/10.21125/inted.2023.1940

[6] H. I. Naşcu and L. Jäntschi, "Multiple choice examination system 1. database design and implementation for general chemistry," *Leonardo J. Sci.*, vol. 3, no. 5, pp. 18–33, 2004. [Online]. Available: http://ljs.academicdirect.org/A05/18_33.pdf

[7] ——, "Multiple choice examination system 2. online quizzes for general chemistry," *Leonardo Electron. J. Pract. Technol.*, vol. 3, no. 5, pp. 26–36, 2004. [Online]. Available: http://lejpt.academicdirect.org/A05/26_36.pdf

[8] D. Nicol, "E-assessment by design: using multiple-choice tests to good effect," *J. Furth. High. Educ.*, vol. 31, no. 1, pp. 53–64, 2007. [Online]. Available: http://tandfonline.com/doi/10.1080/03098770601167922

[9] L. Jäntschi, "Binomial distributed data confidence interval calculation: Formulas, algorithms and examples," *Symmetry*, vol. 14, no. 6, p. 1104, May 2022. [Online]. Available: http://www.mdpi.com/2073-8994/14/6/1104

[10] P. Holmes, "Multiple evaluation versus multiple choice as testing paradigm - feasibility, reliability and validity in practice," PhD Thesis, University of Twente, Enschede, Netherlands, June 2002, http://doc.utwente.nl/38691/1/t0000017.pdf.

[11] L. Jäntschi, C. E. Stoenoiu, and S. D. Bolboacă, "Linking assessment to e-learning in microbiology and toxicology for undergraduate students," in *EUROCON 2007 - The International Conference on "Computer as a Tool"*, 2007, pp. 2447–2452. [Online]. Available: http://ieeexplore.ieee.org/document/4400369

[12] J. R. Loftis, "Beyond information recall: Sophisticated multiple-choice questions in philosophy," *Am. Assoc. Philos. Teach. Stud. Pedag.*, vol. 5, pp. 89–122, 2019. [Online]. Available: http://pdcnet.org/aaptstudies/content/aaptstudies_2019_0005_0089_0122

[13] R. F. Burton, "Quantifying the effects of chance in multiple choice and true/false tests: Question selection and guessing of answers," *Assess. Eval. High. Educ.*, vol. 26, no. 1, pp. 41–50, 2001. [Online]. Available: http://tandfonline.com/doi/10.1080/02602930020022273

[14] Y. Labrak, A. Bazoge, R. Dufour, B. Daille, P.-A. Gourraud, E. Morin, and M. Rouvier, "FrenchMedMCQA: A French multiple-mhoice 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: Association for Computational Linguistics, Dec. 2022, pp. 41–46. [Online]. Available: http://aclanthology.org/2022.louhi-1.5

[15] J. C. Chang and K. Akahori, "An evaluation of japanese call systems on the www comparing a freely input approach with multiple selection," *Comput. Assist. Lang. Learn.*, vol. 12, no. 1, pp. 59–79, 1999. [Online]. Available: http://tandfonline.com/doi/10.1076/call.12.1.59.5717

[16] A. Bayazit and P. Aşkar, "Performance and duration differences between online and paper–pencil tests," *Asia Pacific Educ. Rev.*, vol. 13, no. 2, pp. 219–226, Jun 2012. [Online]. Available: http://link.springer.com/article/10.1007/s12564-011-9190-9

[17] G. J. Anglin, H. Vaez, and K. L. Cunningham, *Visual Representations and Learning: The Role of Static and Animated Graphics.*, ser. Handbook of research on educational communications and technology, 2nd ed. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 2004, pp. 865–916.

[18] R. B. Dannenberg and N. Hu, "Pattern discovery techniques for music audio," *J. New Music Res.*, vol. 32, no. 2, pp. 153–163, 2003. [Online]. Available: http://tandfonline.com/doi/abs/10.1076/jnmr.32.2.153.16738

[19] S. M. Matteson, "Mathematical literacy and standardized mathematical assessments," *Read. Psychol.*, vol. 27, no. 2-3, pp. 205–233, 2006. [Online]. Available: http://doi.org/10.1080/02702710600642491

[20] D. Hestenes, "Toward a modeling theory of physics instruction," *Am. J. Phys.*, vol. 55, no. 5, pp. 440–454, 05 1987. [Online]. Available: http://doi.org/10.1119/1.15129

[21] L. Jäntschi, *General Chemistry*, 3rd ed. Cluj-Napoca, Romania: AcademicDirect, 2013. [Online]. Available: http://ph.academicdirect.org/GCC_v3.pdf

[22] M. Limón, "On the cognitive conflict as an instructional strategy for conceptual change: a critical appraisal," *Learn. Instr.*, vol. 11, no. 4, pp. 357–380, 2001. [Online]. Available: http://sciencedirect.com/science/article/pii/S0959475200000372

[23] C. G. Lord, L. Ross, and M. R. Lepper, "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence," *J. Pers. Soc. Psychol.*, vol. 37, no. 11, pp. 2098–2109, 1979. [Online]. Available: http://doi.org/10.1037/0022-3514.37.11.2098

[24] L. Jäntschi and S. D. Bolboacă, "Auto-calibrated online evaluation: database design and implementation," *Leonardo Electron. J. Pract. Technol.*, vol. 5, no. 9, pp. 179–192, 2006. [Online]. Available: http://lejpt.academicdirect.org/A09/179_192.htm

[25] A. Omari, "An evaluation and assessment system for online mcq's exams," *Int. J. Electron. Electr. Eng.*, vol. 1, no. 3, pp. 219–222, 2013. [Online]. Available: http://doi.org/10.12720/ijeee.1.3.219-222

[26] P. Jiang, K. Yan, H. Chen, and H. Sun, "Building of online evaluation system based on socket protocol," *Comput. Sci. Inf. Syst.*, vol. 19, no. 1, pp. 185–204, 2022. [Online]. Available: http://doi.org/10.2298/CSIS210201047J

[27] S. Tasdemır, M. Balcı, A. Cabi, M. Altin, and O. Cabi, "The design and application of online exam system supported by database," *Int. J. Appl. Math. Electron. Comput.*, vol. 3, no. 3, pp. 204–207, 2015.

[28] T. Nguyen, T. Bui, H. Fujita, T.-P. Hong, H. D. Loc, V. Snasel, and B. Vo, "Multiple-objective optimization applied in extracting multiple-choice tests," *Eng. Appl. Artif. Intell.*, vol. 105, p. 104439, 2021. [Online]. Available: http://sciencedirect.com/science/article/pii/S0952197621002876

[29] S. Hashemi Hosseinabad, M. Safayani, and A. Mirzaei, "Multiple answers to a question: a new approach for visual question answering," *Vis. Comput.*, vol. 37, no. 1, pp. 119–131, Jan 2021. [Online]. Available: http://link.springer.com/article/10.1007/s00371-019-01786-4

[30] L. Jäntschi, *Lucrări de laborator de chimie generală: ghid practic*, 1st ed. Cluj-Napoca, Romania: AcademicDirect, 2023. [Online]. Available: http://ph.academicdirect.org/llcggp.pdf

[31] C. Coman, L. G. Ţîru, L. Meseşan-Schmitz, C. Stanciu, and M. C. Bularca, "Online teaching and learning in higher education during the coronavirus pandemic: Students' perspective," *Sustainability*, vol. 12, no. 24, p. 10367, Dec 2020. [Online]. Available: http://dx.doi.org/10.3390/su122410367

[32] European Parliament and European Council, "Regulation 679," http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02016R0679-20160504, April 2016, accessed: 2023-09-28.

[33] L. Jäntschi, "Formulas, algorithms and examples for binomial distributed data confidence interval calculation: Excess risk, relative risk and odds ratio," *Mathematics*, vol. 9, no. 19, p. 2506, Oct 2021. [Online]. Available: http://dx.doi.org/10.3390/math9192506

[34] S. D. Bolboacă and A. B. Achimaş Cadariu, "Binomial distribution sample confidence intervals estimation. 6. excess risk," *Leonardo El. J. Pract. Technol.*, vol. 3, no. 4, pp. 1–21, Jun 2004. [Online]. Available: http://lejpt.academicdirect.org/A04/01_21.pdf

[35] ——, "Binomial distribution sample confidence intervals estimation. 5. odds ratio," *Leonardo J. Sci.*, vol. 3, no. 4, pp. 26–43, Jun 2004. [Online]. Available: http://ljs.academicdirect.org/A04/01_21.pdf