Increasing the Efficiency of Teacher’s Work: The Case of Undergraduate Mathematics Mid-Term Assessment

Lenka Viskotová, David Hampel

Department of Statistics and Operation Analysis, Faculty of Business and Economics,
Mendel University in Brno, Zemědělská 1, 613 00, Brno, Czech Republic
lenka.viskotova@mendelu.cz, david.hampel.uso@mendelu.cz

Abstract: *Computer-aided assessment is an important tool that reduces the workload of teachers and increases the efficiency of their work. The multiple-choice test is considered to be one of the most common forms of computer-aided testing and its application for mid-term has indisputable advantages. For the purposes of a high-quality and responsible assessment process, it is necessary to provide a sufficiently extensive databank of test items, especially when a large number of students is involved in the examination. In this paper we deal with the issues of automatic generation of such test items for undergraduate mathematics mid-term assessment. We describe the techniques and further circumstances related to designing test items, including the incorrect answers offered. The text also includes a case study dealing in detail with the creation of test items for a particular type of mathematical problem. Finally, efficiency evaluation of automatic item generation in comparison with paper-based questions is presented. Although the results of the article are based on experience influenced by the local conditions of the institution concerned, the ideas suggested remain generally applicable.*

INTRODUCTION

The expansion of computer technology into an ordinary university facility has brought with it new possibilities in the assessment process. An appropriate information system and a large capacity computer lab at one’s disposal allows the modification of traditional assessment and evaluation approaches and significantly improves the efficiency of teachers’ work. In particular, educational institutions organising courses for hundreds of students make use of the benefits of technology very often, no matter what fields of study they provide, including undergraduate mathematics. It is true that mathematics and applied mathematics have their specifics and this often lead in extensive discussions on computer-based assessing students’ knowledge (see, e.g., Bennet et al., 2008; Clariana and Wallace, 2002; Oates, 2011). Various research papers about the advantages and disadvantages of these methods in mathematics have been written (Croft et al., 2001; Rasila, Malinen and Tiitu, 2015; Rønning, 2017). Computer assisted assessment has also been subject to gender analysis for a long time (Akst and Hirsh, 1991; Bennet et al., 2008; Clariana and Wallace, 2002; Goodwin, Ostrom and Scott, 2009). Lee (2011) is of the opinion that today's teachers must be creative and combine the application of information technology with innovative teaching methods and strategies. It is interesting, but also expected, that the use of technology is much more discussed in tertiary mathematics education than in lower mathematics education. Even the research of Foster and Inglis (2018), which analyses two leading UK mathematics teacher professional journals targeting mainly at secondary education, concluded that there is "a smaller decline in discussions relating to […] technology".

The issue for computer aided assessment (CAA), or also e-assessment, follows the development of technology, to which available forms of the assessment process correspond. Journal papers commonly focus on two formats, multiple-choice (MC) and constructed response (CR). The MC format has a lot of critics, but its advantages are indisputable. Its application does not require students to enter their mathematics in fill-in-the-blank items. Sangwin (2013) admits, that "syntax presents the most significant barrier to students' successful use of computer aided assessment, particularly when the stakes are high". There is also great flexibility in use of MC independent of the learning management system. The strong position of the MC format is evidenced by a long-term research (Goodwin, Ostrom and Scott, 2009; Kosh, 2019; Mitkov, Ha and Karamanis, 2006; Sims Wiliams and Barry, 1999) and the fields of psychometrics associated to it, especially so-called item response theory (IRT) (Andrich and Styles, 2011; Embretson and Kingston, 2018; Hoppe, 2016). On the other hand, the usage of the CR format belongs to an important topic of CAA research. The CR items removes guessing, prevents the "reversible mathematical processes" resulting in back substitution, see Sangwin and Jones (2017), and provides detailed feedback to teachers and students. The applications of IRT to the CR items can be found e.g. in Holling, Bertling and Zeuch (2009) or Maxwell and Gleason (2019). Leaving aside the problems of CR format with entering a mathematical expression by students into a machine, there are requirements for the university learning management system to interact with the computer algebra system used to design the test items.

However, this paper does not aim to analyse the pros and cons of MC and CR but focuses rather on the issue of generating test items, more precisely on automatic item generation (AIG). Kosh (2019) presents AIG as an item development process that can be used to supplement item writing efforts and defines it as a three-stage process. In the first stage the cognitive model is developed, i.e. the mathematical knowledge and skills needed to solving items is characterized. The second stage is developing item model encompassing specifications detailed enough to produce items in the last stage through algorithmic means. Kosh (2019) emphasizes that a crucial stage is "developing item models. Despite item model creation serving as a critical methodological component of AIG, to my knowledge there are no published methods that describe the principles or standards used to create AIG item models. Instead, researchers merely present the cognitive model and item models they used without describing how those item models were created, …".

Kosh (2019) distinguishes five steps in the item model creation process, not necessarily in a linear arrangement, as follows: 1. Identification of the schema representing the desired content of the items, 2. Identification of features drawing on existing theory and impacting cognitive complexity of the schema (i.e., structural complexity, contextual support, extent of generalization), 3. Determining item type (MC, CR, etc.), 4. Specification of so-called item meta-model (a model for the item model), 5. Writing item models. The item meta-model in the fourth step characterizes the structure of the item models and their variability aligned to the given cognitive complexity. It defines components that vary across item models, including a list of typical student mistakes and associated distractors. The variability of item models is given by already predefined combinations of features that maintains the same sufficient level of all test items.

The matter of AIG must be viewed from two aspects, not only quality but also quantity. Quantity is an increasingly important requirement in order to create a large variety of tests (see Sims Williams and Barry, 1999) because of the large numbers of students. A substantial item bank prevents unacceptable student collaboration during a test as well as making it harder to illegally share test items already used. The continual supply of new test items also allows to generate an individual test for each student.

Incorporation of AIG into pedagogical practice has positive impact on a teachers' work efficiency. Efficiency is understood here in the classical sense of meaning as a measure of the amount of resources required in order to meet the goals. De Witte and López-Torres (2017) provide an extensive review of the literature on efficiency in education including the 'economics of education' literature. The review deals with different levels of analysis as university level, school/high school level, district level, national level, student level and classroom level, but there are no research papers on the efficiency of the teacher's work, even though the integration of technologies into the educational process is directly offered for such analyses. In particular, it is assumed that an effective way of assessing students' knowledge saves resources such as time and manpower.

In this context, it is worth mentioning that the issue of efficiency needs to be distinguished from the topic of teacher effectiveness and effective teaching. Teacher effectiveness is defined, in the narrowest sense, as a teacher's ability to produce gains in student achievement scores. It is pertinent to note here that such effectiveness in mathematics education at the undergraduate level is closely related to the effectiveness of secondary schools. In order to achieve sufficient effectiveness, universities are looking for various ways to fill the gaps in students' mathematical knowledge from their previous education, see e.g. Hampel and Viskotová (2021) or Dagan, Satianov and Teicher (2019).

A practical guide to evaluating effectiveness can be found in Little, Goe and Bell (2009), other studies as Gurney (2007) and Yue (2019) describe factors and methods for effective teaching. Burden and Byrd (2019) in their work deal with effective teaching in all its complexity. Harris and Sass (2014) involve principals' evaluation of teachers' cognitive and non-cognitive skills in their analyses and touch on the issue of cost-effective methods of measuring teacher productivity. Research on measurements of teacher productivity can be find also in Sass, Semykina and Harris (2014). Ilkovičová, Ilkovič and Špaček (2017) look at teaching efficiency from three different perspectives, one of which is viewed from a teacher and reflects on the rationality of the pedagogical time given. Levin (1997) approaches the issue of raising school productivity using x-efficiency, see Leibenstein (1966).

Another notion that often appears in the literature is the so-called teachers' self-efficacy as a socio-affective concept, different from the concept of teachers' efficiency. Nevertheless, the connection can be traced, the influence of self-efficacy on teachers' performance is obvious. Alibakhshi, Nikdel and Labbafi (2020) describes all the consequences of self-efficacy and, among other things, point to the impact of self-efficacy on the using computers and technology.

The aim of this paper is to present the case of increasing efficiency of teachers' work in the process of assessment, specifically the methodology of computer-aided generation of MC test items with one keyed answer for the purpose of mid-term assessment in undergraduate mathematics. In the following, we propose selection of mathematics topics suitable for computer processing, suggest useful techniques for creating source text files containing test items, discuss related problems that may occur and briefly evaluate teachers' work efficiency of the proposed solution.

MATERIAL AND METHODS

In this section, we present the entry conditions given by the content and enrolment characteristics of the courses where automated test generation is employed, as well as the conditions given by the university information system environment. The courses of undergraduate mathematics at the Faculty of Business and Economics of Mendel University in Brno cover typical topics, that are also found in classical textbooks such as Sydsæter et al. (2016). Namely, these are the issues of linear algebra (matrices and matrix operations, determinant, the inverse of a matrix, matrix equations, systems of linear equations), differential calculus of function of one or more variables (definition of a function and its properties, limits and continuity, derivatives and their application) and integral calculus of one variable (indefinite, definite and improper integral). These classes have the important characteristic that a large number of students (up to 1,000) are enrolled. Thus computer-based assessment has become the standard means of mid-term assessment and has a tradition of more than fifteen years. For this purpose, a computer lab with a capacity of 80 workstations is used which enables handling of the assessment quite quickly, in two days.

The university information system provides the application, which allows the teacher to create a unique test for each student registered for the exam date. The properties of a test, such as its duration, the number of test items, the number of points assigned for each item, and the link to the repository of items, all depend on application settings given by the examiner. To deter students from guessing, it is recommended that incorrect answers be penalised with negative points. Test items have to be imported in advance from text files of a given structure into the test item bank of the information system.

The importable form of items saved in a text file follows the syntax of the LaTeX typesetting system. Each item consists of several text lines; the first line corresponds to the stem, the second line starts with the plus symbol and the keyed answer follows, the other lines (their quantity is in line with the number of responses being offered) contain the minus symbol and the distractor. Each test item is separated by an empty line. As an example, the problem of calculation of a determinant can be typed as mentioned in Figure 1.

LaTeX form of the item in the source text file	Final form of the item in the test

`Calculate the determinant of the ...`
`   ... matrix $\left(\begin{array} ...`
`   ... {rrr}5&1&0\\7&-1&-2\\1&0&-1 ...`
`   ... \end{array}\right)$.`
`+$10$`
`-$5$`
`-$11$`
`-$-4$`
`-$-1$`
	Calculate the determinant of the matrix
	$\left(\begin{array}{rrr}5&1&0\\7&-1&-2\\1&0&-1\end{array}\right).$
	○ 10 ○ 11
	○ 5 ○ −4 ○ −1

Figure 1: Text format of the test item ready for import and its preview in a test

General approaches as parametrization of the problem are used for development of the resulting automated test item generation methodology. The computational system MATLAB R2020b with Symbolic Math Toolbox was chosen as the tool for generation of the test items.

RESULTS

Within this section, we propose classification of items with respect to possible parametrization approaches and appropriate techniques for generation of stems and distractors. The case study with complete description of the generating procedure is elaborated. Finally, the differences between the efficiency of automatic generation and the paper-based creation of particular item types are presented.

Those issues of undergraduate mathematics are being processed whose solutions are of “a technical nature”, when the complexity of the structure of the meta-model does not outweigh the benefits of AIG, and therefore these issues are suitable for midterm technology-assisted assessment. Computer-aided generation has been used for the following tasks: the product of matrices, the determinant of a square matrix, the eigenvalues of a square matrix, the inverse of a square matrix, solvability of the system of linear equations, the derivative of a function of one variable (product rule, quotient rule, chain rule, derivative at a given point), the indefinite integral of rational functions.

Computer-aided generation of stems

Computer-aided generation of a test item for a given area is based on the idea of parametrization of an example test item, where parameters are being considered instead of numerical values in the stem. It is possible to distinguish two techniques for assigning values to parameters:

(1) The technique based on the pseudo-random generation principle is undoubtedly applicable. Pseudo-random generation means that randomly generated values in the test item must

meet pre-determined conditions, otherwise the non-compliant values are dropped and generated again.

(2) The parameters are gradually assigned all values from the given numeric set, usually all the integers from an interval. It is convenient to check whether the new values of the test item correspond to the requirements for the difficulty of calculation and the final form of the result.

The first approach finds wide application in linear algebra, the second approach is more likely to be applied in calculus to test items concerning the calculation of derivatives and integrals.

Random value generation can generally pose the problem of repeated test items. A way to prevent this is to store already generated values in an auxiliary matrix, where each row of the matrix represents the numeric values of one generated item. If the sequence of new values differs from all rows of the matrix, the corresponding item is included in the test bank and the auxiliary matrix is updated. Otherwise, random generation must be performed again.

Determining the appropriate conditions for generating a new test stem is essential to achieve adequate test difficulty. To illustrate the situation, consider a 3×3 determinant calculation. The conditions might be such that the matrix contains a given number of zeros and the absolute value of the determinant does not exceed the specific number.

The thought-out change of conditions in the generating algorithm creates a new group of test items within one issue. For example, a well-chosen variety of conditions of the eigenvalue problem results in separate groups of test items with different number of unique eigenvalues. Variability of conditions may cause different difficulty level of test items, which can be solved by changing the range of the generating interval from case to case.

Fulfilment of given conditions on issues with many numerical values can cause very long program calculation times. Then, it is advantageous to generate only a part of the parameter values of the test item and verify that they meet the conditions, and in a second step use these values to derive by means of a random variable the values of the remaining parameters. For example, the generation of test items for linear equation systems, depending on the required degree of freedom and number of equations and unknowns, can be accelerated as follows. First, the coefficients of as many equations as the expected rank of the coefficient matrix are generated, and secondly, the remaining equations are given by random linear combinations of previously obtained equations. Finally, to achieve a smarter appearance of the system it is more than convenient to test the linear dependence of all possible pairs of equations and accept the system only in the case of independence. To enlarge the variability of test items it is sometimes feasible to generate coefficients so that the corresponding submatrix formed of the first r columns of the coefficient matrix has the rank lower than r, provided that r is the rank of the coefficient matrix.

There are mathematical problems where it is more convenient to generate test items based on the desired result and derive the stem from the correct answer. Suppose that the goal is to create the test item on integration of a proper rational function involving a quadratic in the denominator. Due to the application of appropriate anti-derivative formula, calculation of square roots is often necessary, which makes some of the results too complicated. Therefore, it is preferable to determine the expected result first and then to calculate the stem.

Computer-aided generation of distractors

The issue of computer-aided generation of distractors is primarily solved using algorithms simulating common student mistakes. Similar to the situation where stems are generated, it is necessary to verify the difference of the new distractor from the keyed answer and from already generated incorrect answers. In the case of a match, the distractor is generated randomly and is verified again. As an example of a good distractor on 3 × 3 determinant calculation seems to be the value which is obtained by exchanging the signs of the part of associated minors.

It sometimes occurs that the aspects of creating proper incorrect answers are more extensive in some areas of mathematics. Distractors on calculation derivatives of functions requires more than just changing constants in the result, it needs in addition a change in the function type in the distractor. To prevent students from guessing the answers, a clever composition of offered answers can be used. The idea is that each test item randomly varies the number of wrong answers differing from the keyed answer only in the constants, and the number of incorrect answers differing fundamentally, for example by a different number of roots in an eigenvalues calculation or a different type of function in an antiderivative calculation.

Another idea to discourage guessing can be presented in the case of calculating the elements of the 3-by-3 inverse matrix. To avoid situations where students guess the correct answer from the first few elements of the matrix, it is advisable to list only 3 selected elements in the offered answers, preferably so that each element lies on a different row and column. The choice of listed elements may be influenced by random index generation, which increases the diversity of the answers being offered.

Strings in the source text files

As mentioned in the Entry conditions section, the source text file for importing test items into the test bank must follow the structure given by the relevant information system administering the tests. This requires construction of variables in the format of a string array, which serves to write each test item in the text file. The circumstances of creating these strings depend on the chosen software, but the principle always lies in concatenating shorter strings. These arise as results (converged into LaTeX format) of the generation and calculation given by software, or represent fixed parts of the test item, such as texts or some mathematical symbols. This manipulation of text strings brings a number of difficulties, for example the problem of the presence of constants 1, -1 or 0 at the beginning of the equation, or the format of generated fractions 3/1, 0/3 etc.

Case study: inverse of 3 × 3 matrix

Let’s deal with the issue of an inverse matrix of order 3

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

The matrix elements are randomly generated integers from the interval $\langle -5, 5\rangle$. The order 3 guarantees calculation by students in a reasonably short time. It is necessary to bear in mind that students may choose from the two most common methods, i.e. the method based on using elementary row operations and the method of calculating the determinant and the adjoint matrix. Considering aspects of these techniques, three types of test items have been generated:

(1) The conditions for generating the items of the first type have been set as follows:

- the number of zero elements is a value from the set $\{1, 2\}$;
- $|a_{11}| = 1$;
- $|\det \mathbf{A}| \in \{2, 3\}$;
- $\left|a_{22} - \frac{a_{12}a_{21}}{a_{11}}\right| = 1$, which ensures that after the first step of Gauss-Jordan elimination the absolute value of the second element of the main diagonal equals 1;
- $\operatorname{mod}\left[a_{13} - \left(a_{23} - \frac{a_{13}a_{21}}{a_{11}}\right)\frac{a_{11}a_{12}}{a_{11}a_{22}-a_{12}a_{21}},\ a_{33} - \frac{a_{13}a_{31}}{a_{11}} - \left(a_{23} - \frac{a_{13}a_{21}}{a_{11}}\right)\frac{a_{11}a_{32}-a_{12}a_{31}}{a_{11}a_{22}-a_{12}a_{21}}\right] = 0$ and $\operatorname{mod}\left[a_{23} - \frac{a_{13}a_{21}}{a_{11}},\ a_{33} - \frac{a_{13}a_{31}}{a_{11}} - \left(a_{23} - \frac{a_{13}a_{21}}{a_{11}}\right)\frac{a_{11}a_{32}-a_{12}a_{31}}{a_{11}a_{22}-a_{12}a_{21}}\right] = 0$, i.e. once row operations give zeros in all non-diagonal elements of the first two columns, the first and the second element of the third column are divisible by the element in the third row of this column.

(2) The second type of the inversion matrix items satisfies these demands:

- the number of zero elements is a value from the set $\{1, 2\}$;
- $|a_{11}| = 1$;
- $\det \mathbf{A} = -1$;
- $\left|a_{22} - \frac{a_{12}a_{21}}{a_{11}}\right| = 0$, which ensures that after the first step of Gauss-Jordan elimination the second element of the main diagonal equals 0;
- $\left|a_{32} - \frac{a_{12}a_{21}}{a_{11}}\right| = 1$, which ensures that after the first step of Gauss-Jordan elimination the absolute value of the last element of the second column equals 1.

(3) The third type of test items corresponds to singular matrices, i.e. $\det \mathbf{A} = 0$. The matrix is generated with exactly one zero element with arbitrary position and its singularity is verified by calculating the determinant of the matrix.

The format of the answers (the list of the three selected elements of the resulting inverse matrix, where each element lies on a different row and column) has been already mentioned above. In the case of regular matrices, the construction of distractors is based on typical student mistakes. The first distractor arises from the omission of the determinant in calculation using an adjoint matrix. The initial conditions mentioned above consider $\det \mathbf{A} \neq 1$, so the keyed answer is always different from this distractor for non-zero elements. The second distractor corresponds to the situation when transposing is forgotten when calculating an adjoint matrix. The third distractor has been realized by replacing one of a_{31}, a_{32}, a_{33} by a random generated value. The fourth incorrect answer is the sentence "The inverse of the matrix does not exist.". Of course, the need for a difference of the answers offered must be verified. As regards test items with singular matrices, distractors are generated randomly. Due to the different selection of elements for each answer, there is no problem with equality of the distractors in the case of singularity.

Efficiency evaluation

Evaluating the effectiveness of AIG depends, of course, on the teachers' ability to algorithmize the problem and master programming techniques and the programming language. The following analysis assumes advanced ICT skills, reported times include the implementation of the correct conversion to LaTeX format with all its difficulties. It should be emphasized that the figures given are average estimates.

Table 1 shows that test item creation using AIG requires different times according to the type of problem. It took approximately 9 hours to prepare and check a script for computer-aided generation of test items corresponding to one type of a problem within the topic of linear algebra and indefinite integral of rational functions. In the case of differentiation, covering the whole range of problems required an adequate amount of scripts (typically related to particular function form) that could not be directly included in a single script. A total of 20 shorter scripts were needed to process one type of question to apply all the elementary functions and the related phenomena. Calculated in time, it took 60 hours to create these scripts for one type of differentiation test item with the appropriate variety.

Topic	The scripts for AIG using MATLAB	Paper-based creation of 20 test items	Paper-based creation of 100 test items
Linear algebra	9 h	11 h 40 min	58 h 20 min
Differentiation	60 h	13 h 20 min	66 h 40 min
Indefinite integral of rational functions	9 h	10 h 0 min	50 h 0 min

Table 1: Estimated number of working hours spent for one type of test item creation.

Considering the experiences with the paper-based test item creation, the linear algebra test item is created and inserted into the information system in about 30 minutes, the test item considering differentiation in about 35 minutes and the test item concerning the indefinite integral of rational functions in about 25 minutes. In all mentioned paper-based cases, another 5 minutes are devoted to the final inspection. In Table 1 we can find working time employed to create 20 paper-based test items as minima per 1 semester for one type of the question.

Regarding linear algebra and indefinite integral, it can be argued that the time savings due to AIG are already apparent during the first semester. As for differentiation, AIG compared to paper-based work places great demands on the spent time and is not be paid until the fifth semester. However, it is necessary to keep in mind, that AIG allows to generate an extensive database of given types of questions and provide individual test assignments for each student. The times needed to accomplish such goal with paper-based test items are given in the last column of Table 1.

In the course of undergraduate mathematics AIG has been used to generate 5 types of test items from linear algebra, 4 types from differentiation and 3 types on indefinite integral. In total, the creation of scripts for generating large databank took 313 hours. Paper-based database creation of the same content for single semester, e.g. 20 items per type, would last 158 hours. It is evident that the time invested in AIG pays off after only two semesters. Likewise, it is easy to see that AIG reduces teacher workload in a particularly fundamental way when processing issues of linear algebra and indefinite integral of rational functions.

DISCUSSION

Our approach to the computer-aided generation of test items for selected mathematical problems seems to be successful and has enabled us to create a large test databank of required quality. Thanks to the selection of appropriate topics and the suggested formulations of tasks, the problem of reversibility of mathematical processes, described e.g. in Sangwin and Jones (2017), was suppressed. It can be stated that back substitution is only possible for the items concerning the calculation of indefinite integrals of rational functions. But even here, the back substitution is very complicated, students have to differentiate up to five times including tough simplification of expressions, which overweighs the difficulty of the direct finding an antiderivative. Back substitution strategy problem is being investigated from different points of view; for example, Goodwin, Ostrom and Scott (2009) focused their attention on gender differences in the tendency to use the back substitution strategy in a multiple-choice mathematical test and found that this relationship was not significant.

The fact that the test items have been generated in the LaTeX format enables – beside primary use with the learning management system – creating pdf files with the test items in an easy way. This is important in cases of a technical failure of the information system or Internet outages. Herbert, Demskoi and Cullis (2019) also take advantages of the independence of the generated test from the learning management system. However, when importing, they do not work directly with the LaTeX format, but already use its conversion to PDF format.

The MATLAB computational system with the Symbolic Math Toolbox has proved to be suitable software for generating test items for mathematics assessment. However, considering availability of software and the various tools on offer, many of these can be used, e.g. Maple (see Herbert, Demskoi and Cullis, 2019) or Maple T.A. (see Jahodová Berková, 2017; Rønning, 2017). Of course, the ability of the software to calculate corresponding mathematical problems to get correct answers and distractors according to a given algorithm is assumed. A tool for generating uniformly distributed random numbers is necessary for setting numeric values in all parts of the test items, as well as for random generation of the type of distractor. Furthermore, it is necessary to handle text strings and write them to source text files. The ability of software to return the LaTeX form of symbolic math expression is certainly an advantage.

Computer-aided generation of test items has a huge benefit that once adequate program code is created in a given software, then a lot of items arise in a simple manner. We observe high efficiency of coding for many topics. Gierl and Lai (2016) arrived at similar results and confirmed the high efficiency of AIG, applying their research to test item generation not only in mathematics but also in health science. Moreover, it is realistic to obtain a sufficiently extensive test bank to guarantee, together with the personalized test items for each student, cheat-resistant assessment. Unfortunately, as Manoharan (2019) confirms, personalization approach does not mitigate contract-cheating or cheating using communication devices.

Responsibly designed technology-assisted testing finds its application during the semester as a time-saving tool that allows only promising students to take the final exam. This approach to applicability just for mid-term assessment is fully in accordance with Hoogland and Tout (2018). They point to the important fact that there are two opposing pressures when considering the assessment in mathematics education. On the one hand, emphasis needs to be placed on the higher-order thinking skills and knowledge, and on problem solving and reasoning. On the other hand, the trend of focusing on technology and efficiency results in risk of an assessment of lower-order goals, based on reproduction of calculation procedures.

The applicability of the presented concepts primarily focused on a summative assessment is also expanded to the form of formative assessment. The provision of dozens of items, randomly drawn from an extensive item bank, provides an efficient mechanism for students to practice skills they have learned from their classes. In their paper, Morphew et al. (2020) prove that more frequent testing schedule (for both MC and CR formats) in undergraduate engineering lead to better student outcomes. Similarly, in the case of formative type of assessment, self-testing improves students' learning. Moreover, students find online formative feedback enjoyable and useful (Acosta-Gonzaga and Walet, 2018). Formative assessment also contributes to their ability of self-assessment as a feedback mechanism for improving own learning strategies (Hosein and Harle, 2018). There may arise a discussion about whether or not to grant students the repository of test items in full. This will undoubtedly provide students with the maximum amount of right material to practice and, in addition, any student ambitions to copy or record test items will have no sense.

One of the limitations of our approach is the fact that there are still mathematical topics convenient for assessment in the form of a multiple-choice test, but their computer generation brings many troubles due to the non-uniform structure of test items. Therefore, such items have been made up "manually". Specifically, these were items related to areas like the function and its properties, the limit of a function, integration by parts, integration by substitution, the definite integral, first-order and second-order partial derivatives. Adji et al. (2018) made research on AIG within high school mathematics and found out that 45 % of mathematical questions for the local National Exam cannot be processed by AIG.

Automatic generation of test items, despite efforts to treat all possible situations, can give rise to problematic items where the solution does not match the required level. In addition to insufficient difficulty of test items, computer-aided generation can bring with it other undesirable effects that cannot always be predicted in advance. It is therefore advisable to check a random sample of test items with students whose feedback is significant. Gierl and Lai (2016) also emphasize the importance of external review. In their case, the review is performed by subject-matter experts who did not develop AIG models and is based on the so-called standardized rating rubric.

Once the generated test items are incorporated into the assessment, the information system provides a basic summary of the average scores relating to particular items. Examiners can view these and distinguish items with outlying scores and assess their inclusion in the test, potentially replacing them with newly created items. Future research could focus on evaluation with respect to item response theory. In mathematical education, the application of this methodology can be found, for example, in Bolondi, Branchetti and Giberti (2018).

It should be taken into account that the computer-based assessment place demands on ICT skills of teachers (see, e.g., Tondeur et al., 2019). TALIS (The OECD Teaching and Learning International Survey) repeatedly reports that teachers themselves are aware of this need. OECD (2020) points out that teachers cite ICT skills as the second most urgent area in further professional development. However, the situation about Covid-19 has an accelerating effect on the growth of teachers' and students' ICT abilities (fully online mathematics teaching is discussed in Trenholm and Peschke, 2020) and increases the importance of research in the field of effective technology integration not only into assessment, but also in education generally (see, e.g., Cardoso Espinosa, Cortés Ruiz and Cerecedo Mercado, 2021).

CONCLUSIONS

Our methodology for the computer-aided generation of test items has proved to be useful and has enabled us to create a large test databank containing questions on undergraduate mathematics of various topics and required difficulty. Further, the MATLAB computational system with the Symbolic Math Toolbox seems to be software entirely suitable for generating test items for mathematics assessment. Of course, it is necessary to use the system in a responsible way and to verify the quality and level of individual questions in real testing of students' knowledge.

Automatic generation of test items has brought a new insight to the mid-term assessment, when the huge variety of test items brings teachers a strong tool for the implementation of a fair automatic assessment process. As part of the institution's teaching innovations, the test bank has recently been expanded by over 1,500 test items, now using automatic test item generation. Through minor adjustments, the system of computer-aided generation of test items so introduced can be utilized repeatedly to construct more items already implemented of a given quality as well as new types of items.

Considering AIG in mathematics, there is an opportunity to expand the issue of AIG by test items accompanied by automatic generated graphs of functions. This issue does not occur in the literature, although its applicability is undoubtedly obvious. This identified literature gap may point out additional research areas.

REFERENCES

[1] Acosta-Gonzaga, E. and Walet, N. R. (2018) ‘The role of attitudinal factors in mathematical on-line assessments: a study of undergraduate STEM students’, *Assessment & Evaluation in Higher Education*, Vol. 43, No. 5, pp. 710–726. https://doi.org/10.1080/02602938.2017.1401976

[2] Adji, T. B., Pribadi, F. S., Prabowo, H. E., Rosnawati, R. and Wijaya A. (2018) ‘Generating Parallel Mathematic Items Using Automatic Item Generation’, *Proceeding Book of 1st International Conference on Educational Assessment and Policy (ICEAP 2018)*, Jakarta, Vol. 1, pp. 89–93. https://doi.org/10.26499/iceap.v1i1.78

[3] Akst, G. and Hirsch, L. (1991) ‘Selected Studies on Math Placement’. *Review of Research in Developmental Education*, Vol. 8, No. 4., pp. 3–6.

[4] Alibakhshi, G., Nikdel, F. and Labbafi, A. (2020) ‘Exploring the consequences of teachers’ self-efficacy: a case of teachers of English as a foreign language’, *Asian-Pacific Journal of Second and Foreign Language Education*, Vol. 5, No. 23, pp. 1–19. https://doi.org/10.1186/s40862-020-00102-1

[5] Andrich, D. and Styles, I. (2011) ‘Distractors with Information in Multiple Choice Item: A Rationale Based on the Rasch Model’, *Journal of Applied Measurement*, Vol. 12, No. 1, pp. 67–95.

[6] Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B. and Yan, F. (2008) ‘Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP’, *Journal of Technology Learning and Assessment*, Vol. 6, No. 9, pp. 4–38.

[7] Bolondi, G., Branchetti, L. and Giberti C. (2018) ‘A quantitative methodology for analyzing the impact of the formulation of a mathematical item on students learning assessment’, *Studies in Educational Evaluation*, Vol. 58, pp. 37–50. https://doi.org/10.1016/j.stueduc.2018.05.002

[8] Burden, P. R. and Byrd D. M. (2019) *Methods for Effective Teaching: Meeting the Needs of All Students*, 8th Edition, New York: Pearson.

[9] Cardoso Espinosa, E. O., Cortés Ruiz, J. A. and Cerecedo Mercado, M. T. (2021) ‘Computational Self-Efficacy and Information Skills in Postgraduate Students in Administration at the Times of Covid-19’, *Journal on Efficiency and Responsibility in Education and Science*, Vol. 14, No. 2, pp. 118–129. https://dx.doi.org/10.7160/eriesj.2021.140205

[10] Clariana, R. B. and Wallace, P. E. (2002) ‘Paper-based versus computer-based assessment: key factors associated with the test mode effect’, *British Journal of Educational Technology*, Vol. 33, No. 5, pp. 593–602. https://doi.org/10.1111/1467-8535.00294

[11] Croft, A. C., Danson, M., Dawson, B. R. and Ward, J. P. (2001) 'Experience of using computer assisted assessment in engineering mathematics', *Computers & Education*, Vol. 37, pp. 53–66.

[12] Dagan, M., Satianov, P. and Teicher, M. (2019) 'Teaching Calculus for Engineering Students Using Alternative Representations of Graph-formula Problems', *Mathematics Teaching Research Journal*, Vol 11, No. 3–4, pp. 12–41.

[13] De Witte, K. and López-Torres, L. (2017) 'Efficiency in education: a review of literature and a way forward', *Journal of the Operational Research Society*, Vol. 68, No. 4, pp. 339–363. https://doi.org/10.1057/jors.2015.92

[14] Embretson, S. E. and Kingston, N. M. (2018) 'Automatic Item Generation: A More Efficient Process for Developing Mathematics Achievement Items?', *Journal of Educational Measurement*, Vol. 55, No. 1, pp. 112–131. https://doi.org/10.1111/jedm.12166

[15] Foster, C. and Inglis, M. (2018) 'Mathematics Teacher Professional Journals: What Topics Appear and How Has This Changed over Time?', *International Journal of Science and Mathematics Education*, Vol. 17, pp. 1627–1648. https://doi.org/10.1007/s10763-018-9937-4

[16] Gierl, M. J. and Lai, H. (2016) 'A Process for Reviewing and Evaluating Generated Test Items', *Educational Measurement: Issues and Practice*, Vol. 35, No. 4, pp. 6–20. https://doi.org/10.1111/emip.12129

[17] Goodwin, K. S., Ostrom, L. and Scott, K. W. (2009) 'Gender Differences in Mathematics Self-Efficacy and Back Substitution in Multiple-Choice Assessment', *Journal of Adult Education*, Vol. 38, No. 1, pp. 22–42.

[18] Gurney, P. (2007) 'Five Factors for Effective Teaching', *New Zealand Journal of Teachers' Work*, Vol. 4, No. 2, pp. 89–98.

[19] Hampel, D. and Viskotová, L. (2021) 'The Effects of Preparatory Course on the Mathematics Exam Results: Case Study', *Proceedings of the 18th International Conference Efficiency and Responsibility in Education 2021*, Praha: Czech University of Live Sciences Prague, pp. 29–35. https://erie.pef.czu.cz/en/r-17410-proceedings-2021/proceedings-2021.html

[20] Harris, D. N. and Sass, T. R. (2014) 'Skills, productivity and the evaluation of teacher performance', *Economics of Education Review*, Vol. 40, pp. 183–204. https://doi.org/10.1016/j.econedurev.2014.03.002

[21] Herbert, K., Demskoi, D. and Cullis, K. (2019) 'Creating mathematics formative assessments using LaTeX, PDF forms and computer algebra', *Australasian Journal Educational Technology*, Vol. 35, No. 5, pp. 153–167. https://doi.org/10.14742/ajet.4539

[22] Holling, H., Bertling, J. P. and Zeuch, N. (2009) ‘Automatic item generation of probability word problems’, *Studies in Educational Evaluation*, Vol. 35, pp. 71–76. https://doi.org/10.1016/j.stueduc.2009.10.004

[22] Hoogland, K., Tout, D. 6 (2018) ‘Computer-based assessment of mathematics into the twenty-first century: pressures and tensions’, *ZDM Mathematics Education*, Vol. 50, pp. 675–686. https://doi.org/10.1007/s11858-018-0944-2

[23] Hoppe, F. M. (2016) ‘Benford’s law and distractors in multiple choice exams’, *International Journal of Mathematical Education in Science and Technology*. Vol. 47, No. 4, pp. 606–612. https://doi.org/10.1080/0020739X.2015.1091515

[24] Hosein, A. and Harle, J. (2018) ‘The relationship between students’ prior mathematical attainment, knowledge and confidence on their self-assessment accuracy’, *Studies in Educational Evaluation*, Vol. 56, pp. 32–41. https://doi.org/10.1016/j.stueduc.2017.10.008

[25] Ilkovičová, L., Ilkovič, J. and Špaček, R. (2017) ‘Ways of rationality and effectivity in architectural education’, *World Transactions on Engineering and Technology Education*, Vol. 15, No. 4, pp. 331–337.

[26] Jahodová Berková, A. (2017) ‘Effect of The Use of Computer-aided Assessment System in The Teaching of Mathematical Analysis with Regard to Students’ Approaches to Learning’, *Journal on Efficiency and Responsibility in Education and Science*, Vol. 10, No. 3, pp. 71–75. https://doi.org/10.7160/eriesj.2017.100302

[27] Kosh, A. E. (2019) ‘A Process for Automatically Generating Algebra Items’, *Journal of Applied Testing Technology*, Vol. 20, No. 1, pp. 16–33.

[28] Lee, Y. J. (2011) ‘A study on the effect of teaching innovation on learning effectiveness with learning satisfaction as a mediator’, *World Transactions on Engineering and Technology Education*, Vol. 9, No. 2, pp. 92–101.

[29] Leibenstein, H. (1966) ‘Allocative Efficiency vs. "X-Efficiency"’, *The American Economic Review,* Vol. 56, No. 3, pp. 392–415.

[30] Levin, H. M. (1997) ‘Raising school productivity: An x-efficiency approach’*, Economics of Education Review*, Vol. 16, No. 3, pp. 303–311. https://doi.org/10.1016/S0272-7757(96)00069-6

[31] Little, O., Goe, L. and Bell, C. A. (2009) ‘A Practical Guide to Evaluating Teacher Effectiveness’, *National Comprehensive Center for Teacher Quality* Report, Washington.

[32] Manoharan, S. (2019) ‘Cheat-resistant multiple-choice examinations using personalization‘, *Computers & Education*, Vol. 130, pp. 139–151. https://doi.org/10.1016/j.compedu.2018.11.007

[33] Maxwell, M. and Gleason, J. (2019) ‘Item efficiency: an item response theory parameter with applications for improving the reliability of mathematics assessment’, *International Journal of Mathematical Education in Science and Technology*, Vol. 50, No. 2, pp. 216–243. https://doi.org/10.1080/0020739X.2018.1492038

[34] Mitkov, R., Ha, L. A. and Karamanis, N. (2006) ‘A computer-aided environment for generating multiple-choice test items’, *Natural Language Engineering*, Vol. 12, No. 2, pp. 177–194. https://doi.org/10.1017/S1351324906004177

[35] Morphew, J. W., Silva, M., Herman, G. and West, M. (2020) ‘Frequent mastery testing with second-chance exams leads to enhanced student learning in undergraduate engineering’, *Applied Cognitive Psychology*, Vol. 34, pp. 168–181. https://doi.org/10.1002/acp.3605

[36] Oates, G. (2011) ‘Sustaining integrated technology in undergraduate mathematics’, *International Journal of Mathematical Education in Science and Technology*, Vol. 42, No. 6, pp. 709–721. https://doi.org/10.1080/0020739X.2011.575238

[37] OECD (2020) *TALIS 2018 Results (Volume II): Teachers and School Leaders as Valued Professionals*, Paris: OECD Publishing. https://doi.org/10.1787/19cf08df-en

[38] Rasila, A., Malinen, J. and Tiitu, H. (2015) ‘On automatic assessment and conceptual understanding’, *Teaching Mathematics and Its Applications*, Vol. 34, No. 3, pp. 149–159. https://doi.org/10.1093/teamat/hrv013

[39] Rønning, F. (2017) ‘Influence of computer-aided assessment on ways of working with mathematics’, *Teaching Mathematics and Its Applications*, Vol. 36, No. 2, pp. 94–107. https://doi.org/10.1093/teamat/hrx001

[40] Sangwin, C. (2013) *Computer Aided Assessment of Mathematics*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199660353.001.0001

[41] Sangwin, C. J. and Jones, I. (2017) ‘Asymmetry in student achievement on multiple-choice and constructed-response items in reversible mathematics processes’, *Educational Studies in Mathematics*, Vol. 94, pp. 205–222. https://doi.org/10.1007/s10649-016-9725-4

[42] Sass, T. R., Semykina, A. and Harris, D. N. (2014) ‘Value-added models and the measurement of teacher productivity’, *Economics of Education Review*, Vol. 38, pp. 9–23. http://dx.doi.org/10.1016/j.econedurev.2013.10.003

[43] Sims Wiliams, J. H. and Barry, M. D. J. (1999) ‘Open Testing with a Large Databank of Multiple Choice Questions’, *Teaching Mathematics and Its Applications*, Vol. 18, No. 4, pp. 159–161. https://doi.org/10.1093/teamat/18.4.159

[44] Sydsæter, K., Hammond, P., Strøm, A. and Carvajal, A. (2016) *Essential Mathematics for Economic Analysis*. Harlow, United Kingdom: Pearson Education.

[45] Tondeur, J., Scherer, R., Baran, E., Siddiq, F., Valtonen, T. and Sointu, E. (2019) 'Teacher educators as gatekeepers: Preparing the next generation of teacher for technology integration in education', *British Journal of Educational Technology*, Vol. 50, No. 3, pp. 1189–1209. http://dx.doi.org/10.1111/bjet.12748.

[46] Trenholm, S. and Peschke, J. (2020) 'Teaching undergraduate mathematics fully online: a review from the perspective of communities of practice', *International Journal of Educational Technology in Higher Education*, Vol. 17, No. 37, pp. 1–18. https://doi.org/10.1186/s41239-020-00215-0

[47] Yue, X. (2019) 'Exploring Effective Methods of Teacher Professional Development in University for 21st Century Education', *International Journal for Innovation Education and Research*, Vol. 7, No. 5, pp. 248–257. https://dx.doi.org/10.31686/ijier.vol7.iss5.1506