DOI: 10.2478/cdem-2013-0005 CHEM DIDACT ECOL METROL. 2012;17(1-2):61-68

Ryszard GMOCH[1]

COMPUTERIZED ADAPTIVE TESTING IN POLAND

KOMPUTEROWE TESTOWANIE ADAPTACYJNE W POLSCE

Abstract: New trends relating to computer-based testing of learners' achievements are presented in the paper. It describes adaptive testing methods and results of studies in this problem area. Essential questions connected with the Item Response Theory (IRT) were also discussed. The presented data indicate that computer-based adaptive testing should be popularized in Poland to its fullest extent.

Keywords: education, adaptive testing, item response theory, item characteristic curve

Applying computers in educating has opened new possibilities in the scope of testing learners' cognitive achievements. Among others, it has become feasible to apply software that makes use of models of sequential adaptive testing, ones that are difficult to apply in the classical 'paper and pen' test. Adaptive testing consists in individualization of selection of test tasks designed to be done by the examined. Adaptive testing makes it possible to create a test, with the aid of a computer, as a certain sequence of tasks which may differ from one another both as regards their ordering and the length of the sequence.

A schema of modern computer-based testing system is shown in Figure 1 [1].

A computer-based testing system is an integrated system, whose aim is to accumulate test tasks, construct tests and secure proper testing with the help of a computer. An important feature of it is the possibility of obtaining feedback of the information being processed from the tasks bank, which opens the chance of updating tasks parameters, and in consequence - raises the quality of constructed tests and procedures of adaptive testing. The tasks bank, which is the core of the testing system, is a vital set of data. The basic elements of the testing system contain **sets of data** which are presented in the schema inside the rectangles (the order of processing, answers to tasks, order of tasks in the test, the content of tasks, tasks bank, successive test tasks, data about the student, information about the test, the test itself, information about adaptive testing), as well as **operations** done on the sets, which have been placed in the schema inside the circles (processing the tasks bank, constructing the test, processing the test results, adaptive testing). Constructing a test by the computer is the act of selecting tasks from the tasks bank in such a way that the tasks of appropriate parameters should appear in the test in a proper sequence. The operations of

[1] Institute of Educational Studies, Opole University, ul. o. J. Czaplaka 2A, 45-055 Opole, Poland, email: R.Gmoch@uni.opole.pl

processing the tasks bank, constructing tests, processing test results, consist of operations realized on the set of data. The flow of information is represented in the schema by the lines pointing from the top downwards. With regard to this, the sets of data found on the top of the schema make the input of the system, whereas the sets of data placed at the bottom make the output of it. From the point of view of informatics, the computer-based testing system is thus one of processing data.

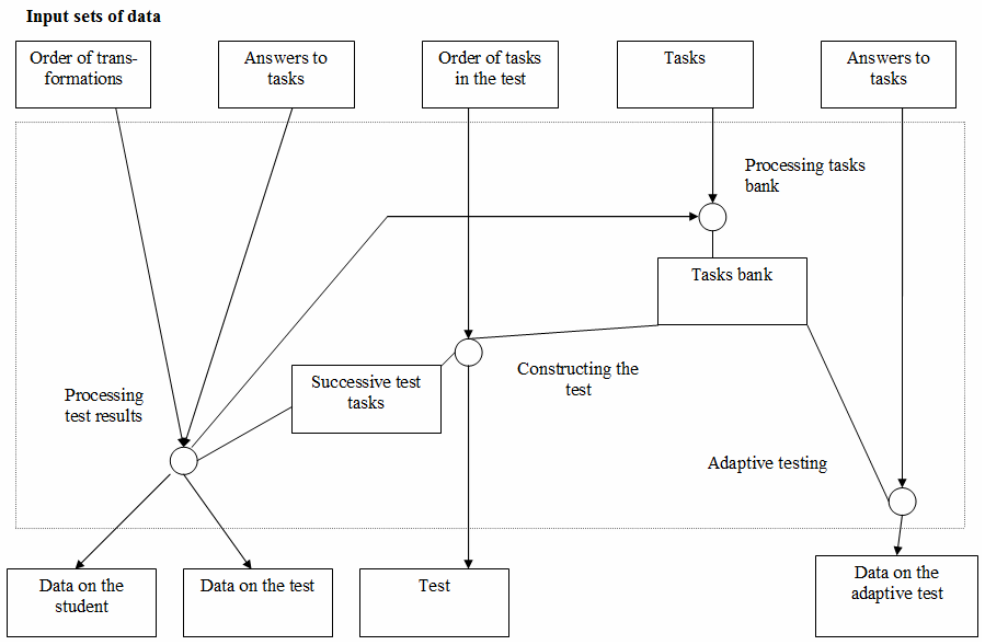

Fig. 1. Schema of computer-based testing system

The software applied for educational purposes to test didactic achievements serves the following goals [2, 3]:

1. Classical computer-based testing (*Computerized Testing* - CT) of didactic accomplishments (the tested subjects are given the same number of the same tasks to solve, arranged in the same order).
2. Computer-based testing of learners' didactic accomplishments, using the method of adaptive testing (*Computerized Adaptive Testing* - CAT).

The following procedures of adaptive testing are distinguished [4]:

- **pyramidal testing** (the tasks bank should have an ordered structure, assignment of the next task is based on the answer to the preceding one, and the number of tasks set to the learners is identical, the first task being of medium level of difficulty),
- **two-stage testing** (the tasks bank does not need to have an ordered structure, assignment of the next task is based on the answer to the preceding one, and the number of tasks set to the learners is identical. The choice of the level of difficulty (low, medium, high) of the test of the second stage depends on the results obtained on that of the first stage),

- **layer testing** (the tasks bank ought to have an ordered structure, assignment of the next task is based on the answer to the preceding one, and the number of tasks set to the learners is identical or changeable. The tasks are assigned to individual layers of the changing level of difficulty, the first task being of the medium level of difficulty),
- **fully adaptive testing** (the tasks bank ought to have an ordered structure, assignment of the next task rests on a mathematical strategy based on information functions of the tasks done by the subject being tested prior to it. The number of tasks set to the learners varies).

Below, the pyramidal procedure of adaptive testing, whose exemplary model is presented in Figure 2 [5].

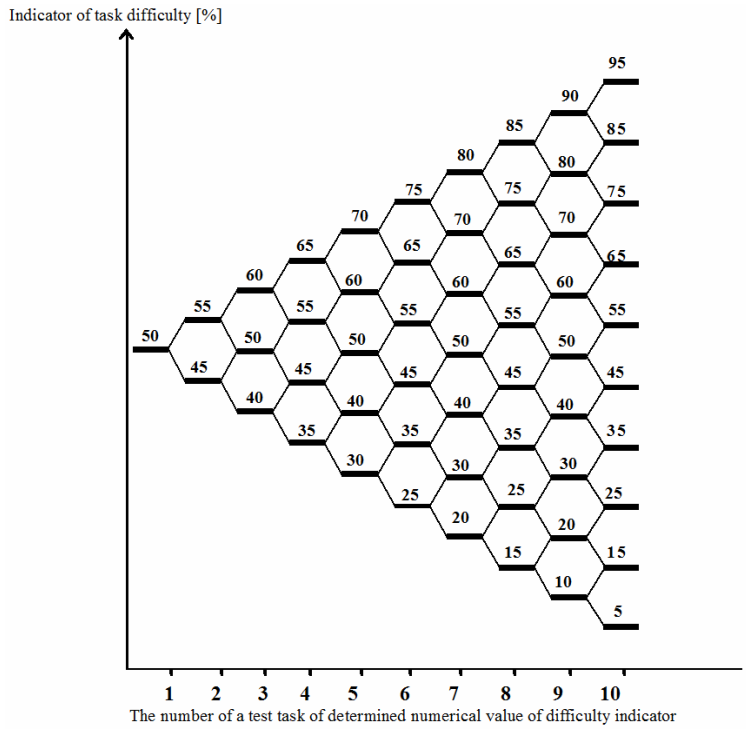

Fig. 2. Model of the pyramidal method of sequential adaptive testing

The name of pyramidal testing derives from the fact that tasks arranged according to indicators of difficulty form the figure of a pyramid with the cross-section of an equilateral triangle. At the same time each step of the pyramid corresponds to a test task of a determined difficulty indicator. The learner begins solving test tasks with one of a medium high difficulty indicator amounting to 50%, which is generated by the computer out of five tasks having this difficulty indicator value. If the learner's answer is right, the person being tested is given a more difficult task to solve, one with the difficulty indicator 45%, selected out of 5 tasks included in the computer program. However, in the case where the learner's answer is wrong the computer assigns to them an easier task to solve as the next one, the

difficulty indicator being 45%. The above procedure is repeated after the learner has provided an answer to each test task set to them, until the moment they have done 10 test tasks out of 55 ones contained in the pyramidal test. The degree of difficulty of the test task with reference to the above-presented procedure remains within the range 5÷95% and changes every 5%. In practice, if we want to elaborate on the so-called 'pyramid', that is a set of test tasks satisfying the given conditions as regards the values of statistical indexes, then we ought to prepare a relatively large set of tasks in order to be able to obtain a suitable number of tasks of determined values of the difficulty indicator.

In the case of the procedures of two-stage testing and layer testing, the task difficulty indicator does not necessarily have to be the criterion of classifying tasks as regards their difficulty level. This role can be played by, for instance, making use of the taxonomy of ABCD educational goals proposed by the creator of the didactic measurement in Poland - B. Niemierko (qualifying a test task by the teacher into the appropriate category of educational goals which the given task aims to test). And thus, tasks that test memorizing of knowledge (the taxonomic category A), as well as those testing the understanding of it (the taxonomic category B) can be classified as ones of the low difficulty level. On the other hand, test tasks relating to skills of applying knowledge in typical situations (the taxonomic category C) can be included among those of the medium difficulty level. The set of tasks of the highest difficulty level should then include tasks relating to the level of skills concerning application of knowledge in problem situations (the taxonomic category D).

The Item Response Theory (IRT), which takes into account responses to individual test tasks, provided by people being tested, is especially useful in interpreting results of adaptive testing [6]. The only parameter that describes the tested subject is the level of their knowledge, which is the measure of achievements. The tasks parameters - ranging from one to three (task difficulty level, differentiating power of the task and possibility of giving the right answer to the task by guessing), depending on the applied IRT mathematical model, characterize these tasks. The theory facilitates determining the probability of supplying the right answer to the given task for each value of the level of the tested subject's knowledge. The interdependence of the probability of providing the right answer to the question and the level of the tested subject's knowledge is called *Item Characteristic Curve* (ICC) in the IRT [7].

One of the most often applied IRT models is Birnbaum's three-parameter mathematical model, described by Equation (1), in which the probability of giving the right answer to the task by the person being tested who represents a given level of knowledge depends on three parameters that are as follows: parameter of the differentiating power of the task, parameter of task difficulty, parameter of guessing [1]:

$$P_{ai} = c_i + (1 - c_i)[1 + \exp(-a_i(\Theta_a - b_i))]^{-1} \tag{1}$$

where: P_{ai} - probability of providing the right answer to task 'i' by person 'a' representing the level of knowledge Θ_a, $a_i \geq 0$ - parameter of the differentiating power of task *i*, $b_i \in [-\infty, +\infty]$ - parameter of difficulty of task *i*, $c_i \in [0, 1]$ - parameter of guessing for task *i*, *ie* probability of giving the right answer when the subject guesses without thinking.

As far as the two-parameter mathematical IRT model (Rasch's model) is concerned, we accept the assumption of zero value of the parameter of guessing. If, additionally, we assume that the parameter of differentiating power of tasks is equal and amounts to one, then we obtain a one-parameter model.

A graphic representation of the IRT model is the ICC [3, 7]. The task parameters determine the shape and location of this curve on the scale of the level of knowledge. In the case of applying the three-parameter IRT model, the shape and the location of the ICC are determined by three above-mentioned tasks. The curve described by means of these parameters takes on the shape of the letter 'S', with the upper probability asymptote equal to 1 and the bottom probability asymptote usually greater than 0 [1, 8].

Application of the IRT models makes it possible to locate tasks in the tasks bank since it is possible to collect tasks parameters and parameters of the learners' levels of knowledge in a mutually independent manner.

Knowledge of the values of the information functions of tasks allows constructing adaptive tests that facilitate determining the level of knowledge of the person being tested.

In adaptive testing, decisions concerning the selection of tasks and the order of their presentation to the subject are taken during the testing. The choice of each successive task is based on the estimated value of the learner's knowledge parameter resulting from the answers given to previous tasks. Initially, the level of the tested subject's knowledge is calculated with approximation, yet with each successive task the gathered information makes it possible to select more and more accurate tasks to be solved.

The computer is able to update the calculated value of the parameter of the level of the learner's knowledge and to search the tasks bank in order to find the appropriate task. Adaptive testing is possible only when we have at our disposal sets of test tasks of well-known parameters and procedures to calculate the parameters of the level of the learners' knowledge.

Below there is a description of selected results of studies in computer-based testing, which are available in the specialist literature.

A computer system was elaborated on at the Gdansk University of Technology [9], which serves the purpose of testing learners and grading their performances. The system is an Internet-based database application. The MS Windows 2000 Professional was chosen as the system platform, and the www server is Apache which satisfies the minimal equipment requirements and is free of charge. In the work on designing and implementation of the system the database platform MySQL was used. The above-mentioned system possesses three basic modules: administration of the system, servicing teachers and servicing students. Interpretation of test results is based on the classical test theory.

The results of our research concerning computer-based testing of learners' accomplishments [5, 10, 11] applying the pyramidal procedure of adaptive testing proved that this method - in comparison with a paper-and-pen test - is a modern testing procedure since it makes it possible to adjust the degree of tasks difficulty to the level of knowledge of the person being tested, raises the reliability of testing results in comparison with the classical paper-and-pen test, shortens the testing time as the person being tested does only 10 tasks generated by the computer out of a 55-task set, recording of full data concerning the test is done by the computer, which facilitates analyzing test results and eventual correction of tasks. The conducted research proved that the computer program PIRAMIDA, elaborated by J. Hurek (Opole University), functions in a proper way.

The literature of the subject [12] describes the concept of a computer-based test adapted to the student's level in compliance with the assumptions of CAT according to an own method. On the basis of the presented concept the authors propose the following two strategies of realization of a test:

- the strategy of the number of test questions established in advance and determined time for supplying the answers;
- the strategy of an unlimited number of test questions and undetermined time set for supplying the answers.

In the opinion of the authors of the concept, the model of a test with adaptation can be used in regular and laboratory classes in various subjects taught.

The literature of the subject [8] presents results of a statistical analysis of a 16-task test applied to 106 group of tested subjects, carried out with the use of the classical and probabilistic test theory (IRT), applying for this purpose respective Statistica 6 packet and RUMM2010 program. It needs to be underlined that the difficulty and the differentiating power, as the parameters of a test task, function both in the classical and the probabilistic test theory, yet the obtained interpretations are indeed different. Results of the study conducted by Ciżkowicz indicate that difficult tasks - in compliance with the IRT - proved difficult also according to the classical theory. On the other hand, in the case of the differentiating power of tasks the obtained dependence was negative, which indicates that tasks regarded as highly differentiating according to one theory are weakly differentiating according to the other one and the other way round.

In order to evaluate the effectiveness of computer-based testing, a computer system was elaborated [13] which makes it possible to dynamically create a test in the course of testing (adaptation). It calculates both the parameters of the classical test theory and the IRT, that is the parameter of the difficulty level of tasks (an estimator of student's knowledge). A bank was prepared of test tasks of different levels of task difficulty (easy, medium and difficult), and the test plans were drawn in such a way that it would be able to present both the possibility of generating parallel versions of the test and also to individualize testing in the form of a two-stage test and a layer one. The test plan in the case of the adaptive testing procedure contained 4 tasks facilitating an initial assessment of the tested person's abilities. In dependence on the number of scored points the subject being tested either finishes doing the test or is given another 4 tasks, whose level of difficulty is pinned on the number of obtained points. The results of the study that are presented in the above-mentioned work, relating to 3 testing procedures point to the fact that there do not occur considerable deviations of the final results of the level of students' achievements when results of classical testing are compared with those of an adaptive (two-stage and layer) one. A comparison of results of a computer-based testing with those of oral answers confirmed the reliability of the computer-based testing.

The comparative analysis of results of their own research relating to the traditional pen-and-paper testing and a fully adaptive one, which was carried out by J. Hurek and A. Sztejnberg [14], proved that the examined group of students (200 subjects) needed, on the average, 30% fewer tasks to solve in the case of an adaptive test in comparison with a traditional written one. The highest number of tasks on the former were done by students of the average level of knowledge, and the number of test tasks necessary to solve was decreasing along with a rise, as well as with a drop in the level of the examined subjects' accomplishments. The correlation between the number of tasks on the adaptive test necessary to do and the level of students' accomplishments was relatively high. The study proved that the benefit of time saving is not identical for all the examined and depends on the levels of their accomplishments.

Application of the IRT in the area of didactic measurement has become possible thanks to the use of computer programs (ASCAL, LOGISTIC, MULTILOG, RASCAL, XCALIBRE, RUMM2010), whose number is systematically growing in the educational market [3].

The results of computer-based adaptive testing, obtained by various authors, which are presented in this paper, clearly point to the fact that adaptive testing is the right direction of development of the didactic measurement system and ought to be popularized to its fullest extent in Poland. This is advisable, among others, because the first step on this road has already been made - analyses of results of resting (with the use of the probabilistic theory of test task - IRT) made by *Okręgowa Komisja Egzaminacyjna* (District Examination Committee) in Krakow are applied by University of Cambridge International Examination [15].

Moreover, there exists a possibility, while mastering teachers' skills of measurement within the IRT, to apply for the aid offered by CITO (National Institute for Educational Measurement established by the Ministry of Education of Holland) which supports examination boards in 30 countries. As far as this scope of activity is concerned the latest monograph dealing with this problem area, which was published by Opole University is worth recommending [4]. The work addresses and covers the following questions:

- foundations of computer-based testing related to educational measurement;
- the modern test theory IRT and its selected models;
- selected applications of the IRT methods to tasks and tests;
- a review of selected studies in the field of adaptive testing;
- strategies of computer-based adaptive testing;
- a comparative analysis of results of application of procedures of a traditional pen-and-paper test and computer-based adaptive testing in measurement of students' cognitive accomplishments;
- computer-based test data bank and its usage for constructing tests.

The above-mentioned publication can complement the extremely valuable work devoted to educational diagnostics, which was elaborated by the founder of the didactic measurement in Poland - B. Niemierko [7].

References

[1] Sztejnberg A, Gmoch R. Z badań nad zastosowaniem teorii wyniku zadania testowego (Item Response Theory) w testowaniu osiągnięć studentów chemii z zakresu elektrochemii. Pregradualni priprava a postgradualni vzdelavani ucitelu chemie. Ostrawa: Sbornik Prednasek; 1999.

[2] Weiss JD. Bibliography on Computerized Adaptive Testing (CAT), (Updated March 26, 2011). http://www.psych.umn.edu/psylabs/ catcentral/bibliographycomplete.htm.

[3] http://en.wikipedia.org/wiki/Computer-adaptive_testing, 2008.

[4] Hurek J, Sztejnberg A. Doskonalenie komputerowego pomiaru testowego. Opole: Uniwersytet Opolski; 2010.

[5] Gmoch R, Sztejnberg A, Hurek J. Checking the knowledge of chemistry students in kinetics and chemical equilibrium by the conventional test and the computer test using the pyramidal method of adaptive test. Science Education and Society. Hradec Kralove: Gaudeamus; 1999.

[6] Wim J, van der Linden C, Glas AW. Computerized Adaptive Testing: Theory and Practice. Kluwer Academic Publishers; 2010.

[7] Niemierko B. Diagnostyka edukacyjna. Podręcznik akademicki. Warszawa: Wyd Nauk PWN; 2009.

[8] Ciżkowicz B. Zastosowanie programów komputerowych w analizie testu. 15 Ogólnopolskie Sympozjum Naukowe „Komputer w Edukacji". Kraków: Akademia Pedagogiczna; 2005. www.ap.krakow.pl/ptn/ref2005.

[9] Choroń O. Komputerowe wspomaganie przeprowadzania i oceniania testów. 15 Ogólnopolskie Sympozjum Naukowe nt. „Komputer w Edukacji". Kraków: Akademia Pedagogiczna; 2005. www.ap.krakow.pl/ptn.
[10] Gmoch R, Hurek J, Sztejnberg A. Komputerowe sprawdzanie osiągnięć szkolnych uczniów. Program PIRAMIDA. Komputer w Szkole. 1993;9:51-60.
[11] Gmoch R, Sztejnberg A, Hurek J. Z badań nad komputerowym sprawdzaniem wiedzy studentów chemii z zakresu elektrochemii przy zastosowaniu piramidalnej metody testowania adaptacyjnego. Komputer w Edukacji. 1996;1-2:43-46.
[12] Klosov O, Pytowski M. Model testu komputerowego z adaptacją do poziomu wiedzy użytkownika. Scientific Bulletin of Chełm. Section of Matemat. Computer Sci. 2008;1:141-149.
[13] Olejarz-Mieszaniec E. Ocena efektywności testowania komputerowego z wykorzystaniem klasycznych i nieklasycznych metod analizy wyników. 16 Ogólnopolskie Sympozjum Naukowe Komputer w Edukacji. Akademia Pedagogiczna: Kraków; 2006. www.ap.krakow.pl/ptn/ref2006/Olejarz.pdf.
[14] Hurek J, Sztejnberg A. Analiza porównawcza wyników zastosowania procedur tradycyjnego testowania pisemnego i procedur komputerowego testowania w pełni adaptacyjnego w pomiarze osiągnięć studentów, Mezinarodni seminar: Soudobe trendy v chemickem vzdelavani. Hradec Kralove: Sbornik Prednasek, Gaudeamus, Univerzita Hradec Kralove; 2006.
[15] Szaleniec H. Dylematy rozwoju systemu egzaminów w Polsce. Egzaminy Naszych Uczniów. 2008;2:5-9.

KOMPUTEROWE TESTOWANIE ADAPTACYJNE W POLSCE

Instytut Studiów Edukacyjnych, Uniwersytet Opolski

Abstrakt: W artykule scharakteryzowano nowe trendy w zakresie komputerowego testowania osiągnięć uczących się. Omówiono metody testowania adaptacyjnego i wyniki badań w tym zakresie. Przedstawiono podstawowe zagadnienia dotyczące teorii wyniku zadania (IRT). Zaprezentowane dane wskazują, że komputerowe testowanie adaptacyjne winno być w pełni upowszechnione w Polsce.

Słowa kluczowe: kształcenie, testowanie adaptacyjne, teoria wyniku zadania, krzywa charakterystyczna zadania