Computer-aided paper-based tests construction process. Experience of the State Students Admission Commission of Azerbaijan Republic.

(Shelaginov O.Y., SSAC)

The paper describes experience of the State Students Admission Commission (SSAC) in automation of process of constructing of paper-based tests, combining separate test items. Developed by SSAC computer-aided system has functioned since 1996 and used for preparation of measuring materials for different exams, conducted by SSAC, including university admission exams.

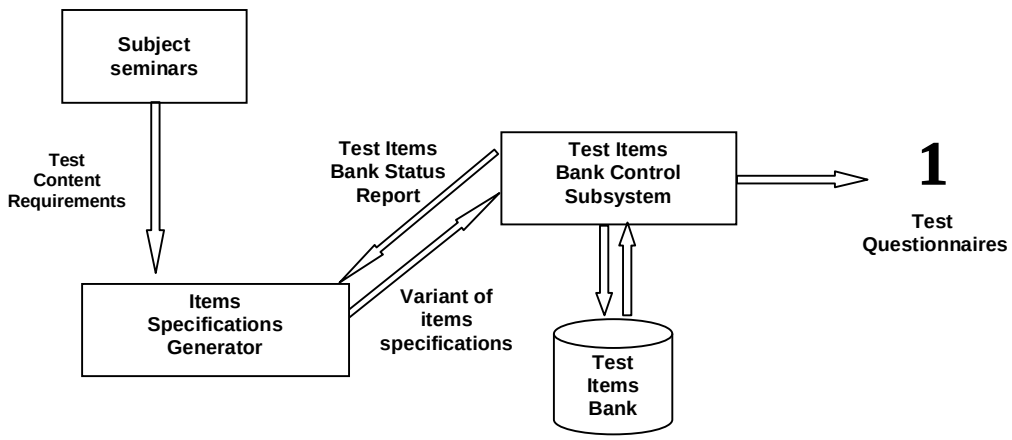
The importance of development of system was conditioned by necessity of
- decreasing time of test construction process
- reducing significance level of “human factor” in items selection process
- ensuring transparency of technological process
- minimizing the risk of error appearances

Basis for the created system is strictly structured Test Items Bank (TIB), where each test item presents with set of characteristics, such as:

- language of test item (Azeri or Russian)
- subject (native language, literature, math and etc.)
- section of subject
- difficulty level
- cognitive category (terminology, factology, calculation, and etc.)
- information about analogues (similar items)
- list of authors
- list of experts
- statistics
- information about its use

The forming of TIB for university entrance examinations has been realized by SSAC since 1994. In current time TIB includes test items on 17 subjects (10 from them has 2 parallel subbases on Azeri and Russian languages). Each subject subbase has 5000-20000 test items.

Test constructing process with the aid of computer-aided system is presented on the following picture:



Picture 1. Test constructing process

Subject seminars functioning under SSAC establishes set of content requirements, which are based on the purpose and structure of test, and predicted ability level of target population group. Accomplishment of those requirements must guarantee the achievement of content validity of test.
Key characteristics of test items are:
- Subject section

- Difficulty level
- Cognitive category

Classification of test items is built on combination of these key characteristics. Formalized requirements for test constructing process use this classification for describing limits of quantity of test items in categories. Constructed test must contain items from whole spectrum of subject themes, different difficulty levels and cognitive categories.
Difficulty level of test item can be characterized by:

1) expert judgment, based on experts' experience and statistics of similar items (e.g. easy, moderate, difficult)
2) objective statistical characteristics, calculated after pretesting

Cognitive categories define quality characteristic of test item. In that sense, item, for example, could be terminological, factological, calculative, explanative, generalizing, predictive or requiring the suggestion of action. Depending on subject this classification can slightly differ.

Requirements of exam content are formalized and presented as data for Items Specifications Generator unit in internal requirements definition language.

Requirements definition language has elements for:

a) defining main test characteristics, such as total number of items, number of versions of test, required number of versions of generated specifications, maximal quantity of items from each of subject section;
b) modification of TIB representation. It means then virtual restructuring of TIB based on key item characteristics, such as combining of subject sections, cognitive categories and etc. For example, if test must have 25 items, but number of sections is 35, then we have to combine some close (similar) sections together and as result the test will contain 1 item from one of several similar sections.
c) defining of restrictions on number of items with specific characteristics or classificational restrictions. Supported quantity restrictions, based on classification by one (e.g. number of factological items), two (e.g. number of simple calculative items) or all three key characteristics (e.g. number of simple calculative items from first 10 subject sections). Available three type of restrictions: 1) number of items not less than; 2) number of items strict equal to; and 3) number of items be in some interval. At the same time restrictions could be applied to whole test or only to some subtests (e.g. separately for botany, zoology, anatomy and etc.)

Furthermore, Test Items Bank Control Subsystem forms data about state of TIB – number of items with different set of key characteristics for appropriate subject subbase.

The aim of Items Specifications Generator unit is finding such set of key characteristics for each item of constructed test, which matches with all test content restrictions. Request could be formed in terms of items difficulty, based on experts' evaluations as well as in terms of indexes of interval statistic of difficulty (e.g. difficulty parameter in Rasch model). The differences in what kind of information the unit receive as data about state of TIB and in which terms content requirements are described.

Items Specification Generator unit must in accordance with some initial random number to choose one (or several) from all possible solutions, satisfying set of formalized content requirements. Consequently, in this case, differently from known optimization models of test constructing (e.g. IRT-based models by Theunissen (1985), Van der Linden and Boekkooi-Timminga (1989), Van der Linden and Luecht (1998)) we don't *optimize* content of test, but are looking for one feasible solution. The key moment is to achieve approximately equal probability of all feasible solutions to be selected, i.e. algorithmically mustn't be any preferences (criteria) to selection of one or another test item, every feasible item must have equal chance to be selected. This approach exclude possibility of preguessing of real test content, even if somebody has full information about TIB and content requirements. Optimization methods allow to find "best" solution in accordance with some criterion, for example, shape of information curve must maximally match desirable. In our case, if we have statistical characteristics of all items, we can estimate result of constructing at the end of constructing process, plot information curve of test and select one from several suggested solutions.

Consider the example of content requirements, formed by biology subject seminar.

Test content requirements on biology (extract, according to the decision of biology seminar)

For **25** test tasks on biology are provided 38 minutes (1.5 minute for each task in average) of overall examination time. Participants of workshop on biology have decided biology test block consist of **25** questions, including **5 questions on botany**, **5 – on zoology**, **7 – on anatomy** and **8 – on general biology**. Distribution of test tasks, depending on difficulty levels and ability indicators is presented in below table.

Sub-Subjects	*Number of questions*	*Easy*	*Medium*	*Difficult*
Botany	5	1	2	2
Zoology	5	1	1	3
Anatomy	7	2	3	2
General Biology	8	1	4	3
Overall	25	5	10	10

According to the decision of biology seminar participants, **1 terminological** and **3-4 factological** tasks (including 1 about authors of scientific discoveries) must be included into the test block. **Calculative tasks** must **not be more than 6**. Number of tasks related to each of remaining ability indicators (**explanation, generalization, predictive, suggested action**) can be **about 3-5**.

Picture 2. Example of test content requirements on biology

Subject Biology contains 30 sections. In order to ensure the test to cover content area evenly, developer can indicate which sections could be united, and resulting test will contain only one item from one of these united sections (e.g. T(2,3), T(8,9), T(17,18), T(27,30), T(28,29)). In accordance with initial random number, unit defines item from which section will be included to resulting test. Sections (1-6) belong to botany, (7-12) – zoology, (13-20) – anatomy, (21-30) – general biology. Requests for each group of sections could be composed separately, as you can see on the picture. For example, we can demand even distribution of factological tasks by all 4 groups of sections and etc.

Picture 3. Example of formalized content requirements

Resulting specification is the set of combinations of key characteristics, where each item is represented by one line.

```
001  T(1) L(1) R(2)          010  T(12) L(3) R(5)          019  T(22) L(2) R(7)
002  T(2,3) L(3) R(6)        011  T(13) L(3) R(5)          020  T(23) L(3) R(7)
003  T(4) L(3) R(7)          012  T(14) L(3) R(3)          021  T(24) L(3) R(5)
004  T(5) L(2) R(5)          013  T(15) L(2) R(6)          022  T(25) L(2) R(4)
005  T(6) L(2) R(3)          014  T(16) L(2) R(7)          023  T(26) L(2) R(3)
006  T(7) L(3) R(7)          015  T(17,18) L(1) R(3)       024  T(27,30) L(1) R(1)
007  T(8,9) L(3) R(6)        016  T(19) L(2) R(3)          025  T(28,29) L(3) R(6)
008  T(10) L(1) R(2)         017  T(20) L(1) R(2)
009  T(11) L(2) R(4)         018  T(21) L(2) R(5)
```

Picture 3. Example of resulting test specifications file

After completion of Items Specifications Generator unit's work, resulting test specifications file enters to input of Test Generation and Assembly Unit, which being initialized by random number selects from TIB the items with specific characteristics, prescribed in test specifications file. Unit forms necessary number of versions by a) scrambling, b) using analogues or c) using other items with the same specifications. Practically, the work of this unit is selection of necessary number of specific test items with characteristics, defined on previous stage. Furthermore, unit can modify placement of distracters and correct answer for each item. But for all that, it takes into account frequency and sequences of symbols in resulting key pattern. Test assembly unit forms resulting test versions as MS Word documents for make-up and publishing.

Must be mentioned, that all stages of test constructing process are thoroughly recorded (including operations of initial random numbers assignment) and possibility of repeat (full review) of all test constructing operations is ensured. This provides transparency of technological process and excludes possibility of unauthorized operations during the test constructing process.

Described test constructing approach, used by SSAC for universities entrance examinations in Azerbaijan allows to develop tests with proper level of reliability and ensure fairness and transparency.

Literature:

van der Linden, W. (1996). Assembling tests for the measurement of multiple abilities. Applied Psychological Measurement, 20, pp. 373-388

van der Linden, W. And Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. Psychometrika, 54, pp. 237-247.

Verschoor, A. (2004). IRT Test Assembly Using Genetic Algorithms, Arnhem: Cito.