A NEW APPROACH TO THE ACHIEVEMENT TEST ITEMS EVALUATION: THE CORRECTNESS COEFFICIENT

Foltýnek, T.

Mendel University of Agriculture and Forestry in Brno,
foltynek@pef.mendelu.cz

Abstract

This paper deals with the problem set of an achievement test item scoring. The scoring process is generalized with the help of correctness coefficient – the new concept set up by the author. The paper describes complexly formalization of the scoring process, contextualizes the contemporary used methods to the general context and brings new methods as well. The scoring methods of sorting items and guessing penalty are described in detail. Observations described in this paper can help examiners with more accurate assessment of achievement test results.

In the first part, the theoretical basics of the test item scoring are given. We are going to find out that whole scoring process depends on the teaching objectives, test item types, curriculum taxonomy and achievement test objectives. Then the theory of the test item types is described. After this theoretical introduction the concepts of the total achievement test score and correctness coefficient are set up. Let's emphasize that using of the correctness coefficient is the new contribution of the author. Than the correctness coefficient is used to express the measure of examinee's answer accuracy within the different test item types. Using the correctness coefficient for evaluation of closed multiple-choice items, injective and general relational items, narrow open items, joining items and sorting items are deeply examined and described. Various scoring method for these item types are discussed, especially for the sorting and joining items. Afterwards the theory of penalty guessing is expressed with the help of the correctness co efficient, which strengthens the ability and universality of theory being described.

The main goal of this research paper is to provide the complex theoretical overview of the test item scoring problems, which can be useful for pedagogues, examiners and testing application (or e-learning system) developers to provide more accurate and clear evaluation process of the achievement test.

Key Words

Evaluation Methodologies, Intelligent Tutoring Systems, Media in Education

Introduction

Achievement test is considered to be the most objective tool of pedagogical evaluation (Foltýnek 2005), Davis (Davis 1993) adds that *"tests are powerful educational tools that serve at least four functions"* (to evaluate students, to motivate and help them, to give a sort of feedback to teachers and reinforce learning). Particular test questions – items – can be however evaluated – scored – with various methods giving various results. This problem appears especially while working with items of more complicated types, where scoring methods aren't obvious at first sight (Foltýnek 2006). As *the assessment is the integral part of the learning process* (Booth et al. 2003), we are going to generalize the scoring methods and analyze possibilities of their concretization in the following text. The analysis will be done with respect to their methodical suitability depending on the type of examined curriculum, expected deepness of students' knowledge, goal of the test and other characteristics (Payne 1968). The paper doesn't suppose just separate evaluation of particular items, but also *weighting of the items* among themselves (Linrace&Wright 1995). Automated, computer supported testing is gaining importance and achievement test results are considered to be an absolutely reliable indicator of the knowledge of examinees (Segall et al. 2005). A technology usually could be *further dissected into finer grain techniques and methods, which correspond to different variations of this functionality and different ways of its implementation* (Brusilovsky 1999). Therefore it is applicable to devote adequate part to the scoring methods for saving the achievement test objectivity.

In the following text, we are going to set up the formalization of achievement test items with the help of correctness coefficient. This coefficient expresses – depending on the test item type – the rate of students' fault. The test score assessment is therefore much more accurate than in case of considering the boundary values (correct – incorrect) only.

Materials and Methods

Let's consider four different bases:

- Teaching objectives taxonomy (Bloom 1956);
- Curriculum taxonomy (Foltýnek 2006);
- Test item types (Foltýnek 2006) and
- Achievement test objectives (Mužić, 1993).

Therefore we have four-dimensional space of indicators, which are helping us to choose the proper scoring method for the specific test. The relations of particular bases among themselves and to the scoring are illustrated at Figure 1. We shall abbreviate Test Item Scoring to TIS in the following text. Teaching targets or curriculum taxonomy impacts – besides TIS – the test item types choosing too. TIS depends on all four bases.

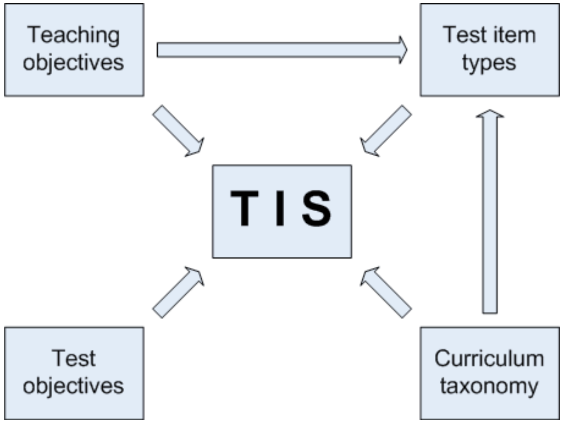

Figure 1: Relation of basis to TIS

Bloom's taxonomy

Bloom distinguishes six levels of knowledge deepness (Hogbood et al. 2004). Construction style of the test items and TIS methods should be adapted to the expected knowledge deepness.

- *Knowledge* – students are asked to remember pieces of information, specific terms and techniques;
- *Comprehension* – students are asked to grasp meaning and to demonstrate understanding by summarizing or explaining;
- *Application* – students are expected to take what they have learned and apply it in a new, real-life situation;
- *Analysis* – breaking down of knowledge into parts and the relation of those parts to the whole concept;
- *Synthesis* – assembling knowledge into a new whole. This means collecting information, then creating a new insight;
- *Evaluation* – students judge the value of the information for a specific purpose.

Knowledge taxonomy

We can classify knowledge according to the way of processing in the human brain and according the application to:

- *Encyclopedic* – knowledge of (isolated) data without deeper relations is important;
- *Relational* – relations between objects are important;
- *Deductive* – knowledge of principles and deduction ability is required;
- *Language* – knowledge and competences about language are important;

Test item types

Test items (more commonly called *questions*) can be divided to many types (Mužić 1993). The crudest division criterion is to *closed* and *open* items. Using closed items, the examinee chooses from given alternatives, events and sorts or joins them. Using open items, the examinee doesn't have any possibilities and he/she is forced to create the answer himself/herself.

Open items are further divided to *wide* and *narrow* distinguished by the length of an answer. Closed items are divided to multiple-choice, where generally the examinee is asked to choose some of offered alternatives and relational, where the examinee looks for the relation between offered objects, respectively sets of objects. The scheme of test items division is illustrated at Figure 2.

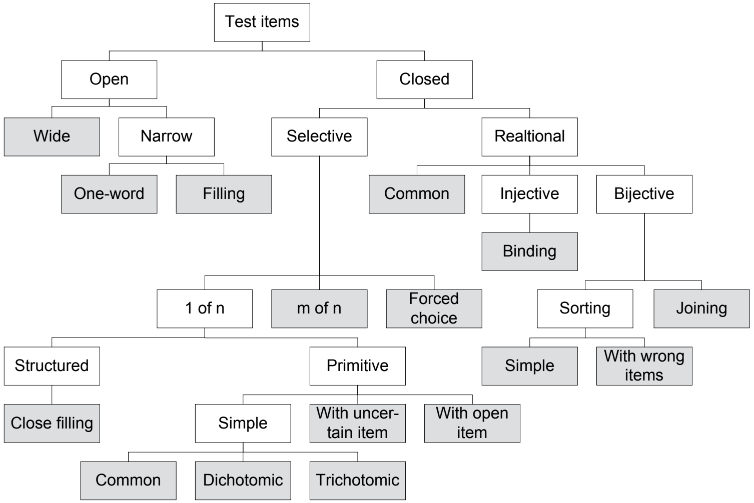

Figure 2: Division of test items types

Achievement test objectives

An achievement test can have many objectives (Foltýnek, 2006):

- *Administrative* – indicates if the student has met given requirements or not;
- *Advisory* – indicates special abilities or disabilities of the student, recommends suitable fields of future personality development;
- *Informational* – indicates performance, output, success, failure of the student, parent or teacher;
- *Motivational* – stimulates the student to increase his/her output;
- *Achievement* – fixes and deeps his/her knowledge;
- *Educational* – forms the student's approach to learning and knowledge;
- *Reflexive* – provides a feedback to the teacher.

It is obviously possible that one achievement test can has more targets. Hardly ever we have a test with just one of the mentioned targets. Always is important to notice that TIS method depends on the test targets.

Total achievement test score

Let's consider test **T**. Let the test be composed of *n* items denoted $p_1, p_2, \ldots, p_n$. Then

$$\mathbf{T} = \{\mathrm{p}_1, \mathrm{p}_2, \ldots, \mathrm{p}_n\}$$

Let $b_1, b_2, \ldots, b_n$ is a point evaluation of items, thus their weights. During the scoring process, we will find so called particular item score, thus the real number from the closed interval <-1,1>, which will be denoted σ_i. This number is serving as a base for computation of item score, which will be denoted s_i. The weights are taken into account during computation of the item score. Obviously is

$$\mathrm{s}_\mathrm{i} = \sigma_\mathrm{i} \cdot \mathrm{b}_\mathrm{i}$$

Therefore total test score S_T is

$$S_T = \sum_{i=1}^{n} s_i = \sum_{i=1}^{n} \mathrm{s}_i \cdot b_i$$

Values $b_1, b_2, \ldots, b_n$ assesses examiner during the preparation of the test. Values σ_i for $i = 1, 2, \ldots, n$ are computed from the examinees´ answers during the scoring process.

Results and Discussion

Correctness coefficient κ

To the most common generalization of TIS formalism, we have to decompose the scoring of each type to the most fundamental elements.

In open items, let's consider text strings.

In multiple-choice items, let's consider particular alternatives

In injective and general relational items let's consider members of Cartesian product A × B (thus ordered pair (*a*,*b*), where $a \in$ A and $b \in$ B).

In pairing items, let's consider particular one-to-one pairings between the sets A and B. In the following text, we will denote the set of all one-to-one pairings by the symbol B_{ij}(A,B).

For each of these elements, let's define its correctness coefficient, denoted by κ, which is going to be the base for the particular item score computation. The correctness coefficient has the same range as the particular item score, and, as we will show later, in some cases these values are equal.

From the methodic aspect, in correctness coefficient the rate of correctness (event. the rate of incorrectness) of every particular answer is hidden. Let's explain the semantics of correctness coefficient in particular groups of test item types.

In open items, the correctness coefficient is assigned to every text string. It is obvious that for most of the strings κ = 0, these are generally incorrect or nonsense answers. For absolutely correct answer κ = 1. In the case of existence of more correct answers, all of them will have the correctness coefficient equal to one. There can also be even partly correct answers. In this case, the values from the inside of interval (0, 1) gain their importance. These values indicate how close the examinee's answer was to the correct one. We can also consider existence of answers indicative of a blunder. In this case, and supposing much restrictive TIS methods, we can also consider values κ ∊ <-1, 0), which will indicate the rate of incorrectness of these answers. Open items serve as an example of items where correctness coefficient and particular item score are equal.

In multiple-choice items, the correctness coefficient is assigned to each alternative. This is independent from the multiple-choice item subtype. We have to notify that multiple-choice items type *m* of *n* are de facto sets of dichotomy items, which are special cases of multiple-choice items type 1 of *n*. We need to elaborate the numeration of TIS just for the multiple-choice items type 1 of *n* and apply gained knowledge to the other types.

Let's have the set of alternatives. The examinee has to choose just one of them. In the most simple case, we assign to the only correct alternative κ = 1 and to other alternatives κ = 0. Now we have to consider the case, where distinguishing the mistake following from choosing particular incorrect alternatives is suitable to consider. Not all of them can be equally incorrect and sometimes some of them can be "almost correct". In that case, similarly as in the open items, we should consider even values from the inside of the interval (0, 1). Moreover, some of alternatives can indicate rough ignorance which should be penalized by the negative score. Therefore, we can consider values from the interval <-1, 0).

Negative values of the correctness coefficient gain their importance also when guessing penalty (penalization of incorrect answers) is employed. In that case let's set κ = 1 for correct alternative and

$$\mathrm{k} = \frac{-1}{n-1}$$

for the incorrect one.

Due to the fact that the items type *m* of *n* are – as we have mentioned above – sets of the items type 1 of *n* with two (eventually three) constant alternatives, in items of this type we have to define the correctness coefficient both for the case if the examinee has chosen the given alternative and for the case if not.

Injective items are de facto sets of items type 1 of *n* with constant set of alternatives equal to B. Correctness coefficient is defined for every pair (*a*, *b*) from the Cartesian product A × B and its semantics lies in the evaluation of validity of formatting object a from the set A to the group *b* ∊ B. Values 0 or 1, respectively values from the inside of interval (0, 1), respectively negative values, have the same meaning as in the multiple-choice items type 1 of *n*.

We can look at the general relational item as if it was a set of multiple-choice items type *m* of *n*, where the examinee decides in every object of the set A to which groups of the set B given objects belong. As we know that items type *m* of *n* are sets of dichotomy items, we can look at the general relational item as if it was a set of |A|·|B| dichotomy items. Thus we assess the correctness coefficient particularly for each pair (*a*, *b*), where *a* ∊ A and *b* ∊ B and each of cases chosen or not. This practice is consistent with necessary thinking processes in the examinee's brain, because he/she has to decide whether every independent given object can be member of a given set or not. The examinee considers independently whether to check each pair or not.

According to the correctness coefficient assessment, the most complicated are pairing (connecting and sorting) items. We can't look to the pairing items as if they were a set of multiple-choice items, because these items weren't independent. To place a given object to the proper group is the task of the examinee, equally as in all other relational items, in the case of pairing items we additionally require using of each group just once. The sets A and B are in this case equally potent. The smallest independent element we can evaluate with the correctness coefficient is thus whole one-to-one pairing. Due to their big amount (k! for k = |A| = |B|) we have to choose (besides the correct pairing) those, which indicate certain specific mistakes and to assess them individual correctness coefficient. We have to assess a suitable implicit value to the other pairings (e.g. κ = 0).

Closed Multiple-Choice Items Scoring

At first, let's look at the most simple case of closed multiple-choice items, items type 1 of *n*, where the examinee chooses just one. We are working with the set of alternatives M. The correctness coefficient κ(m) is defined for each *m* ∊ M. Let m_o ∊ M is the examinee's answer, thus chosen alternative. Particular item score is given by formula

$$\sigma = \kappa(m_o)$$

Equal formula is good for singular cases of closed multiple-choice items, thus for dichotomy and trichotomy items.

We can look at the multiple-choice items type *m* of *n* as if it was set of dichotomy items. Let M is the set of alternatives, let's define correctness coefficient for each alternative and each possibility chosen or not (0 – 1). Let Z ⊆ M is the set of items chosen by the examinee and Ž ⊆ M is the set of not chosen items. Inevitable sets Z and Ž are the decomposition of the set M. Thus Z ∪ Ž = M a Z ∩ Ž = ∅. Particular item score is therefore given by formula

$$\mathsf{s} = \frac{\sum\limits_{m\in Z}\mathsf{k}\,(m,1)+\sum\limits_{m\in \check{Z}}\mathsf{k}\,(m,0)}{|M|}$$

Injective and general relational items scoring

The task of the examinee is to find a relation between two sets. In the relation as in the subset of Cartersian product is true that a given pair either belongs or doesn't belong to the relation. If we decompose the wanted relation to the particular members and consider each member extra as an elementar dichotomy item, we can apply methods of closed multiple-choice items during the TIS.

Let A, B are sets. In binding items we usually call A as the set of objects and B as the set of groups. Then the answer O ⊆ A × B = {(a, b) | a ∊ A, b ∊ B}. In the most general case we define for each member of Cartesian product A × B and each case chosen – not chosen, the correctness coefficient κ, thus mapping κ: A × B × {0, 1} → <-1,1>.

Particular closed relational item score is in this most general case given by formula

$$\mathsf{s} = \frac{\sum\limits_{(a,b)\in O}\mathsf{k}\,(a,b,1)+\sum\limits_{(a,b)\in A\times B\setminus \mathrm{O}}\mathsf{k}\,(a,b,0)}{|A\times B|}$$

Narrow Open Items Scoring

Even in the case of open items, we can widen the TIS problems from the classic aspect "correct answer / incorrect answer" to the more general case, distinguishing correctness of particular possible answers. We aren't able to cover the whole set of possible answers, thus all words of given language with the correctness coefficient. Thus we have to assess an implicit value which will be used in cases where correctness coefficient is not defined for the examinee's answer.

The final evaluation of the narrow open item, answered by the examinee with the word *s*, is therefore given by this simple formula

$$\mathrm{s} = \mathrm{k}(s)$$

Joining Items Scoring

Let's now deal with pairing items except the sorting ones. The examinee has to find a one-to-one pairing between two sets of equal cardinality. Because solving of the pairing items is from the examinee's aspect quite a difficult mental operation and due to the relation of one member of pairing to another isn't possible to decompose it to the fundamental cases we are able to assess the correctness coefficient for, we have to assess it for each pairing.

Let's remind that A and B are the sets of objects and the examinee looks for a relation between them. The correctness coefficient is defined for every pairing, thus for every member of the set *Bij*(A,B). Then, if the answer of examinee is O ∈ *Bij*(A,B), the particular joining item score is given by formula

$$\mathrm{s} = \mathrm{k}(O)$$

We didn't solve the problem of joining items scoring; we just moved it one level down. Let's now think about counting κ for given one-to-one pairing, which represents the examinee's answer and explicates various methods of computation. In the rest of this part, we will denote the correct answer, thus the correct one-to-one pairing, with symbol P. Then P ∈ *Bij*(A,B) and especially κ(P) = 1.

Number of correct pairs

Let's denote the number of members of the set A, respectively B, by the symbol *k*. Thus *k* = |A| = |B|. The examinee has to find *k* pairs. If we use the number of correct pairs as the only criterion, we can assess the correctness coefficient according to the formula

$$\mathrm{k}(O) = \frac{|O \cap P|}{k} \quad ,$$

thus as the ratio of the cardinality of intersection of pairings (thus sets) O and P representing the number of correct pairs, and number of all pairs.

Other methods

The number of correct pairs is the mostly implemented method for correctness coefficient assessment. Furthermore, we can look at the joining items as if they were a special case of relational items and take advantage of the assessment of correctness coefficient for each pair extra. The fact, that it is a pairing item, will serve just to the examinee telling him/her that each member of the group set has to be used just once.

In case of finding the suitable sort of the set A we can look to the pairing items as if they were sorting items and use methods explained in following part.

Sorting Items Scoring

As we mentioned above we assess the correctness coefficient for each permutation of alternative set extra. Let *Per*(M) be the set of all permutation of the set M and O ∈ *Per*(M) be the examinee's answer. Particular item score is given by obvious formula

$$\sigma = \kappa(O)$$

Then, let P ∈ *Per*(M) be the correct answer, thus the permutation representing the proper ordering. Then obviously

$$\kappa(P) = 1$$

Let's now deal with methods for the assessment of κ for various permutation of the sorted objects set.

Permutation difference operation

Now our goal is to define such operation with the set of all permutations of a finite subset of naturals, which expresses the rate of their difference and corresponds with the rate of difference between the examinee's answer and the correct one. It could be suitable to consider such definition of this operation, which satisfies the metric axioms and therefore expresses the distance between two permutations. The test item score was counted from the distance between the examinee's answer and the correct one in this case. Let's call this operation *difference* and denote it by the symbol *d*.

Let's leave the definition of the difference operation now and deal with conversion of its result to the interval <-1, 1>, thus assessment of particular item score. For a correct answer, which isn't distinct of the correct answer, and the operation result will be zero, the wanted function has to be return value 1. Value 0 should represent the maximal possible mistake, thus maximal possible difference of permutations O and P. So we're gaining formula

$$s = 1 - \frac{d(O,P)}{d_{\max}(P)},$$

where *d*(O, P) is difference of permutations O and P and d_{max}(P) is the difference of the most different permutation from the permutation P.

Now, to define the mentioned operation remains. We can define it by various ways which will correspond with various ways of evaluation. All of them will be suitable for substitution in an equation mentioned above.

Sum method

Let A = (a_1, a_2, …, a_n) and B = (b_1, b_2, …, b_n) are two permutations of the set $\mathbf{N}_n$. Their difference *d* is defined by formula

$$d = \sum_{i=1}^{n} |a_i - b_i|$$

For B = P = (*1, 2, …, n*) we're gaining more simple formula

$$d = \sum_{i=1}^{n} |a_i - i|$$

This method counts – for each member of the set of alternatives – how far is this member from its correct position. The sum of these differences (distances) then expresses the total rate of correspondence or difference of whole permutations.

We can gain maximal possible value d_{max} for opposite permutations and its value is equal to

$$d_{\max} = \sum_{i=1}^{n} (n - (n - i + 1)) \quad ,$$

and then

$$d_{\max} = \frac{n^2 - (n \bmod 2)}{2}$$

Sequential method

Let A = (a_1, a_2, … a_n) is permutation of the set $\mathbf{N}_n$. Its difference d from the permutation P = (1, 2, …, n) is defined by formula

$$d = (n-1) - \sum_{i=1}^{n-1} j(a_{i+1} - a_i) \quad ,$$

where

$$j(x) = \begin{cases} 1 \ for \ x = 1 \\ 0 \ \ for \ x \neq 1 \end{cases}$$

The result of this method is the sum of lengths of correct subsequences regardless of their position in the sequence. Because maximal possible subsequence length is n – 1, this expression appears at the beginning of the right side of the equation and ensures that for maximal possible subsequence, the result of operation d will be zero. For permutation without any correct subsequence, we'll gain the result d_{max} = n – 1.

Euclid method

If we look at the permutation of n naturals as if they were points in the n-dimensional Euclid space, we can express the difference of two permutations as the Euclid distance of appropriate points.

Let again A = (a_1, a_2, …, a_n) and B = (b_1, b_2, …, b_n) are two permutations of the set $\mathbf{N}_n$. Let's define their difference d using the Euclid method by formula

$$d = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

Maximal possible distance between points having no coordinate greater than n, is the diagonal of a n-dimensional cube with the edge of length $n - 1$. To simplify, let's omit the fact that opposite vertexes don't represent points appropriate to the sequences O and P. Then we'll gain

$$d_{\max} = \sqrt{n \cdot (n-1)^2} = (n-1)\sqrt{n}$$

Greatest mistake method

This method is based on the difference of position in the sequences in such member, whose position difference is biggest. Let A = (a_1, a_2, …, a_n) and B = (b_1, b_2, …, b_n) are two permutations of the set $\mathbf{N}_n$. Their difference d is then defined by the formula

$$d = \max \left\{ |i - j|;\, a_i = b_j;\, a_i, b_j \in \mathbf{M} \right\}$$

The maximal difference in members' position is obviously

$$d_{max} = \mathrm{n} - 1$$

Individual evaluation method

This method is based on the supposal that each permutation represents a specific mistake of different kind and so unique function formula assessing the score just based on the permutations doesn't exist. The evaluation function is in this case defined by the enumeration and is based on semantics of the members of the set **M**.

Penalty guessing in multiple-choice items

One of the most important properties of the multiple-choice items scoring is their restrictivity, thus penalizing incorrect or not totally correct answers. Typical restrictive action is penalty guessing in multiple-choice items, thus simple penalization of incorrect answers. Sometimes it is necessary to distinguish the rate of mistake and decrease (or increase) the penalization.

Using multiple-choice type 1 of n brings the risk that the examinee is guessing correct answers. This risk can be taken out by both sufficient number of alternatives in each test item and sufficient number of items in the test. Using any number of alternatives and any number of items, we can use penalty for guessing during the scoring. This means to adjust the score in such way, than eventual guessing of the examinee isn't displayed in the result, or is displayed as less as possible (Davis, 1993).

During the penalty guessing, we give to the examinee points according to the number of mistakes he/she has made. We result from the fact that those who guess, make mistakes more often than those who really solve the task and answer only in case of knowing the answer (Mužić, 1993). Correction of score can be done according to the formula

$$S_o = S_n - \frac{N}{n-1},$$

where S_o is so called corrected score, S_n is original score, N is number of incorrect answers in the test and n is number of alternatives offered in one item (Stalker, 1968). The mentioned formula is valid for posterior adjusting of the total test score. We can reformulate it for usability during assessment of a particular item score:

$$\mathrm{s} = \begin{cases} 1 & \text{for correct alternative} \\ -\dfrac{1}{n} & \text{for incorrect alternative} \end{cases}$$

If a particular item score is equal to the correctness coefficient of a chosen alternative, we are gaining the formula for counting correctness coefficient of the alternative m ◎ M.

$$\mathrm{k}(m) = \begin{cases} 1 & \text{for correct alternative} \\ -\dfrac{1}{n} & \text{for incorrect alternative} \end{cases}$$

Changing the correctness coefficient value in alternatives representing exceedingly rough mistake (or in alternatives whose choosing is not serious problem to respect assessed targets of achievement test) is the right of the examiner.

Let's stress that penalty guessing can be done even in dichotomy items as a special case of polytomy multiple-choice items. The denominator is then 2 – 1 = 1 and the way of evaluation matches positive points for the correct answer and negative points for the incorrect one.

If we realize the guessing penalization, we have to call examinees' attention to it. In that case, if they don't know the correct answer or they aren't sure, omitting the answer is advantageous for them. If we don't warn the examinees, they can have valid objections. The scoring rules should be known before the test in detail.

The experts dispute about guessing penalty. Theoretical pedagogues feel the base contradiction in the behavior of the examinee who doesn't know the correct answer. Is better to try it or to confess the ignorance? Not even in real life the clear answer doesn't exist. This is the examiner's decision to judge if guessing of ignorance confession is better (Stalker, 1968).

Using correctness coefficient: An example

Due to the length limitations of the paper there is not enough space to describe differences between classical scoring and scoring with the help of the correctness coefficient properly. Let's illustrate new concept on a short test dealing with set theory and propositional and predicate logic, containing 4 simple closed multiple choice questions (type 1 of n).

1. How many elements the empty set contains?
   a. 0
   b. 1
   c. 2
   d. sometimes 0, sometimes 1
2. Choose the set operation which is *not* commutative:
   a. union
   b. intersection
   c. symmetric diference
   d. Cartesian product
3. Which of following sentences can be considered as a *proposition*?
   a. Is this true?
   b. Send this letter!
   c. I'm at school right now.
   d. What time is it?
4. A symbol ∀ denotes
   a. universal quantifier
   b. existential quantifier
   c. common quantifier
   d. logical compound

Let's now discuss suitable values of the correctness coefficient of proposed possibilities.

Question no. 1: Answer "a" is the correct one. Answer "b" represents typical student mistake flowing from the confusion between the number of elements and number of subsets. Answer "c" is total nonsense as well as answer "d" which is nondeterministic. Some pedagogues with tolerance to some type of confusions can then propose κ values for example as (1, ½, 0, 0).

Question no. 2: Answer "d" is the correct one. The uncorrectness of other answers is evident and there is no possibility to lower the penalization in case of choosing one of them. There for κ values will be (0, 0, 0, 1)

Question no. 3: The correct answer is "c", some examiner may be tolerant if an examinee chooses answers "a" or "b". Then s/he may set κ as (¼, ¼, 1, 0)

Question no. 4: The difference between "universal" and "common" may not be clear enough. Therefore the answer "c" can be considered as "almost correct, whereas the answer "a" is totally correct. Answer "b" may show the temporary confusion of an examinee ad answer "d" is nonsense. Therefore we can set κ as (1, ¼, ½, 0)

Let's imagine examinee's answers 1b, 2a, 3c, 4c. Without using the correctness coefficient the total score would be 0+0+1+0 = 1 out of 4. With using the correctness coefficient the total score rises to ½+0+1+½ = 2 out of 4. The difference would be higher in case of students choosing "almost correct" answers.

Let's emphasize that the assessment of the correctness coefficient values is the examiner's responsibility and that there can be arguments about the values. This was not the goal of the paper, the paper just provide the tool to more accurate achievement test scoring.

The example from mathematics was chosen just due to the author's affiliation. It is obvious that in exact sciences (like mathematics) there is almost no space for middle values. However, in social and human sciences where correct answers are not so clear, the potential is much bigger.

Conclusion

In this contribution, we set up the division of commonly used achievement test items types and summarized theoretical basics of test item scoring (TIS). From this base, we built the formalization of test items scoring based on the correctness coefficient.

Although the found way is very general and provides huge freedom do the examiner, hard work to assess the correctness coefficient in all relevant entities and related error risk is connected with this freedom. Only future pedagogy practices will show if the scoring methods described in this paper are proper or not.

Acknowledgements

The paper is supported by the Ministry of Education of the Czech Republic (the Research Proposal VZ MSM 6215648904/03/03/01).

References

Bloom, B. S. (ed.) (1956). *Taxonomy of Educational Objectives*. Vol. 1: Cognitive Domain. New York: McKay, 1956.

Booth, R.; Clayton, B.; Hartcher, R.; Hungar, S.; Hyde, P.; Wilson, P. (2003): *The development of quality online assessment in vocational education and training*. Leabrrook: Australian National Training Authority, 2003. ISBN 1-74096-155-2. Online http://www.ncver.edu.au/research/proj/nr1F02_1.pdf [cit. 2006-09-15].

Brusilovsky, P. (1999). *Adaptive and intelligent technologies for web-based education*. In C. Rollinger, & C. Peylo (Eds.). Intelligent systems and teleteaching [Special issue]. Künstliche Intelligenz, 4 (pp. 19–25). Available: http://www2.sis.pitt.edu/~peterb/papers/KI-review.html.

Davis, B.G. (1993): *Tools for Teaching*. San Francisco: Jossey-Bass, 1993. ISBN: 1-55542-568-2.

Foltýnek, T. (2005): *The electronic test results as a feedback for teachers*. In Trends in e-learning – Belcom'05 proceedings of papers. Prague, CVUT, 2005. ISBN 80-01-03203-5.

Foltýnek, T.; Motyčka, A. (2006): On *the achievement test items scoring*. In Efficiency and responsibility in education – proceedings of papers 2006. Praha: ČZU, 2006. ISBN 80-213-1509-1.

Hobgood, B.; Thibault, M.; Walbert, D. (2004): *Kinetic connections: Bloom's taxonomy in action*. Online http://www.learnnc.org/articles/bloom0405-1. [cit. 2006-08-31]

Linrace, J.M.; Wright, B.D. (1995): *How to Assign Item Weights*. In Rasch Measurement Transaction, Vol. 8:4. Washingto, American Educational Research Association> 1995. ISSN 1051-0796.

Mužić, V. (1993): *How to Outwit a Test*. Zagreb: Školske novine, 1993. ISBN: 953-160-002-3.

Payne, D.A. (1968): *The specification and measurement of learning outcomes*. Waltham: Blaisdell Publishning Company, 1968.

Segall, N.; Doolen, T.L.; Porter J.D. (2005): *A usability comparison of PDA-based quizzes and paper-and-pencil quizzes*. In Computers & Education, 45(4), pp. 417 – 432. Elsevier, 2005. ISSN 0360-1315.

Stalker, M.J. (1968): *The Penalty for Not Guessing*. In Journal of Educational Measurement, Vol. 5, No. 2 (Summer, 1968), pp. 141-144. JSTOR, 1968. ISSN 00220655.

Legends to figures

Figure 1: Relation of bases to the TIS

Figure 2: Division of test items types