

The New Features of the ExaMe Evaluation System and Reliability of Its Fixed Tests

P. Martinková¹, K. Zvára jr.², J. Zvárová¹, K. Zvára³

¹EuroMISE Centre, Department of Medical Informatics, Institute of Computer Science AS CR, Prague, Czech Republic

²EuroMISE s.r.o., Prague, Czech Republic

³Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Summary

Objectives: The ExaMe system for the evaluation of targeted knowledge has been in development since 1998. The new features of the ExaMe system are introduced in this paper. Especially, the new three-layer architecture is described.

Besides the system itself, the properties of *fixed tests* in the ExaMe system are studied. In special detail, the *reliability* of the fixed tests is discussed. The theory background is explained and some limitations of the reliability are pointed out.

Methods for Estimation of the Reliability: Three characteristics used for estimation of reliability of educational tests are discussed: *Cronbach's alpha*, *standardized item alpha* and *split half coefficient*. The relation between these characteristics and reliability and between characteristics themselves is investigated. In more detail, the properties of Cronbach's alpha, the characteristics mostly used for the estimation of reliability, are discussed. A confidence interval is introduced for the characteristics.

Results and Conclusions: Since 2000, the serviceability of the ExaMe evaluation system as the supporting evaluation tool has been repeatedly shown at the courses of Ph.D. studies in biomedical informatics at Charles University in Prague. The ExaMe system also opens new possibilities for self-evaluation and distance learning, especially when connected with electronic books on the Internet.

The estimation of reliability of tests contains some limitations. Keeping them in mind, we can still get some information about the quality of certain educational tests. Therefore, the estimation of reliability of the fixed tests is implemented in the ExaMe system.

Keywords

Education, evaluation, Internet, reliability, bioinformatics

Methods Inf Med 2006; 45: 310–5

1. Introduction

Training and education in medical informatics, statistics and epidemiology provided by the EuroMISE Centre originated in the joint European project titled "Education in the Methodology Field of Health Care" of the Tempus-Phare program which ran in the years 1993-1995. The teaching scheme has been developed in cooperation among eleven EU universities, the Charles University in Prague and the Institute of Computer Science of the Academy of Sciences of the Czech Republic [1, 2]. The EuroMISE guidelines place considerable emphasis on the need for a wide range of knowledge, including a thorough understanding of health concepts [3-6]. Moreover, education and training in this field have supported bridging the gap between technological innovation and health care practice. European experience using new information technologies in education and training was described in [7], where the experience with the EuroMISE courses running in the years 1993-1995 was also given in [8].

The ExaMe system for the evaluation of targeted knowledge has been developed since 1998 [9]. The idea of the system is based on multiple-choice questions, but with no prior restrictions on the number of selected answers. The only restriction is that at least one answer is correct and at least one wrong. This new idea has led to new concepts of standardization of test results and many new theoretical developments. The ExaMe evaluation system is an important part of education and training at the EuroMISE Centre. The ubiquity of the

Internet and its World Wide Web applications made it possible to realize the new educational goals in an innovative and creative way.

The International Medical Informatics Association (IMIA) agreed on international recommendations in medical/health informatics education. The IMIA recommendations centre on educational needs for health care professionals to acquire knowledge and skills in information processing and information and communication technology. The educational needs are described as a three-dimensional framework. The dimensions are:

- professionals in health care (physicians, nurses, HMI professionals, ...),
- type of specialization in health and medical informatics (IT users, HMI specialists), and
- stage of career progression (bachelor, master, ...).

The recommendations were developed by the IMIA Working Group 1: Health and Medical Informatics Education and published in *Methods of Information in Medicine* in 2000 [10]. These recommendations have been translated into several other languages. Original documentation is available at <http://www.imia.org/>. The Czech translation was published in the Czech journal *Physician and Technology* in 2001 [11]. Recommendations are given for different types of courses/course tracks of educational programs in medicine, nursing, health care management, dentistry, pharmacy, public health, health record administration, and informatics/computer science as well as for dedicated programs with bachelor, master or doctorate degrees.

2. ExaMe System

The first version of the ExaME system was applied in the frame of the European project IT EDUCTRA (<http://www.cordis.lu>). The second version of the system ExaMe was developed in the period 2001-2002. The system is in routine use in the courses organized by the EuroMISE Centre. Also, the ExaME system is applied as a supporting evaluation tool at courses of Ph.D. studies in Biomedical Informatics at Charles University in Prague [2].

2.1 The Evaluation by the ExaMe System

The evaluation by the system ExaMe is based on the knowledge base created for a specified target, mostly for knowledge covered by a special course. The *knowledge base* consists of *generalized multiple-choice questions* (number of answers is not limited, at least one answer is true and at least one false). For each question, scores define its importance and difficulty. The *importance* is given in five categories (very important, important, moderately important, little importance, not important). Similarly, the *difficulty* is given in five categories (very easy, easy, standard, difficult, very difficult). For each answer, its weight is defined. The *weight* is given by number of points (integers) from the range -5 to $+5$, except 0. Weights should be positive for true answers and negative for false answers. Moreover, *explanations* of true or false answers are formulated.

When a student answers a question (by marking the answers he supposes to be right), the system calculates his *standardized score* for the question (in the range -1 and 1), using the weights of the offered answers. After a student has answered the whole test, the *total standardized score* (also in the range -1 and 1) is computed by the system, using the standardized scores for each question and questions' importance. A student should not pass the exam unless his/her total standardized score is higher than zero.

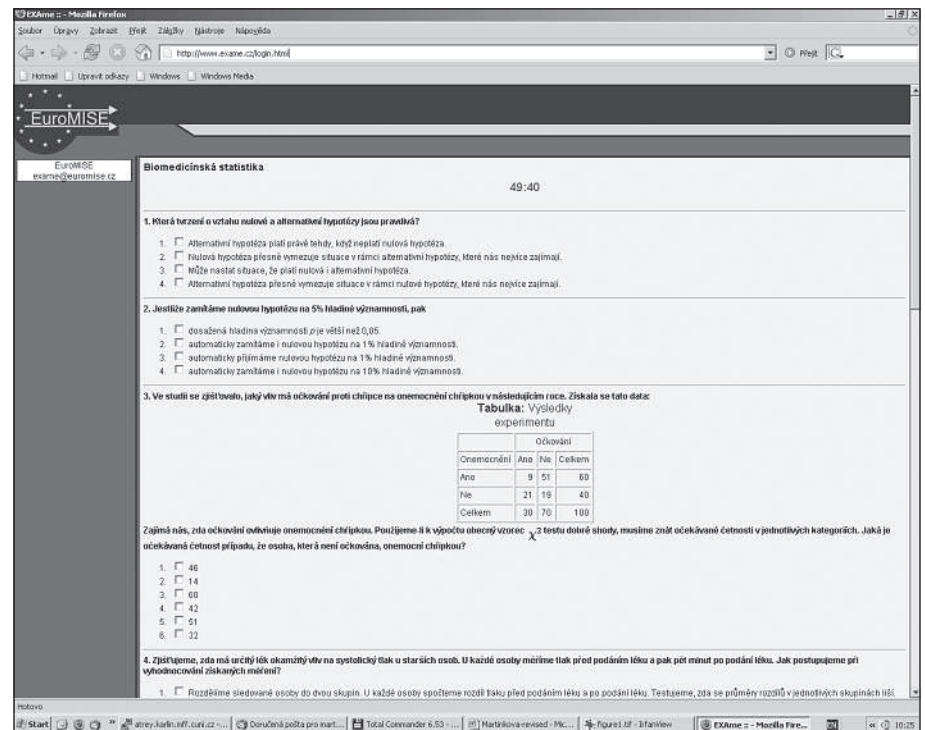


Fig. 1 ExaMe: student's screen (fixed test for course of biomedical statistics, in Czech)

2.2 The New Three-Layer Architecture

Nowadays, the system ExaME works at three levels. The first level is a *supervisor level*, where a supervisor (e.g. group of teachers, defined body) creates generalized multiple-choice questions for a given target, marks true and false answers and formulates explanations. Up to now four knowledge bases have been created. They cover the targeted knowledge of corresponding courses based on Czech electronic books on the Internet (<http://www.euromise.cz>).

The second level is a *specification level*. In this level, the questions' importance and difficulty and answers' weight for a knowledge base are specified. For example, a different specification can be used in the evaluation of medical informatics knowledge for medical students and students of informatics. Automated tests can be created from any knowledge base with specification.

The third level is a *teacher level*. A teacher can create a *fixed test* from a knowl-

edge base with specifications and make the evaluation based on her/his needs.

2.3 Fixed and Automated Tests

From the knowledge bases, two types of tests may be created: the automated and the fixed test. The *fixed test* is appropriate for the evaluation of the group of students in computer classrooms connected to the Internet. In the third (teacher) level of the ExaMe system the fixed test can be easily created by a teacher simply by choosing appropriate questions and answers from a specified knowledge base.

The *automated test* is appropriate for self-evaluation on remote places. The student can ask for an evaluation by an automated test on chosen topics and with chosen overall difficulty. Afterwards, the test is immediately generated by the system out of the appropriate specified knowledge base.

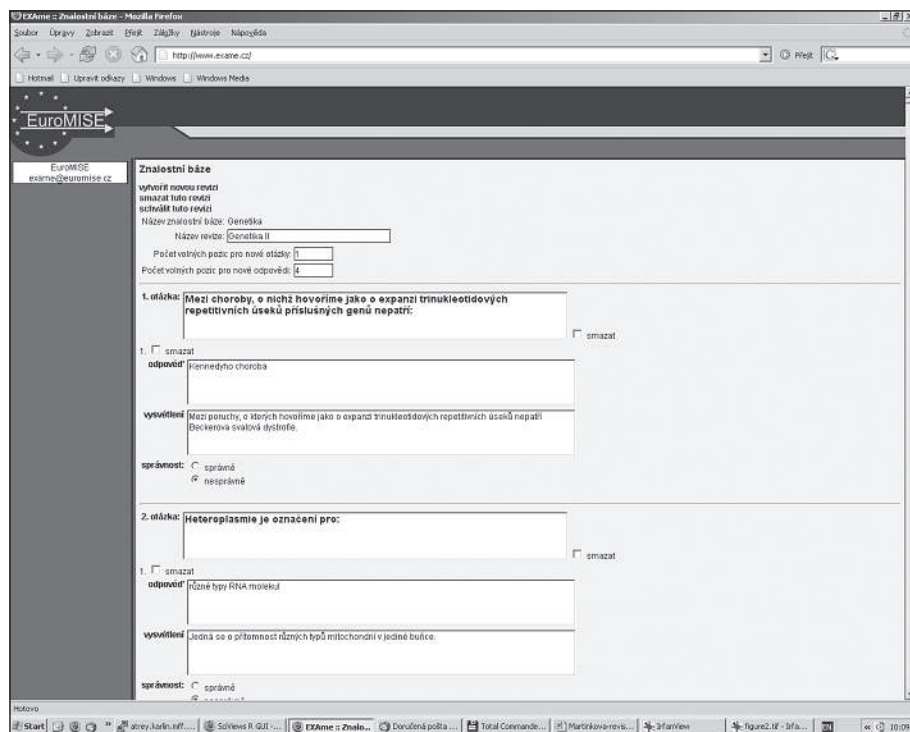


Fig. 2 ExaMe: administrator's screen (supervisor level)

2.4 Main Benefits of the ExaMe System

A quick administration of knowledge evaluation is the main strength of the ExaMe system. From the knowledge base with specified importance of questions and weights of answers, it is very fast and easy for a teacher to prepare a test optimal for his/her special group of students. After a student answers the whole test, the system automatically evaluates the test and a teacher can print a corrected test for each student, where the mistakes are highlighted, explanations of correct answers are given, and the standardized score for each question and the total standardized score are computed.

The ExaMe system is also a modern tool for self-evaluation and self-study. Besides the correction of automatically generated tests, the system gives suggestions on which subject matters should be studied more deeply, and offers links to the electronic textbooks.

3. Objectives – Reliability of the Fixed Tests in the ExaMe

Besides developing the system itself, our next goal is to suggest and research the statistical tools for analyzing the quality of a fixed test and quality of its items. We attempt to suggest the tools, which would help the teacher create a fixed test optimal for the needs of his/her course.

During the last decades the reliability of educational tests has often been examined. In connection with reliability, the characteristics for its estimation, such as Cronbach's Alpha, have often been mentioned. To fulfill our goal, in the first part of our research we have critically examined the statistical background of the reliability and of characteristics used for its estimation. We have investigated the relationship between these characteristics and reliability and between characteristics themselves. Shortcomings and limitations of the theory are discussed in this article. We have also implemented the researched estimates of reliability into the ExaMe system.

4. Methods for Estimation of the Reliability of the Fixed Tests

4.1 Definitions of the Reliability

When examining the reliability of an educational test, we assume that the *observed score* – the score reached by a student – is composed of two components: the *true score* and an *error term*. Mathematically written:

$$X = T + e.$$

The error is assumed to be unbiased, $E(e) = 0$, the random variables T and e are assumed to be independent. From these conditions, we can deduct about variations that

$$\text{var}(X) = \text{var}(T) + \text{var}(e).$$

Obviously we prefer tests whose error variance is small relative to the observed score variance. The reliability of measurement X can further be defined as

$$R_m = 1 - [\text{var}(e)/\text{var}(X)] = \text{var}(T)/\text{var}(X).$$

The reliability is a number between 0 and 1. Measurement is considered to be *Sreliable* when the value of reliability is close to 1.

It is easy to show (see for example [13]) that reliability can be expressed as the squared value of the correlation between the observed score and the true score, $\text{corr}^2(X, T)$. This means that the reliability of educational test can be understood as the strength of the relationship between the score reached by a student and his true knowledge.

Using a similar approach (see again [13]), we can show that if we would have two independent measurements X, X' of the same property on the same object, the reliability could be expressed as the correlation between these two measurements, $\text{corr}(X, X')$. As a result, the reliability reflects to what extent the educational test gives the same result when taken repeatedly by the same person under the same conditions.

These representations are very intuitive but they are not useful when estimating

the reliability of educational tests because they cannot be directly estimated from the observed data. We can't estimate the variance $\text{var}(e)$, the true score T , nor the knowledge of a student by the same educational test twice and independently. With regard to this fact, when estimating the reliability, the fact that the educational test is composed of m items is very often taken into account.

4.2 Reliability of Composite Measurement

Let's study the situation when the observed measurement (the total score) is composed of m measurements, $X = X_1 + \dots + X_m$, and also the true score is composed of m true scores, $T = T_1 + \dots + T_m$. We may assume that when measuring knowledge of any student, for the score observed on the j -th item it holds:

$$X_j = T_j + e_j, \quad j = 1, \dots, m,$$

where the error terms e_j are mutually independent and independent on the true scores T_k for $k = 1, \dots, m$, and having the same variance σ_e^2 . Using these assumptions, the reliability of the composite measurement X can further be expressed as:

$$R_m = [\text{var}(T)/(\text{var}(T) + m\sigma_e^2)].$$

An important, often used, but very rarely satisfied assumption for an educational test is that its items are *T-equivalent*, which means by the definition that for the items' true score the following holds simultaneously:

$$\text{var}(T_1) = \dots = \text{var}(T_m)$$

$$\text{corr}(T_j, T_k) = 1, \quad j, k = 1, \dots, m.$$

When the items of an educational test are T-equivalent, then all the reliabilities R_1 of items do equal and the expression of reliability of the whole test can be simplified (for detailed proof see [14]) in a well-known *Spearman-Brown formula*:

$$R_m = m R_1 / [1 + (m - 1)R_1].$$

4.3 Characteristics Used for Estimation of the Reliability

As noted above, in the case of educational tests, the reliability cannot be estimated from its definition. With regard to this fact, some characteristics were introduced and reliability is often estimated by estimation of these characteristics. The most used are *Cronbach's alpha*, *standardized item alpha* and *split half coefficient*.

Split half coefficient is based on the interpretation of reliability as the correlation of two independent (or parallel – see [13]) measurements: we split the test (assuming it has even number of items m) into two halves, either randomly or having the corresponding items as much alike as possible. Let's label the first half (the $m/2$ -tuple of items) s and the complement half of items $-s$. We can then represent the reliability by the correlation of the total observed scores of the two parts $\rho_{s, -s}$. Or even better we can represent it by the average of all such correlations ρ . Taking into account the number of items (see the Spearman-Brown formula), we get the characteristics for reliability called (*average*) *split-half coefficient*:

$$R_{SH} = \frac{2\rho}{(1 + \rho)}.$$

Standardized item alpha uses the Spearman-Brown formula in a different way. The educational test is split into single items and for each couple of items k, l the correlation between their observed scores $\rho_{k, l} = \text{corr}(X_k, X_l)$ is considered. The reliability is then represented by the average of all such correlations $\bar{\rho}$, which is corrected by the Spearman-Brown formula. We get the characteristics of reliability, which is called *standardized item alpha*:

$$\alpha_N = \frac{m\bar{\rho}}{1 + (m - 1)\bar{\rho}}.$$

And finally, *Cronbach's alpha* is defined as:

$$\alpha = \frac{m}{m - 1} \frac{1 - \sum_j \text{var}(X_j)}{\text{var}(X)}.$$

What is common for these three characteristics is the fact that it is easy to estimate them from the data, simply by using *sample*

variances and *sample correlations* instead of their theoretical versions found in the definitions. Also a common property of these characteristics is the fact (for detailed proof see [14]) that when the items of an educational test are T-equivalent, they are equal to the reliability. While the T-equivalency is often not satisfied, it is substantial to critically research the properties of these characteristics and relationships between these characteristics and reliability before accepting them as suitable for estimating the reliability. This was the goal of our study.

5. Results

5.1 Relationships between Reliability and the Characteristics

First we have examined the relationships between the three mentioned characteristics and the reliability. As written above, with the assumptions of T-equivalence, all three characteristics equal the reliability. This is not true in a general case.

As proved in [15], the Cronbach's alpha is generally the lower bound of the reliability, $\alpha \leq R_m$. Which gives a conclusion that getting a low estimate of reliability by using Cronbach's alpha does not necessary mean that the educational test is not reliable. The possible reason for low estimate could be the internal "inconsistency" of items.

Unfortunately, no similar relationship is known between reliability and split half coefficient or between reliability and standardized item alpha. We have nevertheless corrected the false statement about the equality of Cronbach's alpha and the average split half coefficient [13]. We proved in [14] that in the general case, Cronbach's alpha is a lower bound of the average split half coefficient, $\alpha \leq R_{SH}$. This gives us a conclusion that the estimate using Cronbach's alpha is more "conservative". In spite of getting a higher estimate of reliability when using the split half coefficient, the question remains, whether in general cases do we have the right to use R_{SH} at all.

The relationship between Cronbach's alpha and the standardized item alpha can

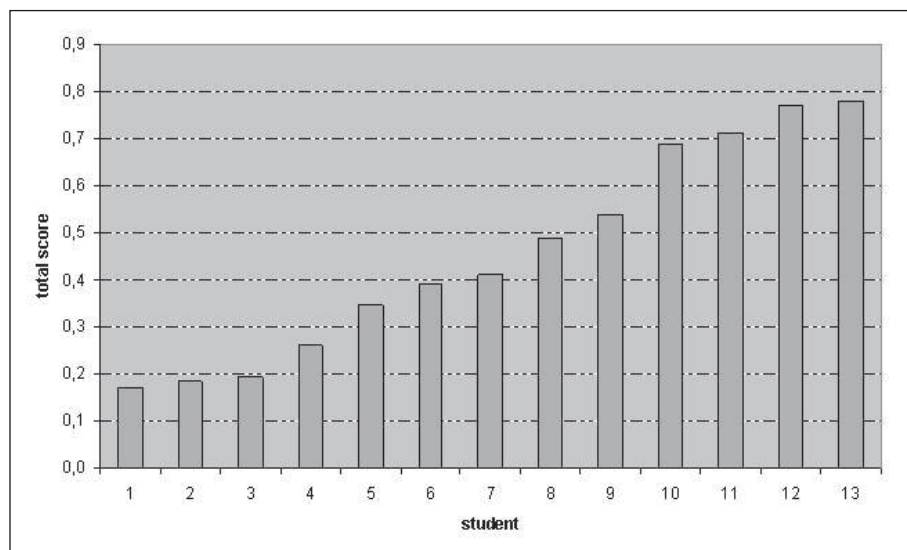


Fig. 3 The test results (total scores) of 13 students of the biomedical statistics course, May 2004

already be deducted from their names, and it can be proven very simply when rewriting the formula for the standardized item alpha: α_N is actually Cronbach's alpha used on data where the scores for each item are standardized by their standard deviations. When a big difference between the sample estimate of Cronbach's alpha and of standardized item alpha occurs, it should be a warning sign for us that the assumptions of T -equivalence are violated.

To summarize, when studying the relationships between reliability and characteristics used for its estimation, in general cases the only clear relationship was found between reliability and Cronbach's alpha.

5.2 Other Properties of Cronbach's Alpha

We have further studied Cronbach's alpha using the tools of two-way ANOVA. As shown more closely in [14], the sample estimate of Cronbach's alpha (and thus an estimate of the reliability) can be written as:

$$\hat{\alpha} = \hat{R}_X = \frac{MS_T - MS_e}{MS_T} = 1 - \frac{1}{F_T},$$

where MS_T and MS_e are the mean sums of squares and F_T is statistics widely used for testing the hypothesis $\text{var}(T) = 0$, when additionally assuming the normal distribution

of all variables. This notation gives us two very important properties of our estimate:

- The greater the estimate of reliability is the better the educational test can distinguish between the students.
- The estimate equals one, if and only if there exist constants $a_i, b_j, i = 1, \dots, n, j = 1, \dots, m$, so that the score reached by the i -th student in the j -th item can be written as $a_i + b_j$. This means that in this case, to get all the information about the students, one item would be enough.

The first founding shows that Cronbach's alpha should rather be understood as a coefficient of internal consistency (this was actually the purpose for which it was constructed by Cronbach). The second statement says that when getting too high an estimate of Cronbach's alpha, we could actually think of lowering the number of items.

When assuming the normality of all variables, the notation gives another important result we came with: the confidence interval for Cronbach's alpha. When the hypothesis $\text{var}(T) = 0$ holds, it is well known that the statistic F_T has the Fisher-Snedecor distribution $F_{n-1, (n-1)(m-1)}$. As shown in detail in [14], the $(1-\gamma)\%$ confidence interval for R_m is $I(1-\gamma) = [R_{min}, R_{max}]$, where

$$R_{min} = \max \left(0, 1 - \frac{F_{n-1, (n-1)(m-1)} \left(1 - \frac{\gamma}{2} \right)}{F_T} \right),$$

$$R_{max} = \min \left(1, 1 - \frac{F_{n-1, (n-1)(m-1)} \left(\frac{\gamma}{2} \right)}{F_T} \right),$$

and where $F_{n-1, (n-1)(m-1)}(1-\gamma/2)$, $F_{n-1, (n-1)(m-1)}(\gamma/2)$ are the critical values of the Fisher-Snedecor distribution with $(n-1)$ and $(n-1)(m-1)$ degrees of freedom.

5.3 Properties of Reliability Itself

Finally, we have studied the properties of reliability itself. We have pointed out two main shortcomings of the reliability:

- Reliability is sample-dependent.
- Reliability is dependent on the number of items.

The first statement can already be seen from the definition of the reliability. It implies that a certain test can have different reliability when given to a population with high variability of tested knowledge and when given to a population with low variability of the knowledge.

The second statement is touched by the Spearman-Brown formula. It says that by adding suitable items to the test, the reliability could approach 1 as close as we would want to. When comparing reliabilities of two educational tests, which in principle cannot have the same number of items (for example a language test and a math test), we should bear this property of reliability in mind.

5.4 Implementation into the ExaMe Evaluation System

Keeping all the limitations of the reliability and its estimates in mind, the mentioned characteristics can still tell us something about properties of an educational test. Therefore, we have implemented the esti-

mation of reliability into the ExaMe system. Procedures were programmed in free statistical software R (see [16]). Estimates are done from data – scores obtained on certain fixed tests by certain group of students – chosen by the teacher.

6. Conclusions

The serviceability of the ExaMe evaluation system as the supporting evaluation tool has been repeatedly shown since 2000 at the courses of Ph.D. studies in biomedical informatics at Charles University in Prague [12]. The system offers the quick and convenient administration of evaluation of groups of students. The ExaMe system also opens new possibilities for self-evaluation and distance learning, especially when connected with electronic books on the Internet.

In the last years, the reliability of the educational test and its estimation using Cronbach's alpha was (at least in the Czech Republic) given more interest than the strength of this tool deserves. In this article we have discussed some shortcoming and limitations of the reliability. When using the tools for estimations of the reliability, we should keep these limitations in mind. Nevertheless when understanding the theoretical background, the characteristics as Cronbach's alpha can give us valuable information about properties of an educational test. Therefore, the estimation of reliability of the fixed tests is implemented in the ExaMe system.

7. Further Goals

Besides evaluation of *the quality of a fixed test* we would like to implement the evalu-

ation of *the quality of single items* into the ExaMe system. Two main frameworks deal with item analysis: classical test theory (CTT) and item response theory (IRT). In CTT, the characteristics such as item difficulty, item discrimination and probability of guessing are defined for multiple-choice question with only one true answer (see [17]). The IRT estimates discrimination, difficulty and probability of guessing of these items using logistic models. First we would like to generalize the characteristics for our type of multiple-choice questions where more than one answer may be true and thus not only scores of 0 or 1 can be obtained. Further we would like to compare the classical theory with the IRT. Finally we attempt to implement the estimation of characteristics into the ExaMe evaluation system. Latest results can be found in [18].

Acknowledgment

The work was partly supported by the Institutional Research Plan AV0Z10300504 of the Institute of Computer Science AS CR, partly by the research project MSM0021620839 of the Ministry of Education CR and partly by the ESF project CZ.04.3.07/4.2.01.1/0013.

References

- Zvárová J. Education in Methodology for Health Care – EuroMISE. *Methods Inf Med* 1994; 3.
- Zvárová J, Engelbrecht R, van Bommel JH. Education in medical informatics, statistics and epidemiology. *International Journal for Medical Informatics* 1997; 45: 1/2, 3-8.
- Van Bommel JH, Zvárová J. Knowledge, Information and Medical Education. Amsterdam: North Hollands Publ Comp; 1991.
- Haux R, Leven FJ, Möhr J, Protti DJ, editors. Special Issue on Health and Medical Informatics Education. *Methods Inf Med* 1994; 3.
- Hasman A. Recommendation for Medical Informatics Training in The Netherlands. *Methods Inf Med* 1994; 3.
- Haux R, Leven FJ. Twenty Years Medical Informatics Education at Heidelberg/Heibronn: Evolution of Specialized Curriculum for Medical Informatics. *Methods Inf Med* 1994; 33 (3); 285-9.
- Hasman A, Albert A, Wainwright P, Klar R, Sosa M, editors. Education and Training in Health Informatics in Europe. IOS Press; 1995.
- Haux R, Swinkels W, Ball M, Kanup P, Lun KC, editors. Special Issue on Health and Medical Informatics Education. Transformation of Healthcare through Innovative Use of Information Technology. *Int J Med Inf* 1997; 44.
- Zvárová J, Zvára K. Evaluation of Knowledge using ExaMe program on the Internet. In: Iakovidis I, Maglavera S, Trakatellis A, editors. User Acceptance of Health Telematics Applications. Amsterdam: IOS Press; 2000. pp 145-51.
- Haux R, et al. Recommendations of the International Medical Informatics Association (IMIA) on Education in Health and Medical Informatics. *Methods Inf Med* 2000; 39: 267-77.
- Preckova P, Straka L. Doporučení mezinárodní asociace pro medicínskou informatiku (IMIA) pro vzdělávání v lékařské a zdravotnické informatice. *Lékař a technika*; 2001. (Translation of the IMIA recommendations [9]).
- Zvárová J, Svačina S: New Czech Postgraduate Doctoral Program in Biomedical Informatics. In: Surjan G, Engelbrecht R, Nair P, editors. Health Data in the Information Society. Amsterdam: IOS Press; 2002, 90. pp 766-9.
- Sireci SG, Wainer H, Braun H. Psychometrics, Overview. In: *Encyclopaedia of Biostatistics*. New York: Wiley; 1998. pp 3577-3602.
- Rexová P. Spolehlivost měření. [Reliability of measurements. In Czech.] Diploma thesis. Department of Probability and Mathematical Statistics, Charles University, Prague 2003.
- Novick MR, Lewis C. Coefficient alpha and the internal structure of tests. Technical report, Education Testing Service. Princeton, New Jersey; 1966.
- The R Project for Statistical Computing. Available from: <http://www.r-project.org/>
- Suen HK. Principles of Test Theories. LEA, New Jersey; 1990. pp 71-129.
- Rexová P. Item Analysis of Educational Tests in System ExaME. In: Hák F, editor. Doktorandský den '04, Institute of Computer Science AS CR, MatfyzPress, Prague; 2004.

Correspondence to:

Patřicia Martinková
EuroMISE Centre
Institute of Computer Science AS CR
Pod Vodřenskou věží 2
182 07 Prague 8
Czech Republic
E-mail: martinkova@euromise.cz