OPEN ACCESS

EURASIA Journal of Mathematics Science and Technology Education
ISSN: 1305-8223 (online) 1305-8215 (print)
2017 13(10):6779-6788
DOI: 10.12973/ejmste/77042

The Development of Model and Measuring Tool for Specialists Accreditation in Area of Public Health Services

Zhanna M. Sizova
Sechenov First Moscow State Medical University, Moscow, RUSSIA

Tatyana V. Semenova
Ministry of Health, Moscow, RUSSIA

Victor I. Zvonnikov
The State University of Management, Moscow, RUSSIA

Alfiya R. Masalimova
Kazan (Volga region) Federal University, Kazan, RUSSIA

Zehra N. Ersozlu
Graduate School of Education, The University of Western Australia, Nedlands, WA, AUSTRALIA

Received 16 April 2017 ▪ Revised 31 August 2017 ▪ Accepted 15 September 2017

ABSTRACT

The main purpose of the paper is to present some theoretical approaches and some methods providing assessment optimization in specialists' accreditation in area of public health services. The results of research presented in this paper, include the model of multistage adaptive measurements and two methods for reliability and validity analysis, providing high justice decisions in accreditation and corresponding to requirements in High-Stakes Testing procedures. The assessment optimization intends for minimization time of assessment and for reliability and validity data increasing. For optimization the special model of measurements based on multistage adaptive testing is offered. The using of offered model in assessment design allows to realize the advantages of traditional adaptive testing and linear testing, while minimizing their disadvantages. So, this model is recommended as dominating for assessment in accreditation. For validity increasing in assessment in accreditation the approach based on Structural Equation Modeling is offered. This method allows to analyze the significance of relations between observed and latent variables that have any interpretation as causal effects, and to construct the model of their relations. The example of model of casual relations between disciplines, latent variables (competencies) and factors is offered. The model helps to increase construct and content validity of measuring tool using in public health services accreditation. The methods of reliability estimation in multistage measurements, offered in paper, has innovative character. It has branching structure as the value of reliability in multistage measurements depends not only on reliability of separate stages, but also from correlations between them. The presented approaches allow to increase validity and reliability of decisions in public health services specialists' assessment or in other spheres of assessment during accreditation.

Keywords: adaptive measurement, assessment, reliability, structural equation modeling, validity

Correspondence: Zhanna M. Sizova, *the Director of Federal Methodical Centre for Accreditation of Specialists, Head of the Department of Medical and Social assessment, Emergent and Outpatient Care, Sechenov First Moscow State Medical University, Moscow, Russia.*
sizova-klinfarma@mail.ru

Contribution of this paper to the literature

- The model of multistage adaptive testing providing efficiency increasing of assessment in the conditions of high reliability and high validity measurements during assessment in accreditation is offered.
- The approach to construct validity increasing in measurements on the basis of Structural Equation Modeling is developed.
- The method of reliability estimation of multistage measurements adequate to the offered model is described.

INTRODUCTION

Relevance of the Problem

The first experience of specialists' accreditation in area of public health services in Russia on 2016 and 2017 has shown that assessment processes require optimization. Thousands of specialists of public health services as examinees, thousands of items and long time, necessary for multistage procedures of measurement, are involved in accreditation process. According to this problem of optimization some questions have defined the paper purposes in which the attempt to give the answers for these questions is presented. These questions are:

1. How to optimize the model of measurements for the purpose of assessment time minimization?
2. How to increase content and construct validity of measurements on the basis of optimum set variables choice?
3. How to estimate the reliability of multistage measurements results?

The decision of these questions was the objective of this research, and the hypothesis includes the assumption that the answer to questions will allow to optimize assessment procedures in accreditation.

Literature Review

In Russian theoretical educational researches, the topics in area of adaptive measurements are presented fragmentary. Except the fundamental monograph of M. Chelyshkova (2000), devoted to the theory of adaptive testing in education, and the dissertation of A.A. Malygin (2011), the publications are concentrated around separate applied problems (Dorozhkin et al., 2016; Ke, Borakova & Valiullina, 2017; Fu, Kayumova & Zakirova, 2017). Basically, its consider the possibilities of adaptive testing in dialogue computer testing which changes an order of test items administration depending on examinee performance of previous item (Oparina et al., 2007; Ushakov & Romanova, 2010; etc.).

The base ideas of adaptive testing confirm the possibility of standard scales construction for testing data interpretation if examinees were carrying out various on difficulty and length adaptive tests. The minimization of adaptive tests length is reached by optimization of item difficulty which is defined for each examinee individually, but in full conformity with the uniform content specification. By individual selection every examinee is not administrated too easy items which he can carry out correctly certain, or too difficult items in which it is waited for certain by failure. So, this base idea of adaptive testing helps to reduce the number of each adaptive test items without loss of reliability, validity and comparability data (Chelyshkova, 2000).

The level of development in international researches in area of educational measurements considerably differs from the level of science development to this problematic in Russia. Numerous scientists publish hundreds of articles and books on problems of substantiation of measurement results quality by estimation of their reliability and validity. The possibilities of Item Response Theory for increasing of objectivity in longitudinal study are considered (Baig & Violato, 2012). Some problems of formative and programmatic assessment and their influence on quality of medical students training are discussed (Heeneman, Oudkerk & Schuwirth, 2015; McKinley et al., 2000). The considerable attention in articles is given to questions of competence approach in medical training and scoring its results (Hawkins et al., 2015). The problems of scaling and aggregations of testing students are analyzed with reference to assessment of medical students (McLachlan & Whiten, 2000). On the beginning of the second

decade in XXI. Century the theory of multistage adaptive testing began to develop intensively at Educational Testing Service (Yan, von Davier & Lewis, 2014).

For the purpose of optimization some models of assessment for specialists' accreditation or certification in other countries, for example, in the Netherlands, Israel, the USA, are under construction in a mode of adaptive testing (Crocker & Algina, 2010; Hambleton, 2000; Yan, von Davier & Lewis, 2014).

But, completely there are no researches devoted to questions of reliability estimation in multistage adaptive measurements, and also there are no models with reference to multistage adaptive testing in accreditation. Despite the highest level of development in different Russian and international articles the problems validity increasing in educational measurements on the basis of Structural Equation Modeling are not considered in Russian and international publications. The decisions of these topics are offered in this paper.

MATERIALS AND METHODS

The Model of Multistage Adaptive Measurements

The simplified model of adaptive multistage testing technology assumes item selection optimization on difficulty not for separate examinees, but for subgroups into which all group of examinees is divided. The results of general test performance are the basis for dividing examinees at first stage of measurements. After that there is a further allocation of examinees subgroups to which adaptive tests are administrated within the model of multistage measurements (Chelyshkova & Zvonnikov, 2013; Yan, von Davier & Lewis, 2014).

In particular, as represented by **Figure 1**, three-stage adaptive measurements are shown. Each stage from second demands the construction some adaptive tests (modules) different on difficulty which optimizes for each examinee subgroup. As a rule, in multistage measurements every stage correlates to the separate range of scale showing the levels of measured construct development (knowledge, abilities, competencies or performance professional functions). Accordingly tests for first stage usually have multiple-choice items with four or more response options. The second stage includes measuring instruments with performance or competencies items demanding free constructed answers. And at the third stage mini-cases having creative problem character use.

For example, for the stages, represented by **Figure 1**, the range of the minimum competence can be correlated with the first stage, the range of base competence corresponds second stage and a range of high competence corresponds third stage.

The adaptability is shown by means of the modules which number increases from stage to stage. At the first stage test administration includes only one module. The results of first stage are the base for dividing group into two subgroups: better and worse prepared group of examinees. As the rule, as threshold for such division 50

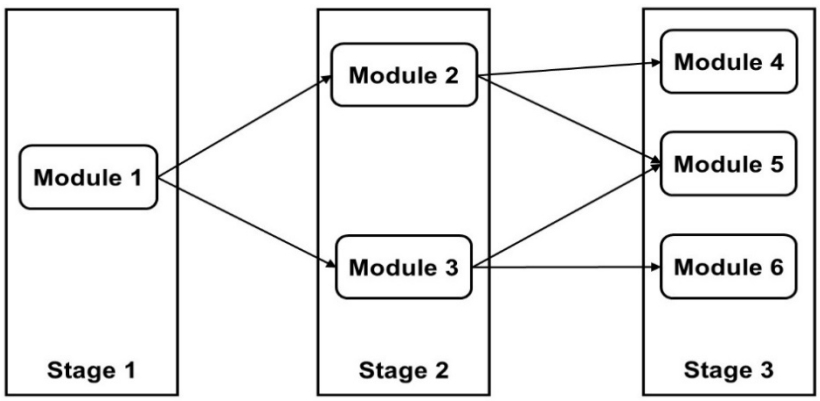

Figure 1. Three-stage adaptive measurements

% level of test performance is established. If the test includes multiple-choice items at the first stage which intend for scoring knowledge underlying professional functions performance, between first and the second stages the organizational break is not necessary. The automated check and number assigned to item responses as the scores (as the rule, 1 or 0) allow to divide group on two subgroups for few minutes after end of the first stage. The complicated variant of group dividing is possible, when the threshold point at first stage is established only for those whose raw scores exceed cut-point for minimum level of achievements or competency.

At the second stage measuring tool includes two modules: second and third. The second module contains more difficult items with free constructed answers and easier than the same form in third module. By results of modules performance examinees are divided by three subgroups. First subgroup includes all examinees which have successfully item performances from second module (exceeded 50 % as threshold or other threshold point). At the third stage they receive the most difficult items in the form of mini-cases from fourth module. Second subgroup combines the worst examinees from second module (they did not overcome the threshold point) and the best examinees from third module (they had passed the threshold point). These examinees receive the mini-cases of average difficulty included in the fifth module. At last, third subgroup intends for weakest examinees. They could not make successfully items from third module. At the third stage, they will receive most easy mini-cases from sixth module. As some experts are needed for checking items after second stage between second and third stages the organizational break is necessary. After checking the decisions about examinees allocation between subgroups are made.

Owing to adaptability each examinee in subgroups does not carry out too easy items or too difficult items. The contribution of such too easy or too difficult items to general reliability of measurements is insignificant. Therefore the optimization of items selection on difficulty and minimization of their number for each examinee will not lower the general reliability. Thus, the general high reliability of measurements will be provided despite minimization of items number in adaptive tests. The model has perspective character and has not found the realization in specialists' accreditation in area of public health services. Realization demands the existence of bank with calibrated items which is not created yet in Russia.

Structural Equation Modeling

The choice of independent variables and their number is the first step on the way of tool constructing for valid measurements (Chelyshkova, 2002; Klein, 1996). Structural Equation Modeling (SEM) is a general approach for the analysis of dependencies or independencies in a set of measured variables and common factors (Kramer, 2007). From the beginning the target model defining assumed structure of variables relations is constructed. The example of this model (path diagram) is represented by **Figure 2** and shows four competencies. As example it has simplified hypothetical character. With path diagram, it is possible to analyze the network of causal processes in terms of direct and indirect effects to check up causal hypotheses about connections between latent variables, factors of influence and results of training on the basis of competence approach (Zvonnikov & Chelyshkova, 2012).

The symbols «D» with a corresponding index designate disciplines, and each arrow specifies that discipline brings the contribution to competencies creation. In simplified variant the base for each professional competence consists from different disciplines. However, in real practice of training one and the same discipline can participate in creation some competencies then the arrows connecting small squares and ovals with competencies numbers will be repeatedly crossed.

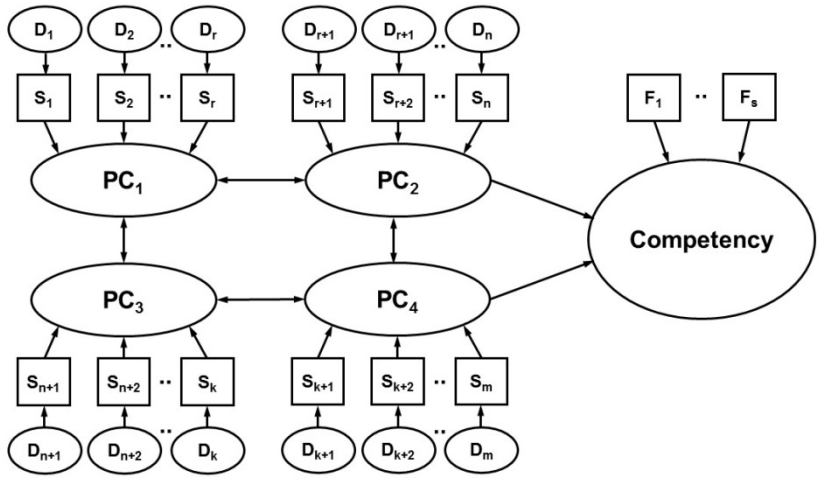

Figure 2. The model of casual relations between disciplines, latent variables (competencies) and factors

Symbol «S» is intended for designation of that contribution which is brought by corresponding discipline with the same index in professional competence creation. The symbol "PC" designates the professional competence, and symbols F_1 and Fs are chosen for the factors influencing in the process of competencies formation.

In SEM the simple logic model including only a few of latent variables and factors is constructed. If sufficient acknowledgement exists the model gradually becomes complicated the by additional variables or relations between them which, as a rule, have cause character. Otherwise, if acknowledgement is not observed, initial logic model must be changed by other variables or connections. As a result of SEM application the optimum set of variables for measurement is created.

The application of SEM is very important for optimization and validation of variables set for constructing measuring tools in accreditation. For practical application of SEM standard statistical packages of type LISREL or EQS are usually used (Joreskog & Sorbom, 2007). The application of SEM to measurement tools constructing for specialists accreditation in area of public health services are discussed in section "Results".

The Reliability of Multistage Measurements

In multistage measurements the estimation of reliability has special difficulty. Such estimation is carried out during some steps. From the beginning the reliability of separate parts of measuring tool is estimated by classical methods, and then the general estimation of reliability for multistage measurements is spent (Chelyshkova, 2002; Gates, 2005). The situation with reliability estimation becomes more complicated when multistage measurements are intended for acceptance high stakes decisions about examinees in accreditation. For accreditation measuring tools are developed within criterion-referenced approach and special methods of reliability estimation are demanded (Berk, 1980). The value of percent performance from full set of requirements in professional standards, and comparison this percent with threshold point (usually 70 % and more) allow to classify examinees on 2 groups: mastery and non-mastery.

For criterion-referenced approach in measurements intended for classification the reliability can be defined as relative stability of examinees groups' classification. According to method offered in the literature, one and the same test is administrated to group of examinees twice after small time interval and the associativity matrix is made. In this matrix there are four groups of examinees: group «a» - proportion of examinees who have done the pass through threshold point in both measurements (did not demonstrate the necessary level of competence or mastery), group «d» - proportion of examinees who have passed through threshold point in both measurements

(demonstrated the high level of competence or mastery), groups "c" and «b» - proportions of examinees which can be carried to classification errors as these examinees have not confirmed the results at double measurements, having changed them to the opposite.

Thus, groups "a" and «d» can be considered as area of classification decisions stability, and groups "c" and «b» can be considered as area of instability classification. Undoubtedly, that the values of proportions will depend not only on quality of measurements, but also from features of examinees samples. If group of examinees can be divided on two subgroups one of which has high level of competence or mastery, and another has the lowest then the minimization of errors classification will be the consequence of examinees distributions instead of measurement quality. Therefore estimations of reliability should be evaluated by representative sample of examinees.

Assuming that sample of examinees is representative, by associativity matrix the formula for reliability estimation in criterion-referenced measurements can be written as L. Crocker & J. Algina (2010).

$$\varphi = \frac{ad - bc}{\sqrt{(a+c)\cdot(b+d)\cdot(a+b)\cdot(c+d)}} \tag{1}$$

where symbol φ is chosen for designation of reliability estimation and other symbols were discussed above. It is coefficient of reliability for criterion-referenced measurements in the form as measure of qualifying decisions consistency (Berk, 1980). There is a normalizing factor in denominator, and numerator can be simply interpreted. From product of proportions reflecting the measure of qualifying decisions stability the measure of instability is subtracted. Accordingly, the greatest values of first product and the least of the second the best reliable of criterion-referenced measurements results corresponds.

The approach to reliability estimation in multistage criterion-referenced measurements is more difficult. The basic component of reliability estimation method for multistage measurements is the assumption that the compound score *C*, received by results *k* stages in measurements, can be presented in the form of *C* = *A* + *B* + … … where *k* composed scores correspond to *k* stages of measurements. If the compound score includes *k* components, the variance σ_C^2 of this observed score will be the sum *k* terms of variance and *k (k* - 1) terms of covariance, i.e.

$$\sigma_C^2 = \sigma_A^2 + \sigma_B^2 + \cdots + \sigma_K^2 + \sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j$$

where $\sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j$ is the sum of *k {k* - 1) terms of covariance, and symbols *i* and *j* designate any pair of components of measuring tool.

By analogy the variance of true score $\sigma_{T_C}^2$ for compound score *C* will be

$$\sigma_{T_C}^2 = \sigma_{T_A}^2 + \sigma_{T_B}^2 + \cdots + \sigma_{T_K}^2 + \sum_{i \neq j} \rho_{T_iT_j}\sigma_{T_i}\sigma_{T_j}$$

where $\sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j$ is the sum *k {k* - 1) terms of covariance, and symbols *i* and *j* designate any pair of components of a measuring tool.

On the basis of the entered equalities and of some the hypotheses which have mathematical character and do not represent interest for developers of tools in education, the algebraic transformations are carried out. As the result the key inequality for value of reliability estimations in multistage criterion-referenced measurements can be written.

$$\rho_{CC'} \geq \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_C^2}\right) \tag{2}$$

where in the right part of inequality there is the expression known as coefficient alpha. It represents reliability of any component from *k* of stages of measurements.

According to this inequality the reliability of a compound score has the least reliability of components of a measuring tool as the lower limit. However, the reliability of compound score depends not only on reliability of each component of measuring tool, but also it depends from value of correlation between measurements results which are collected on separate components. In this connection, the method for reliability estimation in multistage measurements should reflect value of correlation between results in separate components (low or high) and must have branching character.

RESULTS

The Results of Structural Equation Modeling Application for Specialists' Accreditation in Area of Public Health Services

The application of SEM to the analysis of relations between disciplines and labour functions in professional standards has given the chance to choose optimum proportions of items which have formed the basis for measurement tools specifications for specialties: General medicine, Pediatrics, Dentistry. Pharmacy. Preventive medicine. Medical Biochemistry. Medical Biophysics. Medical Cybernetics. This specifications are presented by site

The Results of Reliability Estimation Methods Application in Multistage Measurements for Specialists' Accreditation in Area of Public Health Services

For approbation of methods for reliability estimation in multistage measurements 258 examinees were chosen from population examinees participated in approbation on 2017 in First Moscow State Medical University of I.M. Sechenov.

In connection with model of tools for multistage measurement in accreditation the forms included three stages: first - 60 multiple-choice items with one correct answer, second – 5 practical items for scoring practical skills, third – 7 mini-cases for scoring abilities to make decision in problem situations. For approbation of methods 5 parallel forms were used. All scores of examinees were presented in scales for criterion-referenced approach: pass or non-pass. The cut point was defined by the level of 70%.

As the base, the method of reliability estimation in multistage measurements for assessment in examinees accreditation is represented by **Table 1**. All steps of tool construction for multistage measuring are included in this **Table**, because the performance of each step influences the value of measurement reliability.

The performance of all steps presented by the **Table 1** will provide the professional approach to assessment in accreditation necessary at acceptance of high-stakes decisions.

For estimation of data reliability for every stage the Cronbach's alpha formula was chosen (Crocker & Algina, 2010). It does not demand parallel forms or double test administrations. Coefficient alpha is computed by the formula which is presented by right part in inequality (2). It allows to estimate an internal consistency of items which are dichotomously scored or scored by scoring rubrics with different weights.

The results of reliability estimations are presented by **Table 2.**

For reliability estimates it is necessary to compute correlation between results received by approbation of measuring tool components which include three stages. For correlation estimation the well-known formula of Pearson was used (Chelyshkova, 2002; Crocker & Algina, 2010). The results of application are presented by **Table 3.**

Table 1. The sequence of steps for tool construction with high reliability in multistage measurements

Number of step	Steps and rules for tool construction	
1	To develop the planned model of multistage measurements including number of stages and form of components for different stages in measuring tool, and to define the number of qualifying levels on scale of examinees scores	
2	To apply SEM, and to define the number of variables and common factors of influence	
3	To develop specifications of components content in measuring instrument for representation of professional standards requirements in the form of professional functions or actions	
4	To develop items according to specifications of measuring tool components and the rules for scoring answers in different form items	
5	To execute expertise of items content quality, and correction of their content by results of expertise. To estimate content validity of a measuring tool	
6	To spend approbation of tool by representative sample of examinees	
7	To process data of approbation on each component of a measuring tool by Classical Test Theory or by corresponding models of Item Response Theory, and to analyze test statistics for item calibration. To correlate the results of analysis with different components of a measuring tool and with planned skill levels. To correct item difficulty, to make cleaning and correction of a measuring tool by removal or addition items	
8	To spend repeated approbation of multistage measuring tool by representative sample of examinees	
9	To carry out the factorial and correlation analysis for optimization of number stages in measuring tool, to correct model and a measuring tool by results of the analysis, to estimate construct validity	
10	To define threshold points for each component of measuring tool and corresponding skill level and to spend its empirical validation	
11	To estimate reliability of each component in measuring tool by formulas 1 and 2	
12	To estimate correlation between results on measuring tool components	
13	Case of low correlation (not above 0,3)	Case of high correlation (above 0,3)
14	To choose minimum reliability of measurement results using reliability estimations on separate stages of measurements	To calculate average reliability of results on separat stages of measurements
15	To establish value of minimum reliability as the lower limit of reliability for compound score in multistage measurements	To establish size of average reliability as the lower limit of reliability for compound score in multistage measurements
16	To calculate average reliability of results on separate stages of measurements and to accept it as reliability of a compound score in multistage measurements	To calculate the reliability of all measuring tool by methods of correlation
17	To collect the data by external criteria about quality of graduates' work	
18	To estimate predictive validity of measuring instrument by similar samples of examinees and graduates	

Table 2. The results of reliability estimations for all stages

First stage	Second stage	Third stage
α_1 =0.72	α_2 =0.68	α_3 =0.63

Table 3. The correlations between stages

ρ_{12}	ρ_{23}	ρ_{13}
0.27	0.21	0.23

As **Table 3** shows there is the case of low correlation (not above 0.3). In accordance with methods for this case it is necessary to choose minimum reliability of stages as the *lower limit* of reliability in multistage measurements and then to calculate average reliability of results on separate stages of measurements and to accept it as reliability of a compound score in multistage measurements. So, the value of *lower limit* of reliability in multistage measurements is equal 0.63 and the value of reliability in multistage measurements is equal 0.68.

DISCUSSIONS AND CONCLUSION

The problems of efficiency increasing in assessment by multistage adaptive testing are new to scientists in the area of educational measurements. Basically, these problems began to be considered in the second decade of

XXI century and there are only few books in this area. As a rule, all publications about this topic are concentrated around technologies of adaptive testing (Yan, von Davier & Lewis, 2014). In them all questions are not considered in complex: from the moment of model measurement creation till the moment of quality analysis on the base of tests performance during its application as it is presented in given article.

As a whole, it is possible to notice that such innovative technologies, as multistage adaptive measurements, which applied in accreditation, and have been tested on reliability and validity, allow to raise efficiency of assessment and to motivate of examinees to performance items. Though the interest to using of multistage adaptive measurements worldwide grows, in Russia such researches practically are absent, despite a high urgency in connection with intensive development of qualifications independent assessment system. The reason of such backlog is quite clear: in our country there are no structures possessing banks of calibrated items with stability scores of parameter difficulty.

Using technology of adaptive testing at accreditation of public health services specialists and creation such banks with calibrated items are perspective directions for development of Methodical Centre for Specialists Accreditation in area of public health services (https://fmza.ru). The employees of Methodical Centre gradually develop theoretical base and software and analyze foreign experience. Though the decision of these problems is difficult and expensive, it the future of accreditation system.

As a whole, it is possible to draw the conclusion that carrying out researches on problems of adaptive testing opens new possibilities in creation of effective technologies and tool for assessment in accreditation.

ACKNOWLEDGEMENT

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

REFERENCES

Baig, L. A., & Violato, C. (2012). Temporal stability of objective structured clinical exams: a longitudinal study employing item response theory. *BMC Medical Education, 12*(121), 1-6.

Berk, R. A. (1980). *Criterion-referenced measurement: The state of art*. Baltimor, MD: Johns Hopkins University Press.

Chelyshkova, M. & Zvonnikov, V. (2013). The optimization of formative and summative assessment by adaptive testing and zones of students' development. *Journal of Psychosocial Research, 8*(1), 127-132.

Chelyshkova, M. (2000). *Adaptive testing in education. The monography*. Moscow: Logos.

Chelyshkova, M. (2002). *Theory and practice of educational tests construction: the manual*. Moscow: Logos.

Crocker, L., & Algina, J. (2010). Introduction to classical and modern test theory. Under the editorship of V.I. Zvonnikov and M.B. Chelyshkova. Moscow: Logos Publ.

Dorozhkin, E. M., Chelyshkova, M. B., Malygin, A. A., Toymentseva, I. A. & Anopchenko, T. Y. (2016). Innovative approaches to increasing the student assessment procedures effectiveness. *International Journal of Environmental and Science Education, 11*(14), 7129-7144.

Fu, L., Kayumova, L. R. & Zakirova, V. G. (2017). Simulation Technologies in Preparing Teachers to Deal with Risks. *EURASIA Journal of Mathematics, Science and Technology Education, 13*(8), 4753-4763.

Gates, S. (2005). *Measuring more than efficiency. Report No. R-1356-04-RR*. New York: Conference Board.

Hambleton, R. K., & Zaal, J. (2000). Computerized adaptive testing: Theory, applications, and standards, in: R. K. Hambleton, J. Zaal (Eds.). *Advances in educational and psychological testing: Theory and applications*. Boston: Kluwer Academic Publishers, p. 341-366.

Hawkins, R., Welcher, C., Holmboe, E., Kirk, L., Norcini, J., Simons, K., & Skochelak, S. (2015). Implementation of competency-based medical education: are we addressing the concerns and challenges? **Error! Hyperlink reference not valid.***Medical Education, 49*(11), 1086-1102.

Heeneman, S., Oudkerk, P. A., & Schuwirth, L. W. T. (2015). Department of Pathology, Maastricht The impact of programmatic assessment on student learning: theory versus practice. *Medical Education, 49*(5), 487-498.

Joreskog, K. C, & Sorbom, D. (2007). LISREL 17, A guide to the program and applications. Chicago: SPSS.

Ke, Z., Borakova, N. U., & Valiullina, G. V. (2017). Peculiarities of Psychological Competence Formation of University Teachers in Inclusive Educational Environment. *EURASIA Journal of Mathematics, Science and Technology Education, 13*(8), 4701-4713.

Klein, A. L. (1996). *Validity and reliability for competency-based systems: Reducing litigation risks.* Compensation and Benefits Review, Springer-Verlag, New York.

Kramer, D. (2007). Mathematical data processing in social sciences: modern methods: studies. The grant for students of higher educational institutions / Dunkan Kramer; the translation from English by Timofeeva I. V., Kiseleva J. I., M: Publishing Centre "Academy".

Malygin, A. A. (2011). *Adaptive testing students' educational achievements in distance learning* (PhD Thesis). Moscow.

McKinley, R. K., Fraser, R. C., Van Der Vleuten, C., & Hastings, A. M. (2000), Formative assessment of the consultation performance of medical students in the setting of general practice using a modified version of the Leicester Assessment Package. *Medical Education, 34*(7), 573–579.

McLachlan, J. C., & Whiten, S. C. (2000). Marks, scores and grades: scaling and aggregating student assessment outcomes. *Medical Education, 34*(10), 788–797.

Oparina, N. M., Polina, G. N, Fayzulin, R. M., & Shramkova, I. G. (2007). Adaptive testing. Habarovsk: "DVGUPS".

Ushakov, A. N., & Romanova, M. L. (2010). Adaptive testing in the structure of educational control. *Scientific Notes of P.F. Lesgaft University, 5*(63), 87-93.

Yan, D., von Davier, A. A., & Lewis, C. (2014). Computerized multistage testing: Theory and applications. New York, NY: CRC Press.

Zvonnikov, V. I., & Chelyshkova, M. B. (2012). *Assessment of training results quality at certification: competence approach (the second edition)*. Moscow: Logos.