MODESTUM
OPEN ACCESS

EURASIA Journal of Mathematics, Science and Technology Education, 2019, 15(9), em1749
ISSN:1305-8223 (online)
Research Paper https://doi.org/10.29333/ejmste/108485

Fairness Principle in Accreditation of Health Specialists: The Differential Item Functioning Method

Zhanna M. Sizova [1], Tatyana V. Semenova [2*], Natalia N. Naydenova [3], Victoria V. Narbut [4], Marina B. Chelyshkova [1], Alfiya R. Masalimova [5]

[1] I.M. Sechenov First Moscow Medical University (Sechenov University), Moscow, RUSSIA
[2] Ministry of Health, Moscow, RUSSIA
[3] Institute for Strategy of Education Development of the Russian Academy of Education, Moscow, RUSSIA
[4] Financial University under the Government of the Russian Federation, Moscow, RUSSIA
[5] Kazan (Volga region) Federal University, Kazan, RUSSIA

Received 19 November 2018 ▪ Revised 25 February 2019 ▪ Accepted 15 March 2019

ABSTRACT

The main purpose of this article is to present a Differential Item Functioning method of item analysis. It is designed to minimize the discriminatory effect of individual items in the accreditation of graduates of medical universities with different training programs. The one-parameter Item Response Theory model is used to align graduates' rights in accreditation. The measure of the difference in the location of the item characteristic curves constructed for different graduates' samples is presented using the difficulty estimates. For the development of the methodology, a number of research questions are posed, the solution of which made it possible to analyze the items bias for the "Polyclinic" discipline and the range of item difficulties from 1.5 to 2, 5 logits. The item bank was cleaned and an interpretation of decisions on items release and correction for the cases of different arrangement of their characteristic curves was presented.

Keywords: differential item functioning, item response theory, systematic measurement error, equity principle, accreditation

INTRODUCTION

Starting in 2016, large-scale assessment has been launched in Russia to introduce the autonomous procedure for admitting health professionals to occupational activities on the basis of their accreditation, which is understood as a procedure for determining the readiness of the received medical or pharmaceutical education persons to the implementation of an independent professional activity at a certain position. Accreditation is carried out for all graduates of medical universities in Russia by comparing assessments of the professional readiness of individuals with established criteria and indicators, the role of which is fulfilled by the requirements of professional standards.

When creating the toolkit, the principles governing the development of accreditation were used:

- High fairness of accreditation decisions;

- Sustainable objectivity (reliability) and accurate reasonableness (validity) of accreditation results within the framework of the theories in educational measurements under the quality control of the tools and accreditation procedures;

- Wide interaction between professionals and the public in the health care accreditation field.

To improve the objectivity and validity of accreditation results, it was decided to use multi-stage measurement procedures, combining different evaluation tools for the three stages of accreditation. The first stage includes traditional testing. The second stage is the evaluation of practical skills with the help of special simulation stations as objective structured clinical exam (OSCE) (Baig & Violato, 2012). The third stage is the assessment of the level of

✉ Sizova-klinfarma@mail.ru ✉ tsemenova123@mail.ru **(*Correspondence)** ✉ naydenova@my.com
✉ vvnarbut@fa.ru ✉ mchelyshkova@mail.ru ✉ alfkazan@mail.ru

Contribution of this paper to the literature

- The authors have analyzed the graduates' results of medical universities from their accreditation in Russia. They have considered to take the implementation standpoint of fairness principle that determine the including only the equal items in the accreditation multidisciplinary test.
- The authors determined the differences of the educational programs in medical universities that should be to have the different results in different universities at the executing of some items. The approach to leveling graduates' rights during accreditation is based on the Differential Item Functioning and Item Response Theory.
- An experiment based on the general population of graduates at 2018 year which was divided to four groups of universities with different results using ANOVA method. For the analyzing of results there was build the sample of items from multidisciplinary test. DIF analyst of these items determined the candidate of items to deleting from test.

mastering the labor functions of professional standards using situational tasks. Based on the results of the successful implementation of the three stages, a decision is made to issue an accreditation certificate (Aydarov & Krasilnikov, 2017; Semenova et al., 2017).

The basis of the content of the toolkit was based on Clinical Recommendations containing the meaningful analysis results fulfilled with the requirements on Good Clinical Practice (GCP) and Evidence-based Medicine. In the modern period, medicine is rapidly developing, each year new methods of treatment are being developed in the world, scientific medical schools are being created, and newer medicines and modern equipment are coming into clinical practice. Obviously, all these innovations, reflected to varying degrees in the educational programs of medical universities in Russia, make it difficult to use common approaches to treatment and to require adequate reflection of innovations in the content of evaluation tools for accreditation of health professionals (Semenova et al., 2018). Since differences in the content of educational programs are an inevitable source of variation in measurement results, the analysis of the difference influence on the fairness of graduates' assessments is understood as the lack of advantages in the test score assignments among compared groups of graduates. Advantages can manifest, for example, in cases where there are items in the test that reflect the specifics of individual educational programs in medical universities in Russia.

The presence of these sources of variation forces one to turn to a special apparatus of Differential Item Functioning (DIF-analysis), with the help of which it is possible to establish the presence of item bias for samples differing in various signs of differentiation (Cohen, Kim & Baker, 1993; Crocker & Algina, 2010; Dorans & Kulick, 1986; Holland & Wainer, 1993; Penfield & Camilli, 2007; Zumbo, 1999). DIF-analysis is a special method of the test item exploration. This method is based on items identification that function differently in different samples and lead to item bias increasing of measurement results. Its use is mandatory for the implementation of the principle of fairness when conducting mass assessment procedures of high significance. Nevertheless, despite its importance, it is practically not developed either at the level of theory or at the level of methodology in Russia.

RESEARCH METHODOLOGY

The purpose of this article is to present the using DIF-method in accrediting graduates of medical universities to minimize the discriminatory effect of individual items` and item bias that may arise when assessing the professional readiness of graduates due to different training programs.

Background

The problem of ensuring the equal rights of individuals in mass assessment procedures is solved by many test services, including in comparative international studies of the education quality (Crocker & Algina, 2010; Naydenova, Tagunova & Sukhin, 2018; Tumeneva & Valdman, 2012; Yazdani et al., 2018). Since Item Response Theory (IRT) allows to select the items that discriminate particular groups of individuals in the most effective ways, this article has chosen the DIF-analysis within IRT (Baker & Kim, 2004; Chelyshkova, 2002; Chelyshkova et al., 2018; de Ayala, 2009; Dorozhkin et al., 2016; Filaretov, Danilov & Golovchenko, 2014; Gorbatkova & Zabrodina, 2015; Lord, 1980).

Unlike previously conducted similar studies, where groups of individuals are distinguished by gender or ethnicity, this article deals with the problem of violation of individuals rights in graduate accreditation among universities with various educational programs. The variability of the programs is significantly increased by the new Federal State Educational Standards of the third generation, functioning in Russia today and opening up wide opportunities for building individual learning paths. In this regard, the problem considered in the article is highly relevant for improving the quality of medical education in Russia.

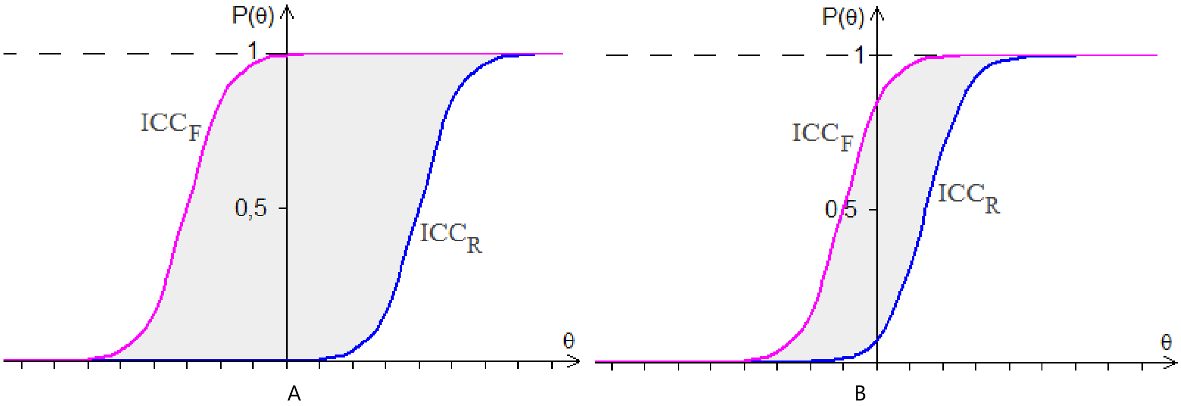


Figure 1. Examples of ICC possible locations of different items for two groups

Methodological Framework

In order to identify test items that function differently among universities with different training programs by DIF-analysis two basic assumptions must be consider. First, the results of the items performed by the individuals may be influenced by sources of variation that differ from those provided for in the general measured construct. Secondly, these extraneous sources of variation have a constant unplanned effect on the test performance by some groups of individuals, leading to a systematic shift of their assessments at the overall assigning the test score.

In the general case, the process of analyzing the item bias involves two stages. First, you should make sure that the same sources of variation affect the assessments of individual groups. Then, it should be determined whether any of these irrelevant sources of variation gives an unfair advantage in the test results for some groups of individuals compared to others. It can be argued that the items do not lead to the appearance of biased estimates, if for individuals with the same score of the latent parameter of the data distribution under the influence of irrelevant sources, variations will be the same. If in the course of the analysis there appear items that work differently for the individuals from a certain group, which, as a rule, represent a minority in the general population of the subjects, they should be removed from the test.

The most promising approach to identifying items that lead to item bias was suggested by Lord (1980). If we assume the possibility that the item will function differently for the central group (general population - F) and some reference group (minority group - R), then in the language of the IRT we can talk about the degree of differences among the Item Characteristic Curves (ICC) which were constructed for these two groups. If the ICC for the two groups completely coincide, then it does not introduce any item bias in measurement results. Thus, the measure of the difference in the location of the ICC, constructed for two groups of individuals, serves as a characteristic of item bias value. An illustration of this statement is shown by **Figure 1**.

Figure 1 shows the ICC$_F$ – the characteristic curves of two items, each of which is constructed for the Central Group, and the curves of the same ICC$_R$ items for the Reference Group. Part A corresponds to the first item, and part B to the second. For the first item, the area between the characteristic curves in the two groups is large; therefore, item 1 introduces a significant item bias in measurement results and it must be removed from the test. For the second item, this area is much smaller, which means a significantly smaller value of item bias, so it can be left in the test.

Thus, the numerically discriminating effect of the item can be expressed in the form of value located between two characteristic curves for two groups of individuals. To determine the area of this area (S_Y^0), the integral of the difference between the functions $P_F(\theta)$ and $P_R(\theta)$ is calculated, which determines the probabilities of the correct answers of the individuals to item using one of the IRT models in the Central and Reference groups (Penfield & Camilli, 2007).

When choosing the analytical form of the characteristic curve of item using the one-parameter model of the IRT, the probability of correct response to the item is represented by the formula (Baker & Kim, 2004; Chelyshkova, 2002; de Ayala, 2009):

$$P(\theta) = \frac{e^{1,7(\theta-\beta)}}{1 + e^{1,7(\theta-\beta)}}$$

where θ is an independent variable, and the item difficulty is chosen as a parameter – β. Since the differences of ICC locations are determined by the distance between their inflection points corresponding to an item difficulty estimates (parameter β) on the horizontal axis of the variable θ, it makes sense to estimate the difference between the values of β in the Central and Reference groups which was calculated by next expression:

$$S_{\gamma}^{0} = \left|\int_{-\infty}^{+\infty}(P_F(\theta)-P_R(\theta))d\theta\right| = \left|\int_{-\infty}^{+\infty}\left(\frac{e^{1,7(\theta-\beta_F)}}{1+e^{1,7(\theta-\beta_F)}}-\frac{e^{1,7(\theta-\beta_R)}}{1+e^{1,7(\theta-\beta_R)}}\right)d\theta\right| = |\beta_F-\beta_R|$$

The module mark is placed in order to identify the value of the overall discriminating effect without taking into account the direction of its action. Thus, when referring to the one-parameter model of IRT, the discriminatory effect of the item is expressed as the absolute value of the difference $|\beta_F - \beta_R|$. Significance values are called a general discriminating effect (Holland & Wainer, 1993).

The sign is chosen in accordance with the location of the curves for the Central and Reference groups, which, in the case of using a one-parameter model, will not have an intersection point. The effect is considered positive in the case when the curve for the central group is to the left, and negative otherwise. The difficulty of interpreting the general effect lies in the fact that its values belong to an unlimited range and allow only to rank the items according to its degree of expression, but do not give a clear answer to the question of which values are significant evidence of the presence of a discriminating effect. This article proposes a criterion for interpreting the overall effect, established empirically and based on the approach of Keeves (1988).

According to his approach, the parameter of the discriminating effect of item (γ) is the ratio of the total effect $sgn_{\gamma}{\cdot}S_{\gamma}$ to the difference $\theta_{MAX} - \theta_{MIN}$, where the difference $\theta_{MAX} - \theta_{MIN}$ geometrically expresses the area of the rectangle between an item curves when calculating the total effect (Keeves, 1988). Therefore, the parameter γ is always modulo less than unity, and its absolute value indicates the proportion of differences in the probability of the item being completed by the subjects in different groups. At the same time, positive values of γ mean that the item gives an advantage to the subjects of the central group, negative ones – the reference one.

In the Keeves' studies, the following critical values were established for the parameter:

– if $|\gamma| < 0,05$ (that is, the parameter deviation from zero is less than 5%), then the item does not have a significant discriminatory effect;

– if $0,05 \le |\gamma| \le 0,1$ (deviation of the parameter from zero from 5% to 10%), then the item has an average level of discriminating effect;

– if $|\gamma| > 0,1$ (the parameter deviation from zero is more than 10%), then the item has a significant level of discriminatory effect.

The considered approach to estimating item bias based on IRT is undoubtedly the most modern. However, it has not only advantages, but also disadvantages. In particular, the approach does not work for the three-parameter logistic model of IRT, the use of which is desirable for items with multiple choice. Since in our study all the test items were with the only correct answer, the preference was given to the one-parameter IRT model.

Research Questions

For the selection of items that violate the principle of fairness in accreditation, a number of issues were raised and resolved:

1. Which method of Differential Item Functioning should be preferred? The answer to the question required a comparative analysis of various approaches of researchers to identify assignments that discriminate against individual groups. Methods considered include: analysis of variance (Cardall & Coffman, 1963), Angoff method (1972), chi-square method (Bishop, Fienberg, & Holland, 1985), standardization method (Dorans & Kulick, 1986) and methods of IRT, based on different models (Kramer, 2007; Lavrakas, 2008; Woods, 2011). Despite the logical and practical simplicity of the Angoff and chi-square methods, a one-parameter model of IRT was chosen. The argument for this choice was the high efficiency of IRT.

2. How to group samples of individuals, to optimize the content of items for the accreditation of health professionals and to determine the items to be analyzed? When answering the first part of the question, the sampling approach was chosen in several stages (Lavrakas, 2008; Naydenova, 2003). In the second part of the question, the priority of choosing a specialty and discipline for analysis was determined by their significance for the professional training of Russian health care professionals (Sizova et al, 2016). Therefore, preference was given to the specialty "General Medicine" and the discipline "Polyclinic".

When considering the third part of the question, it was decided to select assignments with estimates of the difficulty parameter in the vicinity of the threshold score (70% of completion) used to make accreditation decisions. The decisive argument for this choice was the well-known effect of fuzzy decisions, characteristic of 20% of the range surrounding the threshold score (10% to the left and 10% to the right of the threshold score). Due to the existence of a single logit scale for estimating the parameters θ and β in IRT, the values of these parameters can be

Table 1. Characteristics of study samples

Sample	Number of universities	Number of graduates	Number of items	Average frequency
Central group F	76	17960	2592	150
Subject subgroup F	76	17960	328	55
Location subgroup F	76	17960	43	418
Reference group R	8	1516	2363	236
Subject subgroup R	8	1516	326	27
Location subgroup R	8	1516	43	35

compared. Thus, the number of items to be analyzed included those whose difficulty was in the range (1.5; 2.5) logites.

3. How to choose the critical values for the area between the characteristic curves, the excess of which indicates the need to remove the items from the test? The results of the correlation of empirical research data and critical values for the parameter γ (Keeves, 1988) made it possible to determine 0.5 as the critical value for the difference between the estimates of the β parameter. Exceeding this in the case of positive values violates the principle of fairness for representatives of the reference group and requires the removal the item from the test.

The questions posed found their solution in the results of the study presented below.

Methods

The data were analyzed using data grouping, aggregation, representative sampling, the Bonferroni method in ANOVA, frequency and analysis of variance, the Differential Item Functioning method and IRT theory algorithms for constructing ICC implemented using the Testan 6.5 Program (Kramer, 2007; Linacre & Wright,1989).

Sample

The construction of samples of items to be studied was carried out taking into account their belonging to the "Outpatient Care" discipline and the range of difficulty in the range from 1.5 to 2.5 logites. In accordance with the adopted technology of forming options for accrediting health professionals, items in this discipline were scattered throughout the test, so the study used positional sampling to nullify the effect of item location on the analysis results. The results of samples formation for the subjects and items are given in **Table 1**.

The formation of subgroups in both the Central and the Reference groups was due to the interdisciplinary design of the test. To carry out a DIF analysis of test items, it was impossible to carry out without compressing information with a focus on the selection of items. To this end, subgroups were formed in the central and reference groups for a specific discipline "Polyclinic case" in order to remove the analysis of the influence of interdisciplinarity in the test. Also, for a particular discipline, positional subgroups were formed for both groups: Central and Reference. This was done in order to reduce the Bayes positional influence on the curves ICC. In the positional groups, only a few items were selected for TESTAN processing. Detailed information on the composition of the subgroups is given in **Table 1**.

Sampling of 8 universities was conducted on the entire central disciplinary subgroup, starting from a random start and following further with the calculated sampling interval. According to the results, the sample of 8 universities reflected the general population of 99.9%.

DATA ANALYSIS

Data analysis using the Differential Item Functioning method was carried out in two stages, first for the central group and then for the reference group. Estimation of the parameter of difficulty of items and the construction of their characteristic curves preceded the ANOVA dispersion analysis using the Bonferroni method (Kramer, 2007), which showed the presence of statistically significant differences in the frequency of correct execution of tasks in reference subgroups with a probability of 0.95.

To build the characteristic curves using the Testan 6.5 program, 2363 tasks were selected for the "Polyclinic case" discipline and difficulty range in the range from 1.5 to 2.5 logits (**Table 1**). There are shown examples of the resulting location of the curves for the three items from the subgroups F and R by **Figure 2** (item 23), **Figure 3** (item 485) and **Figure 4** (item 172). Item numbers are given by item bank.

```
--------|----ITEM 23  ------------------------------|
  100   |                            ++++
        |                        ++++          :::::
        |                     +++          ::::
   80   |                   ++            :
        |                 ++          ::::
Percent |               ++          ::
   60   |             ++          ::
        |          +++           :
        |      ++++           :::
   40   |   +++            :::
Correct |+++              :
        |              :::
   20   |
        |
        |
    0   |
--------|--1.0---1.5---2.0---2.5---3.0---3.5--------|

”+” - central group
”:” - reference group
```

Figure 2. ICC 23

```
--------|----ITEM 485 ------------------------------|
  100   |                             +++++++
        |                          +   :::
        |                        ++    ::
   80   |                     +++     :
        |                   ++     :::
Percent |                 ++    :::
   60   |               ++    ::
        |            ++    :::
        |        +++   :::
   40   |     +++    ::
Correct | ++++    :::
        |
   20   |
        |
        |
    0   |
--------|--1.0---1.5---2.0---2.5---3.0---3.5--------|

”+” - central group
”:” - reference group
```

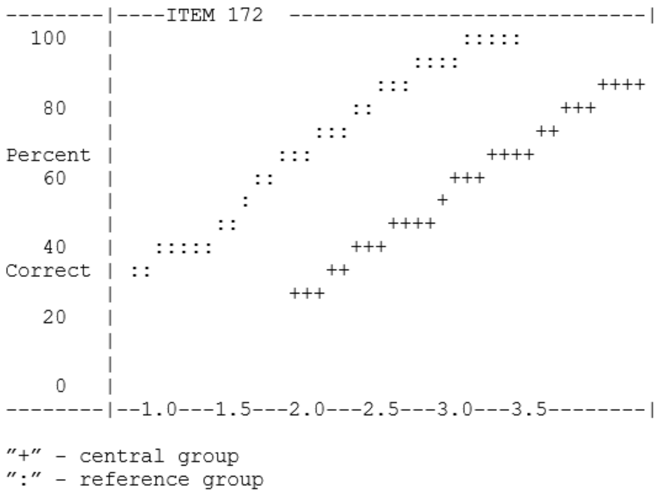
Figure 3. ICC 485

Both of these items are in the positional subgroup of the Central and Reference groups. ICC for empirical settings 23, 172 and 485 as frequencies of its correct execution by individuals of the Central and Reference groups.

DISCUSSION AND CONCLUSIONS

Figure 2 is constructed to the item 23 with difficulty 1.63 logits for the Central group and 2.47 logits for the Reference group of the university with the number 2 out of eight selected universities. It is clearly discriminatory for subjects in a minority group, since the difference between the estimates of difficulty, equal to 0.84 logit, is almost 2 times higher than the criterion (0.5 logit). Therefore, item 23 should be removed from the item bank for accrediting health care professionals.

Figure 3 shows ICC of the item with number 485. Its difficulty is 1.90 logit for the Central group and 2.38 logits for the Reference group of the University at number 2. Difference of difficulty estimates is equal to 0.48 logits. A similar picture was revealed for the other Reference groups of eight universities, so according to the criterion it can be left in the bank. However, this difference is so close to the criterion that it forces the assignment authors to once again analyze the content of assignment 485 for compliance with various educational programs and, possibly, correct it taking into account the minor differences that are most likely to be revealed by careful analysis.



```
--------|----ITEM 172  -----------------------------|
  100   |                            :::::
        |                        ::::
        |                     :::               ++++
   80   |                   ::               +++
        |                ::::               ++
Percent |             :::              ++++
   60   |           ::              +++
        |          :               +
        |        ::            ++++
   40   |   :::::           +++
Correct | ::              ++
        |              +++
   20   |
        |
        |
    0   |
--------|--1.0---1.5---2.0---2.5---3.0---3.5--------|

"+" - central group
":" - reference group
```

Figure 4. ICC 172

Figure 4 shows the characteristic curves of item 172, the location of which is not typical for the **R**eference group of the second University. The item turned out to be easier for a minority group than for a Central group. An analysis of the content of item 172 made it possible to give this location of the curves a completely obvious explanation. In the second University, an internship module "School of Professional Growth" in the amount of 280 academic hours was included in the educational program. The adjustments made to the educational program helped to improve the development of those competencies that were necessary to perform the item 172. In particular, they contributed to the development of skills of independent professional work in the workplace in medical organizations of primary health care, to prepare and adapt senior students to clinic, ensuring that their training is developed the changing conditions of professional activity and social environment.

The use of the Differential Item Functioning apparatus for analyzing the quality of 2363 items of the bank for accreditation of health professionals of the specialty "Medicine" and the "Polyclinic" discipline revealed 17 unsuitable items introducing item bias to the accreditation data. 142 items were corrected due to a slight difference in the difference between the estimates of the difficulty parameter and criterion 0.5.

However, the results require further development and research. The lack of closeness of the distribution of empirical data to the normal law, which is characteristic of accreditation data due to the criterion-oriented approach to the development of accreditation tests, introduces its component into the measurement error (Berk, 1980). In this regard, it is necessary to conduct a study to analyze the impact of the data normalization procedure on the accuracy of measurement results and the correctness of the conclusions about the discriminatory effect of items.

ACKNOWLEDGEMENT

1. The work is performed according to the project of improving competitiveness of the leading Russian universities among the leading world scientific education centers "5-100"of First Moscow State Medical University.

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

REFERENCES

Aydarov, V. I., & Krasilnikov, V. I. (2017). The role of the moral component of personality in ensuring security in medicine. *Vestnik NTSBGD, 2*(32), 117-121.

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. https://doi.org/10.1186/1472-6920-12-121

Baker, F. B., & Kim, S. H. (2004). *Item response theory. Parameter estimation techniques*. New York: Dekker. https://doi.org/10.1201/9781482276725

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

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

Chelyshkova, M. B., Semenova, T. V., Naydenova, N. N., Dorozhkin, E. M., Malygin, A. A., & Akhunov, V. V. (2018). Cross-analysis of big data in accreditation of health specialists. *Electronic Journal of General Medicine, 15*(5), em72. Retrieved from https://doi.org/10.29333/ejgm/93469

Cohen, A. S., Kim, S. H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*, 335–350. https://doi.org/10.1177/014662169301700402

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.

De Ayala, R. (2009). *The Theory and Practice of Item Response Theory*. New York: Guilford Press.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355–368. https://doi.org/10.1111/j.1745-3984.1986.tb00255.x

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.

Filaretov, V. A., Danilov, V. A., & Golovchenko, N. I. (2014). Prevention of burnout syndrome. *Vestnik NTSBGD, 1*(19), 63-66.

Gorbatkova, E. Yu., & Zabrodina, G. Yu. (2015). Student lifestyle and health. *Vestnik NTSBGD, 3*(25), 102-104.

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. London: Routledge.

Keeves, J. P. (1988). *Edukational Reseach Methodology, and Measurement*. New York: Perg. Press.

Kramer, D. (2007). *Mathematical data processing in social sciences: modern methods: studies. The grant for students of higher educational institutions*. Moscow: Publishing Centre "Academy".

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods.* Thousand Oaks. California: Sage Publications, Inc. https://doi.org/10.4135/9781412963947

Linacre, J. M., & Wright, B. D. (1989). Mantel-Haenszel DIF and PROX are Equivalent! *Rasch Measurement Transactions, 3*(2), 51–53.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale: Erlbaum.

Naydenova, N. N. (2003). *Formation of representative samples*. Moscow: Logos.

Naydenova, N. N., Tagunova, I. A., & Sukhin, I. G. (2018). The Problem of Interpretation in Comparative Education (Interdisciplinary Approach). *International Conference "Education Environment for the Information Age" (EEIA-2018)*, 05-06 June 2018, pp. 485-492. https://doi.org/10.15405/epsbs.2018.09.02.56

Penfield, R. D., & Camilli, G. (2007). *Differential item functioning and item bias, in Handbook of statistics*. New York: Elsevier.

Semenova, T., Sizova, Zh., Chelyshkova, M., Dorozhkin, E., & Malygin, A. (2018). Fairness and Quality of Data in Healthcare Professionals' Accreditation. *Modern Journal of Language Teaching Methods, 15*(5), em72. Retrieved from www.mjltm.org

Semenova, T., Sizova, Zh., Zvonnikov, V., Masalimova, A., & Ersozlu, Z. (2017). The Development of Model and Measuring Tool for Specialists Accreditation. *EURASIA Journal of Mathematics, Science and Technology Education, 13*(10), 6779–6788. https://doi.org/10.12973/ejmste/77042

Tumeneva, J., & Valdman, A. (2012). The first data from comparative analysis of the results on TIMSS-2011 and PISA – 2012 tests, administered to the same sample of Russian students. *International Conference "Russian Education in the Mirror of the International Comparative Studies"*. Retrieved from http://centeroko.ru/conf2013/conf2013_eng.html

Woods, C.M. (2011). DIF testing for ordinal items with Poly-SIBTEST, the mantel and GMH tests, and IRT-LR-DIF when the latent distribution is non-normal for both groups. *Applied Psychological Measurement, 35*(2), 145–164. https://doi.org/10.1177/0146621610377450

Yazdani, S., Azandehi, S.K., Ghorbani, A., & Shakerian, S. (2018). Explaining the process of choosing clinical specialties in general medical graduates: A grounded theory. *Electronic Journal of General Medicine, 15*(6), em89. https://doi.org/10.29333/ejgm/93457

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning*. Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense.

http://www.ejmste.com