Zhytomyr Ivan Franko State University Journal.
Pedagogical Sciences. Vol. 2 (101)

Вісник Житомирського державного
університету імені Івана Франка.
Педагогічні науки. Вип. 2 (101)

ISSN (Print): 2663-6387
ISSN (Online): 2664-0155

UDC 372.881.111.1:378.146
DOI 10.35433/pedagogy.2(101).2020.156-165

OPTIMISING ASSESSMENT SYSTEM IN THE ESP COURSE THROUGH THE USE of THE METHODS OF DIFFERENTIAL ITEM FUNCTIONING AND DIFFERENTIAL TEST FUNCTIONING IN FINAL TEST DESIGN

K. I. Shykhnenko[*]

The purpose of the research was to examine how the use of the methods of Differential item functioning and Differential test functioning contribute to the quality of the final assessment (FT-ESP) in the English for Specific Purposes course delivered to the graduate students at tertiary institutions. The study relies on two interventions intended to identify the correlation between the test design and the academic performance of the students in the ESP course through using Pearson's correlation coefficient of the answered versus the unanswered questions. The first intervention test was similarly structured as the one for the second intervention and consisted of the same number of items. In the first intervention, a regular final ESP test was administered. In the second intervention, the originally designed test, which validity and reliability was analyzed using the methods of DIF and DTF, was performed. The test included three sub-domains such as: reading comprehension (15 items), structure (15 items), and compositional analysis (15 items). It has been found that the use of the methods of DIF and DTF boosts the quality of the assessment system in the ESP course delivered to the graduate students at tertiary institutions. It is advisable that the first step in DIF analyses be related to the use of statistical methods to detect the DIF items. It is also advisable to examine the effects of other potential factors on DIF such as item order and mother tongue effects along with unintended content specific factors to explain DIF effect in the context of language testing. The findings also imply that neither of methods addresses the issue of measurement bias, which might occur in tests, because it is complicated and cannot be addressed adequately using simple statistical or classical test theory methods. Further studies are needed to identify the ways of improving the assessment of speaking skills of the graduates of the tertiary institutions.

Key words: *higher education, academic performance, English for Specific Purposes, assessment system, test design, language testing, differential item functioning method, differential test functioning method.*

[*] Candidate of Pedagogical Sciences (PhD in Pedagogy)
(Institute of Public Administration and Research in Civil Protection, Kyiv)
shikhkate@gmail.com
ORCID: 0000-0002-8623-2907

ОПТИМІЗАЦІЯ СИСТЕМИ ОЦІНЮВАННЯ В НАВЧАЛЬНОМУ КУРСІ "АНГЛІЙСЬКА МОВА ЗА ПРОФЕСІЙНИМ СПРЯМУВАННЯМ" ШЛЯХОМ ВИКОРИСТАННЯ МЕТОДІВ ДИФЕРЕНЦІЙОВАНОГО ФУНКЦІОНУВАННЯ ЗАВДАНЬ ТА ДИФЕРЕНЦІАЛЬНОГО ФУНКЦІОНУВАННЯ ТЕСТУ В РОЗРОБЦІ ПІДСУМКОВОГО ТЕСТУ

К. І. Шихненко

Метою дослідження було вивчити, як використання методів диференційованого функціонування завдань і диференціального функціонування тесту сприяє якості підсумкового оцінювання в курсі англійської мови за професійним спрямуванням, що викладається магістрам вищих навчальних закладів. Дослідження складається з двох етапів, завданням яких було з'ясувати співвідношення між тестовою структурою та навчальною успішністю слухачів курсу англійської мови за професійним спрямуванням за допомогою використання коефіцієнта кореляції Pearson між кількістю наданих правильних та неправильних відповідей. Тести для першого та другого зрізу знань були структуровані за аналогічною структурою і наповненням. Під час першого зрізу зі слухачами проводився стандартний підсумковий тест з англійської мови за професійним спрямуванням. Під час другого зрізу використовувався спеціально розроблений тест, достовірність та надійність якого аналізували за допомогою методів диференційованого функціонування завдань і диференціального функціонування тесту. Тест включав три підрозділи, зокрема такі, як розуміння прочитаного (15 запитань), структура (використання мови) (15 запитань) та композиційний аналіз (15 запитань). З'ясовано, що використання методів диференційованого функціонування завдань і диференціального функціонування тесту підвищує якість системи оцінювання в курсі "Англійська мова за професійним спрямуванням". Доцільно, щоб перший крок у аналізі диференційованого функціонування завдань був пов'язаний із використанням статистичних методів для виявлення запитань із високим показником диференційованого функціонування завдань. Також, доцільно вивчити вплив інших потенційних факторів на показники диференційованого функціонування завдань, таких як порядок запитань, вплив рідної мови та непередбачувані змістові фактори для пояснення впливу зазначених показників. Отримані результати передбачають, що жоден з методів не вирішує питання забезпечення об'єктивності вимірювань, що може стати "наріжним каменем" під час виконання тестів, оскільки це завдання є складним і не може бути адекватно вирішено за допомогою простих статистичних методів або класичної теорії методів тестування. Потрібні подальші дослідження, щоб з'ясувати шляхи вдосконалення системи оцінювання мовленнєвих навичок магістрів вищих навчальних закладів.

Ключові слова: *вища освіта, академічна успішність, англійська мова за професійним спрямуванням, система оцінювання, розробка тесту, тестування мовних навичок, метод диференційованого функціонування завдань, метод диференціального функціонування тесту.*

Introduction of the issue. The fairness and credibility of the assessment in the English for Specific Purposes (ESP) course are the major concerns of both students and ESP teachers at higher institutions in Ukraine. The importance of objectivity in test results with regard to different subgroups is emphasized by educators who take part in designing educational tests and developing assessment processes. They attempt to detect irrelevant factors that have an impact on the construct validity of the test. They consider it necessary and important to collect evidence to justify the validity and fairness of the tests, to eliminate the random and systematic errors seen as confounding factors. Moreover, the test is supposed to be valid for examinees of different groups categorized by gender, age, and background. In view of the above mentioned, educators also try to change testing policies. For example, the

European Federation of Psychological Association (EFPA) has introduced a model for collecting evidence of construct validity. It uses DIF (Differential item functioning) as a method to assess the quality of the test. Furthermore, the Test Commission of the Spanish Psychological Association (TCSPA) has supported the EFPA in DIF analysis of the context of test fairness [3].

The commonly used methods of assessing the reliability and validity of test scores are the methods of Differential item functioning (DIF) and Differential test functioning (DTF) which are intended to examine the factorial structure of a test and differential functioning at the item level and test level.

Current state of the issue. The brief outline of DIF and DTF methods is provided below.

The Differential item functioning is commonly defined as a situation when the test item exhibits DIF based on the probability of correct response to the item that differs across subgroups with the same ability level [5]. The types DIF are classified as the uniform DIF and non-uniform DIF. An item referred to as uniform-DIF exhibits the difference consistency in item performance when it favours the certain subgroups across the entire range of ability. The opposite case of DIF is identified as non-uniform [5]. DIF serves as an indicator of item bias. It also indicates that there is the secondary latent trait along with the primary latent trait that an item is supposed to measure. Yet, that secondary latent characteristic does not always indicate bias or a cause biased assessment. For this reason, the item where the secondary latent trait, if it occurs along with the primary trait, is not marked as a cause of biased assessment though it leads to different results across the sub-groups. This means that the results might differ between women and men but it reflects the true ability difference and does not cause unfairness. Additionally, bias is viewed as a systematic error in test administration and contents. This error relies on both statistical tests and expert opinions regardless DIF that only relies on statistical tests [4].

DIF can be detected using parametric and non-parametric methods. They are chosen by the researcher since both have pros and cons [4]. For example, some of them, such as Mantel Haenszel and Rasch methods are effective with a small sample but ineffective to detect non-uniform DIF. The other methods such as IRT based Raju's area method and Lord's Chi-Square method are appropriate for a large sample size but ineffective to detect non-uniform DIF [7]. The above mentioned methods are exploratory ones. They are widely used to detect differential item functioning for categorical variables such as gender, nationality, and age groups. Besides detecting DIF, it is also essential to identify the possible source and cause of occurrence of DIF.

The DIF analyses should be followed by running DTF analyses because the items are insignificant and unreliable compared to the whole test.

The literature review found that Gierl et al [3] emphasised the importance of running the DTF analyses accompanied by the DIF analyses because, according to Hunter, the items are insignificant and unreliable compared to the whole test. The total amount of DIF generally effects the test scores even when there is no item detected as DIF in a test. Moreover, when these DIF items favor different subgroups and the DTF values are negligibly small, DIF effects cancel each out [5]. Since the test administrators' decisions about examinees are not made at item-level, but at the test-level, DTF is also significant. It is interesting that the academic test scores and criteria dependent on such factors as performance in studies, and even if they have been analyzed using DIF and DTF, they still involve bias caused by the subsets of items favoring a particular group [4].

This was supported by findings of Guo and colleagues [4], who examined the performance of gender groups on sentence-completion and reading comprehension questions using standardized DIF and the Mantel-Haenszel methods. It has been found that there was the content specific DIF in sentence-completion items in which 13 non-native test-takers were other factors that might cause DIF. It was caused by foreign language deficiencies of the test-takers as they took a test in a language other than their mother tongue. That meant that the deficiencies in their language skill or failure in wording the content clearly in the item might lead to DIF between sub-groups.

Overall, numerous items that are in favor of a certain group and when unintended construct irrelevant factors are defined as a source of DIF and can result in DIF and DTF item bias and test fairness violation. Additionally, the fairness of test scores is achieved when there is a relatively small number of DIF items and negligibly small DTF effects.

Differential Item Functioning (DIF) analyses and Differential Test Functioning (DTF) analyses are important prerequisites of valid and reliable test results. It is essential when the final test is designed with the methods of DIF and DTF, in particular for the ESP course.

The DIF methods that rely on the item response theory (IRT) are commonly used for a latent variable such as the ability to estimate. This review found several IRT-based methods to distinguish DIF items. Those were as follows: Lord's Chi-square, Raju's area method, likelihood ratio test (LRT method), and item drift method [4], [7]. This study used Lord's Chi-Square DIF method for the reason being that it employs more than one parameter to detect uniform (UDIF) and non-uniform DIF (NUDIF). Below is the formula for Lord's Chi-square DIF methods:

$$Q_j = \left(v_{jR} - v_{jF}\right)' \left(\sum jR - \sum jF\right)^{-1} \left(v_{jR} - v_{jF}\right)$$

Note: V_{jR} = (a_{jR}, b_{jR}, c_{jR}) – the vectors of item parameters attributed to the control (reference) group and V_{jF} = (a_{jF}, b_{jF}, c_{jF}) – the vectors of item parameters attributed to the focal group; $\sum jR$ – the variance-covariance matrices of the control (reference) group; $\sum jF$ – the variance-covariance matrices of the focal group; the Q_j-statistics relies on Chi-square distribution, its degrees of freedom is supposed to be equal to the number of estimated parameters [1].

The values drawn from the DTF analyses should correspond to the total amount of DIF for the entire test. According to Hunter [5], they have to be equal to the sum of item DIF statistics in a test. The review found two major methods to yield DTF data. These are Raju's DFIT [10], and Mantel-Haenszel/Liu-Agresti (MH-LA) method [2]. This study utilised the MH-LA method to calculate DTF attributed to the Final Test in English for Specific Purpose (FT-ESP). The study derives the formula for the method from Camilli & Penfield [2], which is presented below:

$$\tau^2 = \frac{\sum_{i=1}^{I} (\hat{\varphi}_1 - \hat{\mu})^2 - \sum_{i=1}^{I} s_i^2}{I}$$

Note: I – the number of test items; $\widehat{\psi_\iota}$ – MH log-odds ratio statistics; μ – mean; and s_i^2 – the error variance of ψ.

The weighted t^2 formula is as follows:

$$\tau^2 = \frac{\sum_{i=1}^{I} W_i^2 (\hat{\varphi}_1 - \hat{\mu})^2 - \sum_{i=1}^{I} W_i}{\sum_{i=1}^{I} W_i}$$

Note: $w_i = s_i^{-2}$

Aim of research was to examine how the use of the methods of DIF and DTF contribute to the quality of the final assessment (FT-ESP) in the ESP course delivered to the graduate students at tertiary institutions.

Research methods. The study relies on two interventions intended to identify the correlation between the test design and the academic performance of the students in the ESP course through using Pearson's correlation coefficient (r) of the answered versus the unanswered questions. The study attempted to identify what was the factorial structure of ESP test for the entire test and each gender group; whether the items of ESP test functioned differently across gender (Female vs. Male); what was the distribution of DIF items across sub-domains (reading comprehension, structure (use of language), and compositional analysis); when each domain was treated as a separate test; whether the test scores of ESP test exhibit differential test functioning (DTF) across gender (Female vs. Male); whether the scores of ESP test exhibit differential test functioning (DTF) across gender, when each domain was treated as a separate test. The first intervention test was similarly structured as the one for the second intervention and consisted of the same number of items. In the first intervention, a regular final ESP test was administered a month before the course was complete. The test included four skills such as reading, listening, writing, and speaking. Overall, the test consisted of 45 items. To examine the factorial structure of EPT data, it was also necessary to see whether the assumption of unidimensionality is met since the IRT based DIF method will be implemented. A test was supposed to be unidimensional when there was one dominant factor (or latent variable) that underlied the scores obtained from the test. Thus, a one-factor CFA model was tested and fit measures of this one-factor CFA model were compared to see if the one-factor model fits the data. Besides, the one-factor CFA model was tested for both males and females to see whether the factorial structure remained the same across gender.

In the second intervention, the originally designed test whose validity and reliability was analyzed using the methods of DIF and DTF was administered. The data for this study were drawn from the final tests in the ESP administered to 40 (15 males and 25 females) graduate students for the Institute of Public Administration and Research in Civil Protection. The test included three sub-domains such as reading comprehension (15 items), structure (use) of English (15 items), and compositional analysis (15 items). Speaking was administered face-to-face. The DIF method based on IRT was implemented to examine the factorial structure of the test. After that, the Lord's Chi-Square DIF method was utilized to identify the items that exhibit DIF. The significance level was 0.01 with the specification of the threshold that is equal to 9.210. The Mantel-Haenszel/Liu-Agresti DTF method was used to test the effects of DIF items at the test scores that might lead to unfair assessment. The values drawn from the DTF analyses (t2) that were smaller than 0.07 were considered to be negligibly small, the values (t2) between 0.07 and 0.14 were considered to indicate a moderate effect and the values larger than 0.14 indicated a substantial effect. Thus, the values for the DTF analyses that were larger than 0.14 were used as an indicator of considerable DTF for the Final test in ESP. The model based on one-factor confirmatory factor analyses (CFA) was used to see whether the factorial structure of the test was similar for both genders.

The students' results obtained from the regular final ESP test showed that the group was approximately homogeneous (mean = 78 %, ECTS). It suggested the group was a reliable sample for this study.

Results and discussion. The results of the correlation analyses (*r*) between the answered and unanswered questions of the regular final ESP test was –0.3448 which meant that the relationship between the scores for the answered and unanswered questions was weak. This suggested that the test used in the first intervention could be tentatively considered reliable and valid.

The results of the administration of the test used in the second intervention are presented below. The data obtained from the DIF analyses of each subdomain of the FT-ESP that was administered in the second intervention are presented in Table 1.

Table 1

DIF analyses results of each subdomain of the FT-ESP

Reading comprehension			*Structure*			*Compositional analysis*		
Item#	*Statistics*	*p-value*	*Item#*	*Statistics*	*p-value*	*Item#*	*Statistics*	*p-value*
rc1	1.24	0.53	st1	0.54	0.76	ca1	4.35	0.11
rc2	0.08	0.95	st2	7.94	0.01	ca2	0.55	0.75
rc3	1.55	0.45	**st3**	**11.87**	**0.003**	ca3	2.70	0.25
rc4	**24.10**	**0.00**	st4	0.34	0.84	ca4	1.11	0.57
rc5	0.76	0.68	st5	1.95	0.37	ca5	3.52	0.17
rc6	0.15	0.92	st6	4.64	0.09	ca6	4.75	0.09
rc7	**16.81**	**0.00**	st7	1.52	0.59	ca7	2.47	0.29
rc8	6.88	0.03	st8	0.31	0.85	ca8	0.26	0.87
rc9	2.55	0.27	st9	0.16	0.92	ca9	2.06	0.35
rc10	3.28	0.22	st10	4.63	0.10	ca10	4.87	0.08
rc11	5.22	0.07	st11	0.70	0.75	ca11	0.54	0.76
rc12	2.96	0.27	st12	0.41	0.81	ca12	6.22	0.04
rc13	5.67	0.05	st13	7.07	0.02	ca13	3.67	0.16
rc14	2.03	0.36	st14	2.63	0.26	ca14	2.39	0.33
rc15	0.13	0.93	st15	3.28	0.19	ca15	2.72	0.25

The results of DIF analyses in Table 1 suggested that 2 items (rc4, rc7) in reading comprehension and 1 item (st3) in structure domains were identified as DIF. Those values indicated the inconsistency of the interclass correlation coefficient (ICC) between males and females in some abilities.

The Mantel-Haenszel/Liu-Agresti differential test functioning method which is based on variance estimates of DIF items, was used to examine DIF at the test level. The results of the DTF analyses of the test are provided in Table 2.

Table 2

Results of DTF analyses for the entire test and each subdomain

Test/subdomain	*Statistics*	*Value*	*SE*	*Z*
FT-ESP-entire test	t^2	0.06	0.01	5.66
	Weighed t^2	0.06	0.01	6.00
Reading Comprehension	t^2	**0.09**	0.03	3.03
	Weighed t^2	**0.07**	0.02	3.00
Structure	t^2	0.06	0.01	4.18
	Weighed t^2	0.06	0.01	4.00
Compositional Analysis	t^2	0.03	0.01	2.66
	Weighed t^2	0.03	0.01	2.72

As can be seen in Table 2, the values yielded from the DTF analyses of the variance for the entire test (0.06) is less than 0.07 which indicates that the DTF effect is negligibly small. The above figures also indicate that at the test level, test scores do not function differently for males and females. The

DTF results indicate that DIF effect cancels each out at test level, because for some of them females outperform males, while males outperform females for the others. For compensatory DIF, there is a cancellation effect in which the DIF effect may cancel each out in the presence of items favoring different subgroups at test level [8], [9]. These results assure that EPT test scores does not function differently across gender and supports the fairness and validity of the test results at the test level. Although there were three items identified as showing DIF in Table 1, the results of DTF analyses indicated that DIF effect cancels each out at test level. The values for DTF related to reading comprehension (0.09) fell within 0.07 and 0.14 indicating a moderate DTF effect. The relatively larger DTF effect associated with the reading comprehension domain might be an indicator of the existence of a construct-irrelevant latent factor such as the degree of vocabulary knowledge of test takers that have a gentle effect on test results [6]. Moreover, the relatively larger DTF effects associated with reading comprehension and structure domains reveal that the existence of DIF effects at item level influences the DTF results. These results might also imply the existence of content specific DTF effect.

The correlation analyses (*r*) of the answered versus unanswered questions of the second intervention Final ESP test was 0.8951 (the p-value is < .00001; the result is significant at p < .05) which meant that the relationship between the scores for the answered and unanswered questions was good. This suggested that the methods of differential item functioning and differential test functioning significantly improved the quality and the credibility of the final ESP test.

The results of DIF analyses showed that 3 items (rc4, rc7, and st3) in the FT-ESP exhibit DIF regardless the domain of the test. When it comes to the distribution of DIF items across sub-domains, two DIF items were associated with the reading comprehension domain and one item was found within the structure domain. Interestingly, no item from the compositional domain was identified as exhibiting DIF. Additionally, each sub-domain of the test can be treated as an independent test with its parallel results. Besides, this study found the existence of content specific DIF effect for the entire test. The implication on the content effect agrees with the conclusion of Martinková et. al. [8] claiming that unintended latent traits and unintended content-related factors can increase the likelihood of manifestation of DIF when doing the test. It was also discovered that DTF analyses results cancel the DIF effects at test level in cases due to the fact that females outperform males in some questions and males outperform females in the others. The results drawn from this study imply that the FT-ESP does function similarly for both genders and ensures the fairness and validity of the test results at the test level. The validity of the above has been increased by calculation of the Pearson's correlation (*r*) between the number of the answered and unanswered (or answered incorrectly) questions from the Final ESP test. The relatively higher figures for the DTF effect (see Table 2) in the reading comprehension domain might be associated with the existence of a construct-irrelevant latent factor such as the degree of vocabulary knowledge of the students that influence the test results, which complies with findings of Jang & Roussos [6].

The above results are consistent with Chubbuck and his colleagues (2016) who examined the performance of gender groups on sentence-completion and reading comprehension questions using the Mantel–Haenszel and standardized DIF methods. They found out the content specific DIF in sentence-completion items in which males outperformed females in reading comprehension items [12]. The findings of the aforementioned studies support the results of this study concerning the occurrence of content specific DIF. Another factor that might cause DIF is the language skills of non-native test takers that take a test in a language other than their mother tongue. The deficiency in their language skill or failure in wording the content clearly in the item might lead to DIF between sub-groups [12].

The results of DIF and DTF induce item bias and violation of test fairness when a large number of items are in favor of a certain group and when unintended construct irrelevant factors are defined as a source of DIF [11]. Thus, the relatively small number of DIF items and negligibly small DTF effects of the entire test indicate that the fairness of test scores is achieved for the ESP test.

Conclusions and research perspectives. The use of the methods of DIF and DTF boosts the quality of the assessment system in the ESP course delivered to the graduate students at tertiary institutions. These methods help shape the assessment system through putting the learner at the centre of the learning process. The methods seem appropriate to be applied to student summative assessment, student formative assessment and diagnostic assessment. The designed tests based on the use of DIF and DTF can ensure fair measurement of the progress and performance of individual students, plan the further steps aimed at improving teaching and learning.

DIF analysis is one of the most important methods employed to ensure the validity of the test and fairness of test score interpretation [11], [12]. It is advisable that the first step in DIF analyses was related to the use of statistical methods to detect the DIF items. After this, the teacher should decide whether to remove or to revise those items because the statistically significant results of DIF analyses do not always indicate biased items. It requires a comparison of differential functioning results at item and test level and involvement of experts for the final decision. The DIF detected items can be dealt with different approaches. Some test designers suggest removing DIF items to reduce DTF effect and others suggest examining the structure of the test and items before removing DIF items and try to specify the cause of differential functioning [8]. It is also advisable to examine the effects of other potential factors on DIF such as item order and mother tongue effects along with unintended content specific factors to explain DIF effect in the context of language testing. Therefore, items with substantially high DIF values (rc4 and rc7 items) should be examined by content experts. Because, removing DIF items without any evaluation does not ensure the fair test [3], [5], [8], specifically, when DTF effects of test forms are negligibly small and DIF effects cancel each out at test level. There are even some other researchers claiming that removing DIF items may lead to weaker tests (rather than fair test) regarding the representation of construct and variance explained by these items [8]. Therefore, consulting with test developers and content experts before removing the DIF items is suggested. It is also suggested investigating the effects of other potential factors on DIF such as item order and mother tongue effects along with unintended content specific factors to explain DIF effect in the context of language testing. The results also imply that neither of methods addresses the issue of measurement bias, which might occur in tests, because it is complicated and cannot be addressed adequately using simple statistical or classical test theory methods. According to Stark, Chernyshenko & Drasgow, at the item level, bias refers to differences in the probability of correctly answering an item among individuals having the same level of ability but belonging to different groups. At test level, bias refers to differences in the expected total scores for those same individuals [13].

Further studies are needed to identify the ways of improving the assessment of the speaking skills of the graduates of the tertiary institutions.

REFERENCES

1. Camilli, G. (2006). Test fairness. In *Educational Measurement,*4th ed., 221-256. Westport: American Council on Education & Praeger Publishers [In English].

2. Camilli, G., & Penfield, D. (1997). Variance estimation for differential test functioning based on Mantel-Haenszel statistics. *Journal of Educational Measurement*, *34*, 123-139. DOI: https://doi.org/10.1111/j.1745-3984.1997.tb00510.x [in English].

3. Gierl, M., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2005). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences in achievement tests. *Educational Measurement: Issues and Practice, 20 (2),* 26-36. DOI: https://doi.org/10.1111/j.1745-3992.2001.tb00060.x [in English].

4. Guo, H., Robin, F. & Dorans, N. (2017). Detecting Item Drift in Large Scale Testing. *Journal of Educational Measurement, 54 (3),* 265-284. DOI: https://doi.org/10.1111/jedm.12144 [in English].

5. Hunter, C. (2014). A simulation study comparing two methods of evaluating differential test functioning (DTF): DFIT and the Mantel-Haenszel/Liu-Agresti variance. Doctoral Dissertation, Georgia State University, Atlanta, GA, United States. Retrieved from: https://scholarworks.gsu.edu/cgi/viewcontent.cgi?article=1132&context=eps_diss [in English].

6. Jang, E.E., & Roussos, L. (2009). Integrative analytic approach to detecting and interpreting L2 vocabulary DIF. *International Journal of Testing, 9 (3),* 238-259. DOI: https://doi.org/10.1080/15305050903107022 [in English].

7. Kim, S-H. & Cohen, A.S. (1995). A Comparison of Lord's Chi-Square, Raju's Area Measures, and the Likelihood Ratio Test on Detection of Differential Item Functioning. *Applied Measurement in Education, 8 (4),* 291-312. DOI: https://doi.org/10.1207/s15324818ame0804_2 [in English].

8. Martinková, P., Drabinová, A., Liaw, Y., Sanders, E.A., McFarland, J.L., & Price, R.M. (2017). Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments. *CBE-Life Sciences Education, 16 (2),* 1-13. DOI: https://doi.org/10.1187/cbe.16-10-0307 [in English].

9. Penfield, R. & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed-format tests. *Journal of Educational Measurement, 43,* 295-312. DOI: https://doi.org/10.1111/j.1745-3984.2006.00018.x [in English].

10. Raju, N., Van der Linden, W., & Fleer, P. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19 (4),* 353-368. DOI: https://doi.org/10.1177/014662169501900405 [in English].

11. Zhu, X., & Aryadoust, V. (2020). An investigation of mother tongue differential item functioning in a high-stakes computerized academic reading test. *Computer Assisted Language Learning, 33,* 1-24. DOI: https://doi.org/10.1080/09588221.2019.1704788 [in English].

12. Wedman, J. (2018). Reasons for Gender-Related Differential Item

Functioning in a College Admissions Test. *Scandinavian Journal of Educational Research, 62 (6),* 959-970. DOI: 10.1080/00313831.2017.1402365 [in English].

13. Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the Effects of Differential Item (Functioning and Differential) Test Functioning on Selection Decisions: When Are Statistically Significant Effects Practically Important? *Journal of Applied Psychology, 89 (3),* 497-508. DOI:10.1037/0021-9010.89.3.497 [in English].

Received: June 19, 2020
Accepted: July 21, 2020