Cultural Domain in Development and Validation of Tests on Arts to Assess the Skills of Student Graphic Designers

MARIANNA O. NEROZNA[1], IEVGEN P. GULA[1], MYKHAILO F. RUDENKO[1], OKSANA V. MAZNICHENKO[1], OLEKSANDRA P. BASANEC[1], VOLODYMYR F. PRUSAK[2]

[1]Department of Drawing and Painting, Faculty of Design, Kyiv National University of Technologies and Design, Kyiv, UKRAINE

[2]Department of Design, Institute of Woodworking, Computer Technology and Design, Ukrainian National Forestry University, Lviv, UKRAINE

Abstract: - The purpose of the study was to examine whether the presence of items that covered cultural questions in the test on arts for student graphic designers influenced the fairness of the test across different ethnic and social groups. The reason for the study was to address the gap in the development and validation of tests on arts that include a cultural domain to assess the skills of undergraduate graphic designers. The research design was based on DIF and DTF methods to examine the factorial structure of test data and to identify whether the factorial structure of the test was the same across ethnic and social groups. A one-factor CFA model was applied to perform measurements for categorised ethnic and social status groups to identify whether the factorial structure was similar or identical for them. The goodness-of-fit measures were calculated using the chi-square statistics, CFI, TLI, and RMSEA to identify how the obtained data is consistent with the hypothesised model. The inclusion of local culture-related questions in the tests on arts to assess the skills of student graphic designers influences the individuals' latent traits which lead to an increase in DIF values. Though there were detected seven items with the DIF for the whole test, the DTF measurements showed that the DIF effect eliminated each out at test level which is related to the fact the certain ethnic groups performed better in answering some specific test items, while certain status groups performed better in answering some other test items. It is noteworthy that DTF for the Ukrainian art (miscellaneous) section was between 0.07 and 0.14 meaning a moderate DTF effect. However, the DTF variance values for the sections of principles of design and principles of art were lower than 0.07 meaning a small effect. Therefore, it could be concluded that small DTF effects found in both the whole test and each test section separately indicated that DIF effects eliminate each other at the test level.

Key-Words: - Higher education, student graphic designers, test fairness and validity, students' skills assessment, ethnic and social status bias, differential item functioning, differential test functioning

Received: April 17, 2021. Revised: November 2, 2021. Accepted: December 2, 2021. Published: January 3, 2022.

1 Introduction

The development of fair tests on arts containing a cultural domain, specifically the local culture-based one, to assess the skills of student graphic designers is the foremost concern of the relevant university teachers because different student ethnic and social groups perceive cultural questions differently. The concern is related to measurement and cultural biases which are difficult to deal with adequately employing common statistical or classical methods [1-2]. Since the cultural domain relies on latent constructs such as attitudes and perceptions across different groups when common statistical methods are used, there are two causes of measurement bias, first, the item-level bias which is related to variations in answering the question by individuals with the same level of ability but come from different social or ethnic groups, second, the test-level bias which is associated with dissimilarities in the estimated total scores for the test-takers who are found to be homogeneous in their level of ability but belong to different groups. The item response theory (IRT) methods seem effective in dealing with measurement bias through the use of methods of Differential Item functioning (DIF) and Differential Test Functioning (DTF) because these allow identifying whether the developed test provides valid and reliable (unbiased) results [3-4]. In the literature, the above methods are found to be used for checking the reliability and validity of the tests in languages, tests for social studies like psychometrical or psychological tests [5-8]. However, the study found a gap in the development and validation of tests on arts that include a cultural domain to assess the skills of undergraduate graphic designers.

1.1 Problem Formulation

The review found that DIF and DTF methods are often used to validate the tests and they showed that DIF and DTF can occur at the domain level, level of the entire test, level of individual and group [9-10]. The methods are based on the item response function concept which is referred to as a mathematical formula that relies on one or more parameters used to identify how the probability of a specific response to a dichotomous question is related to the level of manifestation of a latent trait [4]. Both DIF and DTF methods fall under the Item Response Theory (IRT) which is described in the literature as a direction of a conventional theory of measurement which is based on three constructs such as the factual score, the observed or actual score, and the coefficient of reliability [11]. The IRT models suppose that unidimensionality at both test and test item levels should be considered together with the latter three constructs as there is the assumption that item scores might be affected by the latent constructs [12]. The above concepts are discussed in greater detail by [13-16].

Since cultural and social issues often cause test bias which leads to unreliable results, this inspired this research and proved it to be feasible.

The purpose of the study is to examine whether the presence of items that cover cultural questions in the test on arts for student graphic designers influence the fairness of the test across different ethnic and social groups.

The research questions were as follows: a) whether the factorial structure of the test on arts consisting of cultural questions meets the assumption of unidimensionality before the DIF method is used; b) whether items of the test on arts consisting of cultural questions function differently across ethnic and social groups; c) whether the distribution of DIF items across the cultural sub-domain is different; d) whether the entire test scores of the test show differential test functioning (DTF) across ethnic and social groups when each domain is treated as a separate test.

2 Methods and Materials

The research design was based on the research questions that were supposed, first, to examine the factorial structure of test data to identify whether the factorial structure of the test was the same across ethnic and social groups. To carry out this, a one-factor CFA model was applied to perform measurements for categorised ethnic and social status groups to identify whether the factorial structure was similar or identical for them. The goodness-of-fit measures were calculated using the chi-square statistics, CFI, TLI, and RMSEA to identify how the obtained data is consistent with the hypothesised model. The items of the test were also examined for facial bias by three experts in linguistic psychology with a Ph.D. degree [2]. When the IRT model was prepared, the test was uploaded to Assess.ai (can be accessed via https://grltd.assess.ai/) which was employed to identify items that exhibited DIF. The reference values were set to be less than 0.05 for the significance level and 9.20 for the detection threshold. Further to this, the differential test functioning (DTF) method (the Mantel-Haenszel/Liu-Agresti method) was used to identify how the DIF items were related to the test scores that seemed to indicate the unfair assessment [17]. This phase relied on the criteria (reference values) to identify the DTF for the Mantel -Haenszel/Liu-Agresti DTF method such as <0.07 is a small DTF effect, from 0.07 to 0.14 is a medium DTF effect and >0.14 is a large DTF effect [18]. Give the above, the values of >0.14 were used as reference ones when calculating DTF statistics.

2.1 Sampling

A single-stage cluster sampling technique was used to hire students majoring in graphic design at Kharkiv State Academy of Design and Arts (KSADA), Ukraine; Trans-Carpathian Academy of Arts (TCAA), Ukraine; Kyiv National University of Technologies and Design (KNUTD), Ukraine; Kyiv State Academy of Decorative and Applied Arts and Design named after Mykhailo Boychuk (KSADAAD), Ukraine; and Lviv National Academy of Arts (LNAA), Ukraine. The sampling procedure was organised as a flow of four steps. First, the population of 278 students was defined. This number of student population seemed representative because it was equal to about 20-25 % of the total number of undergraduates of the graphic design major. Second, the population was divided into clusters of between 50 and 56 people each. Third, the students were informed about the purpose and specifics of the study and they were randomly invited to participate in it. Fourth, those students who agreed formed the cluster-based sample for the study. The number of clusters corresponded to the number of ethnic groups distinguished by nation, religion, culture, and social treatment. The number of subjects in the cluster was supposed to be approximately equal. It was between 20-25 students per cluster which aligned with previous research [19-20]. The key inclusion criteria for the test to have been taken were as follows: a respondent left the information about their gender, age, ethnic group (then categorised regionally as Western Ukrainians, Central and Southern Ukrainians, and Eastern Ukrainians), religious confession belonging (Orthodox, Catholic, Muslim, Other), and social status of their family (categorised as low-income, middle income, and high income).

Table 1. The demographic features of the sampled students

Feature		Institution					***Mean***	***SD***
		KSADA ***n, (%)***	***TCAA*** ***n, (%)***	***KNUTD*** ***n, (%)***	***KSADAAD*** ***n, (%)***	***LNAA*** ***n, (%)***		
Gender	Males, $n = 58$	13, (22.42)	11, (18.97)	12, (20.68)	10, (17.25)	12, (20.68)	11.6	1.019
	Females, $n = 56$	8, (14.28)	11, (19.65)	13, (23.21)	12, (21.43)	12, (21.43)	11.2	1.720
Age	20	4, (15.38)	7, (26.93)	5, (19.23)	5, (19.23)	5, (19.23)	5.2	0.979
	21	7, (20.59)	8, (23.53)	5, (14.71)	6, (17.64)	8, (23.53)	6.8	1.166
	22	10, (18.51)	7, (12.96)	15, (27.77)	11, (20.37)	11, (20.37)	10.8	2.561
Ethnic	Western Ukrainians	1, (2.22)	19, (42.22)	7, (15.56)	5, (11.12)	13, (28.88)	9.0	6.324
	Central and Southern Ukrainians	3, (8.82)	2, (5.89)	11, (32.35)	10, (29.42)	8, (23.52)	6.8	3.655
	Eastern Ukrainians	17, (51.51)	0, (0.00)	8, (24.24)	7, (21.21)	1, (3.04)	6.6	6.086
Religion	Orthodox	13, (35.13)	3, (8.11)	7, (18.92)	9, (24.32)	5, (13.52)	7.4	3.440
	Catholic	4, (8.52)	18, (38.29)	6, (12.77)	5, (10.64)	14, (29.78)	9.4	5.571
	Muslim	3, (15.79)	0, (0.00)	6, (31.58)	7, (36.84)	3, (15.79)	3.8	2.481
	Other	1, (9.09)	1, (9.09)	6, (54.55)	1, (9.09)	2, (18.18)	2.2	1.939
Social	Low-income	3, (20.00)	2, (13.33)	4, (26.67)	3, (20.00)	3, (20.00)	3.0	0.632
	Middle income	15, (21.42)	13, (18.58)	14, (20.01)	11, (15.72)	17, (24.28)	14.0	2.000
	High income	3, (10.34)	7, (24.14)	7, (24.14)	8, (27.59)	4, (13.79)	5.8	1.939

2.2 Ethical Considerations

The ethical considerations were addressed via the voluntary participation of the testees in the pilot study, anonymous testing without collecting the respondents' names and surnames (the participants were given a testee code). The participants were not forced to provide information about their social status or ethnic origin, however, the tests that did not provide these data were eliminated from the study. The questions for the test were formulated in a way so that offensive, bullying, or discriminatory language was avoided. The confidentiality of the research data related to the ethnic and social status of the respondents was ensured.

2.3 Instruments

The test on arts (can be accessed via the link https://forms.gle/KGTV8wrYcnPDRQ3T7) was specifically designed for student designers to cover the principles of design, principles of art, and Ukrainian art (miscellaneous). The test was intended to test students' knowledge in Chromatics, Fundamentals of shaping, Layout and Composition in graphic design, principles of Art and Ukrainian folk arts and crafts, and famous artists. The test consisted of 34 dichotomously scored and multiple-choice questions that were divided into three sections (sub-domains). The first section entitled The Principles of Design included 12 questions. The second section entitled the Principles of Art comprised 12 questions. The third section entitled Ukrainian art consisted of 10 questions.

3 Results

The results of a one-factor CFA model in which each section is considered to be a factor such as the entire test, ethnic groups, and social groups are presented in Table 2. As can be seen in Table 2, the CFA results show the unidimensionality of the test.

Table 2. The results of a one-factor CFA model applied to the test distributed by ethnic groups and social groups

Group	χ^2	**CFI**	**TLI**	**RMSEA**	**90% for RMSEA**		**df**
					LL	**UL**	
Ethnic groups	126.124	0.974	0.972	0.027	0.026	0.028	112
Social groups	127.201	0.953	0.959	0.027	0.026	0.028	112

All	134.119	0.965	0.967	0.031	0.029	0.032	112

The values for CFI and TLI that are >0.95 show a good fit between the model and data for every separate factor [21]. The RMSEA values for the groups and the entire test are lower than 0.06 (reference value) with a 95% confidence interval which also proves a good fit for group factors.

The Chi-square values ($\chi2$) are lower than the critical value of 137.701 [22] which are expected to be lower and this indicated that there is sufficient evidence to state that there is a relationship between the test data and ethnic and social groups. Overall, the results presented in Table 2 show that the one-factor CFA model illustrates a good fit to the data and the test can be regarded as unidimensional.

The descriptive statistics and coefficients of reliability that were drawn from the whole test and each section are presented in Table 3. The Cronbach's reliability coefficient and composite reliability coefficients with factor loadings based on CFA were computed to ensure more reliable results.

Table 3. The descriptive statistics and coefficients of reliability based on the Cronbach's reliability coefficient and composite reliability coefficients and drawn from ethnic groups and social groups

Test section	**Mean**	***SD***	**Cronbach α**	***r***
PD	8.83	2.11	0.87	0.951
PA	8.55	2.06	0.91	0.821
UA	8.11	2.25	0.83	0.781
ALL	25.22	4.56	0.94	0.913

Note: PD - Principles of Design, PA - Principles of Art, UA - Ukrainian art (miscellaneous).

As can be seen in Table 3, the values for Cronbach α and composite reliability coefficients with factor loadings are sufficiently high for the whole test with α= 0.94 for the whole test and r=0.913 for the composite reliability coefficients, respectively. The statistics show that the difference between the coefficients is negligibly small which proves the unidimensionality of the test.

The DIF results drawn from the whole test are presented in Table 4 and these are unrelated to the test sections. The items are abbreviated with 'PD' standing for Principles of Design, 'PA' standing for Principles of Art, and UA standing Ukrainian art (miscellaneous). The DIF values in the second column are obtained from Assess.ai software.

Table 4. Results of DIF computation drawn from the whole test using Assess.ai software

Item	**DIF**	**p-value**	**Item**	**DIF**	**p-value**
PD1	0.033	0.979	PA6	0.322	0.685
PD2	0.873	0.656	PA7	4.471	0.159
PD3	3.217	0.109	PA8	5.921	0.091
PD4	1.276	0.481	PA9	6.578	0.032
PD5	4.617	0.117	PA10	7.943	0.271
PD6	1.497	0.439	PA11	8.832	0.128
PD7	2.298	0.391	PA12	5.891	1.291
PD8	0.791	0.872	UA1	12.795	0.145
PD9	0.041	0.873	UA2	13.021	0.018
PD10	1.881	0.492	UA3	12.747	0.023
PD11	2.561	0.288	UA4	27.242	0.000
PD12	1.379	0.543	UA5	21.438	0.000
PA1	4.432	0.192	UA6	11.043	0.012
PA2	2.947	0.747	UA7	8.654	0.048
PA3	3.937	0.419	UA8	8.885	0.034
PA4	2.442	0.387	UA9	7.579	0.019
PA5	2.464	0.358	UA10	17.229	0.003

As can be noticed in Table 4, the values seven items (UA1, UA2, UA3, UA4, UA5, UA6, UA10) in the test section entitled Ukrainian art (miscellaneous) – these are highlighted bold –substantially exceeded the reference DIF detection threshold of 9.20, particularly UA4, UA5, and UA10. The Mean value for the other four items was12.489 and this tended to be close to the DIF reference value. It is noteworthy that all seven items belong to the Ukrainian culture block.

Each detected DIF item is illustrated in Figure 1 providing item characteristic curves (ICCs) for the focal (ethnic groups) and reference (social status) groups. The straight line is used to illustrate the focal groups, and the dotted line stands for the reference groups. Both lines are used to clarify the probability of giving the correct answer by the testees from different ethnic and social groups. The coloured space between the lines shows the extend of the DIF effect.

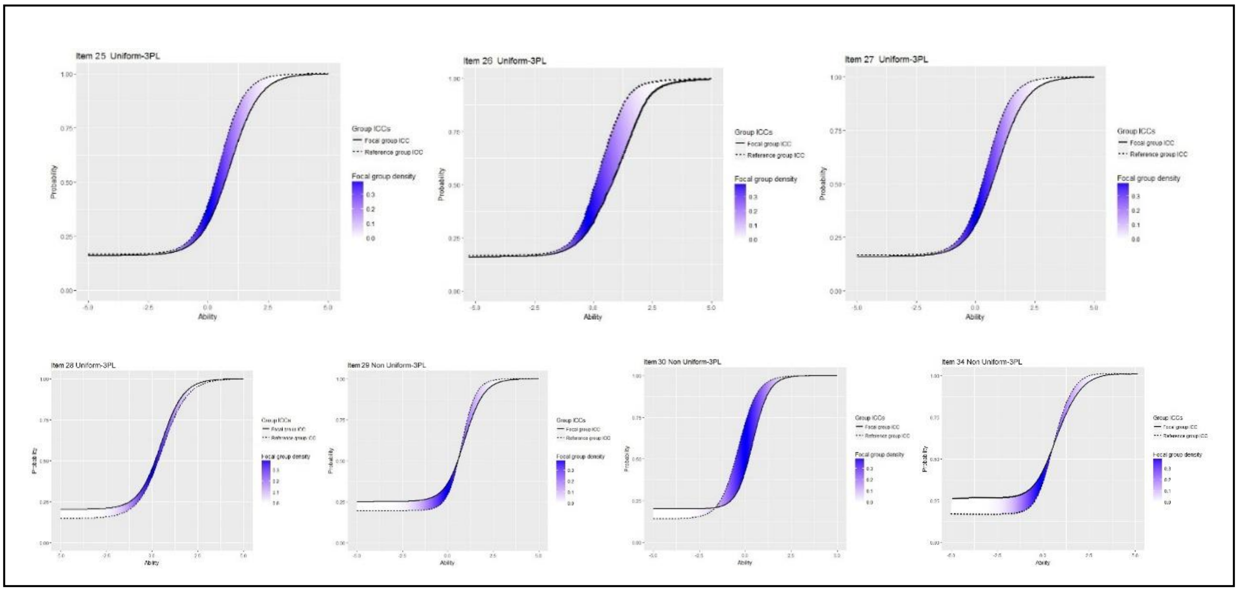

Fig. 1: ICCs for DIF items of the test in focal and reference groups

The DIF results drawn from the analysis of the test sections (principles of design, principles of art, and Ukrainian art (miscellaneous)) separately are provided in Table 5. The DIF results were obtained using the Assess.ai software.

Table 5. DIF results obtained using the Assess.ai software to analyse test sections separately

Principles of design			**Principles of art**			**Ukrainian art (miscellaneous)**		
Item	**DIF**	$p - value$	**Item**	**DIF**	$p - value$	**Item**	**DIF**	$p - value$
PD1	0.471	0.432	PA1	5.222	0.128	UA1	14.522	0.485
PD2	0.791	0.519	PA2	3.741	0.438	UA2	15.887	0.049
PD3	4.172	0.086	PA3	4.351	0.697	UA3	13.229	0.019
PD4	1.224	0.467	PA4	3.239	0.575	UA4	26.873	0.003
PD5	4.455	0.104	PA5	1.659	0.499	UA5	23.963	0.001
PD6	1.976	0.275	PA6	0.489	0.734	UA6	12.227	0.048
PD7	2.854	0.515	PA7	7.781	0.589	UA7	8.669	0.092
PD8	1.072	0.526	PA8	7.557	0.081	UA8	8.982	0.219
PD9	0.067	0.638	PA9	5.678	0.079	UA9	6.922	0.177
PD10	1.717	0.529	PA10	8.112	0.794	UA10	16.775	0.011
PD11	2.916	0.445	PA11	6.275	0.997			
PD12	1.769	0.437	PA12	6.188	1.679			

The DIF values provided in Table 5 shows that the items from the Ukrainian art (miscellaneous) section such as ‘UA1’, ‘UA2’, ‘UA3’, ‘UA4’, ‘UA5’, ‘UA6’, and ‘UA10’ were detected as DIF. It was noteworthy that no items from two other sections were detected as DIF. There were three more items in the test (PA10=8.112, UA7=8.669, UA8=8.982) whose values were close to the DIF detection reference value of 9.20.

The above was followed by the application of the Mantel-Haenszel/Liu-Agresti DTF method to examine DIF at the test level which includes the variance estimates (t^2), weighted variance estimates (Weighted t^2), standard errors (SE), and

Z-scores for the whole test and each test section (see Table 6) [18].

Table 6. Results of the use of a DTF method to examine DIF at test level

Test section	Variance estimates		SE	Z
Principles of design	t^2	0.047	0.016	2.751
	Weighted t^2	0.04	0.013	2.558
Principles of art	t^2	0.063	0.018	4.266
	Weighted t^2	0.06	0.013	4.011
Ukrainian art (miscellaneous)	t^2	0.098	0.041	3.015
	Weighted t^2	0.078	0.032	3.002
Test (integrated)	t^2	0.063	0.017	5.375
	Weighted t^2	0.06	0.01	6.000

The values provided in Table 6 imply that the DTF variance (t^2) for the whole test is <0.07 - (t^2=0.063) which means a small DTF effect. The above results suggested that the scores for the test did not function differently at the test level in both ethnic and social status groups. It meant that the result of the test could be considered fair.

Though there were detected seven items with the DIF for the whole test, the DTF measurements showed that the DIF effect eliminated each out at test level which is related to the fact the certain ethnic groups performed better in answering some specific test items, while certain status groups performed better in answering some other test items. It is noteworthy that DTF for the Ukrainian art (miscellaneous) section was between 0.07 and 0.14 meaning a moderate DTF effect. However, the DTF variance values for the sections of principles of design and principles of art were lower than 0.07 meaning a small effect. Therefore, it could be concluded that small DTF effects found in both the whole test and each test section separately indicated that DIF effects eliminate each other at the test level.

4 Discussion

The attempted to address the research questions such as whether the factorial structure of the test on arts consisting of cultural questions meets the assumption of unidimensionality before the DIF method is used; whether items of the test on arts consisting of cultural questions function differently across ethnic and social groups; whether the distribution of DIF items across the cultural sub-domain is different; whether the entire test scores of the test show differential test functioning (DTF) across ethnic and social groups when each domain is treated as a separate test. The strength of the study is in an attempt to address the issue of the test bias for student graphic designers in Ukraine as well as in the international institutions in other countries.

It was found that the values for CFI and TLI that are >0.95 showed a good fit between the model and data for every separate factor [21]. The RMSEA values for the groups and the entire test were lower than 0.06 (reference value) with a 95% confidence interval which also proved a good fit for group factors.

The Chi-square values (χ2) were lower than the critical value of 137.701 which were expected to be lower and this indicated that there was sufficient evidence to state that there is a relationship between the test data and ethnic and social groups. Overall, the one-factor CFA model illustrated a good fit to the data and the test could be regarded as unidimensional.

The computation of the Cronbach's reliability coefficient and composite reliability coefficients with factor loadings based on CFA showed that the values for Cronbach α and composite reliability coefficients with factor loadings were sufficiently high for the whole test with α=0.94 for the whole test and r=0.913 for the composite reliability coefficients, respectively. The statistics showed that the difference between the coefficients was negligibly small which proved that the test was unidimensional. The use of the Assess.ai software identified the values for seven DIF items (UA1, UA2, UA3, UA4, UA5, UA6, UA10) in the test section entitled Ukrainian art (miscellaneous) – these are highlighted bold – substantially exceeded the reference DIF detection threshold of 9.20, particularly UA4, UA5, and UA10. The mean value for the other four items was 12.489 and this tended to be close to the DIF reference value. It was noteworthy that all seven items belong to the Ukrainian culture block. The DIF items were from the Ukrainian art (miscellaneous) section and it was noteworthy that no items from two other sections were detected as DIF. There were three more items in the test (PA10=8.112, UA7=8.669, UA8=8.982) whose values were close to the DIF detection reference value of 9.20.

The application of the Mantel-Haenszel/Liu-Agresti DTF method to examine DIF at test level which includes the variance estimates (t2), weighted variance estimates (Weighted t2), standard errors (SE), and Z-scores for the whole test and each test section implied that the DTF variance (t2) for the whole test is <0.07 - (t2=0.063) which meant a small DTF effect. The above results suggested that the scores for the test did not function differently at the test level in both ethnic and social status groups.

The study is consistent with the previous research emphasing the importance of DIF analysis in test validation [23]. It goes with [24-25] who claimed that DIF effects could be caused by the content of the test and might take place when latent traits were manifested unintentionally. According to [26], the item bias can take place when the sample is large and the majority of the sample is favoured with certain content which leads to the unfairness of the test. The study aligns with [27] who concluded that the latent factors compensate each other when two different groups are involved.

5 Conclusions

The strength of the study is in an attempt to address the issue of the test bias for student graphic designers. The inclusion of local culture-related questions in the tests on arts to assess the skills of student graphic designers influences the individuals' latent traits which leads to an increase in DIF values. Though there were detected seven items with the DIF for the whole test, the DTF measurements showed that the DIF effect eliminated each out at test level which is related to the fact the certain ethnic groups performed better in answering some specific test items, while certain status groups performed better in answering some other test items. It is noteworthy that DTF for the Ukrainian art (miscellaneous) section was between 0.07 and 0.14 meaning a moderate DTF effect. However, the DTF variance values for the sections of principles of design and principles of art were lower than 0.07 meaning a small effect. Therefore, it could be concluded that small DTF effects found in both the whole test and each test section separately indicated that DIF effects eliminate each other at the test level.

5.1 Recommendations

The practitioners should formulate the questions in a way so that offensive, bullying, or discriminatory language was avoided. The content should be carefully selected or it should be localised taking into account the ethnic and social features of the student population. The researchers should address the issues of tolerating the influences of ethnic and social status latent factors (cultural bias) on test fairness.

5.2 Limitations

Sample size, sampling techniques, and involvement of one major only in the intervention can be considered the limitations of the study.

Acknowledgement:
This study would not have been possible without experts, participants, and all other people who were involved and assisted in the data collection and processing.

Conflicts of Interest:
The authors report the existing no conflict of interest related to financial gains or personal or professional considerations.

References:

[1] Kunnan, A. J., Test fairness, test bias, and DIF, *Language Assessment Quarterly,* Vol. 4, No. 2, 2007, pp. 109-112. https://doi.org/10.1080/15434300701375865.

[2] Kruse, A. J., Cultural bias in testing: A review of literature and implications for music education, *Update: Applications of Research in Music Education*, Vol. 35, No. 1, 2015, pp. 23-31. https://doi.org/10.1177/8755123315576212

[3] Drasgow, F., Nye, D. C., Stark, S. and Chernyshenko, O. S., Chapter 27: Differential Item and Test Functioning, In P. Irwing, T. Booth, and D. J. Hughes (Eds.) *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale, and Test Development*, Hoboken: John Wiley & Sons Ltd., 2018, pp. 885-889. https://doi.org/10.1002/9781118489772.ch27

[4] Nugent, W. R., Understanding DIF, and DTF: Description, methods, and implications for social work research, *Journal of the Society for Social Work and Research,* Vol. 8, No. 2, 2017, pp. 305-334. https://doi.org/10.1086/691525.

[5] Ballangrud, R., Husebø, S. E. and Hall-Lord, M. L., Cross-cultural validation and psychometric testing of the Norwegian version of the TeamSTEPPS® teamwork perceptions questionnaire, *BMC Health Services Research,* Vol. 17, Art. No. 799, 2017. https://doi.org/10.1186/s12913-017-2733-y.

[6] Stark, S., Chernyshenko, O. S. and Drasgow, F., 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,* Vol. 89, No. 3, 2004, pp. 497-508. https://doi.org/10.1037/0021-9010.89.3.497.

[7] Tuncay, O., DIF analysis across genders for reading comprehension parts of English language achievement exam as a foreign language, *Educational Research, and Reviews,* Vol. 10, No.11, 2015, p. 1505-1513. https://doi.org/10.5897/err2015.2284.

[8] Yanlou, L., Hao, Y., Tao, X., Laicheng, S. and Lu, Y., A Comparison of differential item functioning detection methods in cognitive diagnostic models, *Frontiers in Psychology,* Vol. 10, Art. No. 1137, 2019. https://doi.org/10.3389/fpsyg.2019.01137.

[9] Meade, A., A taxonomy of effect size measures for the differential functioning of items and scales, *Journal of Applied Psychology,* Vol. 95, 2010, p. 728-743. http://dx.doi.org/10.1037/a0018966.

[10] Meade, A. and Wright, N., Solving the measurement invariance anchor item problem initem response theory, *Journal of Applied Psychology,* Vol. 97, No. 5, 2012, p. 1016-1031. https://doi.org/10.1037/a0027934.

[11] Finch, W. H., and French, B. F., Item Response Theory (IRT), In *Educational and Psychological Measurement*, New York: Routledge, 2018, pp. 235-276.

[12] Hambleton, R. K., and Zhao, Y., Item Response Theory (IRT) Models for Dichotomous Data, in *Wiley StatsRef: Statistics Reference Online*, Hoboken, NJ: John Wiley & Sons, Ltd., 2014.

[13] Embretsen, S. and Reise, S., *Item response theory for psychologists.* New York, NY: Psychology Press, 2000.

[14] Bonifay, W., Unidimensional item response theory," In *Multidimensional item response theory*, Newbury Park, CA: SAGE Publications, Inc, 2020, pp. 5-26.

[15] Carlson, J. E., Multidimensional Item Response Theory Models, In *Introduction to Item Response Theory Models and Applications*, London: Routledge, 2020, pp. 101-119.

[16] Green, B. F., Book review of Educational measurement by Brennan (Ed.), *Journal of Educational Measurement,* Vol. 45, No. 2, 2008, pp. 195-200. https://doi.org/10.1111/j.1745-3984.2008.00060.x

[17] Fidalgo, A. M., and Madeira, J. M., Generalized Mantel-Haenszel Methods for Differential Item Functioning Detection, *Educational and Psychological Measurement*, Vol. 68, No. 6, 2008, pp. 940-958. https://doi.org/10.1177/0013164408315265

[18] Penfield, R., DIFAS 5.0: Differential item functions analysis system. User's manual, 2013. Available online: https://soe.uncg.edu/wp-content/uploads/2015/12/DIFASManual_V5.pdf (accessed 28.06.2021).

[19] French, B. F., and Finch, W. H., Extensions of the Mantel-Haenszel for multilevel DIF detection, *Educational and Psychological Measurement,* Vol. 73, 2013, pp. 648-671. https://doi.org/10.1177%2F0013164412472341

[20] Hox, J. J. and Maas, C. J. M., The accuracy of multilevel structural equation modelling with pseudobalanced groups and small samples, *Structured Equation Modelling: A Multidisciplinary Journal,* Vol. 8, 2009, pp. 157-174. https://doi.org/10.1207/S15328007SEM0802_1.

[21] Cho, G., Hwang, H., Sarstedt, M. and Ringle, C. M., Cutoff criteria for overall model fit indexes in the generalized structured component analysis, *Journal of Marketing Analytics,* Vol. 8, 2020, pp. 189-202. https://doi.org/10.1057/s41270-020-00089-1

[22] Zach, Chi-square Distribution Table, *Statology*, 2018. Available online: https://www.statology.org/chi-square-distribution-table/ (accessed 28.06.2021).

[23] Hope, D., Adamson, K., McManus, I. C., Chris, L. and Elder, A., Using differential item functioning to evaluate potential bias in high stakes postgraduate knowledge-based assessment, *BMC Medical Education*, Vol. 18, 2018, p. 1-7. https://doi.org/10.1186/s12909-018-1143-0

[24] Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon F., and Lacroix S., Application of think-aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews, *Educational Measurement: Issues and Practice,* Vol. 29, 2010, p. 24-35. https://doi.org/10.1111/j.1745-3992.2010.00173.x

[25] Martinková, P., Drabinová, A., Liaw, Y., Sanders, E. A., McFarland, J. L. and Price, R. M., Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments, *CBE Life Science of Education*, Vol. 16, No. 2, 2017,

pp. 1-13. https://doi.org/10.1187/cbe.16-10-0307

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

[27] Ozdemir, B. and Alshamrani, A. H., Examining the Fairness of Language Test Across Gender with IRT-based Differential Item and Test Functioning Methods, *International Journal of Learning, Teaching and Educational Research*, Vol. 19, No. 6, 2020, p. 27-45. https://doi.org/10.26803/ijlter.19.6.2.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Marianna Nerozna: concept and design, critical revision of manuscript.

Ievgen Gula: data acquisition, data analysis and interpretation, drafting manuscript.

Mykhailo Rudenko: data acquisition, data analysis and interpretation, drafting manuscript.

Oksana Maznichenko: drafting manuscript, critical revision of manuscript.

Oleksandra P. Basanec: data acquisition, data analysis, drafting manuscript.

Volodymyr Prusak: supervision, critical revision of manuscript, final approval.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

There is no source of funding of this research

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)