The role of different test context in computerized scenario-based assessment

Gracheva Daria
grachdasha14@gmail.com
Uglanova Irina
iluglanova@gmail.com
Brun Irina
ibrun@hse.ru
Higher School of Economics

Topic: E-assessments and technology-advanced platforms

Keywords: performance-based assessment, comparability, context

The purpose of the paper is to provide comparability evidence of different forms of Performance-based task aimed at measuring critical thinking skill. PBT "Aquarium" include three texts with the same key-sentences, but additional (non-key) sentences are vary. For the psychometric analysis, we consider three texts as three forms of the same task. Additional sentences and characteristics of the interface are considered as test context. To check the comparability of the forms we use classical equating methods and DIF-analysis. Moreover, we cross-validate the results using item response theory modeling (IRT).

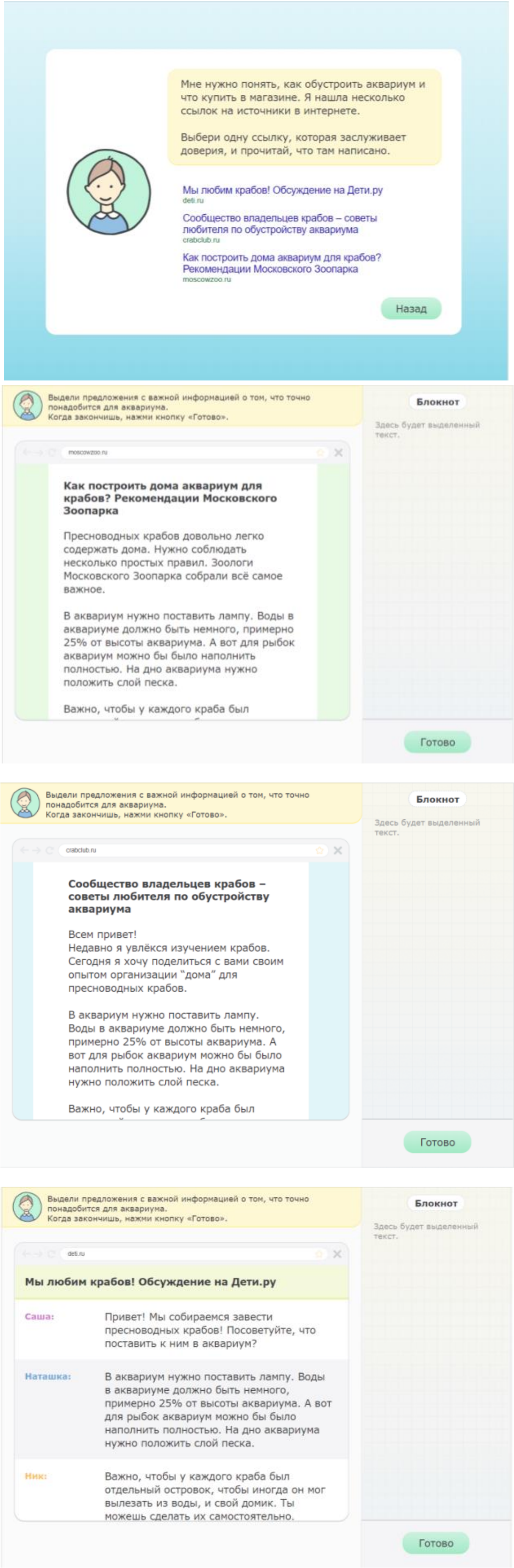
Nowadays performance-based tasks (PBTs) are broadly used in educational assessment (*de Klerk, 2016*). PBTs in computerized form significantly contribute to the assessment of 21st century skills. *Shute et al, 2016* developed competency model for measuring problem-solving skills via video game. Another game, Newton's Playground, was designed to assess Newtonian mechanics as well as creative skills *(Almond et al, 2014)*. Computerized PBTs model wide range of real-life situations, such as conducting a science investigation, searching information on websites or preparing a research paper. Examinees control assessment interface, therefore, different examinees may face different context. However, in order to ensure the fairness of the assessment students' responses must be comparable regardless of context they've chosen.

The **purpose** of the paper is to provide comparability evidence of the scenario-based assessment with respect to different test context.

Following Evidence-Centered Design (*Mislevy, Almond & Lukas, 2003)* we've developed PBT aimed at measuring 21st century skills: creativity and critical thinking. In one task students are asked to find information on how to make an aquarium for crabs. Figure 1 shows the screenshots of links and texts used in the task. Examinees are presented with simulated browser environment with hyperlinks and websites where their critical thinking behavior could be observed. They are asked to choose one link to find information about aquarium: encyclopedia, recommendations, or forum. Figure 1 shows the screenshots of links and texts used in the task. All texts have the same key sentences, but additional (non-key) sentences and the characteristics of the interface vary. For example, forum text is divided into multiple messages, while in the encyclopedia and recommendation texts the whole article is presented. In this work additional sentences and characteristics of the interface are considered as test context.

Each text in the task contains around twenty sentences including nine key sentences aimed to measure critical thinking skill (task-relevant sentences). During the task, test-takers highline sentences within the text they've chosen using computer mouse. For highlighting key sentence, student gets one point. So, all items are dichotomously scored. Overall, maximum 9 points are awarded for the task.

The sample consists of 968 Russian four-grade students (9-11 year-olds).



Мне нужно понять, как обустроить аквариум и что купить в магазине. Я нашла несколько ссылок на источники в интернете.

Выбери одну ссылку, которая заслуживает доверия, и прочитай, что там написано.

Мы любим крабов! Обсуждение на Дети.ру
deti.ru

Сообщество владельцев крабов – советы любителя по обустройству аквариума
crabclub.ru

Как построить дома аквариум для крабов? Рекомендации Московского Зоопарка
moscowzoo.ru

Выдели предложения с важной информацией о том, что точно понадобится для аквариума.
Когда закончишь, нажми кнопку «Готово».

moscowzoo.ru

Как построить дома аквариум для крабов? Рекомендации Московского Зоопарка

Пресноводных крабов довольно легко содержать дома. Нужно соблюдать несколько простых правил. Зоологи Московского Зоопарка собрали всё самое важное.

В аквариум нужно поставить лампу. Воды в аквариуме должно быть немного, примерно 25% от высоты аквариума. А вот для рыбок аквариум можно бы было наполнить полностью. На дно аквариума нужно положить слой песка.

Важно, чтобы у каждого краба был

Выдели предложения с важной информацией о том, что точно понадобится для аквариума.
Когда закончишь, нажми кнопку «Готово».

crabclub.ru

Сообщество владельцев крабов – советы любителя по обустройству аквариума

Всем привет!
Недавно я увлёкся изучением крабов. Сегодня я хочу поделиться с вами своим опытом организации "дома" для пресноводных крабов.

В аквариум нужно поставить лампу. Воды в аквариуме должно быть немного, примерно 25% от высоты аквариума. А вот для рыбок аквариум можно бы было наполнить полностью. На дно аквариума нужно положить слой песка.

Важно, чтобы у каждого краба был

Выдели предложения с важной информацией о том, что точно понадобится для аквариума.
Когда закончишь, нажми кнопку «Готово».

deti.ru

Мы любим крабов! Обсуждение на Дети.ру

Саша: Привет! Мы собираемся завести пресноводных крабов! Посоветуйте, что поставить к ним в аквариум?

Наташка: В аквариум нужно поставить лампу. Воды в аквариуме должно быть немного, примерно 25% от высоты аквариума. А вот для рыбок аквариум можно бы было наполнить полностью. На дно аквариума нужно положить слой песка.

Ник: Важно, чтобы у каждого краба был отдельный островок, чтобы иногда он мог вылезать из воды, и свой домик. Ты можешь сделать их самостоятельно.

Fig. 1. Screenshots of the texts used in the Aquarium PBT.

For the psychometric analysis, we consider three texts as three forms of the same task. Thus, the comparability of these forms could be evaluated with respect to different test context.

Firstly, to prove the comparability of scores across these forms we use classical equating methods.

Linear equating stands that score distributions should be the same besides difference in means and standard deviations. In case of equal standard deviations mean equating is preferred. To determine whether means and standard deviations are the same across test forms two non-parametric criteria were used: Mann-Whitney U-test for equality of means and Kruskal-Wallis test for equality of standard deviations. Moreover, Chi-square test was conducted to test uniformity hypothesis.

The equipercentile procedure for equating is based on the equity of percentile ranks among score distributions of different test forms. Unlike linear equating, equipercentile equating has no assumption about test form distribution *(Kolen, 1988)*. Muraki discourage the use of equipercentile equating as equating method for performance assessment because of limited number of items *(Muraki et al, 2000)*. In the particular case PBTs can be composed of one or two scored responses, so it is almost impossible to equate them. However, in the case of nine scored responses equipercentile equating is appropriate. In our study we can use this classical method of equating without constrains.

To check the comparability percentile-rank score curve is created for all test forms based on percentile ranks determined for each number-correct score. In the paper, for three texts percentile-rank score curves were created and compared to each other.

Secondly, in order to test the hypothesis that the probability of selecting the key sentences depends on the characteristics of context and ability level, we use differential item functioning (DIF) analysis. DIF-analysis is generally known as a key component in the evaluation of the fairness of educational tests *(Zwick et al, 1999)*. DIF occurs if two subgroups have different probabilities of getting a dichotomously scored item correct.

To provide comparability evidence of test forms, it is necessary to make sure that no DIF detected in each item in the test. In our case it means that subgroups of test-takers chosen different links have the same probability of highlighting key sentences in the text regardless test context. Two methods for DIF detection were used: Mantel–Haenszel chi-square test *(Mantel & Haenszel, 1959)* and logistic regression modeling *(DIF-analysis; Zumbo,1999)*. The first was applied to investigate the presence of uniform DIF, the latter – the presence both uniform and non-iniform DIF. Size effect was estimated with the Mantel-Haenszel delta difference statistic *(Zwick, 2012)* and R-square delta difference statistic *(Jodoin, M. G., & Gierl, M. J, 2001).*

All calculations mentioned above were made using MS Excel and SPSS.

Additionally, to assess DIF in test items we use item response theory modeling (IRT). The analysis was conducted using Winsteps software.

Figure 2 represents the distributions of scores for three test forms. Form 1 encyclopedia (N=595), Form 2 recommendations (N=256), Form 3 forum (N=117).

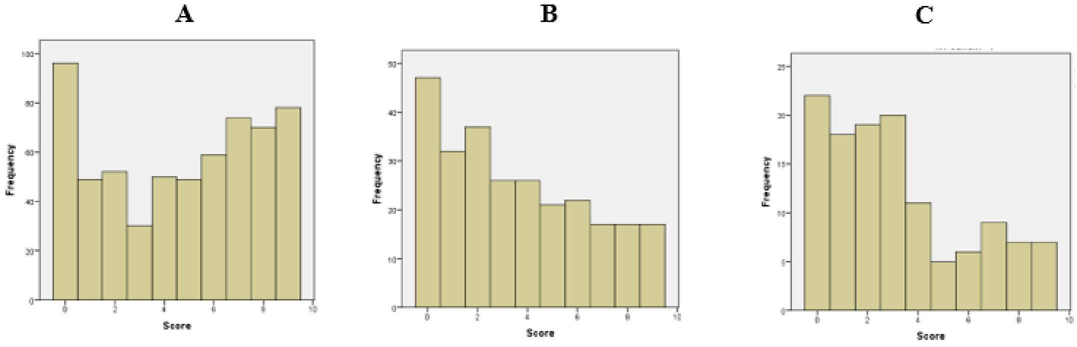


Fig. 2. Score distributions for test Forms. (A) Form 1; (B) Form 2; (C) Form 3.

The data is not normally distributed. Thus, to test hypotheses whether mean and standard deviations are the same across all forms, nonparametric methods are most valid.

Table 1 summarizes values of means and standard deviations for all test forms. As can be seen in the table, Form 1 has greater mean and standard deviation compare to other forms. The values of Form 2 and Form 3 are similar. The results also show significant difference (ρ <0,05) in means between scores from Form 1 and others with the Mann Whitney U test. The mean difference between Form 2 and Form 3 are not significant (p>0,05). Kruskal-Wallis test detect the difference in standard deviations (ρ <0). Moreover, we accept uniformity hypothesis for Form 2 and Form 3 (Chi-square, ρ =0,641).

Table 1

Descriptive statistics of Forms

	Form 1	Form 2	Form 3
Mean	4.65	3.56	3.25
St. Dev	3.87	2.85	2.76

Thus, we can conclude that Form 2 and 3 are very similar in mean and standard deviation values. This is confirmed by nonparametric statistical methods. However, Form 1 is different from other forms. That is why we cannot consider all Forms as comparable.

To collect more evidence for comparability of test scores across Forms 1 to 3, equating was conducted using CCT methods. Figure 3 illustrates the results from mean and linear equating for three Forms in pairs. Red line represents the data, green line and blue line represent linear and mean equating results. From the Figure 3, it is gathered that Form 1 is easier than Form 2 and Form 3. In addition, Form 2 and Form 3 are quite similar in their difficulties.

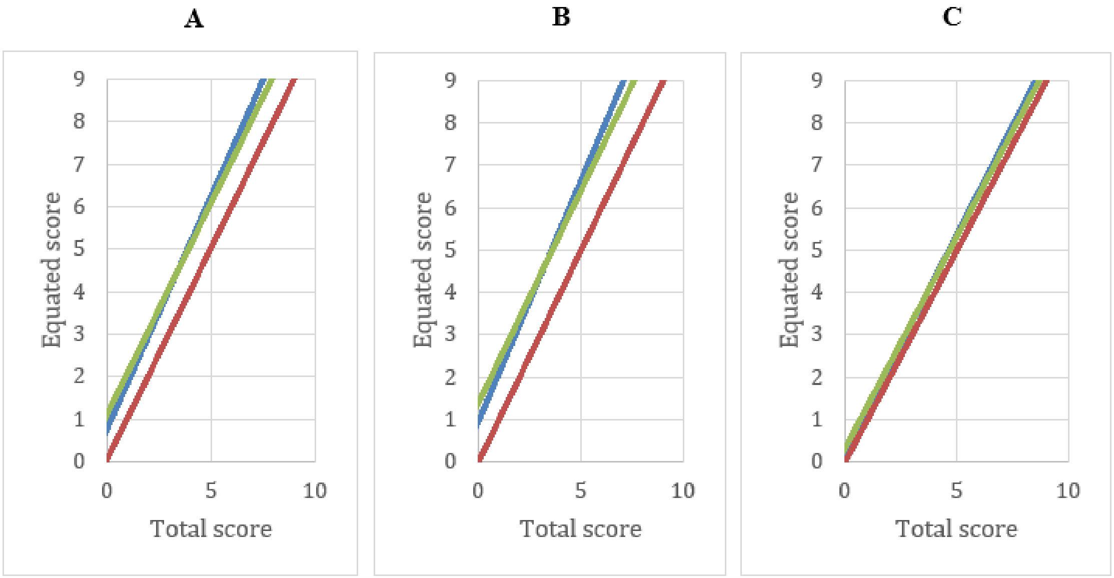


Fig. 3. Linear and mean equating results. (A) Form 1 and Form 2; (B) Form 1 and Form 3; (C) Form 2 and Form 3.

The equipercentile equating results are consistent with the previous findings. The percentile-rank score curves for all Forms are showed in Figure 4. Form 1 is represented by the blue line, Form 2 – by the yellow line and Form 3 – by the red line. To conduct equipercentile equating we should compare scores across forms with a fixed value of percentile-rank score. At 70 percentile-rank score a Form 1 score of 7 corresponds to a Form 3 score of 4 (Figure 4-b) and to a Form 2 score of 5 (Figure 4-a). This is the evidence of Form 1 being easier than Form 2 and

Form 3. In addition, Form 2 scores are very close to Form 3 scores through all sum scores (measure values).

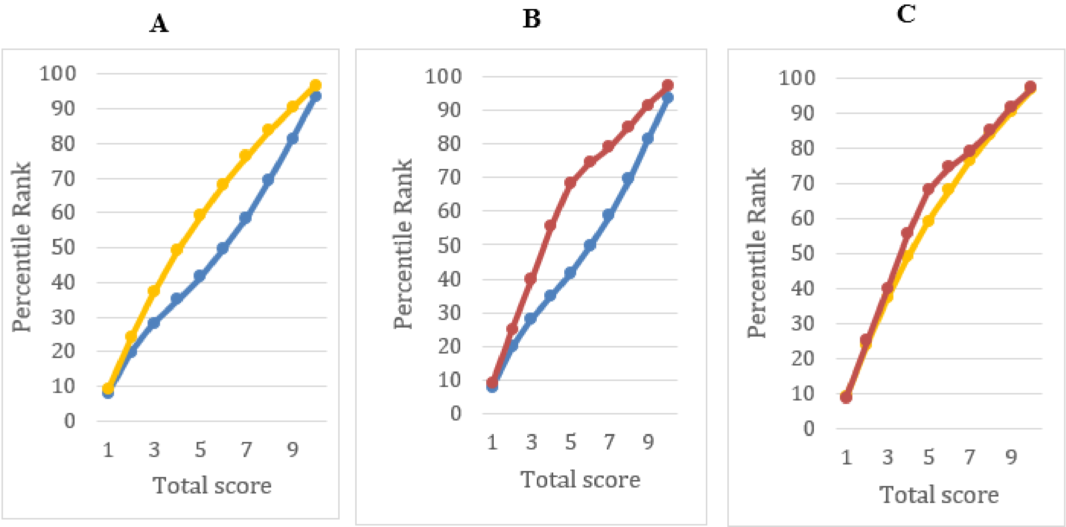


Fig. 4. Equipercentile equating results. (A) Form 1 and Form 2; (B) Form 1 and Form 3; (C) Form 2 and Form 3.

In general, CTT equating methods show that Form 2 and Form 3 are quite comparable to each other. In contrast, Form 1 is easier than other Forms. For the latter analysis we decided to combine Form 2 and Form 3 data together to one Form. Thus, DIF-analysis were conducted using two Forms (Form 1 and united Form 2 and 3 as new Form 2).

Table 2 presents the results of the Mantel–Haenszel chi-square test and its size effect. Here test score for 9 items is used as the proxy for the latent trait being measured. Furthermore, we add three more items aimed at measuring critical thinking to expand test score (11 items).

Table 2

Mantel–Haenszel chi-square test results

	9 items		11 items	
Sentence	ρ (M-H)	Effect size –2.35lnα	ρ (M-H)	Effect size –2.35lnα
1	0,285	-0.731	0.136	0.771
2	0,236	0.545	***0.011***	***1.06***
3	0,285	0.674	***0.001***	***1.802***
4	0,188	-0.75	0.806	-0.179
5	***0,042***	-1.147	0.282	-0.613
6	0,372	0.551	0.163	0.736
7	0,127	0.879	***0.028***	***1.201***
8	0,831	0.165	0.385	0.503
9	0,321	-0.578	0.981	-0.068

As shown in Table 2, there is no relationship between Form and item response, after matching on the total 9-item test score. Sentence 5 is the only exception, but based on the ETS recommendations *(Zwick, 2012)*, effect size less than 1 is not a problem. However, results after matching the total 11-item score differ, The Mantel–Haenszel test p-value in Sentence 2, 3 and 7 resulting in the decision to reject the null hypothesis. There is an evidence that the probability of selecting the key sentences depends on the Form being presented and, therefore, on the context.

In order to check the results, move on to the next detecting DIF method – logistic regression modeling. The logistic regression procedure includes item response (1 – highline key sentence, 0 – no) as the depended variable, grouping variable (1 – Form 1, 0 – Form 2) and total scale score (test score as a measure of critical thinking ability) as independent variable. Moreover, a Form by total score interaction is used as an independent variable. Overall, three logistic regressions were built by adding independent variables step by step for 9-item test and another three for 11-item test including additional items measuring critical thinking in total score. The results are presented in Table 3 and Table 4.

Table 3

Uniform and nonuniform DIF for 9-item test

	Uniform DIF		Nonuniform DIF	
Sentence	ρ (Chi-square)	ΔR^2	ρ (Chi-square)	ΔR^2
1	0,551	0.001	0,421	0
2	***0,013***	0.005	0,255	0.002
3	0,058	0.002	0,291	0.001
4	0,189	0.001	0,586	0
5	***0,033***	0.003	0,282	0.001
6	0,259	0.001	0,611	0
7	0,062	0.002	0,194	0.001
8	0,485	0	0,496	0
9	0,353	0.001	0,700	0

First, we test the models for uniform and nonuniform DIF separately using Chi-square statistic. Based on the Table 3 we can conclude that there is a statistically significant test for uniform DIF for Sentence 2 and 5. However, these sentences demonstrate small size effect and, thus, negligible uniform DIF (based on *Zumbo, 1999* guidelines for interpreting R-square delta difference statistic for detecting DIF). There is no nonuniform DIF detected in each sentence.

Second, we repeat these steps for logistic models with expand total score. The results are presented in Table 4.

Table 4

Uniform and nonuniform DIF for 11-item test

	Uniform DIF		Nonuniform DIF	
Sentence	ρ (Chi-square)	ΔR^2	ρ (Chi-square)	ΔR^2
1	***0.045***	0.004	0.888	0
2	***0.001***	0.01	0.096	0.003
3	***0.002***	0.007	0.874	0
4	0.565	0	0.726	0.001
5	0.164	0.002	0.288	0
6	0.095	0.002	0.295	0
7	0.016	0.004	0.101	0.002
8	0.152	0.001	0.136	0.002
9	0.883	0	0.997	0

As it shown in the Table, there is a statistically significant but negligible uniform DIF in Sentences 1 to 3. Delta R-square values are too small to detect moderate DIF. We can conclude that neither uniform nor nonuniform DIF detected in the test with expand total score.

Overall, each key sentence demonstrates no DIF in case of original and expand test score. The probability of selecting the key sentences does not depends on ability level and the context of the Forms. This is the evidence for their comparability.

Additionally, we cross-validate the results with item response theory modeling (IRT) using Winsteps software. The analysis indicates unexpected response pattern in our data. 46 students with largest outfit values were deleted. Moreover, in order to increase the fit, item 10 (additional item) was also deleted. We've reanalyzed the data after deleting persons and item. The results show that all items share the same dimension, therefore the test might be considered as unidimensional. The first contrast has eigenvalue of 1.9 with its threshold value equal 2.0. The test demonstrates moderate fit with reliability value of 0.75. Cronbach alpha reliability value is 0.84.

The data with 9 items were analyzed separately. The original test also might be considered as unidimensional based on eigenvalue less than 2.0. The reliability value for the original test is slightly smaller than for test with extra items, but still moderate (0.65). The average INFIT statistics for all sentences around 1.0 indicate a little distortion of the measurement.

The presence of DIF in each item could be indicated by DIF Contrast (effect size) and DIF Statistical Significance. Figure 5 displays the results. Form 1 is represented by 1, Form 2 by 0.

The value of DIF Contrast more than 0,7 logits demonstrates moderate to large DIF *(Zwick et al, 1999*). In our case DIF contrast of the item 2 is around 0,7. However, we conclude that there is not a problem. This is the final evidence of comparability of test Forms.

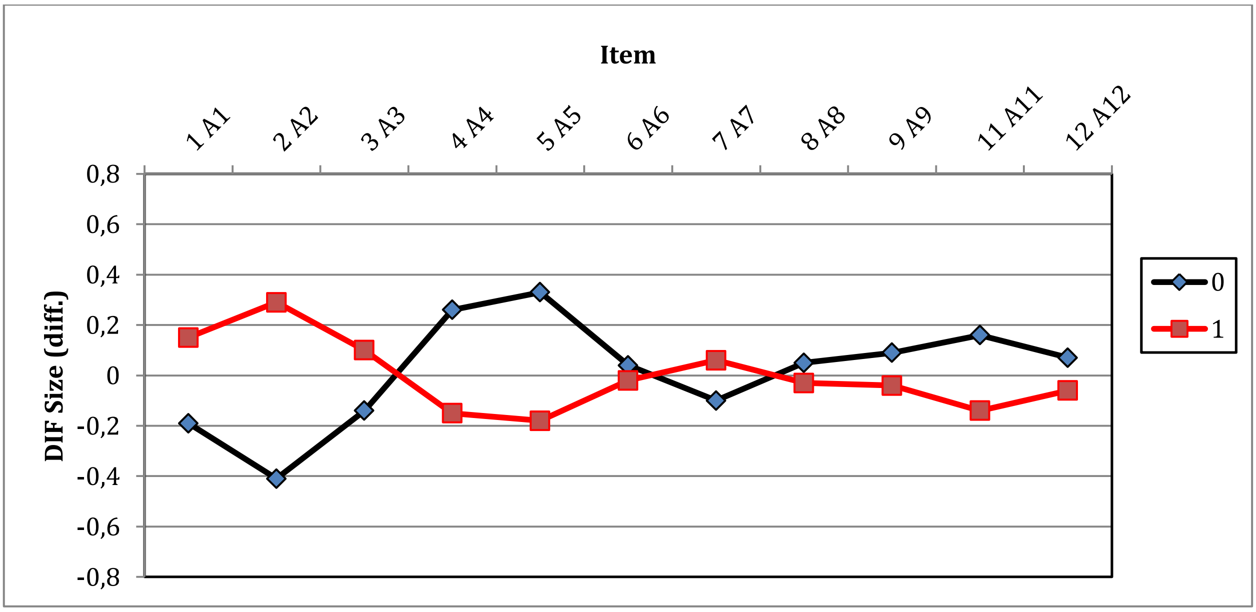


Fig. 4. DIF-Size for test items

In this paper we provide comparability evidence of the Forms from computerized scenario-based task. Each Form represents the text with identical key sentences, but different test context. The CCT equating methods were conducted first. The results show that two out of three Forms are quite comparable. However, Form 1 is easier than others. On the next step we use DIF-analysis to test the relationship between Form and the probability of item response. During the analysis we add three more items aimed at measuring critical thinking skill to expand total score which represents ability level. Despite some items show statistically significant presence of DIF, size effect is negligible. Thus, we made a conclusion that the probability of selecting key-sentence does not depend on non-key sentences or characteristics of the interface provided by different Forms. In order to confirm our conclusions, we made additional analysis with item response theory modeling (IRT).

The current study does not consider the issue of local item dependence, one of the most important cause for concern in performance-based tasks. We realize that test context could be a source of local dependence between the sentences in the texts and therefore become a second dimension. To test this hypothesis multidimensional methods should be applied to the data. This is a basis for further research.

References:

1. Almond, R. G., Kim, Y. J., Velasquez, G., & Shute, V. J. (2014). How task features impact evidence from assessments embedded in simulations and games. Measurement: Interdisciplinary Research & Perspectives, 12(1-2), 1-33.
2. de Klerk, S., Eggen, T. J., & Veldkamp, B. P. (2016). A methodology for applying students' interactive task performance scores from a multimedia-based performance assessment in a Bayesian Network. Computers in human behavior, 60, 264-279.
3. Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. Applied measurement in education, 14(4), 329-349.
4. Kolen, M. J. (1988), Traditional Equating Methodology. Educational Measurement: Issues and Practice, 7: 29-37.
5. Mantel, N., & Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.
6. Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. ETS Research Report Series, 2003(1), 1-29.
7. Muraki E., Hombo C.M., Lee Y-W. (2000). Equating and Linking of Performance Assessments // Applied Psychological Measurement. 2000. №24, pp. 325-227.
8. Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. Computers in Human Behavior, 63, 106-117.
9. Zumbo B.D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores / Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense; 1999. 57 p.
10. Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. ETS Research Report Series, 2012(1), i-30.
11. Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. Journal of Educational Measurement, 36(1), 1-28.