Психологическая наука и образование
2022. Т. 27. № 6. С. 57—67
DOI: https://doi.org/10.17759/pse.2022270605
ISSN: 1814-2052
ISSN: 2311-7273 (online)

Psychological Science and Education
2022. Vol. 27, no. 6, pp. 57—67
DOI: https://doi.org/10.17759/pse.2022270605
ISSN: 1814-2052
ISSN: 2311-7273 (online)

# Analysis of Task Comparability in Digital Environment by the Case of Metacognitive Skills

***Daria A. Gracheva***
National Research University Higher School of Economics
ORCID: https://orcid.org/0000-0002-4646-7349, e-mail: dgracheva@hse.ru

This article discusses the problem of task comparability with the help of scenario-based tasks for metacognitive skills. Using the data of "4C" tool for measuring critical thinking (N=500), the comparability of two scenarios within an identical digital environment with one set of indicators was investigated. The main difference in the scenarios lies in the contextual characteristics. The measurement invariance analysis of the instrument using confirmatory factor analysis was conducted. The results show that even with the equivalent construct structure and tasks' characteristics, the context of the scenario has an effect on the student`'s performance. The main differences in results were recorded for tasks involving interaction with the environment, where the test-taker created an object with elements. Tasks involving working with text in a digital environment can be considered comparable in case of elements content change. The possible reasons behind the observed differences in scenarios are discussed.

***Keywords:*** critical thinking, test comparability, scenario-based tasks, contextualized items, confirmatory factor analysis, measurement invariance.

**Funding.** The reported study was funded by The Ministry of Education and Science of the Russian Federation, project number 075-15-2022-325 from 25.04.2022.

**Acknowledgements.** The author is grateful to Uglanova I.L. for her help and comments on this article.

**For citation:** Gracheva D.A. Analysis of Task Comparability in Digital Environment by the Case of Metacognitive Skills. *Psikhologicheskaya nauka i obrazovanie = Psychological Science and Education*, 2022. Vol. 27, no. 6, pp. 57—67. DOI: https://doi.org/10.17759/pse.2022270605 (In Russ.).

# Анализ сопоставимости измерения метапредметных навыков в цифровой среде

***Грачева Д.А.***
ФГАОУ ВО «Национальный исследовательский университет «Высшая школа экономики» (ФГАОУ ВО НИУ ВШЭ), г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-4646-7349, e-mail: dgracheva@hse.ru

Представлены данные исследования сопоставимости измерения метапредметных навыков с помощью сценарных заданий. На данных инструмента «4К» для измерения критического мышления (N=500) исследована сопоставимость двух вариантов сценариев внутри идентичной цифровой среды, с одним набором индикаторов. Отмечается, что основное различие в сценариях заложено в контекстных элементах. Проведен анализ инвариантности инструмента по вариантам с использованием метода конфирматорного факторного анализа. Установлено, что при эквивалентных характеристиках заданий контекст сценария оказывает эффект на результаты. Различия в оценках зафиксированы для задач, предполагающих более свободное взаимодействие со средой, где тестируемый самостоятельно собирает объект из предложенных элементов. Задания, включающие работу с текстом в цифровой среде, могут считаться сопоставимыми при изменении элементов контекста. Обсуждаются возможные причины, стоящие за различием в оценках по вариантам сценариев.

***Ключевые слова:*** критическое мышление, сопоставимость тестов, сценарные задания, контекст заданий, конфирматорный факторный анализ, измерительная инвариантность.

**Финансирование.** Статья подготовлена в рамках гранта, предоставленного Министерством науки и высшего образования Российской Федерации (соглашение от 25.04.2022 № 075-15-2022-325).

**Благодарности.** Автор благодарит И.Л. Угланову за советы и помощь в подготовке статьи.

**Для цитаты:** *Грачева Д.А.* Анализ сопоставимости измерения метапредметных навыков в цифровой среде // Психологическая наука и образование. 2022. Том 27. № 6. С. 57—67. DOI: https://doi.org/10.17759/pse.2022270605

## Introduction

Assessment of complex constructs is a new trend in educational testing. An example of such a construct is critical thinking, which is referred to as meta-subject skill. However, it is difficult to measure meta-subject skills with traditional item types, such as multiple-choice items. Scenario-based tasks in the digital environment have great potential to solve this problem.

Scenario-based tasks resemble a computer game in which a student is faced with a situation where he needs to solve a number of problems. The student's actions during the test are considered observable evidence of the measured skill — indicators. Scenario-based tasks demonstrated students' behavior that they are likely to perform in similar situations in real life, which is especially

important in the assessment of meta-subject skills [7].

In practice, the use of scenario-based tasks faces many challenges. Among them are low reliability, a small number of tasks, and weak correlation with alternative measurements. In general, the problem with a comparability of measurements is typical for tasks with a focus on the process and product (performance-based tasks): scenario-based tasks, essays, experiments, etc. [6]. Previous attempts to create comparable experiments were unsuccessful, despite the fact that researchers use the same design principles [17].

The first step in the development of a new scenario is the selection of a suitable context. The context is a set of task characteristics that defines the situation where the test-taker will be able to demonstrate the desired skills. The degree of correspondence between the context of scenario-based tasks to each other is directly related to the degree of their comparability. However, the comparability of tasks with context is an underdeveloped area of research [6].

The purpose of this article is to analyze the comparability of scenario-based task forms aimed at measuring critical thinking. Scenario forms contain the same number of indicators and are implemented in an identical digital environment, but differ in contextual elements.

The article is structured as follows: in the first part, previous studies of tasks with context are considered, as well as the methods that are used to analyze the comparability of test forms; the second part presents the results of the analysis of the comparability of scenario-based task forms. The article ends with a discussion of the results, limitations, and further directions of the study.

## Literature Review of Contextualized Tasks

The concept of context and its relationship with the psychometric characteristics of tasks and test results is studied on the example of questionnaires, essays, as well as game-based, and scenario-based tasks.

In a study of personality questionnaires, it was shown that clarification of the context leads to an improvement in psychometric characteristics by reducing the number of interpretations of statements [14].

For essays, the comparability of the tasks with various topics and stimulus materials in the format of pictures was analyzed [9].

In the field of computer games, research on the role of the interface on test results was carried out. For example, in [15] it was found that the choice of a character was associated with the behavior of the test-taker within the game environment.

The idea of the context of the virtual world as a stimulus for creative solutions was studied in [10]. In the study, test-takers "immersed themselves" in different virtual worlds using virtual reality helmets, and then drew a non-existent animal. The ideas of these drawings differed significantly depending on the context presented.

On the example of PISA tasks in science, the characteristics of the context (the degree of abstractness, the purpose of the context, etc.) and their relationship with students' achievements were studied [13].

The use of tasks with context is a promising approach for measuring complex skills. At the same time, the context can be considered as a factor that affects the characteristics of tasks and test results. A range of methods used for comparability analysis will be discussed in the next section.

## Overview of Methods for Comparability Analysis

Comparability of test forms is carried out by qualitative and quantitative methods that can complement each other.

Qualitative methods include the use of test design principles and the involvement of experts to assess the comparability of items.

Test design principles include the use of a test specification to create test forms. However, it has been found that open-ended items created according to the same specification are not always comparable [8].

The opinion of experts is used to assess at what extent the topic of the task covers a general or highly specialized issue [11].

Quantitative methods include the use of statistical methods for comparability analysis. The choice of statistical method depends on the purpose of the study. If the purpose of the study is to evaluate differences between groups, then t-test or ANOVA can be used. For the purpose of predicting the results of future tests, regression analysis is more suitable, and correlation analysis can be considered as a measure of the similarity of results across test forms.

However, the process of analyzing the comparability of test forms goes beyond working with raw test results. To consider test forms comparable, it is necessary to make sure that they measure the same construct, tasks have similar psychometric characteristics [3].

Testing of these assumptions is possible within the methodology of confirmatory factor analysis (CFA) or Item Response Theory (IRT). For example, CFA was used to test the functioning of the tool in different modes [16].

In this article, we focus on the application of CFA to the analysis of test forms comparability. Since data in education is often categorical, the case of CFA for ordinal variables is considered. To analyze the comparability within the framework of CFA the analysis of measurement invariance of the instrument is conducted. Comparability studies usually consider three levels of invariance: configural, metric, and scalar.

At the configural level, the comparability of the construct structure in all groups is checked [12]. At the metric level, the values of factor loadings are assumed to be equal in all groups. At the scalar level, the equality of threshold values is tested (in the case of a categorical CFA). When the level of scalar invariance is reached, it is possible to compare the mean values of latent factors between groups.

Thus, the measurement of complex skills requires the use of statistical methods aimed at studying the structure of the test. For example, CFA is such a method. Further, this method will be used to analyze the comparability of scenario-based task forms.

## Characteristics of the Sample, Methods, Data Collection Procedures and Strategy of Analysis

### *Sample*

The article uses data from 500 fourth grade students who participated in the assessment of 21st century skills in Fall 2020 as part of the project “4K of the modern world. Formation of competencies in the 21st century and assessment of individual progress in their development” with the support of the “Investment to the future” Charitable Foundation.

### *Instrument*

Critical thinking is assessed using computerized scenario-based tasks from the “4C” instrument developed by the staff of the Center for Psychometrics and Measurements in Education (HSE University). The validity of the tool has been proved in multiple test trials [2].

In this work, the comparability of a pair of scenarios for measuring critical thinking, “Aquarium” and “Terrarium”, is analyzed. According to the conceptual framework of the instrument, critical thinking skill includes two components: 1) “Analysis of information” — the skill of working with information in accordance with the goals and conditions of the task; 2) “Making inferences” — the skill of formulating one’s own inference using the results obtained at the stage of working with information [2].

The “Aquarium” task invites test-takers to set up an aquarium for crabs. For the assessment of ability to work with information, a simulation of an Internet browser is used in the task, where the text of the article is presented (Fig. 1). The text of the article includes both relevant and irrelevant sentences. Relevant sentences contain information that will be needed to equip an aquarium for crabs (for example, “Crabs need flagstones to get out of water”). Irrelevant sentences contain information that is not relevant to the task. For each highlighted relevant sentence, 1 point is awarded.

Indicators of the ability to make inferences are evaluated in an interactive environment (constructor), where the test-taker builds an aquarium for the crab from elements based on information from the text (Fig. 2). For each correctly added element, 1 point is awarded.

In the “Terrarium” scenario, test-takers face the same tasks with different content, where the main goal is to build a terrarium for geckos.


*Fig. 1*. Stimulus material (text): a — “Aquarium”, b — “Terrarium”
(in Russian, translation is provided on the example of “Aquarium”)

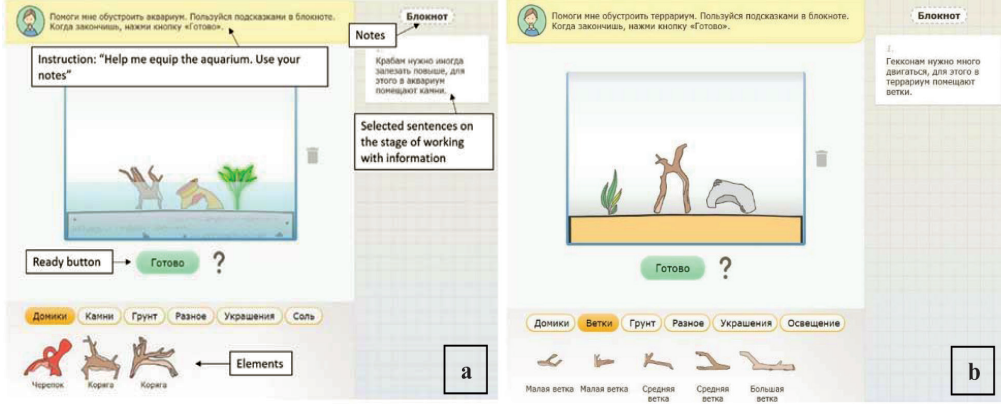


*Fig. 2*. Stimulus material (constructor): a — “Aquarium”, b — “Terrarium”
(in Russian, translation is provided on the example of “Aquarium”)

The skill “Analysis of information” is measured with 14 dichotomous indicators, the skill “Making inferences” with 10 indicators (8 dichotomous and 2 polytomous from 0 to 2 points).

***Procedure***

Testing sessions took place in schools in the presence of a testing administrator. Each participant was provided with a computer with Internet access. At the start of a test session, administrators opened the test website on computers and give individual logins to students to log into the system. All instructions and tasks were presented in computer format.

In the research we used a balanced design, in which both scenarios were performed by the same test takers. The sample was randomly divided into two groups. The first group took the “Aquarium” task first, and then the “Terrarium” task, the second group completed the tasks in the reverse order. This design made it possible to control the effect of the order on the results of the comparability analysis. The break between testing sessions ranged from one day to a week.

***Strategy of analysis***

The study of the comparability of scenario-based tasks forms was carried out using CFA. The analysis included two stages. At the first stage, the structural model of critical thinking was proposed, which was separately tested for scenario forms. At the second stage, the measurement invariance of the general model was tested for two scenarios.

The weighted least squares method (WLSMV) was used as a parameter estimation method, which is most suitable for ordinal and binary data. The quality of the models was assessed by the following indices: CFI>0.90; TLI>0.90; RMSEA<0.05 [12].

The invariance was tested by sequential comparison of three models (configural, metric, scalar). The difference between the fit statistics ($\Delta$CFI within 0.01, $\Delta$RMSEA within 0.015 to confirm invariance) was taken as a comparison criterion [4]. When scalar invariance is achieved, it is possible to compare the mean values of the latent factors of different groups, where the mean values of the factors for one group are equal to zero, and for the other group are freely estimated.

The critical thinking model contains two main related factors — “Analysis” and “Inference”. The model also includes additional orthogonal factors of the stimulus material, which take into account the common source of variance between groups of indicators related to working with text or constructor.

The analysis was carried out in the Mplus program, version 8.3.

## Results

The average score for the ability to analyze information is 5.56 points (sd 3.83) for the “Aquarium” scenario and 5.29 points (sd 3.85) for the “Terrarium” scenario. The average score for the ability to make inferences for the “Aquarium” scenario is 8.2 points (sd 2.72), for the “Terrarium” scenario — 8.25 points (sd 2.67). There were no statistically significant differences between the mean values for both the ability to analyze information ($t(998)=1.11$, $p>0.05$) and the ability to make inferences ($t(998)=-0.29$, $p>0.05$).

Separate models for “Aquarium” ($\chi2(240)=387.691^*$, $p<0.000$; CFI=0.979; TLI=0.976; RMSEA=0.035. 90% CI (0.029;0.041)) and “Terrarium” scenarios ($\chi2(240)=398.031^*$, $p<0.000$; CFI=0.980; TLI=0.977; RMSEA=0.036, 90% CI (0.030; 0.043)) showed good fit with the data. On Fig. 3—4 CFA model and standardized factor loadings for the “Aquarium” and “Terrarium” scenarios are shown.

The results of measurement invariance testing are presented in Table 1. The val-

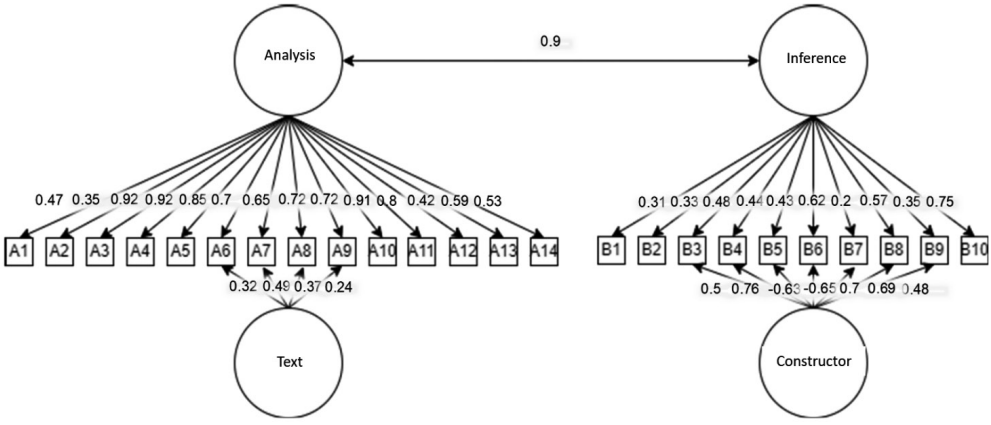


*Fig. 3.* CFA model ("Aquarium"): all parameters of the model are significant p<0.05

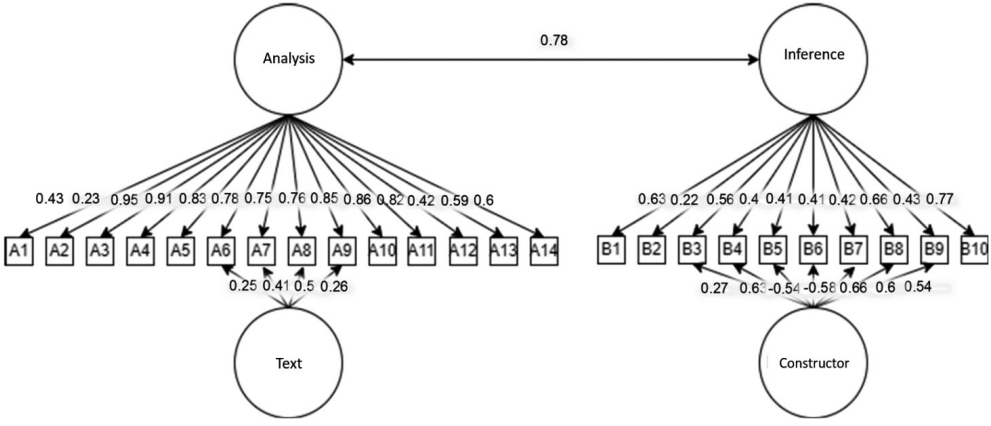


*Fig. 4.* CFA model ("Terrarium"): all parameters of the model are significant p<0.05

ues of fit statistics for the three models are similar, which makes it possible to assume that the scalar invariance is proved. The structure of critical thinking is reproduced in different scenarios, the psychometric characteristics of the indicators do not differ.

Table 1

**Results of measurement invariance testing**

| Model | $\chi^2$ (df) | RMSEA | CFI | TLI |
|---|---|---|---|---|
| Configural | 785.743* (480) | 0.036 (90% CI 0.031; 0.040) | 0.979 | 0.976 |
| Metric | 835.083* (511) | 0.036 (90% CI 0.031; 0.040) | 0.978 | 0.976 |
| Scalar | 915.226* (532) | 0.038 (90% CI 0.034; 0.042) | 0.974 | 0.973 |

*Note:* * p<0.05.

After checking the levels of invariance and achieving scalar invariance, it is possible to compare the mean values of latent factors for tasks "Aquarium" and "Terrarium" (Table 2).

The mean values for the "Analysis" factor did not differ significantly by task forms. That is, on average, the score for the ability to analyze information can be considered interchangeable in two scenarios when the characteristics of the scenario context change. There were also no significant differences in the mean values of the "Text" factor.

Nevertheless, a significant difference in the mean values for the "Inference" factor is evidence that indicators related to the ability to make inferences are easier in "Terrarium" than in "Aquarium" scenario. The differences are preserved in the constructor factor.

A meaningful interpretation of the factors of the stimulus material is often difficult. However, the results obtained allow us to say that the results of students differ significantly in the part of the scenario where they need to demonstrate the ability to make inferences through working with elements in the constructor.

## Discussion

Complex constructs require new measurement approaches. One of the approaches is the use of scenario-based tasks in the digital environment. At the same time, for scenario tasks, the risk of obtaining incomparable results is more pronounced [6].

One threat to comparability is scenario context. In this article, we used the "Aquarium" and "Terrarium" scenarios for measuring critical thinking, which contained the same set of indicators, but differed in contextual characteristics. The analysis of measurement invariance showed that changing the context does not change the theoretical structure of the instrument, and the psychometric characteristics of the indicators did not differ significantly by task forms.

The results of comparing the mean latent factors showed that the test-takers receive lower scores for the ability to make inferences in the "Aquarium" scenario than in the "Terrarium", while scores for the ability to analyze information do not differ by forms.

Due to the data collection design, which respected the random order of forms, we can assume that the differences in the results are not due to the effect of learning in solving similar problems, but due to differences in contextual elements.

Previous research has shown that task context can have an effect on test results. For example, a familiar context can give an advantage in solving problems [5]. In the study of creativity, the context of the "virtual world" was manifested in the drawings of non-existent animals [10].

Another reason for the difference in results could be the type of the tasks within

Table 2

**Mean values of latent factors**

| Factor | Mean values of latent factors for "Terrarium" task | Z-statistic |
|---|---|---|
| "Analysis" | -0.089 (0.066) | -1.353 |
| "Inference" | 0.211 (0.071) | 2.965* |
| "Text" | -0.003 (0.129) | -0.026 |
| "Constructor" | -0.272 (0.079) | -3.433* |

*Note:* The standard errors of measurement are given in parentheses. The mean values of the factors for the "Aquarium" scenario are equated to zero. *p<0.05.

the scenario. It has previously been shown that the multiple-choice item type is less susceptible to fluctuations in difficulty. Larger problems are typical for tasks with an open-ended questions or tasks with a common stimulus material, such as text [3].

However, our results indicate that tasks which include texts as stimulus material can be comparable. In part, this can be explained by the use of the "cloning" approach for test development, which allows us to create the most similar texts in different contexts [1]. Items containing interactive elements are more at risk of incomparability, which could be the reason for the difference in scores by form for the ability to make inferences.

The present research has some limitations. The analysis was conducted on one pair of scenarios to measure one skill, so the results need to be revalidated on other scenarios and skills. In addition, in this work, we analyzed the comparability of forms, based only on the analysis of the data structure and the functioning of indicators.

Further directions for research devoted to the comparability of tasks with context include the use of both quantitative and qualitative methods. Linguistic analysis of task texts and the involvement of experts will allow to gain a deeper understanding of the differences between the scenarios. Another promising direction for future research is to conduct cognitive laboratories and interviews with students to understand the contribution of context to test results. Further application of quantitative methods may be to assess the effect of the interaction of the context of the scenario with other characteristics of tasks.

## Conclusion

Tasks in the digital environment containing interactive elements are a trend in the field of measurements in education. However, it is almost impossible to create comparable tasks "by eye". The variety of situations and greater freedom of action of the test-taker within the digital environment can reduce the comparability of measurements. This is especially important when tasks are used as interchangeable forms, for example, for monitoring studies. The lack of widespread practice of the analysis of forms comparability may create unequal opportunities for test-takers to demonstrate their abilities, and decisions that will be made based on the test results will be invalid.

Our analysis determined that tasks where the test-taker create an object from elements are at a greater risk of incompatibility. Differences in results can be explained by the context of the tasks or the specificity of the task type. The study of the reasons for the results obtained, as well as the revalidation of the conclusions formulated here, can be carried out separately to improve the quality of innovative tasks and explore the possibility of their use for both large-scale and local testing.

***References***

1. Gracheva D.A., Tarasova K.V. Podhody k razrabotke variantov zadanij scenarnogo tipa v ramkah metoda dokazatel'noj argumentacii [Approaches to the development of scenario-based task forms within the framework of evidence-centered design]. *Otechestvennaja i zarubezhnaja pedagogika [Domestic and foreign pedagogy]*, 2022, no. 3(1), pp. 83—97. (In Russ.).

2. Uglanova I.L., Orel E.A., Brun I.V. Izmerenie kreativnosti i kriticheskogo myshlenija v nachal'noj shkole [Measuring creativity and critical thinking in primary school]. *Psihologicheskij Zhurnal [Psychological Journal]*, 2020, no. 6(41), pp. 96—107. (In Russ.).

3. Buerger S. [et al.]. What makes the difference? The impact of item properties on mode effects in reading assessments. *Studies in Educational Evaluation*, 2019. Vol. 62, pp. 1—9. DOI:10.1016/j.stueduc.2019.04.005

4. Chen F.F. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural equation modeling: a multidisciplinary journal*, 2007. Vol. 14, no. 3, pp. 464—504. DOI:10.1080/10705510701301834

5. Crisp V. Exploring features that affect the difficulty and functioning of science exam questions for those

with reading difficulties. *Irish Educational Studies*, 2011. Vol. 30, no. 3, pp. 323—343.
6. Davey T. [et al.]. Psychometric considerations for the next generation of performance assessment. Washington, DC: Center for K-12 Assessment & Performance Management, Educational Testing Service, 2015, pp. 1—100.
7. Kuhn D. A Role for Reasoning in a Dialogic Approach to Critical Thinking. *Topoi*, 2018. Vol. 37, no. 1, pp. 121—128. DOI:10.1007/s11245-016-9373-4
8. Lee H.-K., Anderson C. Validity and topic generality of a writing performance test. *Language testing*, 2007. Vol. 24, no. 3, pp. 307—330. DOI:10.1177/0265532207077200
9. Li J. Establishing Comparability Across Writing Tasks With Picture Prompts of Three Alternate Tests. *Language Assessment Quarterly*, 2018. Vol. 15, no. 4, pp. 368—386. DOI:10.1080/15434303.2017.1405422
10. Nelson J., Guegan J. "I'd like to be under the sea": Contextual cues in virtual environments influence the orientation of idea generation. *Computers in Human Behavior*, 2019. Vol. 90. pp. 93—102.
11. Oliveri M.E. Considerations for Designing Accessible Educational Scenario-Based Assessments for Multiple Populations: A Focus on Linguistic Complexity [Elektronnyi resurs]. *Frontiers in Education*, 2019. Vol. 4. DOI:10.3389/feduc.2019.00088
12. Roos J.M., Bauldry S. Confirmatory factor analysis. SAGE Publications, 2021. 144 p.
13. Ruiz-Primo M.A., Li M. The Relationship between Item Context Characteristics and Student Performance: The Case of the 2006 and 2009 PISA Science Items. *Teachers College Record*, 2015. Vol. 117, no. 1, pp. 1—36.
14. Schmit M.J. [et al.]. Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology*, 1995. Vol. 80, no. 5, pp. 607—620. DOI:10.1037/0021-9010.80.5.607
15. Şengün S. [et al.]. Do players communicate differently depending on the champion played? Exploring the Proteus effect in League of Legends [Elektronnyi resurs]. *Technological Forecasting and Social Change*, 2022. Vol. 177. DOI:10.1016/j.techfore.2022.121556
16. Wang Y., Lu H. Validating items of different modalities to assess the educational technology competency of pre-service teachers [Elektronnyi resurs]. *Computers & Education*, 2021. Vol. 162. DOI:10.1016/j.compedu.2020.104081
17. Wested G.S.-F., Shavelson R.J. Development of performance assessments in science: Conceptual, practical, and logistical issues. *Educational Measurement: issues and practice*, 1997. Vol. 3, no. 16, pp. 16—24.

***Литература***

1. *Грачева Д.А., Тарасова К.В.* Подходы к разработке вариантов заданий сценарного типа в рамках метода доказательной аргументации // Отечественная и зарубежная педагогика. 2022. № 3(1). С. 83—97.
2. *Угланова И.Л., Орел Е.А., Брун И.В.* Измерение креативности и критического мышления в начальной школе // Психологический Журнал. 2020. № 6(41). С. 96—107.
3. *Buerger S.* [*et al.*]. What makes the difference? The impact of item properties on mode effects in reading assessments // Studies in Educational Evaluation. 2019. Vol. 62. P. 1—9. DOI:10.1016/j.stueduc.2019.04.005
4. *Chen F.F.* Sensitivity of goodness of fit indexes to lack of measurement invariance // Structural equation modeling: a multidisciplinary journal. 2007. Vol. 14. No. 3. P. 464—504. DOI:10.1080/10705510701301834
5. *Crisp V.* Exploring features that affect the difficulty and functioning of science exam questions for those with reading difficulties // Irish Educational Studies. 2011. Vol. 30. No. 3. P. 323—343.
6. *Davey T.* [*et al.*]. Psychometric considerations for the next generation of performance assessment. // Washington, DC: Center for K-12 Assessment & Performance Management, Educational Testing Service. 2015. P. 1—100.
7. *Kuhn D.* A Role for Reasoning in a Dialogic Approach to Critical Thinking // Topoi. 2018. Vol. 37. No. 1. P. 121—128. DOI:10.1007/s11245-016-9373-4
8. *Lee H.-K., Anderson C.* Validity and topic generality of a writing performance test // Language testing. 2007. Vol. 24. No. 3. P. 307—330. DOI:10.1177/0265532207077200
9. *Li J.* Establishing Comparability Across Writing Tasks With Picture Prompts of Three Alternate Tests // Language Assessment Quarterly. 2018. Vol. 15. No. 4. P. 368—386. DOI:10.1080/15434303.2017.1405422
10. *Nelson J., Guegan J.* "I'd like to be under the sea": Contextual cues in virtual environments influence the orientation of idea generation // Computers in Human Behavior. 2019. Vol. 90. P. 93—102.
11. *Oliveri M.E.* Considerations for Designing Accessible Educational Scenario-Based Assessments for Multiple Populations: A Focus on Linguistic Complexity [Elektronnyi resurs] // Frontiers in Education. 2019. Vol. 4. DOI:10.3389/feduc.2019.00088
12. *Roos J.M., Bauldry S.* Confirmatory factor analysis. SAGE Publications, 2021. 144 p.
13. *Ruiz-Primo M.A., Li M.* The Relationship between Item Context Characteristics and Student Performance: The Case of the 2006 and 2009 PISA Science Items // Teachers College Record. 2015. Vol. 117. No. 1. P. 1—36.

14. *Schmit M.J.* [*et al.*]. Frame-of-reference effects on personality scale scores and criterion-related validity. // Journal of Applied Psychology. 1995. Vol. 80. No. 5. P. 607—620. DOI:10.1037/0021-9010.80.5.607
15. *Şengün S.* [*et al.*]. Do players communicate differently depending on the champion played? Exploring the Proteus effect in League of Legends [Elektronnyi resurs] // Technological Forecasting and Social Change. 2022. Vol. 177. DOI:10.1016/j.techfore.2022.121556
16. *Wang Y., Lu H.* Validating items of different modalities to assess the educational technology competency of pre-service teachers [Elektronnyi resurs] // Computers & Education. 2021. Vol. 162. DOI:10.1016/j.compedu.2020.104081
17. *Wested G.S.-F., Shavelson R.J.* Development of performance assessments in science: Conceptual, practical, and logistical issues // Educational Measurement: issues and practice. 1997. Vol. 3. № 16. P. 16—24.

***Information about the authors***
*Daria A. Gracheva,* Research Assistant at Center for Psychometrics and Measurement in Education, PhD student, Institute of Education, National Research University Higher School of Economics, Moscow, Russia, ORCID: https://orcid.org/0000-0002-4646-7349, e-mail: dgracheva@hse.ru

***Информация об авторах***
*Грачева Дарья Александровна,* стажер-исследователь Центра психометрики и измерений в образовании, аспирант Института образования, ФГАОУ ВО «Национальный исследовательский университет «Высшая школа экономики» (ФГАОУ ВО НИУ ВШЭ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-4646-7349, e-mail: dgracheva@hse.ru

Получена 22.10.2021
Принята в печать 26.10.2022

Received 22.10.2021
Accepted 26.10.2022