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)

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

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

Daria A. Gracheva
HSE University, Moscow, Russia
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.).

Введение

Одним из направлений развития современного образовательного тестирования является измерение многокомпонентных конструктов. Примером такого конструкта выступает критическое мышление, которое относят к метапредметным навыкам. Однако измерение последних затруднено при использовании традиционных типов заданий. Например, заданий с выбором варианта ответа. Задания же сценарного типа в цифровой среде имеют большой потенциал. Они напоминают компьютерную игру, в которой ученик сталкивается с ситуацией, где ему необходимо решить ряд проблем. Действия ученика при прохождении рассматриваются как наблюдаемые проявления измеряемого навыка — индикаторы. Сценарные задания позволяют приблизиться к поведению ученика, которое он демонстрирует в похожих ситуациях в реальной жизни, что особенно важно при измерении метапредметных навыков [7].

На практике применение такого типа заданий сталкивается с множеством трудностей. Среди них низкая надежность, малое число заданий и слабая корреляция с альтернативными измерениями. В целом, несопоставимость измерений характерна для заданий с фокусом на процесс и продукт

(performance-based tasks): сценарных заданий, эссе, экспериментов и пр. [6]. Прежние попытки создания сопоставимых учебных экспериментов не увенчались успехом, несмотря на то, что исследователи придерживались одних принципов разработки [17].

Основным этапом при разработке нового сценария является подбор подходящего контекста. Последний представляет собой набор характеристик, который задает ситуацию, где тестируемый сможет продемонстрировать нужные навыки. Степень соответствия контекста сценарных заданий друг другу напрямую связана со степенью их сопоставимости. Однако сопоставимость заданий с контекстом является малоизученной областью [6].

Целью нашего исследования являлось установление сопоставимости вариантов сценарных заданий для измерения критического мышления, которые содержат одинаковое количество индикаторов, реализованы в идентичной цифровой среде, но различаются контекстными элементами.

В первой части обсуждения мы рассмотрим предыдущие исследования заданий с контекстом, а также основные методы, которые используются для проверки сопоставимости тестов; во второй части представлены результаты анализа сопоставимости вариантов сценарных заданий. А завершим наш анализ обсуждением результатов, основных ограничений и дальнейших направлений исследования.

Обзор исследований контекста заданий

Понятие контекста и его связь с психометрическими характеристиками заданий и результатами тестирования изучается на примере опросников, эссе, игр и сценарных заданий.

В исследовании личностных опросников было показано, что уточнение контекста ведет к улучшению психометрических характеристик за счет снижения количества интерпретаций утверждений [14].

Для заданий типа эссе проверяется сопоставимость результатов при изменении тематики и стимульных материалов в формате картинок [9].

В области компьютерных игр проводятся исследования роли интерфейса в результатах тестирования. Например, в работе [15] установлено, что выбор персонажа связан с поведением респондента внутри игровой среды.

Идея содержания виртуального мира как стимула креативных решений изучается в работе [10]. В исследовании тестируемые «погружались» в разные виртуальные миры при помощи шлемов виртуальной реальности, а затем рисовали несуществующее животное. Идеи этих рисунков значимо различались в зависимости от предъявляемого контекста.

На примере заданий PISA по естествознанию исследовались характеристики контекста (степень абстрактности, назначение контекста и др.) и их связь с достижениями учащихся [13].

Использование заданий с контекстом является перспективным направлением для измерения комплексных навыков. В то же время контекст можно рассматривать как фактор, который влияет на характеристики заданий и результаты тестирования. Методы анализа сопоставимости будут рассмотрены в следующем разделе.

Обзор методов проверки сопоставимости

Проверка сопоставимости тестов проводится качественными и количественными методами, которые могут друг друга дополнять.

Качественные методы включают определение правил разработки теста и привлечение экспертов для оценки сопоставимости заданий.

К правилам разработки теста относят следование спецификации. Тем не менее установлено, что задания с открытым ответом, созданные по одной спецификации, не всегда сопоставимы [8].

Мнение экспертов, например, используется для оценки того, насколько тема задания покрывает общий или узкоспециальный вопрос [11].

Количественные методы включают использование статистического аппарата для анализа сопоставимости. Выбор статистического метода зависит от цели исследования. Если цель исследования заключается в оценке различий между группами, то в качестве статистических методов используются t-test или ANOVA. Для целей предсказания результатов будущих тестирований применяется регрессионный анализ, а корреляционный анализ может выступать мерой схожести результатов по вариантам.

Однако процесс проверки сопоставимости вариантов теста выходит за рамки работы с сырыми результатами теста. Чтобы считать варианты теста сопоставимыми, необходимо убедиться, что они измеряют один конструкт, задания имеют схожие психометрические характеристики [3].

Проверка этих допущений возможна в рамках методологии конфирматорного факторного анализа (КФА) или современной теории тестирования (Item Response Theory, IRT). Например, КФА использовался для проверки функционирования инструмента в разных форматах [16].

В данной статье мы фокусируемся на применении КФА для проверки сопоставимости. Так как данные в образовании часто являются категориальными, рассматривается случай КФА для порядковых переменных. Доказательство сопоставимости в рамках КФА сводится к проверке инвариантности общей модели инструмента. В исследованиях сопоставимости обычно рассматривают три уровня инвариантности: конфигуральный, метрический и скалярный.

На конфигуральном уровне проверяется сопоставимость структуры конструкта во всех группах [12]. На метрическом уровне значения факторных нагрузок полагаются равными во всех группах. На скалярном уровне проверяется равенство пороговых значений (в случае категориального КФА). При достижении уровня скалярной инвариантности возможно сравнение средних значений латентных факторов между группами.

Таким образом, измерение комплексных навыков требует использования методов анализа, направленных на изучение структуры теста. Например, таким методом выступает КФА. Далее указанный метод будет использован для проверки сопоставимости вариантов сценарных заданий.

Основные характеристики выборки, методов, процедуры сбора данных исследования и стратегия анализа

Выборка

В статье используются данные 500 учащихся четвертых классов, которые принимали участие в исследовании «навыков 21 века» осенью 2020 года в рамках проекта «4К современного мира. Формирование компетенций XXI века и оценка индивидуального прогресса в их развитии» при поддержке благотворительного фонда «Вклад в будущее».

Инструмент

Для оценки критического мышления используются сценарные задания в компьютерной форме из инструмента «4К», разработанного сотрудниками Центра психометрики и измерений в образовании (НИУ ВШЭ). Инструмент прошел ряд апробаций, которые свидетельствуют о его валидности [2].

В этой работе проверяется сопоставимость пары сценариев на измерение критического мышления: «Аквариум» и «Террариум». Согласно концептуальной рамке инструмента, навык критического мышления включает две составляющие: 1) «Анализ информации» — навык работы с информацией в соответствии с целями и условиями поставленной задачи; 2) «Формулирование вывода» — навык формулирования собственного вывода с помощью результатов, полученных на этапе работы с информацией [2].

Сценарий «Аквариум» предлагает тестируемым обустроить аквариум для крабов. Для работы с информацией в сценарии используется симуляция интернет-браузера, где предъявляется текст статьи (рис. 1). Текст статьи включает как релевантные, так и нерелевантные предложения. Релевант-

ные предложения содержат информацию, которая понадобится для обустройства аквариума для крабов (например, «Крабам нужно иногда залезать повыше, для этого в аквариум помещают камни»). Нерелевантные предложения содержат информацию, которая не соответствует поставленной задаче. За каждое выделенное релевантное предложение начисляется 1 балл.

Индикаторы формулирования вывода оцениваются в интерактивной среде (конструктор), где тестируемый обустраивает жилище для краба из элементов на основе информации из текста (рис. 2). За каждый верно добавленный элемент начисляется 1 балл.

В сценарии «Террариум» тестируемые сталкиваются с теми же задачами с другим содержанием, где главная цель — построить террариум для гекконов.

К навыку анализа информации относится 14 дихотомических индикаторов, к навыку формулирования вывода — 10 индикаторов (8 дихотомических и 2 политомических от 0 до 2 баллов).

Процедура сбора данных

Тестирование проходило очно в школах в присутствии администратора тестирования. Каждому участнику предоставили компьютер с доступом в интернет. В начале тестовой сессии администраторы открывали сайт тестирования на компьютерах и раздавали индивидуальные логины учащимся для входа в систему. Все инструкции и

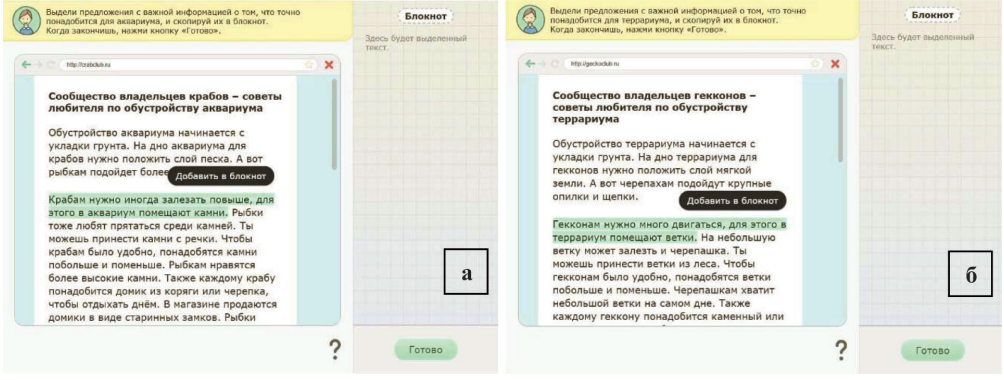

Рис. 1. Стимульный материал (текст): а — «Аквариум», б — «Террариум»

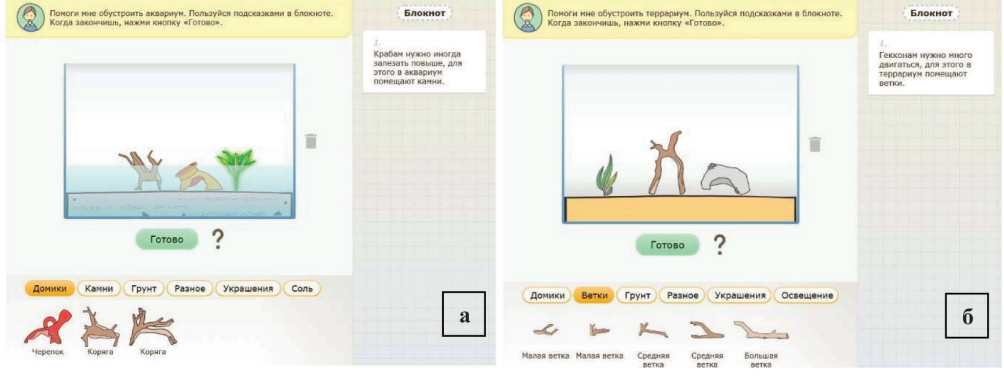

Рис. 2. Стимульный материал (конструктор): а — «Аквариум», б — «Террариум»

задания предъявлялись в компьютерном формате.

В работе использовался сбалансированный дизайн, в котором оба варианта сценария выполнялись одними тестируемыми. Выборка случайным образом была поделена на две группы. Первая группа сначала проходила задание «Аквариум», а потом задание «Террариум», вторая группа проходила задания в обратном порядке. Такой дизайн позволил контролировать эффект порядка предъявления заданий на результаты проверки сопоставимости. Перерыв между тестированиями вариантов сценариев составлял от одного дня до недели.

Стратегия анализа

Исследование сопоставимости вариантов заданий сценарного типа проводилось с применением КФА. Анализ проходил в два этапа. На первом этапе была предложена структура модели критического мышления, которая отдельно проверялась для вариантов сценариев. На втором этапе проводилась проверка измерительной инвариантности общей модели для двух сценариев.

В качестве метода оценки параметров использовался метод взвешенных наименьших квадратов (WLSMV), наиболее подходящий для порядковых и бинарных данных. Качество моделей оценивалось по следующим индексам: CFI>0.90; TLI>0.90; RMSEA<0.05 [12].

Проверка инвариантности осуществлялась путем последовательного сравнения трех моделей (конфигуральная, метрическая, скалярная). В качестве критерия сравнения принята разница показателей статистик согласия (ΔCFI в пределах 0.01, ΔRMSEA в пределах 0.015 для подтверждения инвариантности) [4]. При достижении скалярной инвариантности возможно сравнить средние значения латентных факторов разных групп, где средние значения факторов для одной группы приравниваются к нулю, а для другой группы оцениваются свободно.

Модель критического мышления содержит два главных связанных фактора — «Анализ» и «Вывод». В модели введены дополнительные ортогональные факторы стимульного материала, которые учитывают общий источник дисперсии между группами индикаторов, относящихся к работе с текстом или конструктором.

Анализ проведен в программе Mplus, версия 8.3.

Результаты

Средний балл по навыку анализа информации равен 5.56 балла (ст. отклонение — 3.83) для сценария «Аквариум» и 5.29 балла (ст. отклонение — 3.85) для сценария «Террариум». Средний результат по навыку формулирования вывода для сценария «Аквариум» равен 8.2 балла (ст. отклонение — 2.72), для сценария «Террариум» — 8.25 балла (ст. отклонение — 2.67). Между средними значениями не обнаружено статистически значимых различий как для навыка анализа информации ($t(998)=1.11$, $p>0.05$), так и навыка формулирования вывода ($t(998)=-0.29$, $p>0.05$).

Отдельные модели для сценариев «Аквариум» ($\chi2(240)=387.691$*, $p<0.000$; CFI=0.979; TLI=0.976; RMSEA=0.035. 90% CI (0.029;0.041)) и «Террариум» ($\chi2(240)=398.031$*, $p<0.000$; CFI=0.980; TLI=0.977; RMSEA=0.036, 90% CI (0.030; 0.043)) показали хорошее согласие с данными. На рис. 3—4 приведен общий вид модели и стандартизированные факторные нагрузки для сценариев «Аквариум» и «Террариум».

Результаты проверки измерительной инвариантности представлены в табл. 1. Статистики согласия по трем моделям схожи, что позволяет принять допущение о полной скалярной инвариантности инструмента. Структура критического мышления воспроизводится в разных вариантах сценариев, психометрические характеристики индикаторов не различаются.

После проверки уровней инвариантности и достижения скалярной инвариантности перейдем к сравнению средних значений латентных факторов для заданий «Аквариум» и «Террариум» (табл. 2).

Средние значения для фактора «Анализ» значимо не отличались по вариантам

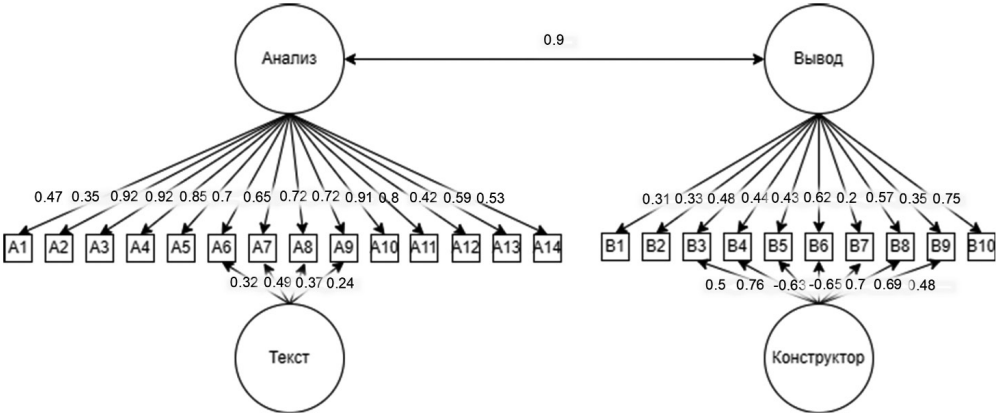

Рис. 3. Модель критического мышления («Аквариум»): все параметры модели значимы на уровне p<0.05

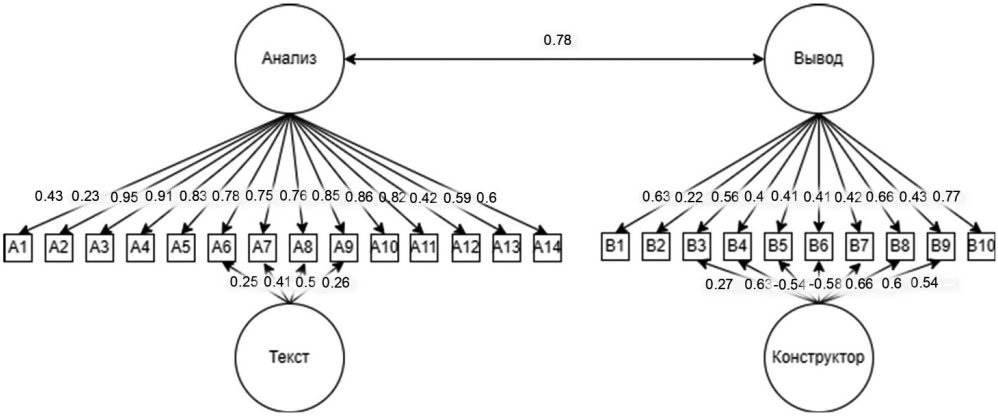

Рис. 4. Модель критического мышления («Террариум»): все параметры модели значимы на уровне p<0,05

Таблица 1

Результаты проверки измерительной инвариантности

Модель	χ^2 (df)	RMSEA	CFI	TLI
Конфигуральная	785.743* (480)	0.036 (90% CI 0.031; 0.040)	0.979	0.976
Метрическая	835.083* (511)	0.036 (90% CI 0.031; 0.040)	0.978	0.976
Скалярная	915.226* (532)	0.038 (90% CI 0.034; 0.042)	0.974	0.973

Примечание: * p<0,05.

заданий. То есть в среднем навык анализа информации одинаково оценивается в двух вариантах сценариев при изменении характеристик контекста сценария. По средним значениям фактора «Текст» также не выявлено значимых различий.

Таблица 2

Средние значения латентных факторов

Фактор	Средние значения для сценария «Террариум»	Z-статистика
«Анализ»	-0.089 (0.066)	-1.353
«Вывод»	0.211 (0.071)	2.965*
«Текст»	-0.003 (0.129)	-0.026
«Конструктор»	-0.272 (0.079)	-3.433*

Примечания: В скобках указаны стандартные ошибки измерения. Средние значения факторов для сценария «Аквариум» приравнены к нулю. * p<0,05.

Тем не менее значимая разница средних по фактору «Вывод» выступает свидетельством того, что в этой части сценарий «Террариум» оказался легче, чем сценарий «Аквариум». Различия сохранились в факторе конструктора.

Содержательная интерпретация факторов стимульного материала часто бывает затруднительна. Однако полученные результаты позволяют говорить о том, что результаты учащихся значимо различаются в части сценария, где им необходимо продемонстрировать навык формулирования вывода через работу с элементами в конструкторе.

Обсуждение результатов

Комплексные конструкты требуют новых измерительных подходов. Таким подходом является применение сценарных заданий в цифровой среде. В то же время для сценарных заданий в большей степени выражен риск получения несопоставимых результатов [6].

Одной из угроз сопоставимости является контекст сценария. В данной статье мы использовали сценарии для измерения критического мышления «Аквариум» и «Террариум», которые содержали один набор индикаторов, но различались контекстными характеристиками. Проведенный анализ измерительной инвариантности показал, что изменение контекста не меняет теоретическую структуру инструмента, а психометрические характеристики индикаторов значимо не отличались по вариантам заданий.

Результаты сравнения средних латентных факторов показали, что тестируемые получают более низкие оценки по навыку формулирования вывода в сценарии «Аквариум», чем «Террариум», в то время как оценки по навыку анализа информации не отличаются по вариантам.

Благодаря дизайну сбора данных, при котором соблюдался случайный порядок предъявления вариантов, мы можем считать, что различия в результатах появляются не за счет эффекта научения в решении подобных задач, а за счет различий в контекстных элементах.

По результатам предыдущих исследований контекст задания может оказывать эффект на результаты теста. Так, знакомый контекст может давать преимущество в решении задач [5]. В исследовании креативности контекст «виртуального мира» проявлялся в рисунках несуществующих животных [10].

Другой причиной различия в результатах мог стать формат заданий внутри сценариев. Ранее было показано, что формат задания с выбором варианта ответа в меньшей степени подвержен колебаниям трудности. Большие проблемы характерны для заданий с открытым ответом или объединенных общим стимульным материалом, например, работа с текстом [3].

Однако полученные нами результаты свидетельствуют о том, что объемные текстовые задания могут быть сопоставимы. Отчасти это можно объяснить использованием метода «клонирования», который позволяет создать максимально похожие тексты в разных контекстах [1]. Задания, содержащие элементы интерактива, в большей степени подвержены риску несопоставимости, что могло стать причиной раз-

личия в оценках по вариантам для навыка формулирования вывода.

Проведенное исследование имеет некоторые ограничения. Оно проводилось на одной паре сценариев для измерения одного навыка, поэтому полученные результаты нуждаются в ревалидизации на примере других сценариев и навыков. Кроме того, в данной работе мы анализировали сопоставимость вариантов, основываясь только на анализе структуры данных и функционировании индикаторов.

Дальнейшие направления исследования сопоставимости заданий с контекстом включают использование как количественных, так и качественных методов. Лингвистический анализ текстов заданий и привлечение экспертов позволят глубже понять различия между вариантами сценариев. Перспективным направлением является проведение когнитивных лабораторий и интервью с учащимися для понимания вклада контекста в результаты теста. Дальнейшее применение количественных методов может заключаться в оценке эффекта взаимодействия контекста сценария с другими характеристиками заданий.

Заключение

Задания в цифровой среде, содержащие интерактивные элементы, являются трендом в области измерений в образовании. Однако создать сопоставимые задания «на глаз» практически невозможно. Разнообразие ситуаций и большая свобода действий тестируемого внутри цифровой среды могут снижать сопоставимость измерений. Это особенно важно в случае, когда задания используются как альтернативные варианты, например, для проведения мониторинговых исследований. Отсутствие проверки сопоставимости вариантов заданий может создавать тестируемым неравные возможности для демонстрации своих способностей, а решения, которые будут приняты по результатам тестирования, окажутся невалидными.

Проведенный нами анализ определил, что большему риску несопоставимости подвержены задания, где тестируемый самостоятельно собирает объект из элементов. Различия в результатах могут объясняться контекстом заданий или особенностью формата заданий. Исследование причин полученных результатов, а также ревалидизация сформулированных здесь выводов могут проводиться отдельно для повышения качества инновационных типов заданий и изучения возможности их использования как для масштабных, так и локальных тестирований.

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.

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.

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

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

Получена 22.10.2021
Принята в печать 26.10.2022

Received 22.10.2021
Accepted 26.10.2022