ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

ИНСТИТУТ ОБРАЗОВАНИЯ

На правах рукописи

Грачева Дарья Александровна

Обеспечение сопоставимости результатов тестирования комплексных конструктов с использованием сценарных заданий

РЕЗЮМЕ ДИССЕРТАЦИИ

на соискание учёной степени

кандидата наук об образовании

Научный руководитель:

Авдеева Светлана Михайловна, к. т. н

Москва – 2024

Оглавление

Используемые понятия и сокращения

Латентный конструкт – это совокупность паттернов и закономерностей проявления какого-либо явления, определенных разработчиком или экспертом на основании известных теорий, представлений и идей, с учетом имеющихся ограничений и поставленных задач.

Комплексный конструкт – конструкт, состоящий из множества элементов, установок, моделей поведения или способов действия и мышления, с фокусом на их применении в жизненных ситуациях.

VPBA – Virtual Performance-Based Assessment, формат тестирования, предполагающий оценку способностей респондента через анализ поведения в заранее смоделированной цифровой тестовой среде.

Поведенческий индикатор – наблюдаемое действие (поведение) в тестовой среде, по которому делается вывод о выраженности латентного конструкта у тестируемого.

Сценарные задания (задания сценарного типа, scenario-based tasks) – тип VPBA, в котором поведенческие индикаторы объединены контекстом.

ECD – Evidence-Centered Design, подход доказательной аргументации к разработке тестов.

КФА – конфирматорный факторный анализ.

Введение

Обоснование актуальности исследования

Современное образование в России ориентировано не только на усвоение предметных знаний, но и на формирование у обучающихся универсальных учебных действий (УУД), которые помогают успешно применять новые знания и навыки в жизненных ситуациях и обеспечивают возможность самостоятельно развиваться на протяжении всей жизни. Федеральные государственные образовательные стандарты (ФГОС) устанавливают требования к формированию УУД (регулярных, познавательных, коммуникативных) на разных ступенях общего образования[1]. В зарубежной литературе вместо совокупности УУД встречаются термины «универсальные навыки» или «навыки 21 века», к которым относятся критическое и креативное мышление, коммуникация, кооперация и другие навыки[2][3]. Важность развития универсальных навыков обучающихся для саморазвития и успешной адаптации во взрослой жизни подчеркивается многими зарубежными экспертами [Ananiadoui, Claro, 2009; Griffin, Care, 2014; Pellegrino, 2017].

Согласно ФГОС начального общего образования (ФГОС НОО), обучающиеся овладевают базовыми логическими и исследовательскими действиями, которые являются частью универсальных учебных познавательных действий и в то же время поведенческими индикаторами критического мышления: «выбирать источник информации», «распознавать достоверную и недостоверную информацию», «выявлять недостаток информации», «анализировать информацию», «формулировать выводы и подкреплять их доказательствами». Во ФГОС основного общего образования (ФГОС ООО) обучающиеся в процессе обучения «самостоятельно формулируют обобщение и выводы», учатся

---

[1] Приказы Министерства просвещения РФ об утверждении федеральных государственных образовательных стандартов начального общего образования, основного общего образования, среднего общего образования. Сайт Министерства просвещения РФ. URL: https://fgosreestr.ru/educational_standard

[2] New Vision for Education. Unlocking the Potential of Technology. World Economic Forum report, 2016.

[3] Partnership for 21st Century Learning (2016). Framework for 21st century learning. URL: http://www.p21.org/our-work/p21-framework

«выбирать, анализировать, систематизировать» информацию, оценивать ее надежность. Также, универсальные учебные коммуникативные действия, описанные во ФГОС НОО и ООО, содержат характеристики коммуникации и кооперации, такие как «проявлять уважительное отношение к собеседнику», «выражать эмоции в соответствии с целями общения» и другие.

Развитие универсальных учебных действий в школе приводит к необходимости их оценивания. Исследователи отмечают сложность оценивания совокупности УУД [Шкерина и др., 2019] и универсальных навыков [Care и др., 2018; Geisinger, 2016] по сравнению с отдельными предметными знаниями. В области измерений универсальные навыки называют комплексными конструктами – «конструкты, состоящие из множества элементов, установок, моделей поведения или способов действия и мышления, с фокусом на их применении в жизненных ситуациях» [Ercikan, Oliveri, 2016]. Другие авторы в качестве основной характеристики комплексного конструкта выделяют наличие множества элементов и связей между ними [Ridley и др., 2021], отмечают, что такие конструкты являются сложными в концептуализации и операционализации [Carneiro, Rocha, Silva, 2009; Gorin, Mislevy, 2013].

В зарубежной литературе термин «комплексный конструкт» преимущественно используется для обозначения универсальных навыков – креативности, критического мышления, коммуникации, навыка совместного решения проблем [Andrews-Todd, Forsyth, 2020; Ercikan, Oliveri, 2016; Hyytinen и др., 2024], однако встречается упоминание иных комплексных конструктов, например, статистическая грамотность [Watson, Callingham, 2003], математическое мышление [Pitta-Pantazi, Sophocleous, 2017] или цифровая грамотность [Авдеева, Тарасова, 2023]. В международном исследовании PISA термин «комплексный конструкт» упоминается в связке с конструктом «глобальная компетентность» [Sälzer, Roczen, 2018]. В России на основе методологии и результатов измерения PISA разработан инструмент оценки комплексного конструкта «функциональная грамотность» [Ковалева, Колачев, 2023].

В отличие от отдельных знаний и навыков измерение комплексных конструктов предполагает выход за пределы традиционных типов заданий, таких как задания с выбором варианта ответа, опросники в формате самоотчета. Альтернативный формат тестирования должен учитывать многообразные составляющие такого конструкта и оценивать разные модели поведения респондентов в некоторой жизненной ситуации [Ercikan, Oliveri, 2016]. Подходящим форматом для оценки комплексных конструктов являются задания в формате *performance-based*, где тестируемые могут продемонстрировать уровень владения навыком через наблюдаемые действия в заранее установленной ситуации (тестовой среде). Ввиду развития технологий появился термин *virtual performance-based assessment* (VPBA) [Andrews-Todd и др., 2021], где поведение респондента фиксируется в цифровой тестовой среде.

К категории *virtual performance-based* относят сценарные задания, в них действия респондентов в цифровой среде объединены контекстом (*scenario-based tasks*), поэтому такие задания принято называть контекстными [Ruiz-Primo, Li, 2015]. Контекст «погружает» респондентов в задание, приближая среду тестирования к реальным задачам, что особенно важно при измерении универсальных навыков. Сегодня задания сценарного типа реализуются в цифровой среде с использованием игровых элементов и симуляций, например, для оценки критического мышления [Braun и др., 2020; Uglanova и др., 2022], навыка совместного решения проблем в рамках международного исследования PISA [Stadler и др., 2020].

Несмотря на преимущества нового формата оценивания, мало внимания уделяется возможности разработки вариантов сценарных заданий. При этом очевидны ограничения, возникающие при использовании только одного варианта задания. В случае повторного тестирования одними заданиями прогресс в результатах может объясняться эффектом практики или научения. Кроме того, частое использование заданий снижает вовлеченность тестируемых в предложенные сценарии, потому что все ситуации кажутся знакомыми. Использование нескольких вариантов заданий сценарного типа позволит не только

справиться с вышеуказанными проблемами, но и откроет возможности к проведению крупных мониторинговых и сравнительных исследований универсальных навыков и других комплексных конструктов.

Эксперты подчеркивают, что сопоставимость результатов, получаемых с помощью вариантов сценарных заданий, должна быть обеспечена на протяжении всего цикла оценивания, начиная с концептуализации и разработки инструмента измерения до этапа работы с данными и анализа последствий тестирования [He, Vijver van de, 2012; Kolen, 1999]. Сопоставимость результатов тестирования часто рассматривается в контексте справедливого оценивания и обеспечения равных возможностей для каждого участника продемонстрировать свои знания и навыки [Gipps, Stobart, 2009; Kunnan, 2004], что является важной ценностью в системе образования.

Настоящее исследование посвящено методологическим вопросам обеспечения сопоставимости результатов тестирования комплексных конструктов с использованием вариантов сценарных заданий. В международной литературе встречаются примеры исследований, где рассматриваются несколько вариантов заданий для измерения универсальных навыков или других комплексных конструктов [Rojas и др., 2021; Wang, Liu, Hau, 2022], однако в них детально не описываются подходы к разработке вариантов или не обосновывается методология работы с данными для доказательства сопоставимости результатов. Разработка сценарных заданий в цифровой среде даже в одном экземпляре является трудоемким процессом [Угланова, Брун, Васин, 2018], поэтому возникает задача формализации подходов для разработки вариантов таких заданий.

Таким образом, актуальность исследования обусловлена важностью объективного и справедливого измерения комплексных конструктов в образовании и необходимостью обеспечения сопоставимости результатов тестирования между вариантами заданий. Нарушение сопоставимости результатов тестирования затрудняет корректное сравнение и интерпретацию результатов и может негативно повлиять на принятие решений по итогам оценки. Разработка теоретически обоснованного и эмпирический доказанного подхода к обеспечению

сопоставимости вариантов сценарных заданий позволит получить более точное сравнение результатов тестирования комплексных конструктов, что, в свою очередь, будет способствовать повышению качества оценки.

Оценивание комплексных конструктов в сценарном формате накладывает дополнительные сложности на процесс обеспечения сопоставимости результатов. Большое внимание в настоящей работе уделяется исследованию контекста как неотъемлемой части сценарного задания. В предыдущих исследованиях разработка вариантов тестов осуществлялась за счет изменения контекстных характеристик, например, в случае измерения навыков письма с помощью эссе [Cho, Rijmen, Novák, 2013] или компетенций в области медицины [Lievens, Sackett, 2007]. Эксперты обеспокоены проблемой экстраполяции (генерализации) выводов, сделанных по результатам тестирования с использованием сценарных заданий с разным контекстом [Andrews-Todd и др., 2021]. Существуют свидетельства того, изменение контекста задачи вызывает различия в результатах тестирования [Nelson, Guegan, 2019; Schliemann, Magalhães, 1990]. Кроме того, выполнение контекстных заданий может требовать от участников использования дополнительных знаний и навыков, которые не являются целью оценивания [Messick, 1994]. В результате, контекст может влиять на измеряемый конструкт, вызывая проблемы с валидностью и надежностью измерений, сопоставимостью результатов [Bond, Moss, Carr, 1996]. Таким образом, подход к обеспечению сопоставимости сценарных заданий должен быть разработан с учетом эффекта контекста – изменения в результатах тестирования (структуре конструкта и/или психометрических характеристиках заданий), вызванного изменением контекстных характеристик сценарного задания. Кроме того, представляется важным оценить масштаб различий в результатах тестирования комплексных конструктов с использованием сценарных заданий, которые вызваны изменением контекста.

Таким образом, **целью** диссертационного исследования является разработка методологического подхода к обеспечению сопоставимости сценарных заданий для измерения комплексных конструктов на протяжении всего цикла оценивания.

Исследовательские вопросы работы поставлены следующим образом:

Каким образом обеспечить сопоставимость вариантов заданий сценарного типа на этапе разработки инструментов измерения?

Какие методологические подходы работы с данными позволят обосновать сопоставимость результатов тестирования между вариантами сценарных заданий?

В какой степени контекст сценарного задания связан с различиями в результатах тестирования комплексных конструктов?

Для достижения поставленной цели решаются следующие задачи:

- Сформировать и обосновать подход к разработке вариантов сценарных заданий для обеспечения сопоставимых результатов между вариантами.
- Сформировать и обосновать методологические подходы работы с данными для проверки сопоставимости вариантов заданий сценарного типа с учетом эффекта контекста.
- Разработать варианты заданий сценарного типа в соответствии с выбранным подходом к разработке.
- Реализовать методологию работы с данными для проверки сопоставимости вариантов заданий сценарного типа.
- Количественно оценить связь контекста с различиями в результатах тестирования между вариантами сценарных заданий.

Выводы исследования подтверждаются на сценарных заданиях для измерения критического мышления у учащихся начальной школы.

Теоретические основания исследования

Работа опирается на концептуальные рамки сопоставимости измерений [He, Vijver van de, 2012; Kolen, 1999] и справедливого оценивания [Kunnan, 2004], которые рассматривают сопоставимость результатов тестирования на протяжении всего цикла оценивания от ранних этапов разработки инструмента до работы с данными тестирования. На их основании были определены два основных критерия сопоставимости: сопоставимость конструкта и сопоставимость статистических

(психометрических) характеристик на уровне всего теста и отдельных индикаторов.

Степень разработанности проблемы

В случае использования нескольких вариантов заданий возникает задача обеспечить сопоставимость результатов тестирования между вариантами. При измерении комплексных конструктов в сценарном формате мы дополнительно сталкиваемся с вызовами, связанными с воспроизведением особенностей тестовой среды, природы конструкта и связей между всеми его составляющими в нескольких вариантах.

Сопоставимость результатов тестирования между вариантами теста должна быть обеспечена на протяжении всего цикла оценивания – от ранних этапов концептуализации и разработки инструмента измерения до работы с данными и интерпретации результатов. Это требует разработки методологического подхода для обеспечения сопоставимости, включая формирование методологических подходов к разработке тестов и количественного анализа данных.

Далее рассмотрим степень разработанности проблемы исследования с двух позиций, наиболее релевантных для данной работы: определение подхода к разработке сопоставимых вариантов заданий сценарного типа и методологии анализа данных с учетом эффекта контекста.

Подходы к разработке сопоставимых вариантов теста

В литературе описаны несколько подходов к разработке сопоставимых вариантов теста (Грачева, Тарасова, 2022).

Субъективный подход предполагает, что сопоставимость тестовых вариантов обеспечивается опытом экспертов. Разработчики создают варианты заданий на основе тестовой спецификации и выносят заключение о их сопоставимости, полагаясь на практический опыт.

Подход на основе банка заданий заключается в случайном отборе заданий из банка заданий для создания новых тестовых вариантов [Irvine, Kyllonen, 2013].

Считается, что случайный отбор нивелирует отличия между вариантами теста, делая их сопоставимыми.

Разработка сопоставимых вариантов заданий активно развивалась в рамках исследований автоматической генерации заданий (АГЗ). Идея автоматической генерации заданий заключается в использовании компьютерных технологий для автоматизации процесса разработки тестов [Gierl, Lai, Tanygin, 2021]. Среди технологий для автоматизации процесса разработки тестов набирает популярность применение моделей машинного обучения, в частности, моделей обработки естественного языка (Natural Language Processing). Например, существуют работы, где модели машинного обучения используются для генерации вариантов заданий по чтению [Attali и др., 2022].

Классические подходы к автоматизации разработки тестов используют шаблоны задания, которые описывались в работах разных исследователей [Bejar, 1991; Haladyna, Shindoll, 1989; Osburn, 1968]. Подход к разработке на основе шаблонов предполагает, что для каждого аспекта области содержания подбирается «задание-родитель», которое затем представляется в виде шаблона. Идея шаблонов используется в подходе логических структур и ключевых элементов [Gierl, Lai, Tanygin, 2021], а также в подходе обязательных и вариативных элементов [Irvine, Dann, Anderson, 1990].

В рамках подхода обязательных и вариативных элементов задания разбиваются на элементы, а затем элементы классифицируются на обязательные и вариативные. Обязательные элементы определяют трудность задания, то есть их замена оказывает влияние на психометрические характеристики задания. Изменение вариативных элементов не оказывает существенного влияния на психометрические характеристики. Таким образом, для создания сопоставимых вариантов теста только вариативные элементы задания подлежат замене. В результате, удается получить задания, которые будут максимально похожи. Такие задания в литературе называют клонами, а вышеупомянутые подходы на основе шаблонов задания можно объединить общим названием – подходы клонирования [Clause и др., 1998; Gierl, Haladyna, 2012].

Проведенный анализ литературы позволил выделить ограничения описанных подходов применительно к разработке вариантов заданий сценарного типа. Прежде всего, выбор подхода к разработке для вариантов заданий сценарного типа должен учитывать особенности тестового формата и измеряемого конструкта.

Разработка заданий сценарного типа является трудоемким процессом (Угланова и др., 2018), поэтому создание банка заданий не является целесообразными. Субъективность экспертов создает риски для сопоставимости результатов между вариантами, а неинтерпретируемые методы машинного обучения не могут с точностью воспроизвести особенности тестовой среды и природу сложного конструкта.

Подходы клонирования позволяют получить более сопоставимые варианты тестов за счет структурирования элементов заданий. Однако среди экспертов нет единого мнения, какие элементы выделять в заданиях, и какие из них относятся к обязательным, а какие – к вариативным [Lievens, Sackett, 2007; Williamson и др., 2002].

В качестве ограничения отметим, что подход клонирования называют упрощенным подходом к разработке вариантов теста, который провоцирует «натаскивание». Воспроизведение большинства элементов задания-родителя приводит к созданию практически идентичных заданий. При этом отмечается, что важным аспектом измерения навыков высокого порядка является возможность их оценивания в более свободной тестовой среде [Поддьяков, 2012]. Возникает противоречие между свободной средой, необходимой для проявления сложного навыка, и справедливым оцениванием.

С развитием технологий АГЗ подходы клонирования активно используются в разных областях. Например при создании симуляций в области сетевых технологий [Fay, Levy, Mehta, 2018], медицинских ситуационных тестов [Lievens, Sackett, 2007], компьютерных тестов для менеджеров [Lievens, Anseel, 2007]. Однако в литературе не было найдено примеров использования подходов

клонирования для разработки вариантов сценарных заданий для измерения комплексных конструктов

Для создания сценарных заданий, оценивающих комплексные конструкты, применяется метод доказательной аргументации (Evidence-Centered Design (ECD) [Mislevy, Haertel, 2006], который рассматривает процесс тестирования как процесс сбора доказательств для вынесения обоснованного вывода о способностях тестируемых. Структура метода доказательной аргументации состоит из нескольких моделей, объединяющих основные процессы при создании инструмента измерения. Такой доказательный подход позволяет обеспечить валидность измерений на протяжении всего цикла оценивания.

Таким образом, подходы клонирования недостаточно изучены в приложении к созданию сопоставимых вариантов сценарных заданий. Подход клонирования в сочетании с методом доказательной аргументации может являться перспективным подходом для обеспечения сопоставимости результатов измерения комплексных конструктов.

Среди первостепенных задач для работы в этом направлении можно выделить следующее: определить конкретные элементы заданий сценарного типа и найти баланс между свободной тестовой средой и правилами клонирования для обеспечения сопоставимых результатов между вариантами сценариев.

Тем не менее, исследования показывают, что обеспечение мер по сопоставимости на этапе разработки не является гарантом сопоставимости на уровне данных [Lee, Anderson, 2007]. Необходимы специальные исследования, доказывающие, что результаты тестирования между вариантами заданий сценарного типа являются сопоставимыми.

Методы работы с данными для обеспечения сопоставимости вариантов теста

Данные, получаемые по результатам тестирования комплексных конструктов с использованием сценарных заданий, имеют особенности, которые определяют методы анализа.

Во-первых, комплексные конструкты включают несколько составляющих, которые являются целью оценивания.

Для анализа комплексных конструктов используются многомерные измерительные модели в рамках подхода латентного моделирования [Levy, 2013]. Например, многомерные модели IRT [Reckase, 2006]. Альтернативой могут являться измерительные модели в методологии моделирования структурных уравнений (Structural equation modeling, SEM) – многомерные модели конфирматорного факторного анализа [Brown, 2006].

Второе отличие оценивания комплексных конструктов с использованием сценарных заданий от «традиционного» оценивания заключается в нарушении допущения о локальной независимости отдельных поведенческих индикаторов. Сценарные задания отличаются от более традиционных заданий, например, с выбором вариантам ответа, более насыщенной тестовой средой, наличием контекста и симуляций, которые создают зависимости между наблюдаемыми действиями респондента. Один из способов учета таких контекстных зависимостей является выделение фактора контекста через построение бифакторных моделей [Levy, 2013; Rijmen, 2010].

Таким образом, анализ результатов измерения комплексных конструктов с использованием сценарных заданий предполагает использование методологии многомерного латентного моделирования с учетом контекстных связей между индикаторами.

В методологии моделирования структурных уравнений сопоставимость на уровне всего теста и психометрических характеристик заданий возможно оценить в мультигрупповой модели КФА. Доказательство сопоставимости сводится к последовательной проверке уровней инвариантности мультигрупповой модели: на уровне структур конструкта (конфигуральная инвариантность) и отдельных психометрических характеристик индикаторов: дискриминативности и трудности (метрическая и скалярная инвариантность) [Brown, 2006].

Несмотря на то, что методология измерительной инвариантности традиционно проверяет сопоставимость инструмента измерения между разными

группами респондентов, существуют исследования, где проверялось функционирования инструмента в разных вариантах теста [Rojas и др., 2021].

В рамках данной диссертационной работы подчеркивается важность исследования контекста сценарного задания для обеспечения сопоставимости измерений. Эффект контекста в вариантах заданий сценарного типа может вызывать различия как в теоретической структуре конструкта, так и характеристиках отдельных индикаторов. Возникает необходимость не только обосновать методологические подходы к обеспечению сопоставимости вариантов сценарных заданий с учетом контекстного компонента, но и количественно определить эффект контекста на результаты тестирования.

В литературе большее внимание уделено исследованию эффекта метода – как результаты тестирования зависят от метода измерения [Eid, Geiser, Koch, 2016]. Для исследования эффекта метода применяется методология на основе матрицы «способности х методы измерения» (Multitrait-Multimethod Matrix, MTMM), которая была предложена Д. Кэмпбелл и Д. Фиске [Campbell, Fiske, 1959].

Другая методология, в рамках которой возможно исследовать эффект метода – Теория генерализации (Generalizability Theory). Основы Теории генерализации описаны в статьях Л. Кронбаха [Cronbach, 1972], и позже были дополнены в работах Р. Шавелсона и Р. Бреннона [Brennan, 1992; Shavelson, Webb, Rowley, 1992]. В рамках данного диссертационного исследования основы Теории генерализации были описаны в статье [Грачева, 2023].

В статьях, где исследуется эффект метода, поднимается вопрос об эффекте взаимодействия метода измерений и респондента. Предполагается, что эффект метода может быть не эквивалентен для всех респондентов [Kroehne и др., 2019; Shavelson, Baxter, Gao, 1993]. Например, одни респонденты лучше справляются с сценарными заданиями в одном контексте и хуже справляются с заданиями в другом контексте, и наоборот для других респондентов. Методы Теории генерализации позволяют не только количественно оценить общий эффект контекста, одинаковый для всех респондентов, но и эффект взаимодействия тестируемого и контекста.

Таким образом, для проверки сопоставимости на этапе работы с данными предлагается использовать мультигрупповые модели КФА, которые позволят одновременно оценить, будет ли достигаться конструктный и психометрический критерий сопоставимости измерений. При этом, использование мультигрупповых моделей должно учитывать особенности данных тестирования комплексных конструктов с использованием сценарных заданий: многомерность и учет контекстных связей между индикаторами. Этим запросам отвечает класс бифакторных моделей КФА. Не менее важной задачей является количественно определились эффект контекста на результаты тестирования комплексных конструктов с использованием сценарных заданий.

Методология и дизайн исследования

Эмпирической базой исследования являются данные по результатам тестирования с применением инструмента «4К» для измерения критического мышления у учащихся начальной школы, разработанного сотрудниками Лаборатории измерения новых конструктов и дизайна тестов Института образования НИУ ВШЭ.

Для формирования подхода к разработке сопоставимых вариантов сценарных заданий использовались принципы метода доказательной аргументации при разработке тестов (Evidence-centred Design, ECD) [Mislevy, Haertel, 2006] и автоматической генерации при разработке заданий [Gierl, Lai, Tanygin, 2021].

Для доказательства сопоставимости на уровне данных применялась методология латентного моделирования и проверки измерительной инвариантности на мультигрупповой модели конфирматорного факторного анализа (КФА) с учетом контекстных связей между индикаторами (бифакторные модели КФА). Ввиду того, что поведенческие индикаторы часто категориальные (дихотомические или политомические), рекомендуется использовать конфирматорный факторный анализ для категориальный переменных (categorical confirmatory factor analysis, CCFA) [Kim, Yoon, 2011]. Дополнительно в рамках исследования методология CFA-MTMM использовалась для оценки степени сопоставимости вариантов сценарных заданий, разработанных в логике подхода клонирования. Для оценки эффекта контекста на результаты оценивания использовались методы Теории генерализации [Cronbach, 1972].

Для сбора данных по вариантам сценарных заданий использовался сбалансированный внутригрупповой дизайн, где варианты теста предъявляются одной группе респондентов во всех возможных порядках (случайным образом). Выбор в пользу внутригруппового дизайна по сравнению с межгрупповым позволяет избежать альтернативных объяснений результатов, связанных с различием в характеристиках респондентов. Исследование проведено на выборках учащихся 4-х классов (выборки от 381 до 1096 учащихся), которые проходили

тестирование универсальных навыков с использованием инструмента «4К» в 2021 г.

Описание инструмента

Инструмент «4К» разработан в логике доказательной аргументации (ECD). В рамках диссертационного исследования подробно рассматриваются два задания сценарного типа для измерения критического мышления у младших школьников: задание «Аквариум» и «Динозавр».

Согласно концептуальной рамке инструмента, навык критического мышления включает навык работы с информацией в соответствии с целями и условиями поставленной задачи и навык формулирования собственного вывода с помощью результатов, полученных на этапе анализа. В инструменте оцениваются следующие составляющие навыка анализа информации: выделение достоверных (надежных) источников информации; выделение релевантной информации для решения задачи. Подробнее концептуальная рамка критического мышления представлена в статье [Uglanova и др., 2022].

Ниже приведено краткое описание заданий сценарного типа в двух вариантах.

Сценарий «Аквариум»

Контекст сценария задает проблемную ситуацию, где тестируемому необходимо обустроить аквариум для крабов.

Сюжет сценария предполагает, что тестируемому сначала нужно определить достоверный источник, содержащий информацию о том, как обустроить аквариум для крабов. Для этого в задании используется симуляция интернет-браузера, где представлено несколько ссылок, только одна из ссылок является достоверной. При выборе наиболее достоверной ссылки тестируемый демонстрирует способность к определению достоверного источника информации и получает 1 балл (Рисунок 1).

Далее тестируемый анализирует текст по выбранной ссылке, чтобы узнать, какие объекты нужны для обустройства аквариума (Рисунок 2). Инструкция к этой части задания звучит следующим образом: «Выдели предложения с важной

информацией о том, что точно понадобится для аквариума». За каждое верно выделенное (релевантное) предложение в тексте статьи тестируемый получает 1 балл. В результате этой части задания тестируемый выборочно сохраняет предложения из текста статьи в блокнот (правая часть Рисунка 2).

Рисунок 1 – Пример экрана сценарного задания «Аквариум» (достоверность источника информации)

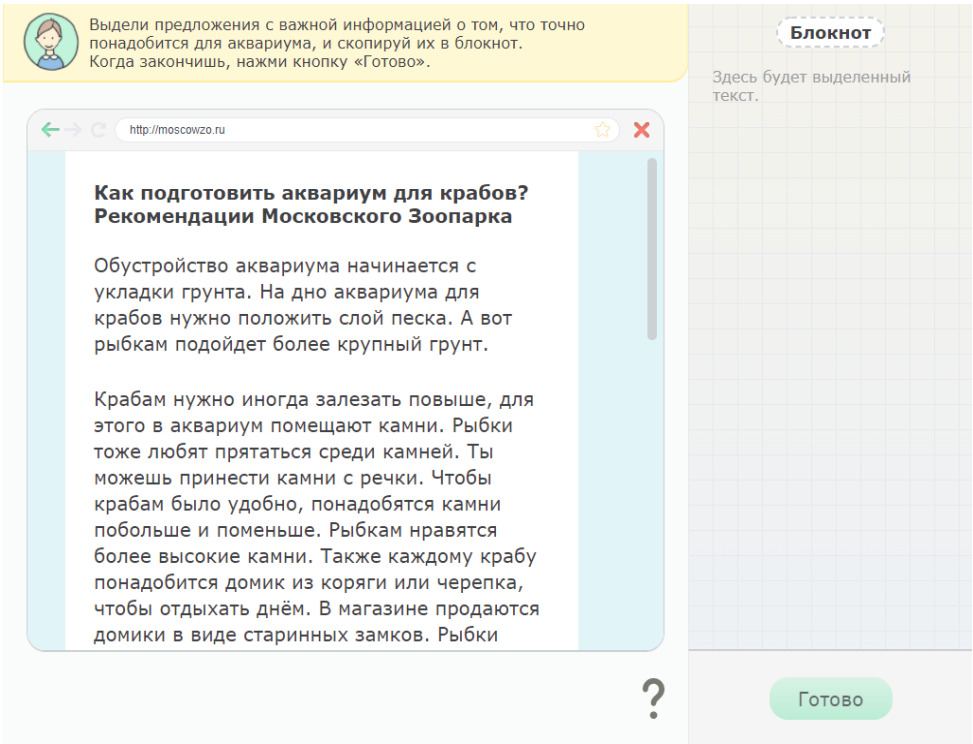

Рисунок 2 – Пример экрана сценарного задания «Аквариум» (релевантность информации)

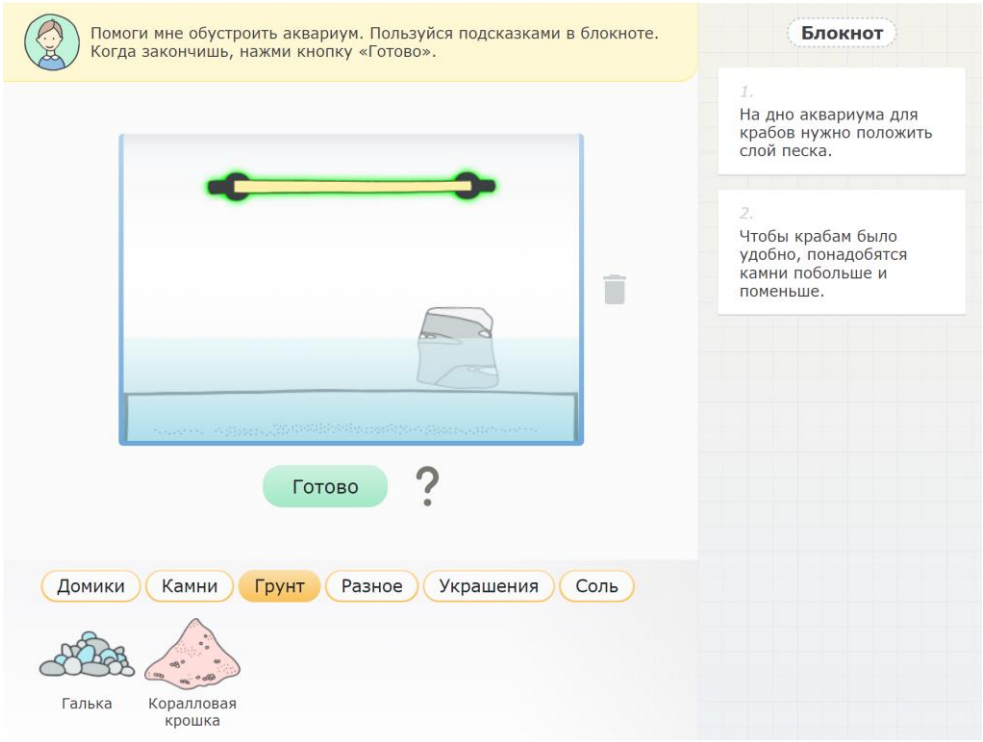

Рисунок 3 – Пример экрана сценарного задания «Аквариум» (формулирование вывода)

Наконец, на основе проанализированного текста тестируемый обустраивает аквариум из объектов в симуляции. Симуляция содержит объекты, которые упоминались в тексте статьи и те, которые не были упомянуты, сгруппированные по категориям: домики, камни, грунт, разное, украшения, соль (Рисунок 3).

Согласно проанализированной информации, тестируемый должен принять решение, какие объекты нужно поставить в аквариум, а какие – нет. За каждый верно поставленный объект тестируемый получает 1 балл.

Сценарий «Динозавр»

Контекст сценария задает проблемную ситуацию, где тестируемого просят помочь подготовить доклад про несуществующего динозавра массоспондила для ответа на ключевой вопрос, на скольких лапах ходил этот динозавр.

Для достижения цели тестируемому необходимо выбрать наиболее достоверную ссылку, где будет содержатся достоверная информация об этом типе

динозавра (Рисунок 4). За выбор наиболее достоверной ссылки тестируемый получает 1 балл.

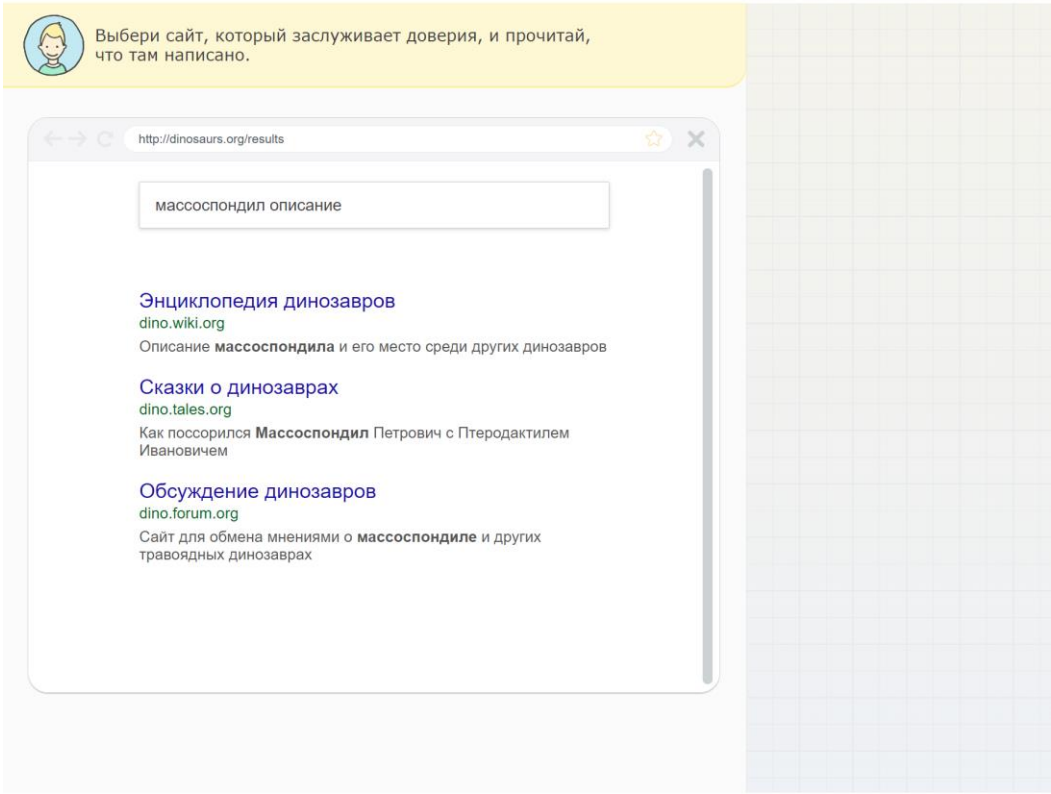

Рисунок 4 – Пример экрана сценарного задания «Динозавр» (достоверность источника информации)

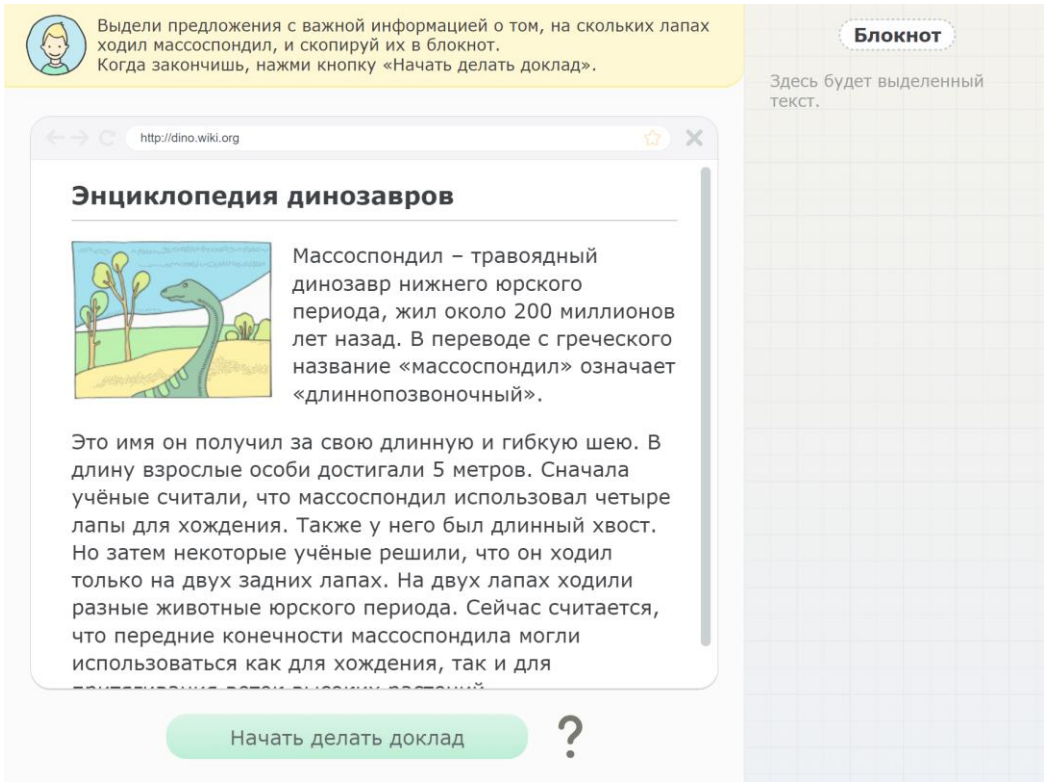

Рисунок 5 – Пример экрана сценарного задания «Динозавр» (релевантность информации)

Далее тестируемому предлагается проанализировать текст электронной статьи, выделяя только важную информацию о том, на скольких лапах ходил массоспондил (Рисунок 5). Тест статьи содержит релевантные предложения, которые содержит три разные точки зрения относительно заданного вопроса (динозавр использовал две лапы для хождения, четыре лапы для хождения, либо точно не известно). В заключение тестируемого просят сделать итоговый вывод, на скольких лапах ходил динозавр массоспондил, опираясь на проанализированную информацию.

Результаты исследования

Результаты исследования представлены в соответствии с исследовательскими вопросами.

Первый этап работы отвечал на вопрос – **каким образом обеспечить сопоставимость вариантов заданий сценарного типа на этапе разработки инструментов измерения?**

В рамках данного диссертационного исследования был сформирован подход клонирования для разработки сопоставимых вариантов сценарных заданий. Основные результаты данного этапа исследования опубликованы в статье [Грачева, Тарасова, 2022] и расширены в тексте диссертационного исследования.

В результате анализа литературы был сделан вывод, что подход клонирования является наиболее подходящим для разработки сопоставимых вариантов сценарного типа, измеряющих комплексные латентные конструкты. Подход клонирования предполагает создание структуры задания на основе «задания-родителя» и воспроизведение этой структуры при разработке вариантов заданий. Для создания такой структуры задание необходимо разделить на элементы, которые могут быть обязательными, оказывающими влияние на трудность задания, или вариативными.

Результатом исследования стало описание подхода клонирования для сценарных заданий. Предложено выделять элементы сценарных заданий на двух уровнях. К элементам верхнего уровня относятся контекст и контент сценарного задания. В рамках исследования были предложены определения контекста и контента как отдельных комплексных элементов, которые являются неотъемлемой частью сценарного формата оценивания.

«*Контекст* сценарного задания - стимульный материал, определяющий основную проблемную ситуацию сценарного задания и развитие ситуации (сюжета) – последовательность действий, отношения между этапами задания, персонажами и пр.».

Согласно данному определению, мы можем воспринимать контекст как комплексный элемент задания, который может быть изменен полностью, либо частично в рамках процедуры клонирования.

При этом одна и та же ситуация может иметь разное тематическое наполнение, то есть отличаться *контентом*. Поэтому мы дополнительно вводим понятие контента сценарного задания - характеристика стимульного материала, которая определяет тематическое наполнение сценария. Контент также будем считать комплексным элементом сценарного задания, который может быть изменен.

К элементам нижнего уровня были отнесены элементы-стимулы, которые мотивируют респондентов на совершение действий, отражающих целевой конструкт (например, выделение релевантной информации в тексте), тематические элементы, которые изменяются под влиянием нового контента, элементы механики, отражающие взаимодействие тестируемого с тестовой средой (например, выделение, выбор, перетаскивание и прочее), и элементы структуры текста.

Разработка сопоставимых вариантов сценарных заданий предполагала разделение выделенных элементов на обязательные и вариативные. В качестве основного обязательного элемента был выбран контекст задания, в качестве вариативного, изменение которого не оказывает существенного влияния на психометрические характеристики задания, - контент (тематическое наполнение). Элементы нижнего уровня могли относиться как к обязательным, так и вариативным.

Для сценарного задания «Аквариум» контекст сценария задает проблемную ситуацию, где тестируемому необходимо обустроить пространство для питомца, используя набор объектов. Контекст сценарного задания «Аквариум» - обустройство аквариума для крабов. При разработке альтернативного варианта сценарного задания «Аквариум» использовался новый контент – обустройство террариума для гекконов. Далее клон сценарного задания «Аквариум» будет называться «Террариум».

Для сценарного задания «Динозавр» контекст сценария задает проблемную ситуацию, где тестируемого просят помочь проанализировать информацию про некоторый объект/субъект для подготовки доклада и ответить на ключевой вопрос доклада. Контекст сценарного задания «Динозавр» – подготовка доклада про несуществующего динозавра массоспондила для ответа на ключевой вопрос, на скольких лапах ходил этот динозавр. При разработке альтернативного варианта использовался новый контент, где тестируемому предлагается проанализировать информацию о том, зачем ежи трутся иголками о предметы (сценарное задание «Еж»).

В статье [Грачева, Тарасова, 2022] было предложено выделять строгое и нестрогое направления клонирования в зависимости от того, какие элементы считаются обязательными. Описание строгого и нестрогого направлений клонирования было расширено и дополнено в настоящем исследовании.

При строгом подходе клонирования структура предложений является обязательным элементом задания, который не может быть изменен. В предыдущих исследованиях грамматические основы предложений тоже выступали обязательным элементом при клонировании текстов ситуационных тестов [Lievens, Sackett, 2007].

При нестрогом подходе клонирования требования к соблюдению структур предложений не предъявляется. В предложениях текстов сценарных заданий также могут выделяться элементы-стимулы, элементы механики и тематические элементы, на которые могут ориентироваться разработчики при создании альтернативного варианта задания. Однако фокус нестрогого подхода клонирования заключается в разработке тестовой среды варианта задания, которая поможет воспроизвести поведение тестируемого без оглядки на строгое соблюдение структур предложений в текстах.

Строгое клонирование было реализовано на паре сценариев «Аквариум» / «Террариум», нестрогое клонирования на паре сценариев «Динозавр» / «Еж». в части заданий для оценки анализа информации как составляющей критического

мышления. Подробные шаблоны заданий для клонирования приведены в тексте диссертационного исследования.

В исследовании были поставлены следующие гипотезы:

- варианты сценарных заданий, созданные по строгому направлению клонирования, будут в большей степени сопоставимы между собой, чем варианты сценарных заданий, разработанные в логике нестрогого клонирования.

- варианты сценарных заданий, имеющие общий контекст, но разный контент (тематическое наполнение), будут в большей степени сопоставимы между собой, чем сценарные задания, разработанные в разном контексте.

В результате использования метода CFA-MTMM было доказано, что нет статистически значимой разницы в согласованности вариантов теста, созданных по разным направления клонирования. При этом, варианты, созданные в одном контексте, демонстрируют большую сопоставимость, чем задания сценарного типа с разным контекстом (например, «Аквариум» и «Динозавр»). То есть оба направления клонирования могут быть использованы для разработки вариантов заданий сценарного типа, измеряющих комплекстные конструкты. Это открывает разработчикам возможности к проявлению большей свободы в разработке сопоставимых вариантов заданий и позволит избежать проблемы излишней схожести клонированных тестов. В качестве ограничения этого вывода стоит отметить, что рассматриваемые части сценарных заданий для оценки навыка анализа информации выстроены линейно и в большей степени посвящены работе с электронными статьями и поиском источников в симуляции браузера. Тестовая среда существенно ограничивает то поведение, которое тестируемый может демонстрировать в похожей ситуации в реальной жизни.

Результаты исследования следует принимать с учетом ограничений. В работе подходы клонирования были опробованы на разных сценариях (по количеству индикаторов, длине, контексту и пр.), поэтому для подтверждения выводов может

быть проведено повторное исследование с использованием нескольких вариантов одного сценарного задания, которые разработаны в логике строгого и нестрогого клонирования.

Тем не менее, для вынесения суждения о сопоставимости измерений, необходимо также доказать сопоставимость структуры конструкта и психометрических характеристик отдельных индикаторов.

Второй этап работы отвечал на вопрос – **какие методологические подходы работы с данными позволят обосновать сопоставимость результатов тестирования между вариантами сценарных заданий?**

При работе с заданиями сценарного типа существует риск того, что результаты тестирования объясняются не только измеряемой латентной чертой, то и контекстным компонентом. Возникает необходимость «очистить» результаты тестирования от эффекта контекста.

В рамках данного диссертационного исследования было предложено расширить методологию проверки сопоставимости результатов тестирования за счет использования бифакторных моделей, которые позволяют воспроизвести многомерную структуру комплексного конструкта и учесть контекстные связи между поведенческими индикаторами. Бифакторные модели предполагают, что различия в ответах на задание могут объясняться общим фактором (латентным конструктом) и специфическими факторами, которые к нему не относятся (факторами контекста).

В статье [Грачева, 2022] показано, что структура измерительной модели с учетом контекстных связей между индикаторами хорошо согласуется с эмпирическими данными. Для получения достоверных результатов о сопоставимости структуры критического мышления и характеристик отдельных индикаторов, измерительная инвариантность инструмента измерения проверялась на модели с учетом контекстных связей (факторов контекста в бифакторной модели КФА).

Обоснование подхода к разработке и методологии работы с данными позволяет реализовать методологию количественного анализа для обеспечения

сопоставимости вариантов заданий сценарного типа, разработанных в соответствии с выбранным подходом к разработке. В рамках исследования была поставлена следующая гипотеза:

- варианты сценарных заданий, созданные по подходу клонирования, будут иметь эквивалентные структуры конструкта и психометрические характеристики индикаторов.

Анализ сопоставимости на уровне данных включает два этапа: проверка измерительной инвариантности инструмента (сопоставимости структур конструкта и характеристик отдельных индикаторов) и сравнение средних результатов по вариантам теста в рамках методологии латентного моделирования.

Измерительная инвариантность между вариантами сценариев проверялась на примере сценария «Аквариум», измеряющего две составляющие критического мышления: анализ информации и формулирование вывода. Индикаторы способности к формулированию вывода в этом задании снимаются в более свободной тестовой среде: на основе прочитанного материала электронной статьи тестируемый обустраивает аквариум из набора объектов (Рисунок 3). Для альтернативного варианта сценария названия объектов и интерфейс были клонированы.

Проведенный анализ измерительной инвариантности в статье [Грачева, 2022] показал, что в обоих вариантах воспроизводится теоретически ожидаемая структура критического мышления (факторы анализа и вывода выделялись отдельно), психометрические характеристики (трудность и дискриминативность) отдельных индикаторов эквивалентны. Сделан вывод, что подход клонирования позволяет получить сопоставимое психометрическое качество инструментов измерения сценарного формата.

Далее сравнивались средние результаты (на шкале факторных баллов, которые оценены моделью КФА) по вариантам. Такое сравнение результатов считается более точным при условии эквивалентности структуры конструкта и характеристик индикаторов. В среднем, результаты тестируемых по навыку анализа информации

не отличались по вариантам. Однако обнаружены статистически значимые различия в средних результатах по навыку формулирования вывода.

В исследовании использовался сбалансированный внутригрупповой дизайн сбора данных, когда оба варианта сценарного задания проходили все тестируемые, варианты предъявлялись в случайном порядке. Поэтому полученные различия могут быть связаны с различиями в контенте (тематическом наполнении сценария), а не с эффектом научения в решении подобных задач или опытом взаимодействия с компьютерным интерфейсом. Альтернативным объяснением может стать формат заданий, подразумевающий большую свободу тестируемого и элемент интерактивности. Можно сделать вывод, что задания, реализованные в более интерактивной среде, подвержены большему риску несопоставимости результатов тестирования.

Исследование имеет ограничения. Анализ сопоставимости проводился на одной паре сценарных заданий, созданных в подходе строгого клонирования, на примере критического мышления как комплексного навыка. Дизайн и методология исследования могут быть применены для анализа сопоставимости других сценариев и конструктов для подтверждения выводов исследования.

Третий этап работы отвечал на вопрос – **в какой степени контекст сценарного задания связан с различиями в результатах тестирования комплексных конструктов?** Целью третьего исследования являлась оценка эффекта контекста и контента на результаты тестирования комплексных конструктов с использованием сценарных заданий. Оба эффекта были количественно оценены в рамках методологии Теории генерализации. Результаты исследования опубликованы в статье [Грачева, 2023].

В статье используются данные, полученные осенью 2021 г. в ходе тестирования учащихся 4-х классов, которые принимали участие в исследовании универсальных навыков (критического мышления, креативности, коммуникации, кооперации) с использованием инструмента «4К». Для оценки критического мышления учащимся предлагалось выполнить три исходных сценарных задания («Аквариум», «Динозавр», «Путешествие»).

Дополнительно учащимся предлагалось выполнить альтернативные варианты исходных сценариев, следуя внутригрупповому сбалансированному дизайну. Однако из-за временных ограничений не представлялось возможным в рамках одного тестирования использовать все сценарные задания в двух вариантах. Поэтому тестируемые случайным образом были поделены на группы. Первая группа проходила варианты сценариев «Аквариум» и «Террариум» (998 респондентов, примерно 2/5 от всей выборки), вторая группа проходила варианты сценариев «Динозавр» и «Еж» (466 респондентов, примерно 1/5 от всей выборки), третья группа проходила варианты сценариев «Путешествие» и «Лабиринт» (1096 респондентов, примерно 2/5 от всей выборки). Сценарные задания «Путешествие»/ «Лабиринт» и «Аквариум»/ «Террариум» предъявлялись большему количеству респондентов, потому что являются ключевыми заданиями для оценки навыков критического мышления, коммуникации и кооперации в инструменте «4К» (содержат больше индикаторов).

Дизайн исследования позволил получить данные по всем заданиям сценарного типа для измерения критического мышления в двух вариантах.

Эффект контекста оценивался при сравнении разных сценарных заданий (с отличным контекстом), измеряющих критическое мышление. Эффект контента оценивался при сравнении сопоставимых вариантов сценарных заданий, разработанных в логике подхода клонирования. В результате анализа было установлено, что эффект изменения контента на результаты тестирования ниже, что эффект изменения контекста, что подтверждает логику подхода клонирования, где контекст отнесен к обязательным элементам сценарного задания, а контент – к вариативным.

Методы Теории генерализации позволили оценить не только общий эффект контекста и контента, но и эти эффекты при взаимодействии с тестируемым. В результате анализа было выяснено, что эффект взаимодействия контекста и тестируемого выше, чем общий эффект контекста. Аналогичные результаты получены для анализа эффекта контента. Иными словами, для одного тестируемого контекст одного сценария оказался проще, чем контекст другого сценария, и

наоборот для другого тестируемого. Можно сделать вывод, что обеспечение сопоставимости контекстов сценарных заданий на этапе разработки инструмента измерения позволит добиться минимизации общего эффекта контекста, однако контекстные задания неизбежно подвержены эффекту взаимодействия тестируемого и контекста. В предыдущих исследованиях на примере заданий типа performance-based эффект взаимодействия тестируемого и задания был преобладающим [Shavelson, Baxter, Gao, 1993]. Также исследования подчеркивают, что в случае использования контекстных заданий степень осведомленности тестируемого о контексте (или контенте) может оказывать влияние на результаты тестирования [Ahmed, Pollitt, 2007].

Для минимизации эффектов, которые нерелевантны измеряемому конструкту, тестирование комплексных навыков должно включить сценарные задания с разным контекстом. В рамках данного исследования эмпирически доказано, что тестирование критического мышления должно проводиться минимум в двух контекстах для того, чтобы достичь удовлетворительных показателей надежности. Увеличение числа контекстов позволит не только повысить надежность измерения, но и валидность выводов, которые сделаны по итогам тестирования критического мышления.

Проведенное исследование имеет ограничения. Во-первых, критическое мышление рассматривается как единый комплексный конструкт. Анализ в разрезе составляющих возможен с использованием методологии многомерной Теории генерализации [Keller, Clauser, Swanson, 2010]. Во-вторых, в данном исследовании использованы классические методы Теории генерализации на сырых данных тестирования, однако существуют расширение этого подхода в методологии моделирования структурных уравнений [Jorgensen, 2021] или байесовских сетей [Jiang, Skorupski, 2018].

Заключение

В работе выполнен анализ предыдущих исследований в области сопоставимости измерений и справедливого оценивания. Было установлено, что существующие концептуальные рамки сопоставимости измерений рассматривают сопоставимость на протяжении всего цикла оценивания [He, Vijver van de, 2012; Kolen, 1999; Kunnan, 2004]. Таким образом, решение о сопоставимости измерений должно базироваться на различных свидетельствах сопоставимости, собранных на этапах разработки и реализации инструмента измерения. На основании этого в работе представлен методологический подход к обеспечению сопоставимости вариантов сценарных заданий для измерения комплексных конструктов (подход клонирования) от этапа разработки инструмента измерения до работы с данными тестирования, связывающий воедино принципы доказательной аргументации при разработке тестов (ECD) (на этапах моделирования области конструкта, создания модели задания и измерительной модели) и принципы автоматической генерации тестовых заданий.

При формировании подхода клонирования были учтены особенности измерения комплексных конструктов с использованием сценарных заданий. Основной особенностью сценарного формата тестирования является наличие контекста, который объединяет поведенческие индикаторы конструкта, приближает задачу к реальной жизни и мотивирует респондента на совершение действий, отражающих латентный конструкт. В работе предлагается разделять контекст (основная проблемная ситуация сценария и ее развитие) и контент (тематическое наполнение) сценарного задания. Для создания сопоставимых вариантов сценарных заданий предложено изменять тематическое наполнение при сохранении контекста сценария.

На основе предложенного подхода разработаны варианты двух сценарных заданий для измерения критического мышления у обучающихся начальной школы. В качестве исходных сценарных заданий использовались задания из инструмента «4К».

В части работы с данными предложено расширить методологию проверки сопоставимости результатов тестирования за счет использования бифакторных моделей, которые позволяют воспроизвести многомерную структуру комплексного конструкта и учесть контекстные связи между поведенческими индикаторами. Эмпирически установлено, что модель оценки критического мышления на основе бифакторных моделей КФА согласуется с данными измерения комплексных конструктов в сценарном формате. Проведенный анализ данных с использованием данной модели показал, что варианты заданий, созданные на основе предложенного подхода клонирования, демонстрируют эквивалентные структуры конструкта (критического мышления), и психометрические характеристики индикаторов (трудностей, дискриминативностей).

В работе стояла задача количественно оценить связь контекста с различиями в результатах тестирования между вариантами сценарных заданий. С использованием методологии Теории генерализации отдельно оценены эффект контекста сценарного задания и эффект взаимодействия контекста и респондента на результаты тестирования критического мышления. Установлено, что эффект взаимодействия тестируемого и контекста оказался выше, чем эффект общего контекста сценарного задания, то есть один контекст может оказаться легче для одного респондента и сложнее для другого респондента. Эмпирически доказано, что тестирование комплексных конструктов в нескольких контекстах (с использованием нескольких сценарных заданий) позволяет снизить контекстные эффекты и повысить надежность измерения.

Проведенный анализ позволил сравнить эффекты контекста, полученные при сравнении разных сценарных заданий, и эффекты контента, полученные при сравнении сопоставимых вариантов заданий сценарного типа, разработанных в логике подхода клонирования. Обнаружено, что эффект контента на результаты тестирования комплексных конструктов ниже, чем эффект контекста сценария. Полученный результат подтверждает, что разработка сопоставимых вариантов сценарных заданий может происходить за счет изменения контента (тематического

содержания), сохраняя основной контекст сценария, который вносит больше различий в результаты измерений.

Теоретическая значимость работы в области измерений заключается в представлении методологического подхода к обеспечению сопоставимости сценарных заданий для оценки комплексных конструктов от разработки инструмента измерения до работы с данными тестирования. В рамках работы были предложены определения контекста и контента сценарного задания, которые могут послужить основой для дальнейшего развития методологических подходов к разработке сопоставимых вариантов сценарных заданий.

Практическая значимость работы состоит в том, что результаты исследования могут быть использованы для разработки сопоставимых вариантов заданий сценарного типа, измеряющих комплексные конструкты, в логике доказательного дизайна к разработке тестов. В контексте российского образования в качестве комплексных конструктов может рассматриваться совокупность учебных универсальных действий, которые зафиксированы во ФГОС. Варианты сценарных заданий инструмента «4К» для оценки критического мышления у учащихся начальной школы, рассмотренные в настоящей работе, могут быть признаны сопоставимыми и использоваться в будущих сравнительных и мониторинговых исследованиях критического мышления, отвечая запросу на справедливое оценивание. Получены новые эмпирические данные о том, как изменение контекста и контента сценарных заданий при оценке критического мышления связано с результатами тестирования и надежностью измерения, что может быть полезно разработчикам при проектировании оценки критического мышления. Кроме того, предложенный подход разработки позволит сократить временные и человеческие ресурсы, требуемые для создания вариантов сценарных заданий без потери их качества.

Исследование имеет ограничения. Во-первых, предложенный подход клонирования разрабатывался для инструментов измерения комплексных конструктов на примере универсальных навыков. Экстраполяция подходов на специализированные навыки и предметные знания требует дополнительной

адаптации модели клонирования и проведения эмпирического исследования сопоставимости. Сопутствующим ограничением является то, что эмпирическая часть исследования проводилась только на данных инструмента «4К» на выборке обучающихся начальной школы. В будущих исследованиях предложенные методологические подходы могут быть апробированы на других инструментах, комплексных конструктах и возрастной группе респондентов.

Во-вторых, методологические подходы для обеспечения сопоставимости разрабатывались для одного формата тестирования – сценарных заданий в цифровой среде как разновидности заданий формата *performance-based.* В отличие от традиционных тестов с набором независимых заданий, тесты в *performance-based* формате не имеют четкой структуры. В данном исследовании мы опираемся на работу [Andrews-Todd и др., 2021], которые ввели термин *virtual-performance-based assessment* и рассматривают три вида VPBA: симуляции, сценарные задания и задания на основе игры (game-based assessment). Однако в реальной практике со стороны исследователей не прослеживается единого понимания видов VPBA, а под термином *performance-based assessment* могут встречаться совершенно разные тестовые форматы. Данная «терминологическая путаница» является ограничением к применению выводов исследования. Кроме того, с развитием технологий сценарные задания могут обладать разной степенью интерактивности и сложности симуляций. Результаты данного исследования показали, что задачи, которые подразумевают большую свободу действий (сбор объекта из элементов) подвержены большему риску несопоставимости результатов. Обеспечение сопоставимости более свободных тестовых сред может требовать новых методологических подходов и стать продолжением данного исследования.

В-третьих, реализация подхода клонирования для разработки сопоставимых вариантов сценарных заданий требует погруженности исследователя или разработчика в теоретическую модель конструкта. Несмотря на то, что подход клонирования основан на принципах автоматической генерации заданий и призван упростить процесс разработки, роль эксперта в процессе разработки по-прежнему является ключевой. Перспективным направлением будущего исследования

является совмещение технологий машинного обучения и заранее разработанных когнитивных моделей и шаблонов заданий для клонирования. Уже появляются исследования, где генеративные модели искусственного интеллекта обучались разрабатывать задания для оценки понимания прочитанного на основе когнитивных моделей заданий [Sayin, Gierl, 2024].

В качестве направлений будущих исследований стоит обратить внимание на изучение эффекта контекста и контента на результаты тестирования комплексных конструктов в образовательном и психологическом тестировании. В данном исследовании удалось установить наличие эффекта контекста на общем уровне и при взаимодействии с респондентом, однако причины возникновения этих эффектов требуют дополнительного изучения. Например, эффект контекста может быть дополнительно исследован в рамках концепции трансфера знаний из одного контекста в другой [Barnett, Ceci, 2002]. Отдельное исследование может быть посвящено анализу связи характеристик контекста (например, нагруженность, абстрактность, приближенность к реальным задачам) с результатами тестирования комплексных конструктов.

В заключение отметим, что собранные свидетельства сопоставимости преимущественно опирались на количественные методы исследования. Создание руководств по проведению когнитивных лабораторий с респондентами для объяснения эффекта контекста и возможных различий в результатах тестирования между вариантами станет важным дополнением к данному исследованию.

Положения, выносимые на защиту

1. Предложенный в работе подход к разработке сопоставимых вариантов сценарных заданий, связывающий воедино этапы моделирования области конструкта и разработки модели задания в парадигме доказательного дизайна, позволил обеспечить эквивалентность структуры комплексного конструкта и психометрических характеристик поведенческих индикаторов между вариантами.

2. Методология работы с данными для обеспечения сопоставимости вариантов сценарных заданий была расширена за счет включения методов, базирующихся на бифакторных моделях, которые позволяют учесть многомерную природу комплексного конструкта и связи между поведенческими индикаторами, обусловленные наличием контекстного компонента.

3. Предложенная методология позволила выявить эффект контекста сценарного задания на результаты тестирования и установить, что эффект взаимодействия контекста и респондента на результаты выше, чем эффект общего контекста сценарного задания.

4. Для снижения эффекта контекста на результаты тестирования и повышения надежности измерений при оценке комплексных конструктов необходимо использовать несколько сценарных заданий с разными контекстами.

Апробация и внедрение результатов исследования

Список публикаций автора диссертации, в которых отражены основные научные результаты исследования:

- Грачева Д.А. Роль контекста в заданиях сценарного типа при измерении универсальных навыков: применение теории генерализации / Д.А. Грачева // Вопросы образования. – 2023. – № 3. – С. 221–230;

– Грачева Д. А. Анализ сопоставимости измерения метапредметных навыков в цифровой среде // Психологическая наука и образование. – 2022. – № 6 (27). – С. 57–67;
– Грачева Д. А. Подходы к разработке вариантов заданий сценарного типа в рамках метода доказательной аргументации / Д.А. Грачева, К.В. Тарасова // Отечественная и зарубежная педагогика. – 2022. – № 3 (1). – С. 83–97.

Дополнительные публикации с участием автора по теме:

– Uglanova I. Computer-based performance approach for critical thinking assessment in children / I. Uglanova, E. Orel, D. Gracheva, K. Tarasova // British Journal of Educational Psychology. – 2023. – №93. – P. 531–544.

Список научных конференций, на которых были представлены результаты исследования:

– Конференция «Quantitative Research Methods Conference (QRM)». Доклад: Testing measurement invariance across alternative test forms? 14-15 июня 2021 г., онлайн конференция.
– Конференция «13th Annual International Conference on Education and New Learning Technologies (EDULEARN21)». Доклад: Investigating the effect of context on comparability of computerized performance-based tasks, 5-6 июля 2021 г., онлайн конференция.
– Конференция «22nd Annual Meeting of the Association for Educational Assessment – Europe (AEA-Europe). Assessment for Changing Times: Opportunities and Challenges». Доклад: Comparability of computerized performance-based assessment for measuring critical thinking, 9-12 ноября 2021 г., Дублин, Ирландия (онлайн выступление).
– Конференция «24nd Annual Meeting of the Association for Educational Assessment – Europe (AEA-Europe 2023). Assessment reform journeys: intentions, enactment and evaluation» Доклад: The application of generalizability

theory to the scenario-based performance assessment of 21st century skills: analysis of task context effect. 6-9 ноября 2023 г., Мальта.

Список литературы

1. Авдеева С. М. Об оценке цифровой грамотности: методология, концептуальная модель и инструмент измерения / С.М. Авдеева, К.В. Тарасова // Вопросы образования. – 2023. – № 2. – С. 8–32.
2. Грачева Д.А. Роль контекста в заданиях сценарного типа при измерении универсальных навыков: применение теории генерализации / Д.А. Грачева // Вопросы образования. – 2023. – № 3. – С. 221–230.
3. Грачева Д. А. Анализ сопоставимости измерения метапредметных навыков в цифровой среде // Психологическая наука и образование. – 2022. – № 6 (27). – С. 57–67.
4. Грачева Д. А. Подходы к разработке вариантов заданий сценарного типа в рамках метода доказательной аргументации / Д.А. Грачева, К.В. Тарасова // Отечественная и зарубежная педагогика. – 2022. – № 3 (1). – С. 83–97.
5. Ковалева Г. С. Функциональность проекта «Мониторинг формирования функциональной грамотности обучающихся» / Г. С. Ковалева, Н. И. Колачев // Отечественная и зарубежная педагогика. – 2023. – Т. 2. № 1 (90). – С. 9–32.
6. Поддьяков А.Н. Решение комплексных проблем в PISA-2012 и PISA-2015: взаимодействие со сложной реальностью / А.Н. Поддьяков // Образовательная политика. – 2012. – № 6 (62). – С. 34–53.
7. Угланова И. Л. Методология Evidence-Centered Design для измерения комплексных психологических конструктов / И.Л. Угланова, И.В. Брун, Г.М. Васин // Современная зарубежная психология. – 2018. – № 3 (7). – С. 18–27.
8. Шкерина Л. В. Метапредметная олимпиада для школьников: новый подход к оцениванию метапредметных универсальных учебных действий обучающихся / Л. В. Шкерина, О. В. Берсенева, Н. А. Журавлева, М.А. Кейв // Перспективы науки и образования. – 2019. –№ 2 (38). – С. 194–211.
9. Ahmed A. Improving the quality of contextualized questions: an experimental investigation of focus / A. Ahmed, A. Pollitt // Assessment in Education: Principles, Policy & Practice. – 2007. – № 2 (14). – P. 201–232.

10. Ananiadoui K., M. Claro. 21st Century Skills and Competences for New Millennium Learners in OECD Countries [Электронный ресурс]. Режим доступа: https://doi.org/10.1787/19939019 (дата обращения: 05.07.2024).
11. Andrews-Todd J. Virtual Performance-Based Assessments Methodology of Educational Measurement and Assessment // Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in R and Python / под ред. A. A. Von Davier, R. J. Mislevy, J. Hao. Springer. – 2021. – P. 45–60.
12. Andrews-Todd J. Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task / J. Andrews-Todd, C. M. Forsyth // Computers in human behavior. 2020. – № 104. – P. 105759.
13. Attali Y. The interactive reading task: Transformer-based automatic item generation / Y. Attali [et al.] // Frontiers in Artificial Intelligence. 2022. – №5. – P. 903077.
14. Barnett S. M. When and where do we apply what we learn?: A taxonomy for far transfer / S.M. Barnett, S.J. Ceci // Psychological bulletin. – 2002. – № 4 (128). – P. 612–637.
15. Bejar I. I. A generative approach to psychological and educational measurement // ETS Research Report Series. – 1991. – № 1. – P 1–54.
16. Bond L., Moss P., Carr P. Fairness in large-scale performance assessment // Technical issues in large-scale performance assessment. – 1996. – P. 117–140.
17. Braun H. I. Performance assessment of critical thinking: Conceptualization, design, and implementation / H.I. Braun [et al.] // Frontiers in Education. – 2020. – №5. URL: https://www.frontiersin.org/articles/10.3389/feduc.2020.00156/full
18. Brennan R. L. Generalizability theory / R.L. Brennan // Educational Measurement: Issues and Practice. 1992. – № 4 (11). – P. 27–34.
19. Brown T. A. Confirmatory factor analysis for applied research. / T.A. Brown. – New York: The Guilford Press, 2006.
20. Campbell D. T., Convergent and discriminant validation by the multitrait-multimethod matrix. / D.T. Campbell, D.W. Fiske // Psychological bulletin. – 1959. – № 2 (56). – P. 81–105.

21. Care E. Education System Alignment for 21st Century Skills: Focus on Assessment / E. Care, H. Kim, A. Vista, K. Anderson // Center for Universal Education at The Brookings Institution. – 2018.
22. Carneiro J. Proposal of a validation framework for a new measurement model and its application to the export performance construct / J. Carneiro, A. Rocha, J. F. Silva // BAR-Brazilian Administration Review. – 2009. – № 6. – P. 331–353.
23. Cho Y. Investigating the effects of prompt characteristics on the comparability of TOEFL iBTTM integrated writing tasks / Y. Cho, F. Rijmen, J. Novák // Language Testing. – 2013. – № 4 (30). – P. 513–534.
24. Clause C. S. Parallel test form development: A procedure for alternate predictors and an example / C.S. Clause // Personnel Psychology. – 1998. – № 1 (51). – P. 193–208.
25. Cronbach L. J. The Dependability of Behavioral Measurements. / L.G. Cronbach [et al]. Hoboken, NJ: Wiley, 1972.
26. Eid M. Measuring Method Effects: From Traditional to Design-Oriented Approaches / M. Eid, C. Geiser, T. Koch // Current Directions in Psychological Science. – 2016. – № 4 (25). – P. 275–280.
27. Ercikan, K. In search of validity evidence in support of the interpretation and use of assessments of complex constructs: Discussion of research on assessing 21st century skills. / K. Ercikan, M.E. Oliveri // Applied Measurement in Education. – 2016. – № 29(4). – P. 310-318.
28. Fay D. M. Investigating Psychometric Isomorphism for Traditional and Performance-Based Assessment / D.M. Fay, R. Levy, V. Mehta // Journal of Educational Measurement. – 2018. – № 1 (55). – P. 52–77.
29. Geisinger K. F. 21st Century Skills: What Are They and How Do We Assess Them? / K. F. Geisinger // Applied Measurement in Education. – 2016. – Vol. 29. № 4. – P. 245–249.
30. Gierl M. J., Haladyna T. M. Using weak and strong theory to create item models for automatic item generation: Some practical guidelines with examples // Automatic Item Generation / под ред. M. J Gierl, T. M. Haladyna. Routledge, 2012. P. 36–49.

31. Gierl M. J. Using Automatic Item Generation to Create Solutions and Rationales for Computerized Formative Testing / M.J. Gierl, H. Lai // Applied Psychological Measurement. – 2018. – № 1 (42). – P. 42–57.
32. Gipps C., Stobart G. Fairness in Assessment // Educational assessment in the 21st century / под ред. C. Wyatt-Smith, J. J. Cumming, Dordrecht: Springer Netherlands, 2009. P. 105–118.
33. Gorin J. S. Inherent measurement challenges in the next generation science standards for both formative and summative assessment / J. S. Gorin, R. J. Mislevy // Invitational research symposium on science assessment. – Citeseer, 2013.
34. Griffin P., Care E. Assessment and teaching of 21st century skills: Methods and approach / P. Griffin, E. Care, Springer, 2014.
35. Haladyna T. M. Item Shells: A Method for Writing Effective Multiple-Choice Test Items / T.M. Haladyna, R.R. Shindoll // Evaluation & the Health Professions. – 1989. – № 1 (12). – P. 97–106.
36. He J., Vijver F. van de Bias and equivalence in cross-cultural research / J. He, F. van de Vijver // Online readings in psychology and culture. – 2012. – № 2 (2). – P. 2307–0919.
37. Hyytinen H. How do self-regulation and effort in test-taking contribute to undergraduate students' critical thinking performance? / H. Hyytinen [et al.] // Studies in Higher Education. – 2024. – Vol. 49. № 1. – P. 192–205.
38. Irvine S. H., Dann P. L., Anderson J. D. Towards a theory of algorithm-determined cognitive test construction / S.H. Irvine, P.L. Dann, J.D. Anderson // British Journal of Psychology. – 1990. – № 2 (81). – P. 173–195.
39. Irvine S. H., Kyllonen P. C. Item generation for test development, Routledge, 2013.
40. Jiang Z. Bayesian approach to estimating variance components within a multivariate generalizability theory framework / Z. Jiang, W.A. Skorupski // Behavior Research Methods. – 2018. – № 6 (50). – P. 2193–2214.
41. Jorgensen T. D. How to estimate absolute-error components in structural equation models of generalizability theory / T.D. Jorgensen // Psych. – 2021. – № 2 (3). – P. 113–133.

42. Keller L. A. Using multivariate generalizability theory to assess the effect of content stratification on the reliability of a performance assessment / L.A. Keller, B.E. Clauser, D.B. Swanson // Advances in health sciences education. – 2010. – №15. – P. 717–733.
43. Kim E.S. Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT / E.S. Kim, M. Yoon // Structural Equation Modeling: A Multidisciplinary Journal. 2011. – № 2 (18). – P. 212–228.
44. Kolen M. J. Threats to score comparability with applications to performance assessments and computerized adaptive tests / M.J. Kolen // Educational Assessment. – 1999. – № 2 (6). – P. 73–96.
45. Kroehne U. Construct Equivalence of PISA Reading Comprehension Measured With Paper-Based and Computer-Based Assessments / U. Kroehne [et al.] // Educational Measurement: Issues and Practice. – 2019. – № 3 (38). – P. 97–111.
46. Kunnan A. J. Test fairness / A.J. Kunnan // European language testing in a global context. – 2004. – №18. – P. 27–48.
47. Lee H.-K., Anderson C. Validity and topic generality of a writing performance test / H.-K. Lee, C. Anderson // Language testing. – 2007. – № 3 (24). – P. 307–330.
48. Levy R. Psychometric and Evidentiary Advances, Opportunities, and Challenges for Simulation-Based Assessment / R. Levy // Educational Assessment. – 2013. – № 3 (18). – P. 182–207.
49. Lievens F. Creating Alternate In-Basket Forms Through Cloning: Some preliminary results / F. Lievens, F. Anseel // International Journal of Selection and Assessment. – 2007. – № 4 (15). – P. 428–433.
50. Lievens F. Situational judgment tests in high-stakes settings: Issues and strategies with generating alternate forms / F. Lievens, P. R. Sackett // Journal of Applied Psychology. – 2007. – № 4 (92). – P. 1043–1055.
51. Messick S. The Interplay of Evidence and Consequences in the Validation of Performance Assessments / S. Messick // Educational Researcher. – 1994. – № 2 (23). V P. 13–23.

52. Mislevy R. J. Implications of Evidence-Centered Design for Educational Testing / R. J. Mislevy, G. D. Haertel // Educational Measurement: Issues and Practice. – 2006. – № 4 (25). – P. 6–20.
53. Nelson J. “I’d like to be under the sea”: Contextual cues in virtual environments influence the orientation of idea generation / J. Nelson, J. Guegan // Computers in Human Behavior. – 2019. – № 90. – P. 93–102.
54. Osburn H. G. Item Sampling for Achievement Testing / H.G. Osburn // Educational and Psychological Measurement. – 1968. – № 1 (28). – P. 95–104.
55. Pellegrino, J. W. Teaching, learning and assessing 21st century skills [Электронный ресурс].–2017. – P. 223 – 251. Режим доступа: https://doi.org/10.1787/9789264270695-12-en
56. Pitta-Pantazi D. Higher order thinking in mathematics: a complex construct / D. Pitta-Pantazi, P. Sophocleous // The 10th Mathematical Creativity and Giftedness. – Nicosia, Cyprus, 2017.
57. Reckase M. D. 18 Multidimensional Item Response Theory Psychometrics // Handbook of statistics / под ред. C. R. Rao, S. Sinharay, Elsevier, 2006. – № 26. – P. 607–642.
58. Ridley C. R. Multicultural Counseling Competence: A Construct in Search of Operationalization / C. R/ Ridley [et al.] // The Counseling Psychologist. 2021. – Vol. 49. – № 4. – P. 504–533.
59. Rijmen F. Formal Relations and an Empirical Comparison among the Bi-Factor, the Testlet, and a Second-Order Multidimensional IRT Model / F. Rijmen // Journal of Educational Measurement. 2010. – № 3 (47). – P. 361–372.
60. Rojas M. Assessing collaborative problem-solving skills among elementary school students / M. Rojas [et al.] // Computers & Education. – 2021. – №175. – P. 104313.
61. Ruiz-Primo M. A. The Relationship between Item Context Characteristics and Student Performance: The Case of the 2006 and 2009 PISA Science Items / M.A. Ruiz-Primo, M. Li // Teachers College Record. 2015. – № 1 (117) – P. 1–36

62. Sälzer C. Assessing global competence in PISA 2018: Challenges and approaches to capturing a complex construct / C. Sälzer, N. Roczen // International journal of development education and global learning. – 2018. – Vol. 10. – № 1. P 5–20.
63. Sayin A. Using OpenAI GPT to Generate Reading Comprehension Items / A. Sayin, M. Gierl //Educational Measurement: Issues and Practice. – 2024. – Vol. 43. – №. 1. – P. 5-18.
64. Schliemann A. D. Proportional reasoning: From shopping, to kitchens, laboratories, and, hopefully, schools / A.D. Schliemann, V.P. Magalhães // Oaxtepec Mexico, 1990. – P. 67–73.
65. Shavelson R. J. Sampling variability of performance assessments / R.J. Shavelson, G.P. Baxter, X. Gao // Journal of educational Measurement. – 1993. – № 3 (30). – P. 215–232.
66. Shavelson R. J., Webb N. M., Rowley G. L. Generalizability theory // Methodological issues & strategies in clinical research / под ред. E. Kazdin, American Psychological Association, 1992.
67. Stadler M. The assessment of collaborative problem solving in PISA 2015: An investigation of the validity of the PISA 2015 CPS tasks / M. Stadler [et al.] // Computers & Education. – 2020. – №157. – P. 103964.
68. Uglanova I. Computer-based performance approach for critical thinking assessment in children / I. Uglanova [et al.] // British Journal of Educational Psychology. – 2023. – №93. – P. 531–544.
69. Wang D. Automated and interactive game-based assessment of critical thinking / D. Wang, H. Liu, K.-T. Hau // Education and Information Technologies. – 2022. – № 4 (27). – P. 4553–4575.
70. Watson J. Statistical literacy: A complex hierarchical construct / J. Watson, R. Callingham // Statistics Education Research Journal. – 2003. – Vol. 2. – № 2. – P. 3–46.
71. Williamson D. M. Hierarchical IRT examination of isomorphic equivalence of complex constructed response tasks. / D.M. Williamson [et al.] // April Paper

presented at the annual meeting of the American Educational Research Association New Orleans, LA, 2002.