Роль контекста в заданиях сценарного типа при измерении универсальных навыков: применение теории генерализации

Дарья Грачева

Статья поступила в редакцию в марте 2023 г.

Грачева Дарья Александровна — младший научный сотрудник Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». Адрес: 101000, Москва, ул. Мясницкая, 20. E-mail: dgracheva@hse.ru. ORCID: https://orcid.org/0000-0002-4646-7349

Аннотация

В современных условиях большое внимание уделяется развитию и оцениванию универсальных навыков у школьников. Для такого оценивания необходимы новые тестовые форматы, основанные на наблюдаемых действиях учащегося в цифровой среде. Один из перспективных вариантов таких инструментов — контекстные задания сценарного типа. Однако контекстное разнообразие таких заданий может затруднять сравнение результатов. В статье анализируется роль контекста сценарных заданий при измерении двух универсальных навыков: критического мышления и коммуникации. С этой целью применяются методы теории генерализации, которые позволяют установить, в какой степени согласованными являются результаты, полученные с помощью разных контекстов сценарных заданий, и как путем изменения количества индикаторов или контекстов сценариев обеспечить достаточную надежность измерения. Исследование основано на данных, которые получены при тестировании учащихся 4-х классов с помощью разных заданий сценарного типа, входящих в состав инструмента «4К». Результаты анализа показали, что поведение тестируемых в сценариях с разным контекстом различается, при этом трудности контекстов практически одинаковы. Для достижения удовлетворительной надежности рекомендуется использовать минимум два сценария с разными контекстами, а использование трех и более сценарных заданий с разными контекстами позволяет существенно сократить количество индикаторов без потери надежности. В исследовании также оценивалась роль контекста при использовании альтернативных вариантов заданий. Альтернативные варианты схожи в основной проблеме и сюжете сценария, но различаются тематическим наполнением (контентом). Изменение только контента сценария позволяет экстраполировать результаты оценивания универсальных навыков на все варианты заданий, т.е. альтернативные варианты могут использоваться как взаимозаменяемые. Проведенное исследование демонстрирует возможности использования методов теории генерализации для оптимизации разработки заданий с учетом требований к надежности измерения.

Ключевые слова теория генерализации, универсальные навыки, задания сценарного типа, контекст задания, психометрика, надежность измерений

Для цитирования Грачева Д.А. (2023) Роль контекста в заданиях сценарного типа для измерения универсальных навыков: применение теории генерализации. *Вопросы образования / Educational Studies Moscow*, № 3, сс. 62–91. https://doi.org/10.17323/vo-2023-16901

The Role of Context in Scenario-Based Tasks for Measuring Universal Skills: The Use of Generalizability Theory

Daria Gracheva

Daria A. Gracheva — Junior Research Fellow at the Center for Psychometrics and Measurements in Education, Institute of Education, National Research University Higher School of Economics. Address: 20 Myasnitskaya St, 101000 Moscow, Russian Federation. E-mail: dgracheva@hse.ru. ORCID: https://orcid.org/0000-0002-4646-7349

Abstract In education, much attention is paid to the development and evaluation of universal skills in schoolchildren. At the same time, the assessment of universal skills requires new test formats based on the observed actions of the student in the digital environment. Scenario-based contextual tasks serve as a promising format. However, the contextual diversity of such tasks can make it difficult to compare results obtained from different scenario tasks. This article aims to analyze the role of scenario task context in measuring two universal skills: critical thinking and communication. The work uses the methods of Generalizability Theory, which allows to analyze to what extent the results can be generalized for other contexts of scenario tasks, and how, by changing the number of indicators or scenario contexts, to ensure satisfied measurement reliability. The study is based on data from fourth-grade students who were tested with various scenario-based tasks of the "4K" instrument. The results of the analysis showed that the behavior of the test-takers differs in scenarios with different contexts, while the difficulties of the contexts are almost the same. To achieve satisfactory reliability, it is recommended to use at least two scenarios with different contexts, and the use of three or more scenarios with different contexts will reduce the number of indicators without loss of reliability. Also, the study evaluated the role of context when using alternative scenario-based tasks forms were used. The alternative forms were similar in the main problem and plot of the scenario, but differed in topic (content). Changing only the content of the scenario makes it possible to generalize the results across scenario forms, that is, alternative forms can be used interchangeably. This study demonstrates how Generalization Theory can be used to optimize the development of tasks, taking into account the requirements for measurement reliability.

Keywords generalizability theory, universal skills, scenario-based tasks, task context, psychometrics, reliability of measurement

For citing Gracheva D.A. (2023) Rol´ konteksta v zadaniyakh stsenarnogo tipa pri izmerenii universal´nykh navykov: primenenie teorii generalizatsii [The Role of Context in Scenario-Based Tasks for Measuring Universal Skills: The Use of Generalizability Theory]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3, pp. 62–91. https://doi.org/10.17323/vo-2023-16901

Современное образование нацелено на развитие и оценку универсальных навыков у школьников. Согласно ФГОС, умения применять средства коммуникации и критически анализировать информацию относятся к метапредметным навыкам, которые школьники осваивают в процессе обучения. Каждый универсальный навык представляет собой не однородную структуру, а несколько связанных между собой составляющих, поэтому их называют сложными конструктами.

Оценивание таких сложных конструктов предполагает выход за пределы традиционных типов заданий, таких как задания с выбором варианта ответа, самоотчетные опросники со шкалами Ликерта. Наиболее подходящим форматом для оценки универсальных навыков являются задания в формате *performance-based*, где оценивание происходит через индикаторы наблюдаемого поведения в тестовой среде, в том числе в цифровой [Davier von, Mislevy, Hao, 2022].

К категории *performance-based* относят задания сценарного типа, в них отдельные индикаторы объединены общим контекстом (*scenario-based tasks*), поэтому такие задания принято называть контекстными [Ruiz-Primo, Li, 2015]. Контекст определяет основную проблемную ситуацию сценарного задания, ее контент (тематическое наполнение) и сюжет (последовательность этапов задания и действий персонажей).

Разработав удачный контекст задания, можно повысить мотивацию тестируемого и вовлечь его в прохождение теста. При этом контекстные задания позволяют увидеть, как респонденты применяют навыки в различных ситуациях, в том числе приближенных к реальной жизни, что особенно важно для измерения универсальных навыков. Например, при измерении коммуникации можно создать ситуацию продолжительного диалога, которая позволит оценить всю динамику коммуникативного процесса. В то же время различия в контекстах затрудняют сравнение результатов, полученных с применением разных сценарных заданий, в том числе альтернативных вариантов сценариев, которые часто используются при повторных тестированиях респондентов [Davier von, Mislevy, Hao, 2022]. Иными словами, в различия в тестовом балле вносит вклад не только оцениваемая характеристика, но и ситуация, и тема сценария.

В данной статье мы анализируем роль контекста сценарных заданий при оценке двух универсальных компетенций: критического мышления и коммуникации. Эмпирической основой исследования послужили данные по инструменту «4К», который состоит из заданий сценарного типа, реализованных в цифровой среде с автоматическим скорингом (без привлечения экспертов) для измерения универсальных навыков у млад-

ших школьников[1]. В работе используются методы теории генерализации (*generalizability theory*) [Cronbach et al., 1972], которые позволяют количественно оценить вклад контекста заданий в результаты. В рамках анализа генерализации возможно проверить предположения о том, в какой степени полученные результаты могут быть экстраполированы на другие контексты сценарных заданий и как можно изменить процедуру тестирования для повышения надежности измерения универсальных навыков. Результаты такого анализа можно применять на практике для проектирования контекстных инструментов оценивания с опорой на эмпирические доказательства их психометрического качества, а также для проверки справедливого и сопоставимого оценивания навыков в разных контекстах.

В статье поставлены следующие исследовательские вопросы.

1. Каков вклад контекста сценарного задания в результаты оценивания универсальных компетенций?

2. Какое количество контекстов сценарных заданий необходимо для надежного измерения универсальных навыков?

Статья построена следующим образом: сначала рассмотрены основные положения теории генерализации, затем представлен обзор исследований, в которых применяются методы теории генерализации. Далее описано эмпирическое исследование с применением методов теории генерализации для оценки надежности результатов измерения универсальных навыков с использованием заданий сценарного типа и вклада контекста сценария в тестовый балл.

1. Основные положения теории генерализации

Основы теории генерализации впервые были изложены в работах Л. Кронбаха и его коллег как расширение представлений о концепции надежности в рамках классической теории тестирования [Cronbach et al., 1972]. Позже идеи теории генерализации подробно раскрывались в работах Р. Шавелсона и Р. Бреннона [Shavelson, Webb, Rowley, 1992; Brennan, 1992]. Согласно определению, теория генерализации — это статистическая теория надежности (*dependability*) инструментов измерения [Shavelson, Webb, Rowley, 1992]. Надежность здесь понимается как точность экстраполяции результатов выборочной процедуры измерений на всю генеральную совокупность измерений.

[1] Инструмент «4К» для оценки универсальных навыков (критическое мышление, креативность, коммуникация и кооперация) разработан сотрудниками Центра психометрики и измерений в образовании Института образования НИУ ВШЭ в рамках договора о научно-исследовательской работе с благотворительным фондом «Вклад в будущее». Сайт инструмента доступен по ссылке: https://ioe.hse.ru/4k_test/

В теории генерализации процедуру измерения раскладывают на компоненты, каждый из которых может быть источником наблюдаемых различий в баллах. Например, различия в баллах могут объясняться способностью респондентов (объектом измерения), трудностью заданий в тесте, временем тестирования, степенью строгости экспертов. Для заданий сценарного типа таким компонентом также может быть контекст сценария. Любой компонент процедуры измерения, отличный от объекта измерения, принято называть фасетом. Каждый фасет является источником различий в баллах, которые относятся к ошибке измерения.

В соответствии с целью статьи предположим, что есть несколько сценарных заданий с разными контекстами, при этом все респонденты проходят все сценарии. Такая процедура тестирования включает следующие компоненты: респонденты (объект исследования, *p*), индикаторы сценарного задания (фасет *i*) и контекст сценария (фасет *c*). В рамках теории генерализации предполагается, что элементы фасета конкретной процедуры исследования выбраны случайно из универсума, или полного множества объектов генерализации (*universe of admissible observations*). Аналогично объекты исследования являются случайной выборкой из популяции.

Пусть любой респондент из популяции выполняет любой индикатор сценария из полного множества возможных индикаторов в любом контексте из полного множества возможных контекстов. Тогда балл респондента *p* по индикатору *i* контекста *c* можно представить в виде:

$$X_{pio} = U + v_p + v_i + v_c + v_{pi} + v_{pc} + v_{ic} + v_{pic,e}, \quad (1)$$

где U — генеральное среднее в популяции; v — независимые эффекты (компоненты), а именно: v_p — эффект респондента, v_i — эффект индикатора, v_c — эффект контекста, v_{pi} — эффект взаимодействия респондента и индикатора, v_{pc} — эффект взаимодействия респондента и контекста, v_{ic} — эффект взаимодействия индикатора и контекста, $v_{pic,e}$ — остаточный компонент, включающий эффект взаимодействия всех фасетов и компонент ошибки, т.е. случайной изменчивости и систематической изменчивости, которая не объясняется фасетами конкретной процедуры измерения.

Тогда дисперсия баллов из выражения (1) по всем респондентам из популяции и элементам фасета из полного множества объектов генерализации имеет вид:

$$\sigma^2(X_{pic}) = \sigma^2(p) + \sigma^2(i) + \sigma^2(c) + \sigma^2(pi) + \\ + \sigma^2(pc) + \sigma^2(ic) + \sigma^2(pic,e). \quad (2)$$

Таким образом, дисперсия баллов может быть разложена на независимые компоненты дисперсии, связанные с различием в истинной способности респондентов $\sigma^2(p)$ и с разными источниками ошибки измерения — различиями в трудности индикаторов $\sigma^2(i)$, контекстов $\sigma^2(c)$, эффектами взаимодействия и остаточной дисперсией $\sigma^2(pic,e)$. Это компоненты дисперсии со случайным эффектом. В теории генерализации возможно оценить компоненты дисперсии с фиксированным эффектом — в этом случае элементы конкретного фасета составляют ограниченный набор и не экстраполируются на полное множество объектов генерализации. Дизайн с фиксированными эффектами не рассматривается в данной статье.

Получить количественную оценку каждого компонента дисперсии можно в рамках дисперсионного анализа (ANOVA). Расчет независимых компонентов дисперсии по фасетам исследования становится результатом первого этапа анализа в рамках теории генерализации — G-исследования (*generalizability study*).

В рамках G-исследования определяются фасеты, которые могут оказывать влияние на результаты оценивания, а также отношения между фасетами (или дизайн G-исследования). Выражение (1) иллюстрирует перекрестный дизайн с двумя фасетами (индикаторы и контексты), который принято обозначать как *p* x *i* x *c*. При таком дизайне каждый элемент одного фасета встречается в комбинации с каждым элементом другого, т.е. каждый респондент выполняет каждый индикатор сценария в каждом контексте.

Вложенные дизайны позволяют моделировать вложенные отношения между фасетами. Фасет называется вложенным в другой фасет, если разные элементы первого фасета встречаются в комбинации с каждым элементом второго фасета. Например, дизайн *p* x (*i* : *c*) предполагает, что каждый респондент выполняет каждый индикатор, вложенный в контекст сценария (*i*, размещенное внутри *c*).

Использование перекрестного G-дизайна предпочтительно, потому что вложенный G-дизайн не позволяет оценить все комбинации фасетов между собой (например, в дизайне *p* x (*i* : *c*) не оценивается эффект взаимодействия контекстов и индикаторов). Однако перекрестные G-дизайны не всегда выполнимы на практике.

Таким образом, цель G-исследования — определить, какие фасеты важны для измерения. Чтобы количественно определить вклад фасетов, результаты G-исследования представляют в виде процентов от общей дисперсии баллов.

На основе результатов G-исследования реализуется второй этап анализа в рамках теории генерализации — D-исследование (*decision study*). Цель D-исследования — определить процедуру тестирования, которая повысит надежность за счет изменения

количества элементов фасета или связей между фасетами. Например, в рамках D-исследования возможно ответить на вопрос, как изменение количества индикаторов или контекстов сценариев скажется на надежности результатов.

Концепция надежности в рамках теории генерализации требует дополнительного пояснения, которое приведено в следующем подразделе.

1.1. Надежность в теории генерализации

В теории генерализации существуют два коэффициента для оценки надежности: коэффициент генерализации (*generalizability coefficient*, E_{p^2}) и коэффициент зависимости (*dependability coefficient*, φ).

Общая формула для коэффициентов надежности имеет следующий вид:

$$C = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(e)}, \quad (3)$$

где C — коэффициент генерализации или зависимости; $\sigma^2(p)$ — дисперсия различий в результатах, связанная с различиями в истинной способности респондентов; $\sigma^2(e)$ — дисперсия ошибки измерения.

Различия между коэффициентами надежности заключаются в определении дисперсии ошибки измерения, которая зависит от дизайна исследования и количества элементов фасета. Для перекрестного дизайна с двумя фасетами из выражения (1) дисперсия ошибки для E_{p^2} (3.1) и φ (3.2) имеет вид:

$$\sigma^2(e) = \frac{\sigma^2(pi)}{n_i} + \frac{\sigma^2(pc)}{n_c} + \frac{\sigma^2(pic,e)}{n_i n_c}; \quad (3.1)$$

$$\sigma^2(e) = \frac{\sigma^2(i)}{n_i} + \frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(ic)}{n_i n_c} + \frac{\sigma^2(pi)}{n_i} + \frac{\sigma^2(pc)}{n_c} + \frac{\sigma^2(pic,e)}{n_i n_c}. \quad (3.2)$$

Для дизайна с вложенными фасетами p x (i : c) дисперсия ошибки для E_{p^2} (3.3) и φ (3.4) имеет вид:

$$\sigma^2(e) = \frac{\sigma^2(pc)}{n_c} + \frac{\sigma^2(pi, pic,e)}{n_i n_c}; \quad (3.3)$$

$$\sigma^2(e) = \frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(i,\ ic)}{n_i n_c} + \frac{\sigma^2(pc)}{n_c} + \frac{\sigma^2(pi, pic,e)}{n_i n_c}, \quad (3.4)$$

где n_i — количество индикаторов; n_c — количество контекстов.

Таким образом, дисперсия ошибки в коэффициенте генерализации учитывает только компоненты, содержащие эффект респондента, поэтому ее называют дисперсией относительной ошибки измерения (*relative error variance*). Дисперсия ошиб-

ки в коэффициенте зависимости дополнительно включает основные эффекты фасетов и взаимодействия между фасетами и называется абсолютной дисперсией ошибки (*absolute error variance*). Коэффициент зависимости всегда будет меньше, чем коэффициент генерализации. Формулы для коэффициентов генерализации и надежности для иных дизайнов исследования можно найти в [Shavelson, Webb, Rowley, 1992].

Выбор коэффициента надежности зависит от типа тестирования. В нормоориентированном тестировании, когда исследователи заинтересованы в упорядочивании респондентов относительно друг друга, рассчитывается коэффициент генерализации. В критериально ориентированном тестировании, когда результаты тестируемых сравниваются с установленным пороговым баллом, рассчитывается коэффициент зависимости.

2. Применение теории генерализации в исследованиях

Методы теории генерализации применяются в исследованиях, ориентированных на изучение фасетов, которые оказывают наибольшее влияние на результаты измерений. Чаще всего это исследования эффектов эксперта и заданий теста [Hild, Gut, Brückmann, 2019].

Для изучения эффектов эксперта обычно используются письменные задания или задания *performance-based*, и эксперты оценивают созданный продукт или поведение респондента в процессе выполнения заданий. В применяемых в рамках оценки персонала заданиях типа «почтовая корзина» (*in-basket*), в которых кандидату требуется проанализировать документы компании, эффект эксперта составляет 3,4% общей дисперсии результатов [Li, Pan, Wang, 2021]. Для заданий на измерение креативности решения задач расхождения в работе экспертов зафиксированы в оценке оригинальности как составляющей креативности (15%), в то время как для других составляющих эффект эксперта ниже: 6% для полноты решения, 2% для практичности решения [Hooijdonk van et al., 2022]. Эффект эксперта на результаты тестирования можно минимизировать за счет обучения экспертов и достижения согласованности между экспертами в отношении критериев оценивания [Hild, Gut, Brückmann, 2019].

В дизайнах, где используется фасет заданий, значительная часть дисперсии результатов связана с различиями в трудности заданий и во взаимодействии тестируемого и задания. Например, Р. Шавелсон, Г. Бакстер и С. Гао рассматривают несколько заданий *performance-based*, в которых эффект взаимодействия тестируемого и задания достигает 48–82% общей дисперсии результатов, т.е. некоторые учащиеся справлялись с одними заданиями лучше, а с другими — хуже [Shavelson, Baxter, Gao, 1993]. В измерении коммуникации между студентами — будущими

стоматологами эффект заданий составлял 35,2%: такой показатель свидетельствует о том, что задания в тесте различались по трудности [Uzun et al., 2018].

При исследовании эффекта заданий в отдельных случаях вводятся дополнительные фасеты: например, оценивается эффект заданий, вложенных в содержательную область (по спецификации теста) [Keller, Clauser, Swanson, 2010], или при исследовании эффекта заданий для измерения читательских навыков оценивается вклад количества абзацев в стимульном материале [Liao, 2023]. Для заданий, предназначенных для оценки навыков письма, анализируют вклад жанра [Bouwer et al., 2015] и темы текста [Wu, Steinkrauss, Lowie, 2023] в баллы тестируемых.

Считается, что наибольший вклад в дисперсию результатов должен вносить эффект респондента. Если это так — значит, инструмент измерения успешно дифференцирует респондентов по уровню способности [Briesch et al., 2014]. На практике из-за превалирующей роли других эффектов (например, из-за сильных различий в трудности заданий) вклад респондентов в результаты оценивания может оказаться небольшим. Например, эффект респондентов составил 2,8% общей дисперсии результатов при измерении коммуникативных навыков студентов [Uzun et al., 2018], около 10% для заданий по чтению и навыку письма [Bouwer et al., 2015; Liao, 2023], 15% для компьютерных симуляций, моделирующих общение врачей и пациентов [Keller, Clauser, Swanson, 2010], от 22 до 28% при измерении креативности решения задач [Hooijdonk van et al., 2022]. Однако существуют примеры, где эффект респондентов превышает 50% [Buyukkidik, Anil, 2015].

Исследования, в которых применяются методы теории генерализации, различаются и по измеряемому конструкту, и по формату инструмента. В данном исследовании используются задания сценарного типа с системой автоматической проверки, в то время как в большинстве описанных выше исследований для оценивания респондентов привлекались эксперты. Кроме того, только в нескольких работах используются инструменты для оценивания универсальных навыков или исследуются фасеты, схожие с контекстом сценариев (например, тема текста в заданиях типа эссе).

3. Описание инструмента

3.1. Контекст как особенность заданий сценарного типа

Особенностью сценарных заданий является наличие контекста. Согласно одному из определений, контекст — это общий стимульный материал для нескольких заданий (тестлеты заданий) [Haladyna, Downing, Rodriguez, 2002]. Другие исследователи отождествляют понятие контекста с понятием сценария (*scenario*) в заданиях в видеоформате [Zhai et al., 2021]. М. Руис-Примо и М. Ли [Ruiz-Primo, Li, 2015] предложили расши-

ренную классификацию характеристик контекста, среди которых: сложность текстовых материалов, степень абстрактности контекста, тип контекста (например, школьный, профессиональный) и др. В той же статье авторы выделяют три уровня контекста: общий контекст (основная проблема сценария, связывающая все индикаторы), контекст для группы индикаторов (например, индикаторы относятся к одному тексту внутри сценария) и индивидуальный контекст индикатора (внутри сценария создается уникальный контекст для одной задачи). Согласно этой классификации индикатор может относиться сразу к нескольким контекстным уровням.

В данной статье мы рассматриваем контекст сценарного задания как стимульный материал, задающий среду тестирования, которая мотивирует респондентов на совершение действий, отражающих конструкт [Davier von, Mislevy, Hao, 2022]. В соответствии с методом доказательной аргументации при разработке тестов (*evidence-centered design*, ECD) [Mislevy, Almond, Lukas, 2003] необходимо разделять свидетельства для оценки выраженности конструкта (например, свидетельством критического мышления может быть действие «выделяет в тексте информацию, релевантную задаче») и контекст, который необходим для стимуляции выполнения нужного действия.

Таким образом, контекст определяет основную проблемную ситуацию сценарного задания (например, тестируемый решил завести домашнего питомца и найти информацию об условиях содержания питомца) и развитие ситуации (сюжета) — последовательность действий, отношения между этапами задания, персонажами и проч. [Грачева, Тарасова, 2022]. При этом одна и та же ситуация может иметь разное тематическое наполнение (например, в качестве питомца может быть кролик или собака), т.е. отличаться контентом (контент стимульного материала, *content of source*) [Homayounzadeh, Saadat, Ahmadi, 2019].

3.2. Инструмент «4К»

Для оценки универсальных навыков используются задания сценарного типа в цифровой среде из инструмента «4К», разработанного в рамках метода доказательной аргументации. Инструмент содержит автоматизированную систему проверки (без привлечения экспертов), он прошел апробацию на выборках из нескольких российских регионов и показал хорошие психометрические характеристики. Результаты валидизации инструмента представлены в нескольких статьях и докладах на научных конференциях [Брун, Орел, Угланова, 2020; Угланова, Жильцова, Лебедева, 2021].

В данной работе рассматриваются сценарии для оценки критического мышления и коммуникации. Согласно концеп-

туальной 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. Подробнее концептуальная рамка критического мышления представлена в [Uglanova et al., 2022].

Коммуникация в данном инструменте понимается как способность успешно общаться на письме и устно, используя как вербальные, так и невербальные средства. Измерение коммуникации происходит в диалоговом общении с симуляционными аватарами или персонажами сценария (*human-to-agent approach*) [Rosen, 2017]. Таким образом, инструмент оценивает коммуникацию не как часть речевой способности, а как способность решать различные жизненные задачи в условиях сотрудничества с другими людьми — взрослыми или сверстниками. Согласно концептуальной рамке инструмента [Угланова, Жильцова, Лебедева, 2021], конструкт «коммуникация в условиях сотрудничества» включает несколько взаимосвязанных составляющих, соответствующих фазам коммуникативной деятельности:

- ориентация — способность анализировать коммуникативную ситуацию, в том числе информацию о собеседнике, чтобы распределять роли в команде, выявлять общую цель общения, а также адаптировать коммуникативные действия к ситуации общения;
- активная фаза коммуникации — способность распознавать и реализовать коммуникативные намерения в соответствии с социальными и языковыми конвенциями;
- регуляция общения — способность распознавать некорректное коммуникативное поведение собеседника и нарушение социальных норм и адекватно реагировать на них.

В исследовании используются пять заданий сценарного типа для оценки критического мышления и коммуникации. Два сценария («Аквариум» и «Динозавр») направлены на оценку критического мышления, два сценария («Спектакль» и «Торт») — на оценку коммуникации. Еще один сценарий («Путешествие») содержит индикаторы, относящиеся как к коммуникации, так и к критическому мышлению. Далее в статье эти сценарии будут упомянуты как оригинальные.

Дополнительно инструмент «4К» включает альтернативные варианты для каждого оригинального сценария. Разработка альтернативных вариантов происходила с использованием процедуры клонирования для создания заданий с единой структурой и эквивалентными психометрическими характери-

стиками в максимально похожем контексте [Грачева, Тарасова, 2022]. Альтернативные варианты отличаются от оригинальных сценариев только контентом при сохранении проблемы и сюжета. Ниже приводится описание сценарных заданий с целью продемонстрировать разнообразие контекстов инструмента.

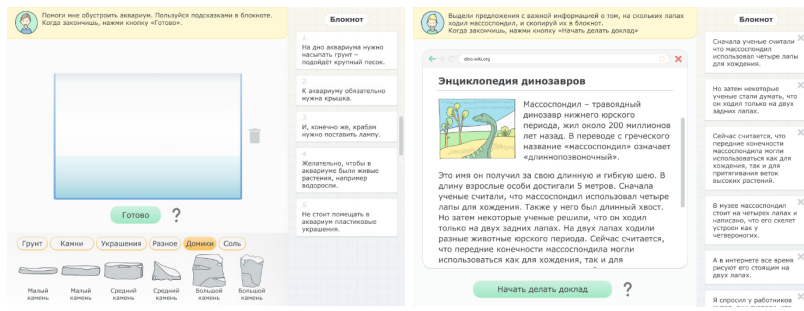
3.3. Описание контекстов сценарных заданий

3.3.1. Сценарное задание «Аквариум» для измерения критического мышления

В основе контекста сценария «Аквариум» — задача обустройства аквариума для крабов. В сценарии моделируется интернет-браузер, в котором тестируемый может изучать тексты электронных статей, выделять и сохранять информацию о предметах, которые понадобятся при обустройстве аквариума. На основе проанализированной информации тестируемый обустраивает аквариум для крабов из ограниченного набора предметов, проявляя при этом способность к формулированию собственного вывода о том, какие предметы должны быть в аквариуме для крабов, а какие нет (рис. 1а).

В альтернативном варианте сценария «Террариум» тестируемые сталкиваются с теми же этапами задания с другим контентом. Здесь главная задача — построить террариум для гекконов. Сценарии содержат по 24 индикатора критического мышления. Подробнее со сценарным заданием «Аквариум» можно ознакомиться в [Грачева, 2022].

Рис. 1. **Примеры экранов из сценарных заданий «Аквариум» и «Динозавр»**



(а) «Аквариум» (б) «Динозавр»

3.3.2. Сценарное задание «Динозавр» для измерения критического мышления

В основе контекста сценария «Динозавр» — подготовка школьного доклада про динозавров. В ходе сценария тестируемому необходимо выбрать наиболее достоверный источник информации (ссылку), проанализировать текст электронной статьи (рис. 1б), сделать вывод о том, на скольких лапах ходил динозавр, и составить слайд для презентации.

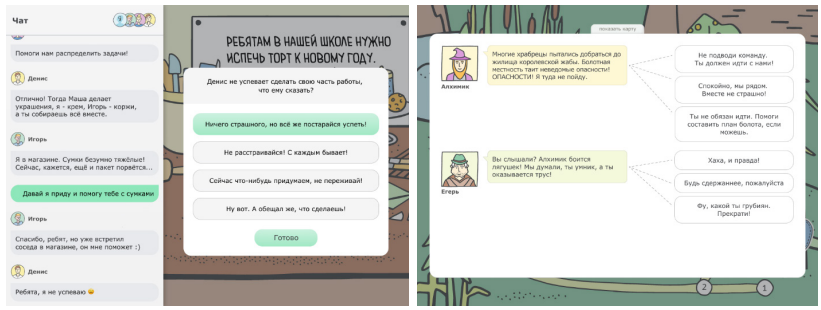
В альтернативном варианте сценария «Еж» тестируемые сталкиваются с теми же задачами с другим контентом. В нем главная цель — подготовка школьного доклада про ежей для ответа на вопрос, зачем ежики трутся иголками о предметы. Сценарии содержат по 6 индикаторов критического мышления.

3.3.3. Сценарное задание «Торт» для измерения коммуникации

Основная проблемная ситуация сценария «Торт» задается стартовым сообщением на экране: «Ребятам в нашей школе нужно испечь торт к Новому году. Помоги им, пожалуйста!». В сценарии используется симуляция чата, где тестируемый общается с персонажами, выбирая сообщение из предложенных (рис. 2а). По сюжету сценария тестируемый сталкивается с разными задачами, например распределяет роли для персонажей для приготовления торта, распознает эмоциональное состояние персонажа, который не справился с задачей, выражает недовольство при нарушении социальных норм (персонаж предлагает съесть торт до праздника).

В альтернативном варианте сценария «Плакат» тестируемые сталкиваются с теми же задачами с другим контентом, здесь главная цель — подготовка плаката к школьному празднику День весны. Сценарии содержат по 15 индикаторов коммуникации.

Рис. 2. **Примеры экранов из сценарного задания «Торт» и «Путешествие»**



(а) «Торт» (б) «Путешествие»

3.3.4. Сценарное задание «Спектакль» для измерения коммуникации

Основная задача сценария «Спектакль» — помочь школьному театру подготовить водолазный костюм для спектакля. В сценарии используется симуляция чата, в котором тестируемый общается с персонажами (аналогично чату на рис. 2а). По сюжету сценария тестируемому необходимо познакомиться с командой, совместно с персонажами выяснить, как выглядит костюм, распределить задачи, задавать уточняющие вопросы, распоз-

нать эмоциональное состояние персонажей и предлагать помощь, где необходимо.

В альтернативном варианте сценария «Конкурс» тестируемому требуется помочь школьному кружку подготовить костюм космонавта для участия в конкурсе. Сценарии содержат по 10 индикаторов коммуникации.

3.3.5. Сценарное задание «Путешествие» для измерения критического мышления и коммуникации

В сценарии «Путешествие» тестируемый оказывается в волшебном королевстве, и местные жители просят его найти ингредиенты для зелья, чтобы вылечить короля. Тестируемый отправляется в путешествие за ингредиентами вместе с командой — Егерем, Алхимиком и Воином. В сценарии тестируемый проявляет свои способности к коммуникации, выбирая реакцию в ответ на сообщение каждого персонажа (рис. 2б). Для проверки навыка критического мышления тестируемого просят сравнить два сообщения, описывающие дорогу к волшебным землям, и выделить предложения, в которых дорога описывается по-разному, либо выбрать нужный ингредиент для зелья, проанализировав информацию от напарников.

В альтернативном варианте сценария «Лабиринт» тестируемые сталкиваются с теми же задачами в другом контенте. Здесь их главная цель — найти волшебные вещи, чтобы помочь расколдовать королеву. Сценарии содержат по 8 индикаторов критического мышления и 22 индикатора коммуникации.

4. Выборка и процедура исследования

В статье используются данные, полученные осенью 2021 г. в ходе тестирования учащихся 4-х классов, которые принимали участие в исследовании универсальных навыков.

Перед началом тестирования администраторам тестирования были высланы руководства с описанием инструмента, этапов подготовки к тестированию и требований к программному обеспечению. Тестирование проходило в школах, в компьютерном классе, каждому респонденту предоставляли персональный компьютер и логин для доступа на сайт инструмента «4К». Для всех учеников получено согласие родителей на участие в исследовании.

В рамках исследования тестируемым предлагалось выполнить пять оригинальных заданий и дополнительно два альтернативных варианта заданий, которые назначались случайным образом. Альтернативный вариант сценария мог предъявляться как до соответствующего оригинального сценария, так и после него. Тестирование проходило в два дня и в сумме занимало около 80 минут.

Описанная процедура исследования позволила получить данные по всем заданиям сценарного типа (ориги-

нальные и альтернативные варианты сценариев) за ограниченное время тестирования. Данные для каждой пары сценариев получены по случайным подвыборкам респондентов: 998 респондентов — «Аквариум»/«Террариум», 1096 респондентов — «Спектакль»/«Конкурс», 1052 респондента — «Путешествие»/«Лабиринт», 868 респондентов — «Торт»/ «Плакат», 466 респондентов — «Динозавр»/«Еж».

5. Методология анализа в рамках теории генерализации

Для анализа вклада контекста сценарного задания в результаты оценивания универсальных навыков предлагаются два дизайна G-исследования. В дизайнах используются фасеты со случайным эффектом.

5.1. Дизайн G-исследования

Первый дизайн — с вложенными фасетами $p \times (i : c)$, где p — респонденты, $(i : c)$ — индикаторы, вложенные в сценарные задания. Он позволит оценить часть различий в баллах, связанных с контекстом сценария, а также с взаимодействием контекста и респондента. В анализе используются данные по оригинальным сценариям. Для оценки критического мышления предлагаются три сценария с разным контекстом: «Аквариум», «Динозавр», «Путешествие», для оценки коммуникации — три сценария с разным контекстом: «Спектакль», «Путешествие», «Торт».

Второй дизайн — перекрестный с двумя фасетами $p \times i \times v$, где p — респонденты, i — индикаторы, v — вариант сценарного задания (оригинальный или альтернативный). Данный дизайн позволит оценить компоненты дисперсии, отражающие различия в контексте вариантов заданий сценарного типа, которые были разработаны как взаимозаменяемые (только с изменением контента). Анализ проводится отдельно для каждой пары вариантов сценариев на подвыборках респондентов. Для оценки критического мышления используются три пары сценариев: «Аквариум»/«Террариум», «Динозавр»/«Еж», «Путешествие»/«Лабиринт», для оценки коммуникации — также три пары сценариев: «Путешествие»/«Лабиринт», «Торт»/«Плакат», «Спектакль/«Конкурс».

5.2. Дизайн D-исследования

Результаты G-исследований используются в D-исследованиях для оценки надежности измерений универсальных навыков. Инструмент «4К» создавался для целей упорядочивания респондентов по уровню развития универсальных навыков, поэтому для оценки надежности измерений в рамках теории генерализации наиболее подходящим является коэффициент генерализации. Тем не менее будет рассчитан и коэффициент зависимости для целей критериального тестирования.

Для первого дизайна G-исследования коэффициенты надежности отражают внутреннюю согласованность заданий

инструмента «4K» при измерении критического мышления и коммуникации. Для второго дизайна G-исследования коэффициенты надежности отражают ретестовую надежность — возможность получения одинаковых результатов у тестируемых по взаимозаменяемым вариантам сценарных заданий. Значения коэффициентов надежности ниже 0,7 свидетельствуют об неудовлетворительной надежности, от 0,7 до 0,8 — об удовлетворительной надежности, значения выше 0,8 — о высокой надежности [Engelhardt, 2009].

Одна из целей D-исследования состоит в подборе такой процедуры тестирования, которая позволит получить надежность измерения не ниже удовлетворительной (0,7). В соответствии с этой целью будет определено количество элементов фасета (индикаторов, контекстов, вариантов сценариев), достаточное для достижения удовлетворительных показателей надежности.

Анализ проведен в программной среде R с использованием пакетов *gtheory* и *lme4* [Huebner, Lucht, 2019].

6. Результаты

6.1. Описательные статистики по сценарным заданиям

Для измерения критического мышления (КМ) используются 37 индикаторов из трех заданий сценарного типа. Средний балл критического мышления на полной выборке (2255 респондентов) равен 21,35 (стандартное отклонение = 7,20). Для измерения коммуникации используются 47 индикаторов из трех заданий сценарного типа. Средний балл коммуникации (КО) на полной выборке (2015 респондентов) равен 30,69 (стандартное отклонение = 7,41). Все индикаторы сценариев дихотомизированы для удобства интерпретации результатов (индикаторы принимают значения 0 или 1).

В табл. 1 приведены описательные статистики для подвыборок респондентов, которые выполняли разные пары вариантов сценариев, с указанием количества индикаторов в каждом сценарии по каждому из навыков.

Таблица 1. **Описательные статистики по сценарным заданиям**

Сценарные задания	*N*	Навык	Среднее (стандартное отклонение)
«Путешествие»/«Лабиринт»	22	КО	15,41 (3,99) / 15,48 (4,22)
«Спектакль»/«Конкурс»	10	КО	6,83 (2,16) / 6,76 (2,06)
«Торт»/«Плакат»	15	КО	9,97 (2,84) / 9,62 (2,89)
«Путешествие»/«Лабиринт»	7	КМ	4,20 (1,63) / 4,21 (1,69)
«Аквариум»/«Террариум»	24	КМ	13,53 (5,61) / 13,40 (5,64)
«Динозавр»/«Еж»	6	КМ	2,76 (1,60) / 3,00 (1,58)

Примечание: ***N*** — количество индикаторов КМ или КО сценарного задания.

6.2. Результаты первого дизайна G-исследования

В табл. 2 представлены оцененные компоненты дисперсии для первого дизайна G-исследования *p* x (*i* : *c*). Основной эффект контекста сценария (*c*) оказался минимальным как для КМ (1%), так и для КО (0%). В среднем результаты оценки универсальных компетенций не различаются в зависимости от предлагаемого контекста сценария, т.е. уровень трудности сценарных заданий с разным контекстом практически одинаков. Тем не менее часть дисперсии результатов связана с эффектом взаимодействия тестируемого и контекста (*p* x *c*: 6,3% для КМ, 4,1% для КО). Например, тестируемый успешно справляется со сценарием «Аквариум», но результат в сценарии «Динозавр» оказывается ниже, а у другого тестируемого — наоборот.

Часть дисперсии результатов связана с различиями в трудности индикаторов внутри сценарного задания, причем индикаторы, предназначенные для оценки КМ, менее гомогенны по трудности (13,9%), чем индикаторы для оценки КО (8,1%). Анализ в рамках теории генерализации предполагает, что все индикаторы сценариев должны быть одинаковой трудности, чтобы эффект индикаторов был минимальным. Однако на практике может стоять задача разработки заданий разной трудности [Arterberry et al., 2014]. Например, легкие задания предъявляются в тестах в первую очередь, чтобы снизить тревожность респондентов перед тестированием. Нивелирование эффекта трудности индикаторов возможно за счет увеличения количества индикаторов (заданий) теста.

Таблица 2. **Оцененные компоненты дисперсии для дизайна *p* x (*i*: *c*)**

Компонент	Коммуникация		Критическое мышление	
	Дисперсия	%	Дисперсия	%
p	0,018	7,8	0,024	9,9
c	0,000	0,0	0,002	1,0
p x *c*	0,009	4,1	0,015	6,3
i : *c*	0,018	8,1	0,034	13,9
e	0,182	80	0,168	68,9

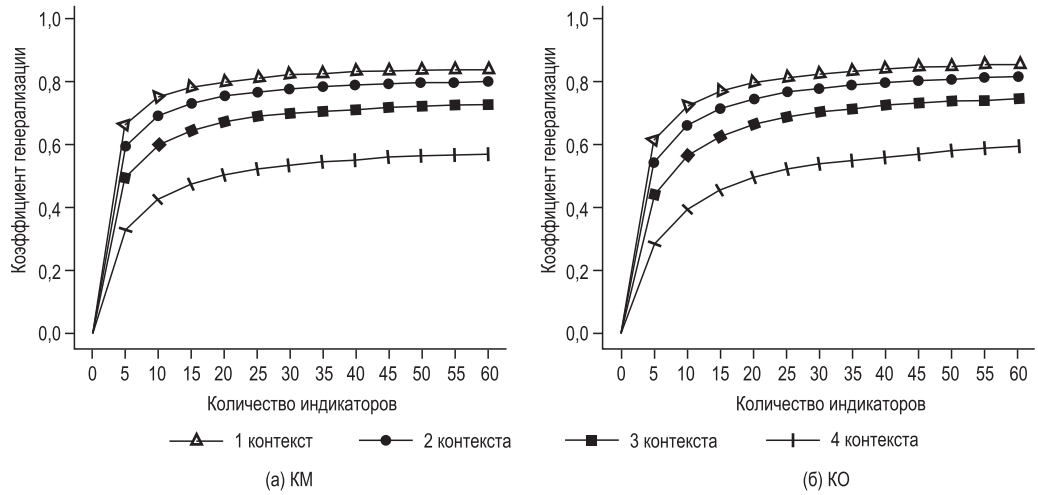
Остаточный компонент дисперсии (*e*), который включает эффект взаимодействия всех фасетов и компонент случайной или систематической ошибки измерения, вкладывается в измерения в большей степени (68,9% для КМ и 80% для КО). Полученный результат может свидетельствовать о наличии фасетов, которые не были учтены в предложенном G-дизайне.

Оцененные компоненты дисперсии по разным источникам позволяют рассчитать коэффициенты генерализации и зави-

симости. Надежность для обоих навыков удовлетворительна и близка к высоким значениям (0,8): коэффициент генерализации (E_{p^2}) для КМ равен 0,79 (коэффициент зависимости φ — 0,76), для КО — 0,80 (коэффициент зависимости — 0,79).

На рис. 3 представлены значения коэффициента генерализации (надежность для нормоориентированного тестирования) при разном количестве индикаторов и контекстов сценариев. Процедура тестирования, в которой все индикаторы критического мышления (38) или коммуникации (47) принадлежат одному контексту, не обеспечивает удовлетворительную надежность для нормоориентированного тестирования (ниже 0,7). При двух контекстах достижение удовлетворительного значения коэффициента генерализации (выше 0,7) возможно при не менее чем 30 индикаторах. Использование сценариев с тремя разными контекстами позволяет достичь удовлетворительных значений коэффициента генерализации при не менее чем 11 индикаторах КМ и 15 индикаторах КО и высоких значений (выше 0,8) — при не менее чем 55 индикаторах КМ и 45 индикаторах КО. Четырех контекстов с 7 индикаторами КМ и 10 индикаторами КО будет достаточно для достижения удовлетворительных значений коэффициента генерализации и с 20 индикаторами КМ и 25 индикаторами КО — для высоких значений.

Рис. 3. **Коэффициент генерализации при разных количествах индикаторов и контекстов сценарных заданий**



Таким образом, использование нескольких контекстов сценарных заданий для измерения универсальных навыков позволяет существенно повысить надежность и степень генерализации результатов, а также сократить время тестирования за счет использования меньшего количества индикаторов.

6.3. Результаты второго дизайна G-исследования

Во втором дизайне исследования анализ проводится отдельно по парам вариантов заданий сценарного типа. В табл. 3 приведены оцененные компоненты дисперсии для сценариев КО.

Таблица 3. **Оцененные компоненты дисперсии для дизайна *p* x *i* x v** (коммуникация)

Компонент	«Путешествие»/«Лабиринт»		«Торт»/«Плакат»		«Спектакль»/«Конкурс»	
	Дисперсия	%	Дисперсия	%	Дисперсия	%
p	0,024	11,3	0,023	10,1	0,025	11,4
i	0,009	4,2	0,022	10,0	0,020	9,3
v	0,000	0,0	0,000	0,1	0,000	0,0
p x *i*	0,036	17,2	0,045	19,6	0,046	20,9
p x *v*	0,004	1,7	0,001	0,6	0,003	1,2
i x *v*	0,002	1,0	0,001	0,6	0,000	0,2
e	0,136	64,7	0,135	59,0	0,125	57,1

Перекрестный дизайн исследования позволил оценить долю дисперсии результатов, связанную со взаимодействием тестируемого и индикаторов сценария: она составляет примерно пятую часть дисперсии баллов (*p* x *i*: от 17,2 до 20,9%). На основании этого показателя можно заключить, что тестируемые непоследовательны в своих ответах на разные индикаторы сценариев. С одной стороны, полученный результат может быть связан с различиями индикаторов по трудности (*i*: от 4,2 до 10%). С другой стороны, согласно теоретической рамке инструмента, коммуникация включает несколько взаимосвязанных составляющих, что могло стать причиной разного восприятия индикаторов отдельными тестируемыми.

Вариант сценария (*v*) в меньшей степени вкладывается в дисперсию результатов. Следовательно, темы сценариев в разных вариантах не различаются по трудности. Эффект взаимодействия варианта и тестируемого/индикаторов присутствует, но он невысокий — до 2%. Таким образом, варианты сценарного типа с изменением только контента могут считаться взаимозаменяемыми.

В табл. 4 приведены оцененные компоненты дисперсии для сценариев, измеряющих критическое мышление.

Процент дисперсии, связанный с взаимодействием респондентов с индикаторами, для компетенции критического мышления находится в диапазоне от 11,5 до 12,6%. Эффект различий в трудности индикаторов различается по сценариям. Например, в сценариях «Путешествие»/«Лабиринт» различия в трудности индикаторов в 2 раза меньше, чем в сценариях «Аквариум»/«Террариум». Одной из причин таких различий может быть разное количество индикаторов.

Таблица 4. **Оцененные компоненты дисперсии для дизайна *p* x *i* x *v* (критическое мышление)**

Компонент	«Путешествие»/«Лабиринт»		«Аквариум»/«Террариум»		«Динозавр»/«Еж»	
	Дисперсия	%	Дисперсия	%	Дисперсия	%
p	0,034	14,9	0,041	16,7	0,036	14,1
i	0,015	6,6	0,038	15,6	0,023	9,1
v	0,000	0,00	0,000	0,00	0,001	0,2
p x *i*	0,027	11,7	0,031	12,6	0,029	11,5
p x *v*	0,003	1,4	0,006	2,3	0,003	1,5
i x *v*	0,002	1,2	0,001	0,5	0,002	0,7
e	0,146	64,1	0,129	52,4	0,161	63,3

Вариант сценария привносит меньше дисперсии в результаты оценивания (*v*: 0–0,2%). Присутствует небольшой (0,5–2,3%) эффект взаимодействия варианта сценария с индикаторами/респондентами.

В рамках D-исследования рассчитаны коэффициенты генерализации (E_{p2}) и зависимости (φ) как меры ретестовой надежности (табл. 5).

Таблица 5. **Ретестовая надежность для двух вариантов сценариев**

Показатель	Коммуникация			Критическое мышление		
	«Путешествие»	«Торт»	«Спектакль»	«Аквариум»	«Динозавр»	«Путешествие»
(E_{p2})	0,78	0,74	0,67	0,86	0,64	0,60
φ	0,77	0,70	0,64	0,83	0,6	0,55
N	22	15	10	24	6	7

Примечание: для удобства в шапке таблицы указаны названия только оригинального варианта сценария. *N* — количество индикаторов в сценарном задании; E_{p2} — коэффициент генерализации; φ — коэффициент зависимости.

Для большинства сценариев коэффициенты ретестовой надежности удовлетворительны (выше 0,7). Неудовлетворительная ретестовая надежность (ниже 0,7) характерна для сценариев с наименьшим количеством индикаторов в каждом сценарии (до 10 включительно).

D-исследование позволило оценить минимальное количество индикаторов, необходимое для достижения удовлетворительной ретестовой надежности при нормоориентированном тестировании (табл. 6).

Судя по полученным результатам, при нормоориентированном тестировании с двумя вариантами сценариев 12–13 инди-

Таблица 6. **Минимальное количество индикаторов, необходимое для достижения удовлетворительной ретестовой надежности (коэффициента генерализации)**

Количество вариантов	Коммуникация			Критическое мышление		
	«Путешествие»	«Торт»	«Спектакль»	«Аквариум»	«Динозавр»	«Путешествие»
2	13	12	12	7	8	12
3	9	10	9	5	6	9
4	8	9	8	4	6	7

каторов в каждом варианте достаточно для достижения нижней границы ретестовой надежности. Повышение надежности возможно за счет увеличения количества вариантов теста. Однако с учетом затрат времени и ресурсов для повышения ретестовой надежности оптимальной стратегией является разработка дополнительных индикаторов (от 2 до 5) в существующие варианты сценарных заданий.

7. Обсуждение результатов

Проведено исследование с целью проанализировать роль контекста сценарных заданий при измерении универсальных навыков. Применение методов теории генерализации позволило оценить вклад контекста в результаты (G-исследование) и установить, какое количество контекстов необходимо для достижения удовлетворительной надежности измерений (D-исследование).

Исследование проводилось на данных, полученных при выполнении 4-классниками заданий методики «4К» — инструмента сценарного типа, предназначенного для измерения критического мышления и коммуникации. По результатам G-исследования для обоих навыков контексты сценарных заданий оказались практически одинаковой трудности, однако результаты тестируемых по разным сценариям различались, что проявилось в наличии эффекта взаимодействия тестируемого и контекста (4,2% для коммуникации, 6,2% для критического мышления). В предыдущих исследованиях в формате *performance-based* эффект взаимодействия тестируемого и контекста заданий был одним из наиболее сильных [Shavelson, Baxter, Gao, 1993; Hild, Gut, Brückmann, 2019]. Контекст может вовлекать и мотивировать одного тестируемого, способствуя успешному выполнению заданий, и в то же время сбить с толку другого тестируемого и исказить его результаты [Messick, 1994].

Для снижения эффекта взаимодействия тестируемого и контекста тестирование универсальных навыков необходимо проводить с использованием нескольких сценариев с разными

контекстами. D-исследование позволило оценить количество контекстов и индикаторов, необходимое для достижения удовлетворительной надежности измерений при нормоориентированном тестировании (с использованием коэффициента генерализации).

При оценивании универсальных навыков с применением одного контекста значения коэффициента генерализации оказались неудовлетворительными (ниже 0,7), при этом увеличение количества контекстов позволяет повысить значение этого коэффициента надежности. Согласно расчетам, 30 индикаторов критического мышления или коммуникации в двух контекстах обеспечивают такую же надежность для нормоориентированного тестирования (0,7), как 15 индикаторов коммуникации или 11 индикаторов критического мышления в трех разных контекстах. Проведенный анализ наглядно демонстрирует возможности D-исследования при проектировании инструмента: для оценки универсальных навыков рекомендуется использовать минимум два сценарных задания с разным контекстом — при таких условиях достигаются удовлетворительные значения коэффициента генерализации, а использование трех контекстов позволит существенно сократить количество индикаторов без потери надежности. Если для инструмента поставлены более строгие критерии надежности (0,8 и выше), рекомендуется использовать четыре сценарных задания с разными контекстами.

Повторное использование контекстных заданий неминуемо ведет к появлению эффекта запоминания — а значит, и к искажениям результатов тестирования. Для решения этой проблемы создаются альтернативные варианты заданий, которые используются как замена оригинальных. Например, в заданиях, оценивающих навыки письма, для создания сопоставимых вариантов и снижения эффекта запоминания изменяют тему стимульного материала, при этом другие характеристики задания, например жанр, сохраняются [Wu, Steinkrauss, Lowie, 2023].

Данные текущего исследования позволили оценить роль контекста для вариантов заданий сценарного типа с максимально похожими контекстами, отличающихся только темой ситуации (контентом). Согласно результатам G-исследования, для всех сценарных заданий, независимо от тестируемого навыка, вклад варианта в тестовые баллы минимален (0–0,2%). Тестируемые показывают стабильные результаты в обоих вариантах сценариев, хотя размер эффекта различается между парами сценариев (эффект взаимодействия тестируемого и варианта находится в диапазоне от 0,6 до 2,3%).

В анализе генерализации нет принятых границ для интерпретации размера эффекта. Исследователи, работающие в рамках данного подхода, рекомендуют сравнивать размеры эф-

фекта, полученного при разных условиях [Briesch et al., 2014]. Результаты текущего исследования позволяют сделать вывод, что фасет варианта сценария в меньшей степени вкладывается в результаты измерений универсальных навыков, однако этот эффект не является нулевым. Таким образом, замена только контента сценария позволяет создать в целом сопоставимые варианты заданий сценарного типа.

Для альтернативных вариантов в рамках D-исследования оценивалась ретестовая надежность — степень, в которой результаты тестируемых воспроизводятся в вариантах заданий, отдельно для пар вариантов сценариев. Ретестовая надежность оказалась неудовлетворительной для тех сценариев, где число индикаторов в каждом варианте меньше 10. Для обеспечения удовлетворительной ретестовой надежности при нормоориентированном тестировании по двум вариантам сценариев достаточно 12–13 индикаторов в каждом варианте.

Помимо надежности, которой в теории генерализации уделяется большое внимание, при оценивании универсальных навыков необходимо обеспечить валидность измерения. Использование нескольких сценариев позволяет не только уменьшить ошибку измерения, но и получить более полную картину поведения тестируемых, которое отражает целевой конструкт. Оценивание сложных навыков должно производиться в разных контекстах, чтобы респонденты могли продемонстрировать свои способности в разных, в том числе незнакомых для них, ситуациях [Wang, Liu, Hau, 2022]. Напротив, разработка альтернативных вариантов заданий сценарного типа подразумевает подбор нового контекста, в котором поведение тестируемого останется стабильным, чтобы выводы по результатам тестирования были справедливы для всех респондентов, независимо от предъявляемого варианта. Использование схожих контекстов, различающихся темами ситуации, представляется оптимальной стратегией для создания сопоставимых вариантов заданий сценарного типа.

Методы теории генерализации и результаты данного исследования могут быть полезны разработчикам при проектировании заданий сценарного типа. Разработка сценарных заданий требует немалых ресурсов [Угланова, Брун, Васин, 2018], в особенности если задания реализованы в цифровой среде. Решение о количестве контекстов и индикаторов должно приниматься с учетом цели тестирования и имеющихся ресурсов.

8. Ограничения и дальнейшие направления исследования

Результаты исследования следует воспринимать и использовать с учетом ограничений.

Рекомендации о количестве индикаторов и контекстов даны для разных дизайнов G-исследования — набора фасетов и отно-

шений между ними. При проектировании инструмента следует брать в расчет рекомендации по дизайну исследования, который будет реализован в конкретной ситуации тестирования. Например, инструмент измерения «4К» для оценки критического мышления содержит три сценария, в этом случае для оценки ретестовой надежности нужно рассматривать три сценария в каждом варианте как единое целое, а не как отдельные пары сценариев.

В данном исследовании изучался эффект общего контекста задания, а не отдельных характеристик контекста, на результаты тестирования. Изучение роли отдельных характеристик контекста (например, необходимости предметных знаний для решения задачи, приближенности контекста к реальности, сложности контекста для респондентов разного возраста, количества персонажей, ветвей сюжета и проч.) позволит глубже понять функционирование контекстных заданий. Также в будущих исследованиях целесообразно проанализировать контекстную нагруженность индикаторов в разрезе трех уровней контекста, предложенных в [Ruiz-Primo, Li, 2015]: общего контекста, контекста для группы индикаторов и индивидуального контекста индикатора. М. Руис-Примо и М. Ли анализировали среднюю трудность для групп индикаторов только с одним контекстным уровнем, двумя и тремя уровнями (например, для двух уровней: индикаторы одновременно объединены общим контекстом и специфическим групповым контекстом). Работы в этом направлении могут быть продолжены анализом эффекта контекста в зависимости от количества уровней (контекстной нагруженности) индикаторов.

С точки зрения доказательства валидности выводов по итогам тестирования особого внимания заслуживает изучение поведения респондента в сценариях, контекст которых приближен к реальной жизни. Например, отличается ли выраженность навыков коммуникации у ученика в тестовой среде, симулирующей учебную ситуацию, от реального поведения в классе? В будущих исследованиях вклад контекста с использованием теории генерализации может быть дополнительно рассмотрен в рамках концепции трансфера знаний и навыков из одного контекста в другой [Barnett, Ceci, 2002].

В данном исследовании используется ограниченное число фасетов для объяснения результатов оценивания. Другие фасеты или отношения между фасетами могут быть протестированы для сравнения результатов и снижения доли дисперсии, связанной со случайной или систематической ошибкой измерения.

В текущем исследовании каждый универсальный навык рассматривается как одномерный конструкт, в то время как теоретическая рамка инструмента подразделяет навыки на несколько составляющих. В предыдущем исследовании, осу-

ществленном в другой методологии, анализ проводился по отдельным составляющим критического мышления и на примере пары сценариев «Аквариум»/«Террариум» было показано, что оценки по вариантам значимо различаются между собой для навыка формулирования вывода и не различаются для навыка анализа информации [Грачева, 2022]. Применение многомерной теории генерализации позволит проанализировать возможности сценарных заданий в измерении отдельных составляющих навыков [Keller, Clauser, Swanson, 2010]. Кроме того, в данной работе применяются «классические» методы теории генерализации для работы с сырыми баллами тестирования. В современных работах в этом направлении предпринимаются попытки переложить идеи теории генерализации на модели структурных уравнений [Jorgensen, 2021] или байесовских сетей [Jiang, Skorupski, 2018].

9. Заключение

Измерение универсальных навыков с использованием сценарных заданий — нетривиальная задача для разработчиков и психометриков. В данной статье показано, как применение методов теории генерализации позволяет оценить вклад контекста задания в результаты тестирования для случаев, когда используются разные сценарии или сценарии с похожими контекстами (альтернативные варианты). Полученные результаты используются для предсказания надежности измерений при разном количестве контекстов или индикаторов сценария.

Таким образом, теория генерализации предлагает гибкий подход к проектированию структуры теста для разных форматов заданий и целей тестирования. Результаты анализа позволяют дать конкретные рекомендации по улучшению организации тестирования и достижению удовлетворительной надежности измерений.

Благодарности

Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ.

Автор благодарит И.Л. Угланову за комментарии и поддержку в подготовке статьи.

Литература

1. Брун И.В., Орел Е.А., Угланова И.Л. (2020) Измерение креативности и критического мышления в начальной школе. *Психологический журнал*, т. 41, № 6, сс. 96–107. https://doi.org/10.31857/S020595920011124-2
2. Грачева Д.А. (2022) Анализ сопоставимости измерения метапредметных навыков в цифровой среде. *Психологическая наука и образование*, т. 27, № 6, сс. 57–67. https://doi.org/10.17759/pse.2022270605
3. Грачева Д.А., Тарасова К.В. (2022) Подходы к разработке вариантов заданий сценарного типа в рамках метода доказательной аргументации.

Отечественная и зарубежная педагогика, т. 1, № 3, сс. 83–97. https://doi.org/10.24412/2224-0772-2022-84-83-97

4. Угланова И., Брун И., Васин Г. (2018) Методология *Evidence-Centered Design* для измерения комплексных психологических конструктов. *Современная зарубежная психология*, т. 7, № 3, сс. 18–27. https://doi.org/10.17759/jmfp.2018070302
5. Угланова И.Л., Жильцова Л.Ю., Лебедева М.Ю. (2021) Измерение навыков коммуникации и кооперации в начальной и средней школе: могут ли школьники договориться с инопланетянином? Материалы *V Международной научной конференции «Информатизация образования и методика электронного обучения: цифровые технологии в образовании» (Красноярск, 20–23 сентября 2022 г.),* Красноярск: Сибирский федеральный университет, сс. 682–686.
6. Arterberry B.J., Martens M.P., Cadigan J.M., Rohrer D. (2014) Application of Generalizability Theory to the Big Five Inventory. *Personality and Individual Differences*, vol. 69, October, pp. 98–103. https://doi.org/10.1016/j.paid.2014.05.015
7. Barnett S.M., Ceci S.J. (2002) When and Where Do We Apply What We Learn?: A Taxonomy for Far Transfer. *Psychological Bulletin*, vol. 128, no 4, pp. 612–637. https://doi.org/10.1037/0033-2909.128.4.612
8. Bouwer R., Béguin A., Sanders T., van den Bergh H. (2015) Effect of Genre on the Generalizability of Writing Scores. *Language Testing*, vol. 32, no 1, pp. 83–100. https://doi.org/10.1177/0265532214542994
9. Brennan R.L. (1992) Generalizability Theory. *Educational Measurement: Issues and Practice,* vol. 11, no 4, pp. 27–34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x
10. Briesch A.M., Swaminathan H., Welsh M., Chafouleas S.M. (2014) Generalizability Theory: A Practical Guide to Study Design, Implementation, and Interpretation. *Journal of School Psychology*, vol. 52, no 1, pp. 13–35. https://doi.org/10.1016/j.jsp.2013.11.008
11. Buyukkidik S., Anil D. (2015) Investigation of Reliability in Generalizability Theory with Different Designs on Performance-Based Assessment. *Education and Science*, vol. 40, no 117, pp. 285–296. http://dx.doi.org/10.15390/EB.2015.2454
12. Cronbach L.J., Gleser G.C., Nanda H., Rajaratnam N. (1972) *The Dependability of Behavioral Measurements*. New York, NY: Wiley.
13. Davier von A.A., Mislevy R.J., Hao J. (eds) (2021) *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment: With Examples in R and Python.* Cham: Springer International. https://doi.org/10.1007/978-3-030-74394-9
14. Engelhardt P.V. (2009) An Introduction to Classical Test Theory as Applied to Conceptual Multiple-Choice Tests. *Getting Started in PER* (eds C. Henderson, K. Harper), College Park, MD: American Association of Physics Teachers, pp. 1–40.
15. Haladyna T.M., Downing S.M., Rodriguez M.C. (2002) A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, vol. 15, no 3, pp. 309–333. https://doi.org/10.1207/S15324818AME1503_5
16. Hild P., Gut C., Brückmann M. (2019) Validating Performance Assessments: Measures That May Help to Evaluate Students' Expertise in 'Doing Science'. *Research in Science & Technological Education*, vol. 37, no 4, pp. 419–445. https://doi.org/10.1080/02635143.2018.1552851
17. Homayounzadeh M., Saadat M., Ahmadi A. (2019) Investigating the Effect of Source Characteristics on Task Comparability in Integrated Writing Tasks. *Assessing Writing*, vol. 41, no 2, pp. 25–46. https://doi.org/10.1016/j.asw.2019.05.003

18. Hooijdonk van M., Mainhard T., Kroesbergen E.H., van Tartwijk J. (2022) Examining the Assessment of Creativity with Generalizability Theory: An Analysis of Creative Problem Solving Assessment Tasks. *Thinking Skills and Creativity*, vol. 43, no 1, Article no 100994. https://doi.org/10.1016/j.tsc.2021.100994
19. Huebner A., Lucht M. (2019) Generalizability theory in R. *Practical Assessment, Research, and Evaluation*, vol. 24, no 5. https://doi.org/10.7275/5065-gc10
20. Jiang Z., Skorupski W. (2018) A Bayesian Approach to Estimating Variance Components Within a Multivariate Generalizability Theory Framework. *Behavior Research Methods*, vol. 50, no 3, pp. 2193–2214. https://doi.org/10.3758/s13428-017-0986-3
21. Jorgensen T.D. (2021) How to Estimate Absolute-Error Components in Structural Equation Models of Generalizability Theory. *Psych*, vol. 3, no 2, pp. 113–133. https://doi.org/10.3390/psych3020011
22. Keller L.A., Clauser B.E., Swanson D.B. (2010) Using Multivariate Generalizability Theory to Assess the Effect of Content Stratification on the Reliability of a Performance Assessment. *Advances in Health Sciences Education*, vol. 15, no 5, pp. 717–733. https://doi.org/10.1007/s10459-010-9233-8
23. Li G., Pan Y., Wang W. (2021) Using Generalizability Theory and Many-Facet Rasch Model to Evaluate In-Basket Tests for Managerial Positions. *Frontiers in Psychology*, vol. 12, July, Article no 660553. https://doi.org/10.3389/fpsyg.2021.660553
24. Liao R.J. (2023) The Use of Generalizability Theory in Investigating the Score Dependability of Classroom-Based L2 Reading Assessment. *Language Testing*, vol. 40, no 1, pp. 86–106. https://doi.org/10.1177/02655322211070840
25. Messick S. (1994) The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, vol. 23, no 2, pp. 13–23. https://doi.org/10.3102/0013189X023002013
26. Mislevy R.J., Almond R.G., Lukas J.F. (2003) *A Brief Introduction to Evidence-Centered design. ETS Research Report Series no 2003(1)*. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x
27. Rosen Y. (2017) Assessing Students in Human-to-Agent Settings to Inform Collaborative Problem-Solving Learning. *Journal of Educational Measurement*, vol. 54, no 1, pp. 36–53. https://doi.org/10.1111/jedm.12131
28. Ruiz-Primo M.A., Li M. (2015) The Relationship between Item Context Characteristics and Student Performance: The Case of the 2006 and 2009 PISA Science Items. *Teachers College Record*, vol. 117, no 1, pp. 1–36. https://doi.org/10.1177/016146811511700118
29. Shavelson R.J., Baxter G.P., Gao X. (1993). Sampling Variability of Performance Assessments. *Journal of Educational Measurement*, vol. 30, no 3, pp. 215–232. https://doi.org/10.1111/j.1745-3984.1993.tb00424.x
30. Shavelson R.J., Webb N.M., Rowley G.L. (1992) Generalizability Theory. *Methodological Issues & Strategies in Clinical Research* (ed. A.E. Kazdin), American Psychological Association, pp. 233–256. http://dx.doi.org/10.1037/10109-051
31. Uglanova I., Orel E., Gracheva D., Tarasova K. (2023) Computer-Based Performance Approach for Critical Thinking Assessment in Children. *British Journal of Educational Psychology*, vol. 93, no. 2, pp. 531–544. https://doi.org/10.1111/bjep.12576
32. Uzun N.B., Aktas M., Asiret S., Yormaz S. (2018) Using Generalizability Theory to Assess the Score Reliability of Communication Skills of Dentistry Students. *Asian Journal of Education and Training*, vol. 4, no 2, pp. 85–90. https://doi.org/10.20448/journal.522.2018.42.85.90
33. Wang D., Liu H., Hau K.T. (2022) Automated and Interactive Game-Based Assessment of Critical Thinking. *Education and Information Technologies*, vol. 27, no 4, pp. 4553–4575. https://doi.org/10.1007/s10639-021-10777-9

34. Wu M.Y., Steinkrauss R., Lowie W. (2023) The Reliability of Single Task Assessment in Longitudinal L2 Writing Research. *Journal of Second Language Writing*, vol. 59, no 4, Article no 100950. https://doi.org/10.1016/j.jslw.2022.100950
35. Zhai X., Haudek K.C., Wilson C., Stuhlsatz M. (2021) A Framework of Construct-Irrelevant Variance for Contextualized Constructed Response Assessment. *Frontiers in Education*, vol. 6, October, Article no 751283. https://doi.org/10.3389/feduc.2021.751283

References

Arterberry B.J., Martens M.P., Cadigan J.M., Rohrer D. (2014) Application of Generalizability Theory to the Big Five Inventory. *Personality and Individual Differences*, vol. 69, October, pp. 98–103. https://doi.org/10.1016/j.paid.2014.05.015

Barnett S.M., Ceci S.J. (2002) When and Where Do We Apply What We Learn?: A Taxonomy for Far Transfer. *Psychological Bulletin*, vol. 128, no 4, pp. 612–637. https://doi.org/10.1037/0033-2909.128.4.612

Bouwer R., Béguin A., Sanders T., van den Bergh H. (2015) Effect of Genre on the Generalizability of Writing Scores. *Language Testing*, vol. 32, no 1, pp. 83–100. https://doi.org/10.1177/0265532214542994

Brennan R.L. (1992) Generalizability Theory. *Educational Measurement: Issues and Practice*, vol. 11, no 4, pp. 27–34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x

Briesch A.M., Swaminathan H., Welsh M., Chafouleas S.M. (2014) Generalizability Theory: A Practical Guide to Study Design, Implementation, and Interpretation. *Journal of School Psychology*, vol. 52, no 1, pp. 13–35. https://doi.org/10.1016/j.jsp.2013.11.008

Brun I.V., Orel E.A., Uglanova I.L. (2020) Izmerenie kreativnosti i kriticheskogo myshleniya v nachal'noy shkole [Measuring Creativity and Critical Thinking in Primary School]. *Psikhologicheskij zhurnal*, vol. 41, no 6, pp. 96–107. https://doi .org/10.31857/S020595920011124-2

Buyukkidik S., Anil D. (2015) Investigation of Reliability in Generalizability Theory with Different Designs on Performance-Based Assessment. *Education and Science*, vol. 40, no 117, pp. 285–296. http://dx.doi.org/10.15390/EB.2015.2454

Cronbach L.J., Gleser G.C., Nanda H., Rajaratnam N. (1972) *The Dependability of Behavioral Measurements.* New York, NY: Wiley.

Davier von A.A., Mislevy R.J., Hao J. (eds) (2021) *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment: With Examples in R and Python.* Cham: Springer International. https://doi.org/10.1007/978-3-030-74394-9

Engelhardt P.V. (2009) An Introduction to Classical Test Theory as Applied to Conceptual Multiple-Choice Tests. *Getting Started in PER* (eds C. Henderson, K. Harper), College Park, MD: American Association of Physics Teachers, pp. 1–40.

Gracheva D.A. (2022) Analiz sopostavimosti izmereniya metapredmetnykh navykov v tsifrovoy srede [Analysis of Task Comparability in Digital Environment by the Case of Metacognitive Skills]. *Psikhologicheskaya nauka i obrazovanie / Psychological Science and Education*, vol. 27, no 6, pp. 57–67. https://doi.org/10.17759/pse.2022270605

Gracheva D.A, Tarasova K.V. (2022) Podkhody k razrabotke variantov zadaniy scenarnogo tipa v ramkakh metoda dokazatelnoy argumentatsii [Approaches to the Development of Scenario-Based Task Forms within the Framework of Evidence-Centered Design]. *Otechestvennaya i zarubezhnaya pedagogika*, vol. 1, no 3, pp. 83–97. https://doi.org/ 10.24412/2224–0772–2022–84–83–97

Haladyna T.M., Downing S.M., Rodriguez M.C. (2002) A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, vol. 15, no 3, pp. 309–333. https://doi.org/10.1207/S15324818AME1503_5

Hild P., Gut C., Brückmann M. (2019) Validating Performance Assessments: Measures That May Help to Evaluate Students' Expertise in 'Doing Science'. *Research in Science & Technological Education*, vol. 37, no 4, pp. 419–445. https://doi.org/10.1080/02635143.2018.1552851

Homayounzadeh M., Saadat M., Ahmadi A. (2019) Investigating the Effect of Source Characteristics on Task Comparability in Integrated Writing Tasks. *Assessing Writing*, vol. 41, no 2, pp. 25–46. https://doi.org/10.1016/j.asw.2019.05.003

Hooijdonk van M., Mainhard T., Kroesbergen E.H., van Tartwijk J. (2022) Examining the Assessment of Creativity with Generalizability Theory: An Analysis of Creative Problem Solving Assessment Tasks. *Thinking Skills and Creativity*, vol. 43, no 1, Article no 100994. https://doi.org/10.1016/j.tsc.2021.100994

Huebner A., Lucht M. (2019) Generalizability theory in R. *Practical Assessment, Research, and Evaluation*, vol. 24, no 5. https://doi.org/10.7275/5065-gc10

Jiang Z., Skorupski W. (2018) A Bayesian Approach to Estimating Variance Components Within a Multivariate Generalizability Theory Framework. *Behavior Research Methods*, vol. 50, no 3, pp. 2193–2214. https://doi.org/10.3758/s13428-017-0986-3

Jorgensen T.D. (2021) How to Estimate Absolute-Error Components in Structural Equation Models of Generalizability Theory. *Psych*, vol. 3, no 2, pp. 113–133. https://doi.org/10.3390/psych3020011

Keller L.A., Clauser B.E., Swanson D.B. (2010) Using Multivariate Generalizability Theory to Assess the Effect of Content Stratification on the Reliability of a Performance Assessment. *Advances in Health Sciences Education*, vol. 15, no 5, pp. 717–733. https://doi.org/10.1007/s10459-010-9233-8

Li G., Pan Y., Wang W. (2021) Using Generalizability Theory and Many-Facet Rasch Model to Evaluate In-Basket Tests for Managerial Positions. *Frontiers in Psychology*, vol. 12, July, Article no 660553. https://doi.org/10.3389/fpsyg.2021.660553

Liao R.J. (2023) The Use of Generalizability Theory in Investigating the Score Dependability of Classroom-Based L2 Reading Assessment. *Language Testing*, vol. 40, no 1, pp. 86–106. https://doi.org/10.1177/02655322211070840

Messick S. (1994) The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, vol. 23, no 2, pp. 13–23. https://doi.org/10.3102/0013189X023002013

Mislevy R.J., Almond R.G., Lukas J.F. (2003) *A Brief Introduction to Evidence-Centered design. ETS Research Report Series no 2003(1)*. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Rosen Y. (2017) Assessing Students in Human-to-Agent Settings to Inform Collaborative Problem-Solving Learning. *Journal of Educational Measurement*, vol. 54, no 1, pp. 36–53. https://doi.org/10.1111/jedm.12131

Ruiz-Primo M.A., Li M. (2015) The Relationship between Item Context Characteristics and Student Performance: The Case of the 2006 and 2009 PISA Science Items. *Teachers College Record*, vol. 117, no 1, pp. 1–36. https://doi.org/10.1177/016146811511700118

Shavelson R.J., Baxter G.P., Gao X. (1993). Sampling Variability of Performance Assessments. *Journal of Educational Measurement*, vol. 30, no 3, pp. 215–232. https://doi.org/10.1111/j.1745-3984.1993.tb00424.x

Shavelson R.J., Webb N.M., Rowley G.L. (1992) Generalizability Theory. *Methodological Issues & Strategies in Clinical Research* (ed. A.E. Kazdin), American Psychological Association, pp. 233–256. http://dx.doi.org/10.1037/10109-051

Uglanova I., Brun I., Vasin G. (2018) Metodologiya Evidence-Centered Design dlya izmereniya kompleksnykh psikhologicheskikh konstruktov [Evidence-Centered Design Method for Measuring Complex Psychological Constructs]. *Journal of Modern Foreign Psychology*, vol. 7, no 3, pp. 18–27. https://doi.org/10.17759/jmfp.2018070302

Uglanova I.L., Zhiltsova L.Y., Lebedeva M.Y. (2021) Izmerenie navykov kommunikatsii i kooperatsii v nachal'noy i sredney shkole: mogut li shkol'niki dogov-

orit'sya s inoplanetyaninom? [Communication and Cooperation Assessment in Primary and Middle School: How Students Negotiate with an Alien?]. Proceedings of the *5th International Conference "Informatization of Education and E-learning Methodology: Digital Technologies in Education" (Krasnoyarsk, 2022, September, 20–23),* Krasnoyarsk: Siberian Federal University, pp. 682–686.

Uglanova I., Orel E., Gracheva D., Tarasova K. (2023) Computer-Based Performance Approach for Critical Thinking Assessment in Children. *British Journal of Educational Psychology*, vol. 93, no. 2, pp. 531–544. https://doi.org/10.1111/bjep.12576

Uzun N.B., Aktas M., Asiret S., Yormaz S. (2018) Using Generalizability Theory to Assess the Score Reliability of Communication Skills of Dentistry Students. *Asian Journal of Education and Training*, vol. 4, no 2, pp. 85–90. https://doi.org/10.20448/journal.522.2018.42.85.90

Wang D., Liu H., Hau K.T. (2022) Automated and Interactive Game-Based Assessment of Critical Thinking. *Education and Information Technologies*, vol. 27, no 4, pp. 4553–4575. https://doi.org/10.1007/s10639-021-10777-9

Wu M.Y., Steinkrauss R., Lowie W. (2023) The Reliability of Single Task Assessment in Longitudinal L2 Writing Research. *Journal of Second Language Writing*, vol. 59, no 4, Article no 100950. https://doi.org/10.1016/j.jslw.2022.100950

Zhai X., Haudek K.C., Wilson C., Stuhlsatz M. (2021) A Framework of Construct-Irrelevant Variance for Contextualized Constructed Response Assessment. *Frontiers in Education*, vol. 6, October, Article no 751283. https://doi.org/10.3389/feduc.2021.751283