

Доказательный дизайн для оценки универсальных компетенций в высшем образовании: преимущества и особенности

Научная статья

DOI: 10.31992/0869-3617-2025-34-1-82-105

Авдеева Светлана Михайловна – канд. тех. наук, руководитель лаборатории измерения новых конструкторов и дизайна тестов Центра психометрики и измерений в образовании Института образования, SPIN-код: 3183-7208, ORCID: 0000-0002-3599-5138, Researcher ID: ABC-6896-2020, savdeeva@hse.ru

Тарасова Ксения Вадимовна – канд. пед. наук, директор Центра психометрики и измерений в образовании Института образования, SPIN-код: 5121-1559, ORCID: 0000-0002-3915-3165, Researcher ID: ABD-3327-2020, ktarasova@hse.ru

Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
Адрес: 101000, Россия, г. Москва, Потаповский пер., 16, стр. 10

***Аннотация.** Статья посвящена проблеме оценки универсальных компетенций, которые приобретают всё большее значение в условиях современных требований к выпускникам вузов. Рассматривается необходимость создания валидных и надёжных инструментов для измерения таких навыков, как критическое мышление, креативность, коммуникация и способность решать комплексные задачи. Приводится анализ доказательного дизайна, который предполагает интеграцию когнитивных теорий, инновационных образовательных технологий и психометрических моделей для достижения объективной оценки универсальных компетенций.*

Особое внимание уделяется методу доказательной аргументации, который опирается на использование сценарных заданий, контекстуально приближённых к реальным жизненным и профессиональным ситуациям. Метод доказательной аргументации акцентирует внимание на построении валидного аргумента, который связывает наблюдаемые действия студентов с заявляемыми выводами об уровне их компетенций. Метод позволяет не только фиксировать проявления знаний и навыков в приближённых к реальной жизни ситуациях, но и структурировать процесс разработки инструментов таким образом, чтобы полученные эмпирические результаты можно было обосновать теоретически. Важную роль в доказательном подходе играет процесс построения аргумента на основе собранных доказательств, что отличает данный подход от традиционных методов, где акцент делается на статистических и психометрических характеристиках тестов. Изложение основ метода проводится на примере инструмента по оценке критического мышления студентов вузов CT Test, который обладает доказанным психометрическим

качеством. Данный инструмент прошёл апробацию более чем на 10 000 студентов российских вузов.

В статье обосновывается, что внедрение инструментов, основанных на доказательном подходе, позволит значительно улучшить качество оценки универсальных компетенций у студентов, что делает данный подход перспективным для применения в образовательном процессе.

Ключевые слова: доказательный подход в оценивании, метод доказательной аргументации, универсальные компетенции, комплексный латентный конструкт, модель измерения, валидный аргумент

Для цитирования: Авдеева С.М., Тарасова К.В. Доказательный дизайн для оценки универсальных компетенций в высшем образовании: преимущества и особенности // Высшее образование в России. 2025. Т. 34. № 12. С. 82-105. DOI: 10.31992/0869-3617-2025-34-1-82-105

Evidence-Based Design Approach for Assessing Universal Competencies in Higher Education: Advantages and Features

Original article

DOI: 10.31992/0869-3617-2025-34-1-82-105

Svetlana M. Avdeeva – Cand. Sci. (Engineering), Head of the Laboratory for Measuring New Constructs and Test Design, Institute of Education, SPIN-code: 3183-7208, ORCID: 0000-0002-3599-5138, savdeeva@hse.ru

Ksenia V. Tarasova – Cand. Sci. (Pedagogical Sciences), Director of the Centre for Psychometrics and Measurement in Education, Institute of Education, SPIN-code: 5121-1559, ORCID: 0000-0002-3915-3165, ktarasova@hse.ru

National Research University Higher School of Economics, Moscow, Russia

Address: 16, bld. 10, Potapovsky lane, 101100 Moscow, Russian Federation

Abstract. The article addresses the issue of assessing universal competencies, which are becoming increasingly important given modern demands on university graduates. It emphasizes the need for valid and reliable tools to measure skills such as critical thinking, creativity, communication, and complex problem-solving. The paper presents an analysis of an evidence-based design approach, which integrates cognitive theories, innovative educational technologies, and psychometric models to achieve objective assessments of these competencies.

Particular attention is given to the evidence-centered design methodology, which relies on scenario-based tasks closely reflecting real-life and professional situations. This method emphasizes constructing valid argument that link students' observable behaviors to claims about their competence levels. The approach not only captures demonstrations of knowledge and skills in realistic contexts but also structures the instrument development process to ensure that empirical results are theoretically sound. A distinctive feature of this approach is the process of building arguments from collected evidence, differentiating it from traditional methods, which focus primarily on the statistical and psychometric properties of tests.

The principles of this method are illustrated using the CT Test for assessing university students' critical thinking, a tool with established psychometric quality, which has been piloted with over 10,000 students in Russian universities. The article argues that implementing evidence-based tools

will substantially enhance the validity and reliability of competency assessments among students, making this approach a promising candidate for integration into educational practice.

Keywords: evidence-based design approach, evidence-centered design, universal competencies, complex latent construct, measurement model, validity argument

Cite as: Avdeeva, S.M., Tarasova, K.V. (2025). Evidence-Based Design Approach for Assessing Universal Competencies in Higher Education: Advantages and Features. *Vysshee obrazovanie v Rossii = Higher Education in Russia*. Vol. 34, no. 12, pp. 82-105, doi: 10.31992/0869-3617-2025-34-1-82-105 (In Russ., abstract in Eng.).

Введение

Современный рынок труда выдвигает требования наличия у человека не только профессиональных знаний, но и развития у него навыков, которые связаны с успешностью решения задач в различных профессиональных и жизненных областях. Также актуализировался вопрос о формировании и измерении разнообразных видов «новых грамотностей» (цифровой, финансовой, информационной, экологической и т. п.). Таким образом, вызовы современного мира ставят перед системой высшего образования новые практические и исследовательские задачи [1].

Универсальные компетенции, закреплённые во ФГОС 3++¹ и обновляемые в рамках разработки ФГОС нового поколения², соответствуют требованиям рынка труда, а их перечень определяется на основе главных целей образования, структурного представления социального опыта и опыта личности, а также основных видов деятельности студентов, позволяющих им формировать и развивать способности будущей практической деятельности и жизни в современном обществе.

Универсальные компетенции, как и новые грамотности, внепредметны [2]. Это означает, что их формирование и проявление не ограничиваются какой-то одной дисциплиной или сферой деятельности, а

применение – решением конкретной задачи с помощью набора нужных инструментов и сервисов. Универсальные компетенции – это «сквозные» способности, необходимые для решения разнообразных задач в различных профессиональных и жизненных контекстах и включающие совокупность паттернов мышления, поведения и реакций, которые формируются и проявляются на самых разных уровнях под влиянием взаимосвязей и характеристик среды [3].

Согласно хорошо зарекомендовавшей себя модели: образовательная программа – обучение – оценка [4–6], описывающей полный цикл обучения, оценка является неотъемлемой частью процесса обучения [7; 8]. Следовательно, оценивание должно быть направлено на улучшение усвоения студентами не только предметных знаний, но также универсальных компетенций, включающих навыки управления проектами, работу в команде, коммуникацию и эффективное взаимодействие, самоорганизацию и саморазвитие и многие другие. В зарубежной литературе важность оценивания этих навыков, которые чаще называют универсальными или гибкими навыками [9], также подчёркивается множеством экспертов [10; 11].

Автором компетентностного подхода считается Д.К. МакКлелланд, который в своей статье [12] описал результаты исследования, доказывающие, что традиционные академи-

¹ Федеральные государственные образовательные стандарты высшего образования. ФГОС 3.0+++. Министерство науки и высшего образования Российской Федерации. URL: <https://fgos.ru/> (дата обращения: 25.10.2024).

² Приказ Министерства науки и высшего образования Российской Федерации от 2 мая 2023 г. № МН-5/169012. URL: <https://fgosvo.ru/fgosvo/index/31> (дата обращения: 25.10.2024).

ческие тесты способностей и тесты на предметные знания не прогнозировали эффективное выполнение работы или успех в жизни тестируемого. Во введении к книге *Competence at Work: Models for Superior Performance*³ Д.К. МакКлелланд изложил основные характеристики компетенций, акцентировав внимание на том, что компетенции должны быть измеримыми и надёжно отличать лучших работников от средних и худших [13].

С того времени подходы и методы измерения компетенций непрерывно развивались и способствовали, в том числе, возникновению аутентичного оценивания, в процессе которого тестируемые должны продемонстрировать достижение требуемых результатов в ситуациях, максимально приближённых к условиям реальной жизни. Но при всей своей привлекательности аутентичное оценивание распространено сегодня достаточно ограничено, в том числе из-за трудностей и высоких затрат на разработку. Однако использование цифровых технологий, искусственного интеллекта и доказательного подхода к оцениванию помогает преодолевать эти трудности.

С точки зрения измерения универсальные компетенции представляют собой комплексные латентные конструкты [14; 15]. Для их оценки необходимо пронаблюдать за тем, как тестируемые принимают решения и действуют в сложных ситуациях в реальной жизни [16]. Доказательный подход или подход с использованием набора наблюдаемых критериев (*criterion-sampling approach*) в последнее время стал использоваться в дополнение к применяемому ранее аналитическому подходу в измерении комплексных конструктов. В этом подходе акцент делается на описании поведения студентов в разнообразных профессиональных и/или жизненных ситуациях и понимании того, что можно предсказать по результатам анализа

поведения студентов в этих ситуациях. За тем на основе данных анализа из этого набора ситуаций формируется тест, результаты прохождения теста анализируются, чтобы в свою очередь предсказать, как тестируемые будут вести себя в похожих ситуациях. Доказательный подход учитывает теоретически предполагаемую сложность и многогранность процессов рассуждения и мышления, развитие которых и является ключевой целью образования.

Цель данной статьи – представить подход к оценке сложных конструктов с использованием метода доказательной аргументации и аутентичных заданий.

Доказательный дизайн: основные свойства и особенности

Доказательный подход к оцениванию базируется прежде всего на работах С. Мессика [17] и Р.Дж. Мислеви [18], в которых было предложено до начала разработки любого инструмента измерения, опираясь на научные методы, сформулировать ответы на следующие вопросы:

- Что следует оценить, почему и зачем?
- Какой вывод планируется сделать о тестируемом?
- На основе каких доказательств мы сможем сделать планируемый вывод?
- Где и как возможно получить эти доказательства?
- Как мы интерпретируем доказательства, как и на основании чего связываем их с предметом оценки?
- Возможно ли распространить вывод, сделанный на основе этих доказательств, на ситуации, выходящие за рамки тех, что были предложены в процессе оценивания?

Научно обоснованные ответы на вышеприведённые вопросы являются ядром аргумента⁴ и позволяют точно понять, что имен-

³ Книга была переведена на русский язык и издана в 2005 г. в издательстве НРРО: *Спенсер-мл. А.М., Спенсер С.М. Компетенции на работе: пер. с англ. М: НРРО, 2005. 384 с. ISBN: 5-98293-066-0.*

⁴ Аргумент в доказательном дизайне – это научно обоснованное подтверждение того, что результаты оценивания могут быть использованы для сделанного заявления или вывода относительно измеряемых характеристик тестируемого.

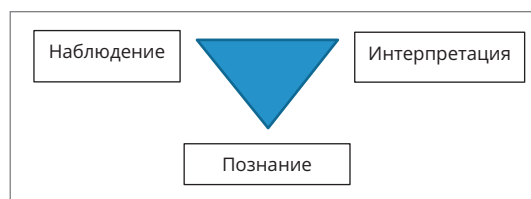


Рис. 1. Схема доказательного мышления (треугольник Пеллигрино)

Fig. 1. Scheme of evidential reasoning (Pelligrino triangle)

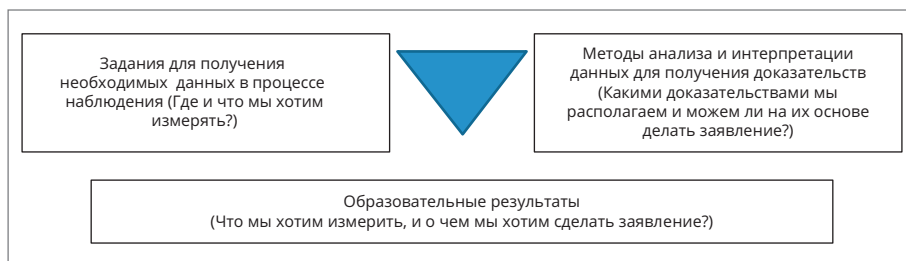


Рис. 2. Модификация треугольника Пеллигрино для разработки инструментов измерения

Fig. 2. Modification of the Pelligrino triangle for the development of measurement tools

но должно оцениваться при помощи разрабатываемого инструмента и доказательно аргументировать сделанные выводы.

Данный подход базируется на правилах формальной логики и аргументации С.Э. Тулмина, который сформулировал идею использования аргумента при доказательствах ещё в середине прошлого века [19]. Он описал структуру аргумента (заявление о тестируемом, данные, полученные при наблюдении за ним, и связи, обеспечивающие переход от данных к заявлению), а С. Мессик [17], Р.Дж. Мислеви [20], Дж.В. Пеллегринно [8] и другие исследователи, опираясь на структуру аргумента, уже представили весь процесс оценивания как процесс аргументации, призванный ответить на вопрос: можем ли мы сделать вывод о компетенциях, тестируемых на основе доказательств, которые мы получили во время тестирования?

Подход, основанный на доказательствах, использовался отдельными опытными разработчиками инструментов измерения и раньше. Однако он не имел прочной связи с развивающимися когнитивными науками и всё более усложняющимся психометрическим аппаратом, применяемым для оцени-

вания, а главное – процесс разработки не был детально описан и задокументирован. Работы Мессика, Мислеви и других исследователей подтолкнули область оценивания к формализации процесса оценивания, основанного на доказательствах. Когнитивные теории, статистический и психометрический аппарат, поведенческая психология и теории обучения при разработке инструментов измерения в этом процессе были связаны воедино.

В 2001 г. группа учёных под руководством Дж. Пеллегринно описали процесс доказательного мышления при обучении в виде простой схемы (Рис. 1) [8].

Для разработки инструментов измерения эту схему можно модифицировать (Рис. 2).

Согласно схеме на рисунке 2, в основе любого инструмента измерения лежат три связанных и обязательных компонента: измеряемые образовательные результаты; совокупность представлений о ситуациях и действиях тестируемого, в которых будут проявляться эти результаты (задания), а также методы и инструменты анализа, объясняющие, почему и как данные в процессе наблюдения за тестируемым связаны с оце-

ниваемыми образовательными результатами и с заявлением о тестируемом.

Таким образом, доказательства, поддерживающие измеряемую универсальную компетенцию и заявление, которое мы хотим сделать о способностях тестируемого на основании данных, имеют в доказательном дизайне ключевое значение. И именно это в первую очередь отличает его от традиционных подходов к разработке инструментов измерения, в которых предпочтение отдаётся психометрическим характеристикам разрабатываемого инструмента. Безусловно, анализ данных и психометрические методы много значат и в доказательном дизайне. Но в нём приоритет отдаётся доказательствам и построению аргумента на их основе. Рассматривать оценку как процедуру построения аргумента становится очевидным, как только осознаётся факт того, что валидность, одна из основных характеристик оценки, заключается в «степени, в которой эмпирические доказательства и теоретические обоснования подтверждают адекватность и уместность выводов и действий, основанных на результатах тестов или других форматах оценки» [21].

Доказательный дизайн – это комплексный подход, на основании которого разработано несколько методов оценивания [22]. Далее подробно рассмотрим метод доказательной аргументации (*Evidence-centered design, ECD*) [23–25]. Причиной выбора именно этого метода является успешный опыт разработки авторами статьи на его основе целого ряда инструментов измерения комплексных конструктов (критического мышления, коммуникации, информационно-коммуникационной компетентности, цифровой и правовой грамотности и др.) с доказанным качеством результатов измерения [26–28].

Метод доказательной аргументации

Метод доказательной аргументации – это систематический подход к разработке инструментов измерения, соответствующий принципам доказательного дизайна и вклю-

чающий набор взаимосвязанных процедур, описывающих оцениваемые образовательные результаты (знания, навыки, способности, компетенции), доказательства, которые подтверждают их наличие или отсутствие, и средства (инструменты среды), которые используются в качестве стимула проявления доказательств, а также методы их последующего анализа.

Метод доказательной аргументации можно использовать для создания различных инструментов измерения: от тестов с несколькими вариантами ответа для оценки знаний и простых навыков до инструментов оценивания комплексных латентных конструктов – критического мышления, решения проблем, креативного мышления и др. Чем более комплексный конструкт требуется оценить, тем более необходимо использование метода доказательной аргументации при разработке инструмента его измерения для получения результатов с доказанным качеством.

Упрощённо метод доказательной аргументации может быть изображён в виде схемы, представленной на *рисунке 3*.

Полная структура метода доказательной аргументации состоит из пяти этапов, которые основоположники метода называют «слоями» (*layers*) [29]. Разделение на этапы способствует тщательной проработке измеряемых конструктов, особенно если это комплексные конструкты с латентными составляющими, и внесению необходимых изменений на ранних этапах разработки инструмента.

Таким образом, разработка любого инструмента оценивания с использованием метода доказательной аргументации состоит из следующих этапов.

1. Анализ области оценивания (*Domain Analysis*): изучение теоретической информации о предметной области; формулирование целей оценивания и условий использования инструмента; описание структуры измеряемого конструкта.

2. Моделирование области оценивания (*Domain Modeling*): выявление отношений



Рис. 3. Схема метода доказательной аргументации

Fig. 3. Scheme of evidence-centered design

между составляющими измеряемого конструкта и видами деятельности, в которых они проявляются, ситуациями выполнения задач и доказательствами, необходимыми для достижения целей оценки; определение уровня достижений измеряемого конструкта и его составляющих.

3. Теоретическая (концептуальная) рамка оценивания (*Conceptual Assessment Framework*): разработка моделей, позволяющих выявлять, собирать и накапливать доказательства для формирования заявления об уровне сформированности исследуемого конструкта у тестируемого.

4. Реализация инструмента измерения (*Assessment Implementation*): разработка инструмента на основе описания паттерн-дизайнов, спецификаций, уровней достижений, моделей концептуальной рамки и другой информации, полученной на предыдущих этапах.

5. Применение инструмента измерения (*Assessment Delivery*): взаимодействие тестируемых с разработанным инструментом измерения для получения доказательств, необходимых, чтобы сформировать заявление и предоставить обратную связь о результатах оценивания.

Хотя выше эти этапы перечислены последовательно, важно отметить, что, как правило, работа над несколькими этапами ведётся одновременно, изменения на одном этапе часто приводят к изменениям в другом, что особенно часто происходит на ранних эта-

пах реализации метода доказательной аргументации. Рассмотрим их подробнее.

Анализ области

На этом этапе тщательно собирается, исследуется и документируется существующая информация о конструкте, который мы хотим измерить. Описываются цели оценки, как с точки зрения причин собственно процедуры оценки, так и с учётом всех аспектов использования результатов оценивания. Затем, исходя из целей оценки, изучается теоретическая литература из области когнитивных наук, психологии, теории обучения и развития [30]. Цели оценки не только служат основой для разработки инструмента, на них опирается построение аргумента для обоснования интерпретации и использования результатов оценивания. Далее выявляются компоненты, из которых состоит исследуемый конструкт, определяются способности, которые нужны для его проявления [31]. На этом этапе очень важно взаимодействие разработчиков инструмента измерения со специалистами в области исследуемого конструкта, так как важно не просто детально исследовать «содержание» области исследуемого конструкта, а понять, как студенты используют это содержание при решении задач в реальной жизни.

Моделирование области

На этом этапе информация, полученная в результате анализа, структурируется и формализуется. Ставятся следующие

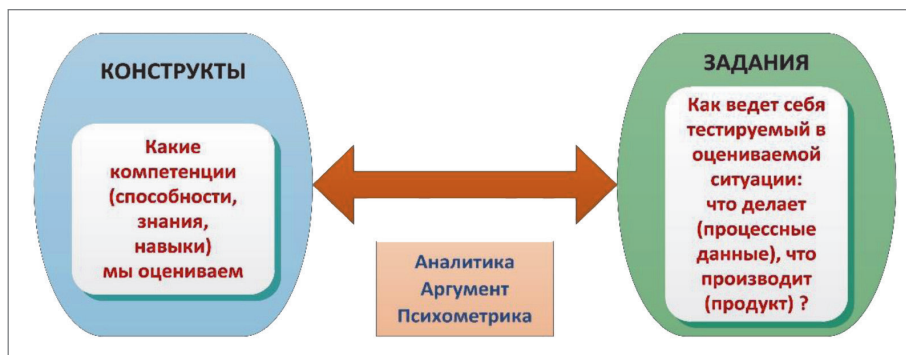


Рис. 4. Теоретическая рамка оценивания

Fig. 4. Conceptual Assessment Framework

вопросы: в каких отношениях находятся компоненты интересующего нас явления; в каких ситуациях каждый из них может проявиться; в каких отношениях эти проявления находятся друг с другом; какие из этих компонентов и их проявлений важны с учётом целей, которые мы ставим; как будут использоваться и интерпретироваться результаты оценивания. Для поиска и формулировки ответов на эти вопросы могут создаваться концептуальные карты (онтологии), учебные прогрессии, иерархические схемы, строится модель аргумента. Также проводится процедура, имеющая название «дизайн паттернов». Суть этой процедуры состоит в составлении таблицы, в которой указаны принципиальные при оценивании знания, навыки, способности, для которых необходимо создать возможность их проявления при разработке тестовых заданий. Также на этом этапе принимаются решения о том, как результаты оценивания будут интерпретироваться. Как уже упоминалось, один из способов принятия и реализации этих решений – разработка показателей достижения уровня (*Achievement Level Descriptors / Performance Level Descriptors, ALDs, PLDs*). Результаты, полученные в процессе работы на этом этапе, тщательно описываются в наиболее удобной форме. Например, результаты дизайна паттернов и показатели достижения уровней обычно представляются в виде таблиц.

Теоретическая рамка (структура) оценивания

Теоретическая рамка оценивания призвана соединить результаты анализа и моделирования с прикладными процессами, которые происходят при разработке и сдаче в эксплуатацию конкретного инструмента измерения [22; 23]. Аргумент оценки, изложенный в повествовательной форме на предыдущем этапе, при разработке теоретической рамки приобретает форму спецификации – операционального проекта инструмента, включающего создание основных четырёх моделей: модель конструкта (*Construct Model*); модель задания (*Task Model*); модель доказательств (*Evidence Model*) и модель сборки инструмента (*Assembly Model*).

Обобщённо теоретическая рамка оценивания может быть представлена в виде схемы на рисунке 4.

Рассмотрим основные характеристики и структуру моделей теоретической рамки оценивания.

Модель конструкта

При разработке модели конструкта (Рис. 5) определяется, какие из элементов обобщённой модели, сформированные на этапе моделирования области компетенций, будут подлежать оцениванию разрабатываемым инструментом и, при необходимости, корректируется степень их детализации. Кроме того, при создании модели конструк-

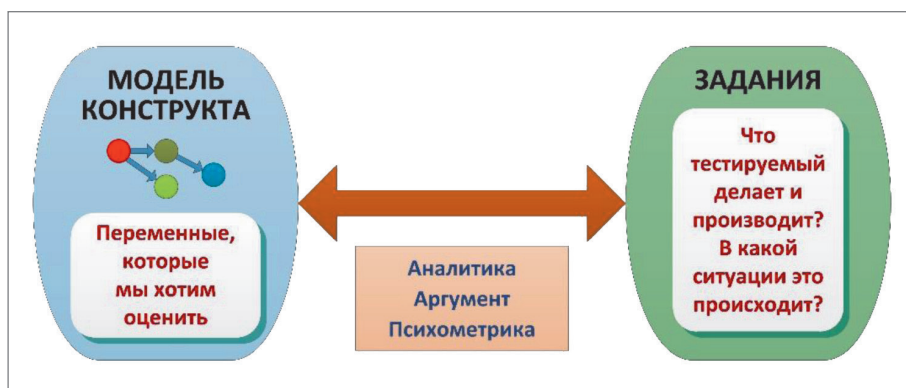


Рис. 5. Модель конструкта

Fig. 5. Construct Model

та учитывается дополнительная информация о целевой аудитории, для которой разрабатывается инструмент оценивания. Модель конструкта также учитывает, как студенты продвигаются в его освоении от поверхностных знаний к более глубокому пониманию и применению освоенных знаний.

Здесь и далее реализация основных моделей теоретической рамки оценивания будет приводиться на примере разработанного инструмента оценки критического мышления (КМ) студентов вузов *CT Test*, который обладает доказанным психометрическим качеством.

На основе анализа существующих инструментов измерения и синтеза основных теоретических и эмпирических исследований в области КМ [32–34] было определено, что критическое мышление является комплексным латентным конструктом, и для его описания в зарубежной и отечественной традиции выделяются схожие элементы, даже если для их описания используют разные термины.

Изначально исследования критического мышления опирались на философскую традицию изучения познания, что позволяло выделить чёткую теоретическую структуру исследуемого конструкта. Этим обусловлен тот факт, что большинство стандартизированных методик (*California Critical Thinking Skills Test (CCTST)*, *Cornell Critical Thinking Test (CCTT)*, *Ennis — Weir Critical Thinking Essay Test (EWCIET)*, *HEIghten Critical Thinking*

Assessment и др.) для оценки критического мышления созданы на основе философского подхода. В образовательной практике наибольшее внимание уделялось формированию критического мышления, а в рамках психологического подхода – теоретической проработке и ориентации на описание и объяснение глубинных процессов, определяющих критическое мышление. Учитывая вышеперечисленные подходы, при операционализации конструкта КМ и разработке инструмента его измерения было принято решение проводить оценку критического мышления студентов через релевантное наблюдаемое поведение. Это позволило использовать компьютерную форму, а на этапе проектирования модели конструкта сосредоточиться на таких аспектах критического мышления, которые могут быть интегрированы в цифровую среду посредством наблюдаемых индикаторов.

Таким образом, на этом этапе было конкретизировано определение критического мышления – последовательность когнитивных действий, направленных на: оценку качества исходной информации с целью определения проблемы; поиск возможных решений и выбор наилучшего из них; обоснование собственного вывода и выявление его ограничений.

На основе определения с привлечением экспертов в данной области были выделены следующие ключевые элементы.

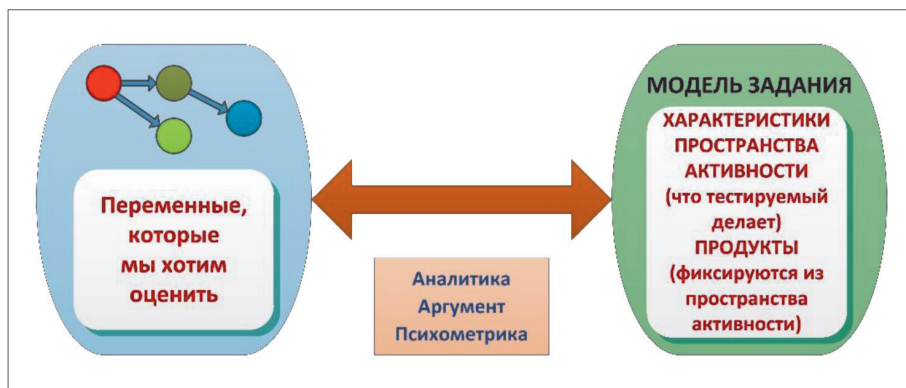


Рис. 6. Модель задания

Fig. 6. Task Model

1. Проверка исходной информации:

- определение различных категорий информации:
 - факты, данные, описание событий;
 - мнения, интерпретации, спекуляции;
- выделение ключевых терминов/понятий/определений;
- определение актуальности и релевантности информации;
- оценивание компетентности и авторитетности источников информации.

2. Суждение и выдвижение гипотез:

- выделение основной мысли автора или проблемы (суждение автора);
- подвержение сомнению аргументированности утверждений (выводов) автора:
 - выявление явных и неявных предположений автора и недостаточности исходной информации;
 - оценивание логичности и последовательности причинно-следственных связей, предложенных автором;
 - определение неоднозначности выводов автора;
 - демонстрация, если возможно, абсурдности аналогичных аргументов;
- выдвижение новых (альтернативных, в т. ч. радикальных) гипотез.

3. Анализ:

- формулирование анализируемой гипотезы;
- применение моделей, концепций, подходов, методов;
- обоснование валидности выводов⁵;
- формулирование надёжных (истинных)⁶ выводов.

4. Рефлексия:

- выявление ограничений вывода вследствие допущенных предположений (авторских и собственных);
- оценивание степени неопределённости и с учётом этого корректирование выводов;
- оценивание преимуществ и недостатков вывода перед альтернативными.

Также для критического мышления на этапе моделирования конструкта было определено три возможных уровня проявления – базовый, высокий, продвинутый, и описаны их характеристики.

Модель задания

Модель задания (Рис. 6) описывает ситуацию, в которую разработчики инструмента измерения собираются поместить тестируемого для того, чтобы он продемонстрировал навыки, отражённые в модели конструкта.

⁵ Валидные выводы – это логические непротиворечивые умозаключения, основанные на сделанных предпосылках.

⁶ Надёжные выводы – это валидные выводы, основанные на проверенных (истинных) предположениях (предпосылках).

тонн. В Астраханской области, лидере по выращиванию хлопка в России, производство только начинает демонстрировать рост.

Дополнительной проблемой для текстильщиков стал рост затрат на комплектующие. "Рынок материалов с октября прошлого года повысился на сегодняшний день процентов на 100–150, а местами и на 200%. Возьмем даже простейшую вещь – металлическую фурнитуру, октябрь прошлого года – однозамковая молния стоит 60 рублей, сейчас – плюс 250%: молниины – рост цены до 100%, пуговицы – удорожание на 100–120%", – сказала основательница собственного стартапа дизайнерской одежды из Иваново Ирина Карпова. "Подобный всплеск цен на комплектующие был, на моей памяти, лишь в 2015 году, когда резко подорожали гвозди".

Ситуация с ростом цен на сырье приведет к повышению цен на одежду, уверены производители. «С начала года хлопок подорожал приблизительно на 100%, а готовое полотно – на 40–50%, поскольку его стоимость определяет не только сырье, но расходы на оплату труда, которые не сильно изменились», – сказала Анна Лебедева, генеральный директор Fashion Consulting Group, подчеркивая, что в цене растут не только хлопок, но и смешанные ткани, а также ткани из синтетических и искусственных волокон.

Заместитель генерального директора по правовому регулированию и коммерческому развитию ООО «Трейд Менеджмент» Игорь Гришин подтвердил, что с октября прошлого года компания повышала расценки на производимую одежду единожды – на 18%, на при

Задание 1.

Выделите в предложенном фрагменте статьи ТРИ предложения, содержащие **нерелевантную контексту статьи** информацию. Перенесите их в правую часть экрана.

*ОБРАТИТЕ ВНИМАНИЕ, что под релевантностью информации в данном случае понимается степень соответствия основному содержанию текста. Релевантная информация необходима для понимания смысла статьи, тогда как **нерелевантная информация является избыточной**.

Нужно выбрать в левой части предложение и кликнуть на него, чтобы оно попало в правую часть

Дальше

Рис. 7. Пример задания СТ Test

Fig. 7. Example of a CT Test task

Как представлено на рисунке 6, в модели задания описывается среда, в которой будет производиться оценивание, например, программа создания презентаций для подготовки выступления, статья, которую надо прочитать и на основании прочитанного текста выполнить какие-то действия, садовый участок, который надо обустроить и засадить нужными растениями и т. п. Перечень действий, которые надо выполнить тестируемому в процессе выполнения задания составляют характеристики пространства активности.

Как правило, эта модель включает следующие элементы [22]:

1) подробно описанные свидетельства, детализированные до нужной степени, – что разработчики инструмента измерения хотят наблюдать;

2) тип стимула или материалы, которые будут использоваться, чтобы вызвать необходимое поведение;

3) описание того, что именно предложат выполнить тестируемому в ходе оценки;

4) описание элементов или технических характеристик, которые должны присутствовать в задании, чтобы тестируемый мог продемонстрировать те действия, которые предполагается наблюдать;

5) элементы, которые повлияют на сложность и комплексность заданий;

6) примеры заданий, которые могут быть созданы с использованием этой модели.

Модель задания помогает разработчикам тестовых заданий определить, задания какого типа смогут «выявить» оцениваемые компетенции в наилучшей форме.

Модель задания для оценки критического мышления студентов создавалась на основе шаблонов проектирования (*Pattern Design*) [35; 36]. Разработанный шаблон помог заполнить пробелы аргумента в отношении измеряемого конструкта. При разработке также учитывались многообразные составляющие такого комплексного конструкта, как критическое мышление и необходимость оценивать разные модели поведения тестируемых [14], поэтому при разработке *CT Test* в каче-

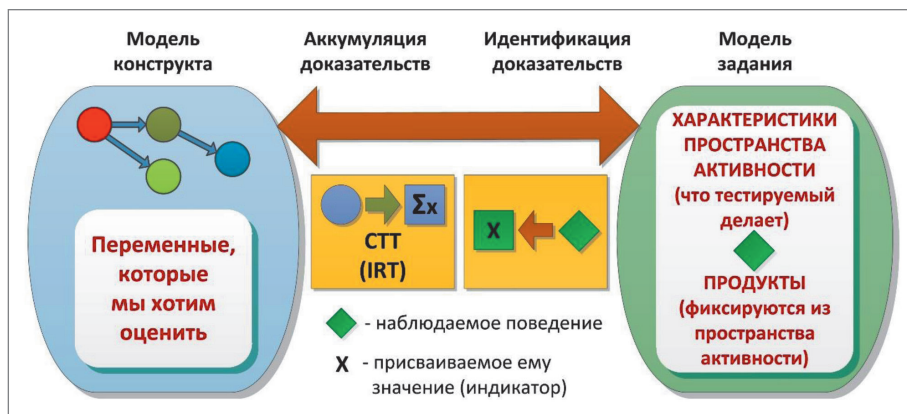


Рис. 8. Модель сбора свидетельств

Fig. 8. Evidence Model

стве модели задания была выбрана форма, в которой респонденту следует выполнить те или иные действия (*performance tasks*), причём эти действия комплексные, происходят в максимально аутентичном контексте [37].

Стимульным материалом стала доработанная новостная заметка из реально существующего издания (пример представлен на рисунке 7). В процессе работы с заметкой необходимо было произвести первичный анализ информации, выдвинуть свои гипотезы, сделать выводы в соответствии с заданиями, которые последовательно представлялись респондентам. Выбор механики каждого задания (выделение фрагментов текста, выбор утверждения, классификация, сопоставление) был обусловлен тем, насколько хорошо производимое действие отражает заложенную переменную измеряемого конструкта. Такая модель задания является подходящей средой для стимуляции поведения тестируемого, соответствующего измеряемому конструкту [32].

Модель сбора доказательств

Модель сбора доказательств (Рис. 8) содержит компоненты, отвечающие за идентификацию и накопление доказательств.

Каждый набор последовательности действий имеет определённое значение для процедуры оценивания. Объединение набора действий и их анализ с целью получения

доказательств для заявления, которое мы хотим сделать, относится к психометрической части разработки инструмента, и представляет собой процесс аккумуляции доказательств.

Для *CT Test* модель сбора доказательств представлена ниже на примере субконструкта критического мышления «Проверка исходной информации».

Каждый продукт деятельности, представленный в таблице 1, в последующем при непосредственной разработке заданий *CT Test* служил источником для определения индикаторов. Они оценивались дихотомически (0/1) или полиномически, предполагая частичное выполнение (0/1/2). Правила начисления баллов по каждому индикатору устанавливались при участии экспертов. А при выборе статистических методов анализа данных и формирования обратной связи разработчики инструмента ориентировались как на выбранную модель задания, так и на тип индикаторов. Затем для оценки степени согласия теоретически ожидаемой факторной структуры и структуры эмпирически полученных данных в рамках структурного моделирования был выбран конфирматорный факторный анализ (КФА). Этот метод достаточно гибкий, чтобы учитывать как композитный характер латентного конструкта (КМ), так и возможные до-

Таблица 1
Table 1

Пример модели сбора доказательств для CT Test

Example of an Evidence Model for CT Test

Составляющая КМ	Свидетельство	Кодировка	Связь с уровнями КМ			Продукты деятельности (<i>work products</i>)
			базовый	высокий	продвинутый	
1. Проверка исходной информации	Определение различных категорий информации	Verification text VT1	–	Может различить факты и мнения, позитивные и нормативные суждения	Может различить факты и мнения, позитивные и нормативные суждения	Выделяет в тексте (механика: <i>highlight/move/underline</i>) информацию, относящуюся к определённой категории (факты, интерпретации, спекуляции, отношения (эмоция) / оценочное суждение, нормативные суждения (мнения), позитивные суждения (прогнозы)
	Выделение ключевых терминов/понятий/определений	VT2	Может определить основную мысль и ключевые термины статьи	Может определить основную мысль и ключевые термины статьи	Может определить основную мысль и ключевые термины статьи	Верно выделенные в тексте ключевые термины, понятия, определения. Используются в поисковом запросе ключевые слова (понятия, определения, термины) относятся к теме задания в эссе/ для поиска информации
	Уточнение ключевых терминов/понятий/определений	VT3	–	Может уточнять термины	Может уточнять термины	Уточнённые ключевые термины/понятия/определения. Возможно использовать в заданиях в открытой среде, предполагается использование браузера
	Определение актуальности информации	VT4	Может оценить актуальность исходной информации	Может оценить актуальность исходной информации	Может оценить актуальность исходной информации	Проведена оценка информации из источника(ов) по степени актуальности. Приводит актуальные ссылки. Проверяет актуальность ссылок, которые даны в задании. Верно ранжирует и оценивает свои ссылки по степени актуальности

Продолжение таблицы 1 / Continuation of table 1

Составляющая КМ	Свидетельство	Кодировка	Связь с уровнями КМ			Продукты деятельности (<i>work products</i>)
			базовый	высокий	продвинутый	
	Определение достоверности информации	VT5	-	-	Может оценить достоверность информации и степень её непредвзятости	Оценивает информацию из источника(ов) по степени достоверности (утверждение). Выделяет в тексте предположения, требующие дополнительной проверки. Проверяет достоверность информации (как в <i>իրի</i> тексте, так и в Интернете), переходя по нескольким ссылкам. Верно ранжирует и оценивает свои ссылки по степени достоверности
	Определяет релевантность информации	VT6	-	-	Может оценить релевантность источников	Оценивает информацию на предмет релевантности – приводит релевантные ссылки (по теме данного парадокса), ранжирует их
	Оценивает компетентность источников информации	VT7	-	Может оценить компетентность источников	Может оценить компетентность источников	Оценивает источники по степени компетентности. Выделяет компетентные источники. Приводит ссылки, после завершения эссе необходимо их оценить. Верно оценивает компетентность источников – смонтирует на сайт и делает выбор, остаются на нём или закрыть.
	Оценивает авторитетность источников	VT8	-	Может оценить авторитетность источников	Может оценить авторитетность источников	Оценивает источники информации по степени авторитетности (и автор, и ресурс). Выделяет авторитетные источники

полнительные источники корреляций между индикаторами [38].

В некоторых случаях при создании инструментов оценивания детализация доказательства должна быть ещё более подробной, чем в представленном примере выше, чтобы оценить конструкт в полной мере. Например, в случае оценки конструктов «Коммуникация», «Коллаборация», «Совместное решение проблем» респонденту необходимо предоставить возможность продемонстрировать свои навыки эффективного взаимодействия и коммуникации с другими людьми. Для этого создаётся специальная среда, в которой респонденты взаимодействуют друг с другом или с компьютерными агентами (аватарами) в синхронном или асинхронном режиме [39; 40]. В такой среде можно фиксировать все действия респондентов, а не только финальные ответы.

Суммируя описание и пример выше, можно сказать, что модель сбора свидетельств связывает модель конструкта – то, что мы хотим оценить, и модель задания – то, где мы хотим это оценить. В итоге модель сбора свидетельств предоставляет инструкции для последующей интерпретации действий тестируемого и результатов оценивания и состоит из следующих компонент.

1. Результаты, действия, объекты (продукты, *work product*), которые тестируемый произведёт в ходе оценки и которые позволят сделать желаемое заявление;

2. Правила доказательства (*evidence rules*) – каждый продукт будет содержать некоторое количество индикаторов, которым приписаны определённые значения. Правила доказательств указывают, как идентифицировать и оценивать эти индикаторы.

3. Измерительная модель (*measurement model*), которая, собственно, и отвечает за аккумуляцию доказательств. Она связывает модель конструкта с получаемыми индикаторами. Самая простая модель измерения – это сумма или процент ответов, соответствующих ключу в тесте с несколькими вариантами ответа. Более сложные

модели – это модели современной теории тестирования (*Item Response Theory – IRT*) или сети Байеса. Как уже было сказано выше, выбор измерительной модели зависит от целей оценивания и типа создаваемого инструмента, для комплексных конструктов предпочтительней использовать модели *IRT* (политомические, многомерные, многофасетные и т. п.) или сети Байеса.

Модель сборки

Модель сборки включает в себя правила и критерии, которые определяют, какие задания должны быть включены в тест, и как они должны соотноситься с целями оценивания. Это помогает равномерно распределить задания по различным аспектам составляющих оцениваемого конструкта и обеспечивает справедливость оценивания, а также позволяет создать тест, где все задания логически связаны и соответствуют общей структуре и целям.

Рассмотрим пример, основанный на представленном выше *CT Test*. Целью теста является оценка способности студентов анализировать качество исходной информации для определения проблемы, искать возможные решения и выбирать наилучшее из них, обосновывать собственный вывод и выявлять его ограничения.

Следуя модели сборки, в тестовый вариант были включены задания, которые относятся к каждому из четырёх субконструктов критического мышления, определённых на этапе моделирования. Включены задания различного уровня сложности, чтобы обеспечить возможность дифференциации студентов по уровню критического мышления. Кроме того, в структуре предъявления заданий реализуется теоретически заложенная модель конструкта – линейная последовательность когнитивных действий от первичного анализа информации в заметке до этапа рефлексии. Таким образом, все задания логически связаны и соответствуют общей цели теста – оценке критического мышления.

Модель сборки обеспечивает структурированность, согласованность, сбалансированность инструмента измерения, что особенно важно при разработке инструментов, направленных на оценку комплексных конструкций, а также способствует получению валидных и надёжных результатов оценки.

Реализация инструмента оценивания

На этом этапе на основе моделей, описанных выше, создаётся работающий прототип инструмента. Затем происходит апробация с участием выбранной целевой аудитории, включая проведение когнитивных лабораторий и пилотных тестирований. В ходе апробации проверяется, насколько созданные ранее модели соотносятся с реальностью, соответствует ли работа инструмента целям оценивания. В том случае, если работа инструмента не соответствует установленным требованиям, начинают вносить изменения в задания или дорабатывать алгоритмы оценивания, например, удалять слишком трудные или слишком простые задания.

Применение инструмента тестирования

Этот этап предназначен для финальной подготовки инструмента измерения к передаче в эксплуатацию. Если результаты, полученные в ходе апробации, подтверждают качество разрабатываемого инструмента, он начинает полноценно использоваться на реальной целевой аудитории, результаты, полученные в ходе измерения, интерпретируются и документируются, студенты получают обратную связь по результатам тестирования. Доставка инструмента реализуется посредством системы, состоящей из четырёх модулей (*Delivery Model*): выбор активности (*Activity Selection*), презентация (*Presentation*), обработка ответа (*Response Processing*) и финальная обработка результатов (*Summary Scoring*) [41]. Процессы, происходящие в рамках этой системы, цикличны и образуют круг администрирования при проведении тестирования.

1. Выбор активности – выбор задания (вопроса, действия и т. д.), с которым будет работать тестируемый. Выбор осуществляется в соответствии с моделью сборки. При применении традиционного «линейного» инструмента, выбор всех заданий может происходить до того, как тестируемый начнёт работу с ними. В случае, если инструмент «адаптивный», компьютер обращается в созданную библиотеку тестовых заданий и извлекает из неё наиболее подходящее данному тестируемому задание. Например, если тестируемый правильно ответил на вопрос ранее, то следующий извлечённый вопрос будет более трудным или комплексным.

2. Презентация – задание предоставляется тестируемому, и ответ, который дал тестируемый, или действие, которое он выполнил, фиксируются. Ответ (или действие), которое совершает тестируемый, передаётся для обработки ответа.

3. Обработка ответа – это первый этап определения и выставления баллов (*Scoring Process*). На этом этапе идентифицируются и оцениваются продукты деятельности, зафиксированные ранее. Они будут предоставлять доказательства о навыках, способностях и знаниях, которые проверяются посредством инструмента.

4. Итоговая обработка результатов тестирования – разделение процессов, о котором говорилось выше, позволяет предоставлять тестируемому разные виды обратной связи – обратную связь по результатам выполнения одного задания и итоговую обратную связь (уровень сформированности измеряемого конструкта и его составляющих) по результатам выполнения всех заданий. Итоговая обратная связь – это заявление, основанное на аккумуляции всех доказательств, полученных из всех заданий. Валидная и надёжная обратная связь по результатам обучения необходима на всём процессе освоения образовательной программы, но особенно важным качеством измерения становится при оценивании «с высокими ставками», когда результаты имеют

существенные последствия, например, при поступлении на новый уровень образования, получении профессионального сертификата или победе в конкурсе при приёме на работу.

Обсуждение

Выше была описана базовая структура метода доказательной аргументации. В первом приближении она может показаться достаточно сложной для освоения, а разработка инструментов измерения в парадигме доказательного оценивания – ресурсозатратной. С одной стороны, этот метод доказательной аргументации действительно подразумевает разработку большого количества документации, соблюдение процедур, требующих значительного объёма работ на каждом из этапов разработки инструмента. Однако благодаря этим особенностям вышеописанный метод имеет и ряд значимых преимуществ.

Основное преимущество состоит в том, что он позволяет построить прочные и обоснованные связи между целями, для которых используются результаты оценивания, тестовыми заданиями разработанного инструмента, а также применяемым психометрическим аппаратом, то есть помогает встраивать валидность в процесс проектирования и разработки теста. Инструмент измерения разрабатывается в процессе осознанного взаимодействия между экспертами из предметных областей, разработчиками заданий и психометриками. С самого начала работы над инструментом устанавливается, что именно он будет оценивать, каким образом и какие выводы можно сделать по результатам оценивания об уровне исследуемых компетенций. Таким образом метод доказательной аргументации помогает гарантировать, что тестовые задания измеряют релевантные для конструктора компетенции.

Ещё одно преимущество заключается в том, что применение таких элементов доказательного дизайна, как разработка моделей, паттерн-дизайнов и применение поэтапной архитектуры позволяют создавать

параллельные формы заданий [42] и ускорять процесс формирования банка заданий за счёт их автоматической генерации.

Разработка инструментов измерения в парадигме доказательного подхода предполагает соблюдение на каждом этапе следующих основополагающих принципов.

- Сбор и накопление доказательств валидности инструмента на протяжении всех этапов его разработки – главный, «системообразующий» принцип. Он основывается на понимании оценивания как построения аргумента. То есть интерпретация результатов оценивания, психометрические характеристики инструмента, контент тестовых заданий, заявления, которые разработчики инструмента измерения делают об уровне компетенций, – все эти параметры, полученные из разных источников, должны поддерживаться доказательствами для создания единого аргумента.

- Чёткое формулирование цели разработки инструмента и соответствие всех процессов его разработки заявленной цели. То есть измеряемые компетенции, которые планируется оценить при помощи разрабатываемого инструмента, должны быть чётко определены и детально описаны в самом начале работы над инструментом.

- Использование или разработка когнитивных моделей, моделей обучения и/или моделей демонстрации компетенций, которые планируются к оценке. Они помогут обосновать выбранный способ оценивания и связать поведение студента в процессе оценки и его поведение в реальной жизни, то есть позволят правильно интерпретировать и использовать результаты оценивания.

- Использование психометрического аппарата, который поможет достичь заявленной цели оценивания, выбрать нужный способ интерпретации и использования результатов. Психометрический аппарат должен помогать выявлять, извлекать, накапливать и интерпретировать доказательства, он может варьироваться от классических психометрических моделей до многомерных

современных моделей в случае аутентичного и адаптивного оценивания [43].

- Соответствие тестовых заданий, стимульного материала и разнообразных видов активности разрабатываемого инструмента сформированным целям, способу интерпретации и использования результатов измерения. Для рассматриваемых комплексных конструкторов предпочтительней использовать аутентичные задания сценарного типа (*performance based tasks*) с реальным профессиональным или жизненным контекстом [44].

Как было показано выше, при доказательном оценивании не применяется ничего, что бы противоречило традиционному процессу разработки инструментов измерения. Однако при использовании доказательного дизайна все аспекты разработки определяются более детально и структурированно и направлены на формирование валидного аргумента.

Доказательный дизайн разработки инструментов измерения продолжает развиваться и использоваться за пределами традиционных форм образовательного оценивания [45], расширяя концепции и методы на новые формы, которые включают взаимодействие тестируемых в процессе оценки в имитационных и игровых средах, автоматизированную оценку результатов и многие другие. Однако все появляющиеся новые формы должны учитывать следующие основополагающие аспекты: социокогнитивные теории развития способностей и то, как они применяются для успешной жизни в реальном мире, а также создание аргументов оценки, построенное на основе выявленных доказательств.

Заключение

Перечень универсальных компетенций претерпевает изменения в связи с разработкой ФГОС нового поколения. Поскольку формирование универсальных компетенций является обязательной целью обучения, сформулированной в ФГОС высшего обра-

зования, то для их оценивания необходимы инструменты измерения с доказанным качеством. Использование таких инструментов в качестве элементов образовательных систем позволит целенаправленно формировать и развивать эти компетенции у студентов и выпускников университетов, а также установить стандарты, сопоставимые на международном уровне.

Применение доказательного подхода к оценке универсальных компетенций позволит обеспечить высокую степень точности и объективности оценок, а также улучшить качество образовательного процесса, так как результаты оценки предоставляются студентам и преподавателям в виде обратной связи, что позволит студентам понять свои сильные и слабые стороны, а преподавателям – скорректировать свои методы обучения для достижения обучающимися лучших результатов. Кроме того, этот подход позволит не только оценить текущий уровень их подготовки, но и выявить пробелы, требующие дальнейшего развития, что в итоге будет способствовать повышению качества образования и подготовке специалистов, востребованных современным рынком труда.

Литература

1. Разумова Т.О., Телешова И.Г. Трансформация системы высшего образования: вызовы и перспективы // Уровень жизни населения регионов России. 2023. Т. 19. № 3. С. 338–349. DOI: 10.52180/1999-9836_2023_19_3_3_338_349
2. Care E., Kim H., Vista A., Anderson K. Education System Alignment for 21st Century Skills: Focus on Assessment. The Center for Universal Education at the Brookings Institution, 2018. URL: https://www.researchgate.net/publication/330740772_Education_system_alignment_for_21st_century_skills_Focus_on_assessment (дата обращения: 25.10.2024).
3. Mislevy R.J. Sociocognitive Foundations of Educational Measurement. Routledge, 2018. DOI: 10.1111/jedm.12255
4. Шмигирилова И.Б., Рванова А.С., Григоренко О.В. Оценивание в образовании: современные тенденции, проблемы и противоречия

- (обзор научных публикаций) // Образование и наука. 2021. Т. 23. № 6. С. 43–83. DOI: 10.17853/1994-5639-2021-6-43-83
5. *Achtenbagen F.* The curriculum-instruction-assessment triad // *Empirical Research in Vocational Education and Training*. 2012. No. 4. P. 5–25. DOI: 10.1007/BF03546504
 6. *Biggs J., Tang C.* Teaching for Quality Learning at University: What the Student Does. New York: McGraw-Hill Education, 2011. 480 p. ISBN 10: 0335242758.
 7. *Mislevy R.J.* Four Metaphors We Need to Understand Assessment. The Gordon Commission on the Future of Assessment in Education, 2013. 36 c. URL: https://www.ets.org/Media/Research/pdf/mislevy_four_metaphors_understand_assessment.pdf (дата обращения: 25.10.2024).
 8. *Pellegrino J.W., Chudowsky N., Glaser R.* Knowing what students know: The science and design of educational assessment. Washington, DC: National Academy Press, 2001. URL: https://www.researchgate.net/publication/270584995_Knowing_What_Students_Know_The_Science_and_Design_of_Educational_Assessment (дата обращения: 25.10.2024).
 9. *Lamri, J., Lubart T.* Reconciling Hard Skills and Soft Skills in a Common Framework: The Generic Skills Component Approach // *Journal of Intelligence*. 2023. No. 11. Article no. 107. DOI: 10.3390/jintelligence11060107
 10. Education for Life and Work: Developing transferable knowledge and skills in the 21st century: ed. by *Pellegrino J.W., Hilton M.L.* Washington, DC: National Academies Press, 2012. 256 p. DOI: 10.17226/13398
 11. *Shavelson R.J., Zlatkin-Troitschanskaia O., Mariño J.P.* International performance assessment of learning in higher education (iPAL): Research and development // *Assessment of Learning Outcomes in Higher Education*. 2018. P. 193–214. DOI: 10.1007/978-3-319-74338-7_10
 12. *McClelland D.C.* Testing for competence rather than for “intelligence” // *American Psychologist*. American Psychological Association. 1973. Vol. 28. No. 1. P. 1–14. DOI: 10.1037/h0034092
 13. *Spencer L.M., Spencer S.M.* Competence at Work: Models for Superior Performance. New York: John Wiley & Sons, 1993. ISBN: 0-471-54809-x. URL: <https://www.wiley.com/en-us/Competence+at+Work%3A+Models+for+Superior+Performance-p-9780471548096> (дата обращения: 25.10.2024).
 14. *Ercikan K., Oliveri M.E.* 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 // *Applied Measurement in Education*. 2016. No. 29(4). P. 310–318. DOI: 10.1080/08957347.2016.1209210
 15. *Zlatkin-Troitschanskaia O., Jitomirski J., Happ R., Molerov D., Schlax J., Kübling-Thees C., Förster M., Brückner S.* Validating a test for measuring knowledge and understanding of economics among university students // *Zeitschrift für Pädagogische Psychologie*. 2019. No. 33 (2). P. 119–133. DOI: 10.1024/1010-0652/a000239
 16. *Shavelson R.J., Zlatkin-Troitschanskaia O., Beck K., Schmidt S., Mariño J.P.* Assessment of university students’ critical thinking: Next generation performance assessment // *International Journal of Testing*. 2019. Vol. 19. No. 4. P. 337–362. DOI: 10.1080/15305058.2018.1543309
 17. *Messick S.* The Interplay of Evidence and Consequences in the Validation of Performance Assessments // *Educational Researcher*. 1992. Vol. 23. No. 2. P. 13–23. DOI: 10.3102/0013189X023002013
 18. *Mislevy R.J.* Evidence and inference in educational assessment // *Psychometrika*. 1994. Vol. 59. No. 4. P. 439–483. URL: <http://www.springerlink.com/content/161116h6652714714> (дата обращения: 12.10.2024).
 19. *Toulmin S.E.* The Uses of Argument. Updated edition. Cambridge University Press, 2008. 247 p. DOI: 10.1017/cbo9780511840005
 20. *Mislevy R.J.* Substance and Structure in Assessment Arguments // *Law, Probability and Risk*. 2003. No. 2. P. 237–258. DOI: 10.1093/lpr/2.4.237
 21. *Messick S.* Validity // *Educational Measurement: 3rd ed.*, ed. by R.L. Linn. New York: American Council on Education/Macmillan, 1989. P. 13–103. URL: <https://psycnet.apa.org/record/1989-97348-002> (дата обращения: 25.10.2024).
 22. *Ferrara S., Lai E., Nichols P.* Principled Approaches to Assessment Design, Development, and Implementation // *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*. 2016. P. 41–74. DOI: 10.1002/9781118956588.ch3
 23. *Mislevy R.J., Almond R.G., Lukas J.F.* A Brief Introduction to Evidence-centered Design. Princeton: Educational Testing Service, 2003. 37 p. URL: <https://files.eric.ed.gov/fulltext/ED483399.pdf> (дата обращения: 25.10.2024).

24. Угланова И.А., Брун И.В., Васин Г.М. Методология Evidence-Centered Design для измерения комплексных психологических конструкторов // Современная зарубежная психология. 2018. № 3 (7). С. 18–27. DOI: 10.17759/jmfp.2018070302
25. Ефремова Н.Ф. Приёмы доказательной аргументации оценки компетенций // Инновационная Наука: Психология, Педагогика, Дефектология. 2020. Т. 3. № 2. С. 112–124. EDN: UPOODB.
26. Uglanova I., Orel E., Gracheva D., Tarasova K. Computer-based performance approach for critical thinking assessment in children // British Journal of Educational Psychology. 2023. Vol. 93. No. 2. P. 531–544. DOI: 10.1111/bjep.12576
27. Абдеева С.М., Руднев М.Г., Васин Г.М., Тарасова К.В., Панова Д. Оценка информационно-коммуникационной компетентности учащихся: подходы, инструмент, валидность и надёжность результатов // Вопросы образования. 2017. № 4. С. 104–132. DOI: 10.17323/1814-9545-2017-4-104-132
28. Абдеева С.М., Тарасова К.В. Об оценке цифровой грамотности: методология, концептуальная модель и инструмент измерения // Вопросы образования. 2023. № 2. С. 8–32. DOI: 10.17323/1814-9545-2023-2-8-32
29. Mislevy R.J., Bebreus J., DiCerno K.E., Levy R. Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining // Journal of Educational Data Mining. 2012. No. 4. P. 11–48. URL: http://www.educationaldatamining.org/JEDM/images/articles/vol4/issue1/MislevyEtAlVol4Issue1P11_48.pdf (дата обращения: 12.10.2024).
30. DiBello L.V., Roussos L.A., Stout W. Review of cognitively diagnostic assessment and a summary of psychometric models // Handbook of Statistics. 2007. No. 26. P. 979–1030. DOI: 10.1016/S0169-7161(06)26031-0
31. Tjoe H., de la Torre J. The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework // Mathematics Education Research Journal. 2014. No. 26. P. 237–255. DOI: 10.1007/s13394-013-0090-7
32. Тарасова К.В., Орёл Е.А. Измерение критического мышления студентов в открытой онлайн-среде: концептуальная рамка и типология заданий // Вопросы образования. 2022. № 3. С. 187–212. DOI: 10.17323/1814-9545-2022-3-187-212
33. Lai E.R. Critical Thinking: A Literature Review // Pearson's Research Reports. 2011. No. 6. P. 40–41. URL: http://paluchja-zajecia.home.amu.edu.pl/seminarium_fakult/sem_f_krytyczne/Critical%20Thinking%20A%20Literature%20Review.pdf (дата обращения: 25.10.2024).
34. Liu O.L., Frankel L., Roobr K.C. Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessment // ETS Research Report Series. 2014. No. 1. P. 1–23. DOI: 10.1002/ets2.12009
35. Riconscente M., Mislevy R., Hamel L. An introduction to PADI task templates // PADI Technical Report. 2005. Vol. 3. URL: https://padi.sri.com/downloads/TR3_Templates.pdf (дата обращения: 25.10.2024).
36. Mislevy R.J., Riconscente M.M., Rutstein D.W. Design patterns for assessing model-based reasoning // Large Scale Assessment Technical Report 6. Menlo Park, CA: SRI International, 2009. URL: http://ecd.sri.com/downloads/ECD_TR6_Model_Based_Reasoning.pdf (дата обращения: 25.10.2024).
37. Zlatkin-Troitschanskaia O., Shavelson R.J. Advantages and Challenges of Performance Assessment of Student Learning in Higher Education // British Journal of Educational Psychology. 2019. Vol. 89. No. 3. P. 413–415. DOI: 10.1111/bjep.12314
38. Wang W.C., Su C.M., Qiu X.L. Item response models for local dependence among multiple ratings // Journal of Educational Measurement. 2014. Vol. 51. No. 3. P. 260–280. DOI: 10.1111/jedm.12045
39. Тарасова К.В., Грачева Д.А. Вычислительная психометрика: ближайшее будущее или уже реальность. Рецензия на книгу “Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment” // Вопросы образования. 2023. № 3. С. 221–230. DOI: 10.17323/vo-2023-17938
40. Andrews-Todd J., Forsyth C.M. Exploring Social and Cognitive Dimensions of Collaborative Problem Solving in an Open Online Simulation-Based Task // Computers in Human Behavior. 2020. Vol. 104. Article no. 105759. DOI: 10.1016/j.chb.2018.10.025
41. Almond R., Steinberg L., Mislevy R. Enhancing the design and delivery of assessment systems: A four-process architecture // Journal of Technology, Learning, and Assessment. 2002. Vol. 1.

- No. 5. URL: <https://ejournals.bc.edu/index.php/jtla/article/view/1671> (дата обращения: 25.10.2024).
42. Грачева Д.А., Тарасова К.В. Подходы к разработке вариантов заданий сценарного типа в рамках метода доказательной аргументации // Отечественная и зарубежная педагогика. 2022. Т. 1. № 3. С. 83–97. DOI: 10.24412/2224-0772-2022-84-83-97
 43. Yan D., Rupp A., Foltz P. The Handbook of Automated Scoring: Theory into Practice. CRC Press, 2020. ISBN: 9781351264808. DOI: 10.1201/9781351264808
 44. De Klerk S., Eggen T.J.H.M., Veldkamp B.P. A Methodology for Applying Students' Interactive Task Performance Scores from a Multimedia-based Performance Assessment in a Bayesian Network // Computers in Human Behavior. 2016. Vol. 60. I. С. P. 264–279. DOI: 10.1016/j.chb.2016.02.071
 45. Mislevy, R. Sociocognitive and argumentation perspectives on psychometric modeling in educational assessment // Psychometrika. 2024. Vol. 89. No. 1. DOI: 10.1007/s11336-024-09966-5
- Благодарности.** Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ в 2024 г.
- Статья поступила в редакцию 06.11.2024
Принята к публикации 21.12.2024

References

1. Razumova T.O., Telekhova I.G. (2023). Transformation of the System of Higher Professional Education: Challenges and Perspectives. *Standard of Living of the Population of the Regions of Russia*. Vol. 19, no. 3, pp. 338–349, doi: 10.52180/1999-98362023193333834 (In Russ., abstract in Eng.).
2. Care, E; Kim, H; Vista, A; Anderson, K. (2018). *Education System Alignment for 21st Century Skills: Focus on Assessment*. The Center for Universal Education at the Brookings Institution. Available at: https://www.researchgate.net/publication/330740772_Education_system_alignment_for_21st_century_skills_Focus_on_assessment (accessed: 25.10.2024).
3. Mislevy, R.J. (2018). *Sociocognitive Foundations of Educational Measurement*. Routledge, doi: 10.1111/jedm.12255
4. Shmigirilova I.B., Rvanova A.S. Grigorenko O.V. (2021). Assessment in education: Current trends, problems and contradictions (review of scientific publications). *Education and Science Journal*. Vol. 23, no. 6, pp. 43–83, doi: 10.17853/1994-5639-2021-6-43-83 (In Russ., abstract in Eng.).
5. Achtenhagen, F. (2012). The curriculum-instruction-assessment triad. *Empirical Research in Vocational Education and Training*. No. 4, pp. 5–25, doi: 10.1007/BF03546504
6. Biggs, J., Tang, C. (2011). *Teaching for Quality Learning at University: What the Student Does*. New York: McGraw-Hill Education, 480 p. ISBN 10 : 0335242758.
7. Mislevy, R.J. (2013). *Four Metaphors We Need to Understand Assessment*. The Gordon Commission on the Future of Assessment in Education, 39 p. Available at: https://www.ets.org/Media/Research/pdf/mislevy_four_metaphors_understand_assessment.pdf (accessed: 25.10.2024).
8. Pellegrino, J.W., Chudowsky, N., Glaser, R. (2001). *Knowing What Students Know: The science and Design of Educational Assessment*. Washington, DC: National Academy Press. Available at: https://www.researchgate.net/publication/270584995_Knowing_What_Students_Know_The_Science_and_Design_of_Educational_Assessment (accessed: 25.10.2024).
9. Lamri, J., Lubart T. (2023) Reconciling Hard Skills and Soft Skills in a Common Framework: The Generic Skills Component Approach. *Journal of Intelligence*. No. 11: 107, doi: 10.3390/jintelligence11060107, doi: 10.1007/s11135-021-01149-z
10. Pellegrino, J.W., Hilton, M.L. (Eds.) (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington, DC: National Academies Press, 256 p., doi: 10.17226/13398

11. Shavelson, R.J., Zlatkin-Troitschanskaia, O., Mariño, J.P. (2018). International performance assessment of learning in higher education (iPAL): Research and development. *Assessment of Learning Outcomes in Higher Education*. P. 193-214, doi: 10.1007/978-3-319-74338-7_10
12. McClelland, D.C. (1973). Testing for competence rather than for "intelligence". *American Psychologist*. *American Psychological Association*. Vol. 28, no. 1, pp. 1-14, doi: 10.1037/h0034092
13. Spencer, L.M., Spencer S.M. (1993). *Competence at Work: Models for Superior Performance*. New York: John Wiley & Sons. ISBN 0-471-54809-x. Available at: <https://www.wiley.com/en-us/Competence+at+Work%3A+Models+for+Superior+Performance-p-9780471548096> (accessed: 25.10.2024).
14. Ercikan, K., Oliveri, M.E. (2016). 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. *Applied Measurement in Education*. Vol. 29, no. 4, pp. 310-318, doi: 10.1080/08957347.2016.1209210
15. Zlatkin-Troitschanskaia O., Jitomirski J., Happ R., Molerov D., Schlax J., Kühling-Thees C., Förster M., Brückner S. Validating a test for measuring knowledge and understanding of economics among university students // *Zeitschrift für Pädagogische Psychologie*. No. 33 (2), pp. 119-133, doi: 10.1024/1010-0652/a000239
16. Shavelson R.J., Zlatkin-Troitschanskaia O., Beck K., Schmidt S., Mariño J.P. (2019). Assessment of university students' critical thinking: Next generation performance assessment. *International Journal of Testing*. Vol. 19, no. 4, pp. 337-362, doi: 10.1080/15305058.2018.1543309
17. Messick, S. (1992). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*. Vol. 23, no. 2, pp. 13-23, doi: 10.3102/0013189X023002013
18. Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika*. Vol. 59, no. 4, pp. 439-483. Available at: <http://www.springerlink.com/content/l6116h6652714714> (accessed: 12.10.2024).
19. Toulmin, S.E. (2008). *The Uses of Argument*. Updated edition. Cambridge University Press, 247 p., doi: 10.1017/cbo9780511840005
20. Mislevy, R.J. (2003). Substance and structure in assessment arguments. *Law, Probability and Risk*. No. 2, pp. 237-258, doi: 10.1093/lpr/2.4.237
21. Messick, S. (1989). Validity. In R.L. Linn (Ed.). *Educational Measurement*: 3rd ed., pp. 13-103. New York: American Council on Education/Macmillan. Available at: <https://psycnet.apa.org/record/1989-97348-002> (accessed: 25.10.2024).
22. Ferrara, S.; Lai, E.; Nichols, P. (2016). Principled Approaches to Assessment Design, Development, and Implementation. *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*. Pp. 41-74, doi: 10.1002/9781118956588.ch3
23. Mislevy, R.J., Almond, R.G., Lukas, J.F. (2003). *A Brief Introduction to Evidence-centered Design*. Princeton: Educational Testing Service, 37 p. Available at: <https://files.eric.ed.gov/fulltext/ED483399.pdf> <https://psycnet.apa.org/record/1989-97348-002> (accessed: 25.10.2024).
24. Uglanova, I.L., Brun, I.V., Vasin, G.M. (2018). Evidence-Centered Design method for measuring complex psychological constructs. *Journal of Modern Foreign Psychology*. Vol. 7, no. 3, pp. 18-27, doi: 10.17759/jmfp.2018070302 (In Russ., abstract in Eng.).
25. Efremova, N.F. (2020). Techniques of Evidence-Based Argumentation for Competency Assessment. *Innovatsionnaya nauka: Psikhologiya, Pedagogika, Defektologiya = Innovative Science: Psychology, Pedagogy, Defectology*. Vol. 3, no. 2, pp. 112-124. Available at: https://elibrary.ru/download/elibrary_44268469_64255654.pdf (accessed: 25.10.2024) (In Russ., abstract in Eng.).

26. 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, doi: 10.1111/bjep.12576
27. Avdeeva, S.M., Rudnev, M.G., Vasin, G.M., Tarasova, K.V., Panova, D.M. (2017). Assessing Information and Communication Technology Competence of Students: Approaches, Tools, Validity and Reliability of Results. *Educational Studies Moscow*. No. 4, pp. 104-132, doi: 10.17323/1814-9545-2017-4-104-132 (In Russ., abstract in Eng.).
28. Avdeeva, S.M., Tarasova, K.V. (2023). On measuring digital literacy: methodology, conceptual model and measurement tool. *Educational Studies Moscow*. No. 2, pp. 8-32, doi: 10.17323/1814-9545-2023-2-8-32 (In Russ., abstract in Eng.).
29. Mislevy, R.J., Behrens, J., DiCerbo, K.E., Levy, R. (2012). Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. *Journal of Educational Data Mining*. No. 4, pp. 11-48. Available at: http://www.educationaldatamining.org/JEDM/images/articles/vol4/issue1/MislevyEtAlVol4Issue1P11_48.pdf (accessed: 12.10.2024).
30. DiBello, L.V., Roussos, L.A., Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of Statistics*. No. 26, pp. 979-1030, doi: 10.1016/S0169-7161(06)26031-0
31. Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal*. No. 26, pp. 237-255, doi: 10.1007/s13394-013-0090-7
32. Tarasova, K.V., Orel, E.A. (2022). Measuring Students' Critical Thinking in Online Environment: Methodology, Conceptual Framework and Tasks Typology. *Educational Studies Moscow*. No. 3, pp. 187-212, doi: 10.17323/1814-9545-2022-3-187-212 (In Russ., abstract in Eng.).
33. Lai, E.R. (2011). *Critical Thinking: A Literature Review Research Report*. London: Parsons Publishing, Available at: http://paluchja-pajecia.home.amu.edu.pl/seminarium_fakult/sem_f_krytyczne/Critical%20Thinking%20A%20Literature%20Review.pdf (accessed: 25.10.2024).
34. Liu, O.L., Frankel, L., Roohr, K.C. (2014). Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessment. *ETS Research Report Series*. No. 1, pp. 1-23, doi: 10.1002/ets2.12009
35. Riconscente, M., Mislevy, R., Hamel, L. (2005). An introduction to PADI task templates. *PADI Technical Report*. Vol. 3. Available at: https://padi.sri.com/downloads/TR3_Templates.pdf (accessed: 25.10.2024).
36. Mislevy, R.J., Riconscente, M.M., & Rutstein, D.W. (2009). Design patterns for assessing model-based reasoning. *Large Systems Technical Report 6*. Menlo Park, CA: SRI International. Available at: http://ecd.sri.com/downloads/ECD_TR6_Model Based_Reasoning.pdf (accessed: 25.10.2024).
37. Zlatkin-Troitschanskaia, O., Shavelson, R.J. (2019). Advantages and Challenges of Performance Assessment of Student Learning in Higher Education. *British Journal of Educational Psychology*. Vol. 89, no. 3, pp. 413-415, doi: 10.1111/bjep.12314
38. Wang, W.C., Su, C.M., Qiu, X.L. (2014). Item response models for local dependence among multiple ratings. *Journal of Educational Measurement*. Vol. 51, no. 3, pp. 260-280, doi: 10.1111/jedm.12045
39. Tarasova, K.V., Gracheva, D.A. (2023). Computational Psychometrics: The Near Future or Already a Reality. Review of the Book "Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment". *Educational Studies Moscow*. No. 3, pp. 221-230, doi: 10.17323/vo-2023-17938

40. Andrews-Todd, J., Forsyth, C.M. (2020). Exploring Social and Cognitive Dimensions of Collaborative Problem Solving in an Open Online Simulation-Based Task. *Computers in Human Behavior*. Vol. 104, article no. 105759, doi: 10.1016/j.chb.2018.10.025
41. Almond, R., Steinberg, L., Mislevy, R. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment*. Vol. 1, no. 5. Available at: <https://ejournals.bc.edu/index.php/jtla/article/view/1671> (accessed: 25.10.2024).
42. Gracheva, D.A., Tarasova, K.V. (2022). Approaches to the development of scenario-based task forms within the framework of evidence-centered design. *Otechestvennaia i zarubezhnaia pedagogika = Domestic and Foreign Pedagogy*. Vol. 1, no. 3(84), pp. 83-97, doi: 10.24412/2224-0772-2022-84-83-97 (In Russ., abstract in Eng.).
43. Yan, D.; Rupp, A.; Foltz, P. (2020). *The Handbook of Automated Scoring: Theory into Practice*. CRC Press. ISBN: 9781351264808, DOI: 10.1201/9781351264808.
44. De Klerk, S., Eggen, T.J.H.M., Veldkamp, B.P. (2016). A Methodology for Applying Students' Interactive Task Performance Scores from a Multimedia-based Performance Assessment in a Bayesian Network. *Computers in Human Behavior*. Vol. 60, i. C, pp. 264-279, doi: 10.1016/j.chb.2016.02.071
45. Mislevy, R. (2024). Sociocognitive and argumentation perspectives on psychometric modeling in educational assessment. *Psychometrika*. Vol. 89, no. 1, doi: 10.1007/s11336-024-09966-5

Acknowledgement. This research was supported by the National Research University Higher School of Economics under the Fundamental Research Program in 2024.

*The paper was submitted 06.11.2024
Accepted for publication 21.12.2024*



Science Index РИНЦ-2023

СОЦИОЛОГИЧЕСКИЕ ИССЛЕДОВАНИЯ	10,445
ВОПРОСЫ ФИЛОСОФИИ	9,735
ВОПРОСЫ ОБРАЗОВАНИЯ	9,251
ВЫСШЕЕ ОБРАЗОВАНИЕ В РОССИИ	9,187
ЭПИСТЕМОЛОГИЯ И ФИЛОСОФИЯ НАУКИ	8,248
ОБРАЗОВАНИЕ И НАУКА	8,032
ПСИХОЛОГИЧЕСКАЯ НАУКА И ОБРАЗОВАНИЕ	7,998
УНИВЕРСИТЕТСКОЕ УПРАВЛЕНИЕ: ПРАКТИКА И АНАЛИЗ	6,973
ИНТЕГРАЦИЯ ОБРАЗОВАНИЯ	6,567
АЛМА МАТЕР (ВЕСТНИК ВЫСШЕЙ ШКОЛЫ)	3,681
ПЕДАГОГИКА	3,606
ВЫСШЕЕ ОБРАЗОВАНИЕ СЕГОДНЯ	3,308