

**Ссылка для цитирования этой статьи:**

Султанаев Я.Т., Абзалилов Р.Р., Вильданова В.Ф. Верификационная тестология: формализованная система оценки тестовых заданий для образовательных систем с искусственным интеллектом // Human Progress. 2025. Том 11, Вып. 9. С. 16. URL: [http://progress-human.com/images/2025/Tom11\\_9/Sultanayev.pdf](http://progress-human.com/images/2025/Tom11_9/Sultanayev.pdf) DOI 10.46320/2073-4506-2025-9a-16.

## **ВЕРИФИКАЦИОННАЯ ТЕСТОЛОГИЯ: ФОРМАЛИЗОВАННАЯ СИСТЕМА ОЦЕНКИ ТЕСТОВЫХ ЗАДАНИЙ ДЛЯ ОБРАЗОВАТЕЛЬНЫХ СИСТЕМ С ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ**



**Султанаев Юдат Талгатович**

доктор физико-математических наук,  
профессор кафедры математики и статистики,  
Башкирский государственный педагогический университет им. М.  
Акмуллы  
г. Уфа, Российская Федерация



**Абзалимов Рамиль Рафикович**

кандидат физико-математических наук, доцент,  
Директор ГАУ РБ Центр оценки профессионального мастерства и  
квалификаций педагогов,  
г. Уфа, Российская Федерация



**Вильданова Венера Фидарисовна**

кандидат физико-математических наук,  
доцент кафедры математики и статистики,  
Башкирский государственный педагогический университет им. М.  
Акмуллы  
г. Уфа, Российская Федерация

**Аннотация.** Статья посвящена преодолению системного кризиса доверия к качеству тестовых материалов, генерируемых искусственным интеллектом (ИИ), для педагогических измерений. Вводится и обосновывается концепция верификационной тестологии - методологии доказательного обеспечения качества на основе многоуровневой автоматизированной проверки. Представлена комплексная система формализованных критериев верификации, включающая оценку семантической ясности, дидактической релевантности, прогнозируемой эффективности и устойчивости заданий. Особое внимание

уделено принципу «ИИ проверяет ИИ» для обеспечения внутренней согласованности. Описана успешная практическая реализация методологии в платформе «ИИ-эксперт» для оценки управленческих компетенций, показавшая уровень качества генерации 0,97. Делается вывод о том, что верификационная тестология закладывает основу для новой парадигмы образовательной диагностики, обеспечивающей надежность, объективность и персонализацию в условиях массового использования ИИ.

**Ключевые слова:** верификационная тестология, искусственный интеллект, педагогические измерения, качество тестов, генерация тестовых заданий.

## 1. Введение

Современное образование претерпевает фундаментальную трансформацию, движимую стремительным развитием цифровых технологий. Одной из наиболее перспективных, но одновременно и проблемных областей применения ИИ в педагогике является генерация тестовых материалов. Способность алгоритмов быстро создавать огромные массивы персонализированных заданий потенциально способна вывести образовательную диагностику на качественно новый уровень.

Однако за этой кажущейся простотой скрывается системный вызов, ставящий под сомнение надежность всей системы автоматизированного оценивания. Речь идет о проблеме неопределенности и непредсказуемости качества тестовых заданий, сгенерированных ИИ. Доверяя искусственному интеллекту создание измерительного инструмента, педагогическая система делегирует ему ответственность за валидность и надежность получаемых данных о знаниях и компетенциях обучающихся. В отличие от человека-эксперта, современные генеративные модели не обладают ни глубинным пониманием предметного контекста, ни педагогическим опытом, создавая тексты на основе статистических закономерностей в данных. В результате сгенерированный вопрос может оказаться семантически двусмысленным, не соответствующим учебной программе, иметь несколько правильных ответов или не иметь ни одного, что в условиях «высоких ставок» (например, вступительные экзамены или профессиональная сертификация) приводит к серьезным последствиям - несправедливой оценке и эрозии доверия ко всей системе образования.

Сложность усугубляется в реальном времени, когда система адаптивного тестирования генерирует вопросы «на лету», и у педагога-эксперта физически отсутствует возможность проверить каждый из них. Традиционная тестология, с ее отработанными процедурами экспертизы, пилотажного тестирования и статистического анализа [1], оказалась не готова к вызовам, порожденным масштабируемым ИИ. Ее методы, доказавшие эффективность в

условиях ручной разработки тестов, становятся практически неприменимыми из-за фундаментального противоречия между скоростью создания контента и требованиями к его качеству.

Таким образом, возникает острая необходимость в новой парадигме, которая органично интегрировала бы незыблемые педагогические принципы с технологическими возможностями, обеспечивая доказательное качество каждого тестового задания еще до момента его предъявления обучающемуся. Ответом на этот вызов и призвана стать верификационная тестология.

Целью настоящей статьи является теоретическое обоснование и методологическое проектирование верификационной тестологии как новой научно-практической дисциплины, направленной на обеспечение надежности и валидности педагогических измерений в условиях массовой генерации тестовых материалов системами искусственного интеллекта.

## **2. Анализ ограничений традиционной тестологии в условиях ИИ-генерации**

Современная тестология, обладая хорошо разработанным методологическим аппаратом, столкнулась с системным кризисом, вызванным стремительным развитием технологий искусственного интеллекта. Этот кризис можно охарактеризовать как разрыв между технологическими возможностями и педагогическим качеством, когда скорость и масштабируемость генерации тестового контента вступают в противоречие с фундаментальными принципами обеспечения его валидности и надежности. Анализ литературы и практики позволяет выявить несколько ключевых ограничений традиционных подходов в новых условиях. Классическая методология разработки тестов, регламентированная, в том числе, и российскими стандартами [1], предполагает многоэтапную экспертизу каждого задания человеком-специалистом. Этот процесс включает проверку на соответствие программе, однозначность формулировок, адекватный уровень сложности и способность дифференцировать учащихся. Однако в контексте ИИ-генерации, когда система может производить сотни и тысячи уникальных вариантов заданий в реальном времени для множества пользователей, традиционная экспертиза становится не только экономически нецелесообразной, но и физически невозможной. Человек-эксперт не в состоянии проверить такой объем материалов с необходимой тщательностью, что приводит к вынужденному отказу от проверенных процедур в угоду оперативности.

Существующие стандарты и методы тестологии в основном ориентированы на постфактум анализ, отвечая на вопрос «насколько хорошим оказался тест?» после его проведения и статистической обработки результатов. Они не предоставляют инструментов для априорной оценки каждого конкретного задания до момента его использования. В

условиях ИИ-генерации уникальных заданий, которые могут быть применены единожды, запоздалый анализ неприменим. Необходимы формализованные, измеримые и алгоритмически проверяемые показатели, позволяющие оценить качество задания в режиме реального времени, до его предъявления тестируемому. Как отмечается в исследованиях, нейросетевые языковые модели по своей природе стохастичны и непредсказуемы [6]. Они могут генерировать грамматически правильные, но семантически ошибочные формулировки; создавать вопросы, имеющие несколько возможных интерпретаций; производить контент, который лишь поверхностно соответствует теме. Без системы оперативного выявления таких дефектов возникает риск использования внешне убедительных, но педагогически несостоятельных тестовых материалов. Эта проблема усугубляется тем, что современные методы машинного обучения часто работают как «черный ящик», и даже разработчики не всегда могут точно объяснить, почему модель сгенерировала тот или иной конкретный вопрос [7].

Традиционные методы оценки психометрических характеристик заданий, такие как расчет дискриминативности и трудности, требуют значительного объема данных о реальных ответах испытуемых. В контексте ИИ-генерации уникальных заданий это означает, что мы можем получить статистические характеристики только после того, как вопрос уже был использован и потенциально нанес ущерб качеству измерений. Возникает классическая проблема «курицы и яйца»: чтобы оценить качество задания, нужно его использовать, но использовать его нельзя, не оценив качество. Это ограничение особенно критично в системах адаптивного тестирования, где один некачественный вопрос может исказить не только текущий результат, но и всю последующую траекторию тестирования.

Таким образом, традиционная тестология оказалась в методологической ловушке: ее классические инструменты не адаптированы к скорости, масштабу и природе ИИ-генерации. Преодоление этого кризиса требует не модернизации отдельных процедур, а пересмотра фундаментальных принципов обеспечения качества тестовых материалов, когда их создание делегировано интеллектуальным системам.

### **3. Верификационная тестология: сущность и методологический фундамент**

В ответ на системные ограничения традиционной тестологии, выявленные в предыдущем разделе, предлагается концепция **верификационной тестологии**. Данная концепция представляет собой не просто набор новых технических приемов, а целостную методологическую платформу, обеспечивающую доказательное качество педагогических измерений в эпоху искусственного интеллекта.

## Определение понятия.

**Верификационная тестология** – это методология доказательного обеспечения качества тестовых материалов, генерируемых системами искусственного интеллекта, основанная на многоуровневой системе автоматизированной проверки соответствия заданий установленным дидактическим, психометрическим и содержательным критериям до момента их использования в учебном процессе. Ключевым аспектом здесь является именно **доказательность** - каждое задание, допущенное к применению, должно иметь формальное, алгоритмически полученное подтверждение своего соответствия всем необходимым педагогическим и метрологическим требованиям.

В контексте верификационной тестологии качественный тестовый **вопрос** понимается как задание, удовлетворяющее двум фундаментальным требованиям: во-первых, оно адекватно измеряет заявленные знания и компетенции (обладает содержательной валидностью), а во-вторых - обеспечивает воспроизводимость результатов при различных условиях предъявления (обладает надежностью). Это означает, что вопрос должен быть семантически однозначным, соответствующим учебной программе, иметь прогнозируемые психометрические характеристики и сохранять свои измерительные свойства при допустимых вариациях формулировок. Верификационная тестология операционализирует различие между верификацией и валидацией, критически важное для сложных систем [7]. Если **валидация** в традиционной тестологии отвечает на вопрос «измеряем ли мы правильный конструкт?» и часто проводится постфактум, то **верификация** отвечает на вопрос «корректно ли работает наш измерительный инструмент на уровне каждого отдельного задания?». Таким образом, верификационная тестология не отменяет валидацию, а создает для нее надежный фундамент, гарантируя техническую корректность каждого элемента теста до его использования.

## Методологические принципы.

Методологический каркас верификационной тестологии базируется на трех взаимосвязанных принципах:

**1. Принцип приоритета превентивного контроля.** Данный принцип предполагает кардинальный сдвиг от реагирования на проблемы к их системному предупреждению. Система оценки качества должна работать до момента предъявления задания тестируемому, а не после анализа результатов тестирования. Такой подход позволяет предотвратить потенциальный ущерб от использования некачественных измерительных материалов, а не констатировать этот ущерб постфактум, что особенно важно в условиях высоких ставок.

**2. Принцип формализации критериев качества.** Этот принцип требует перевода традиционных экспертных требований к тестовым заданиям, закрепленных, в том числе, в стандартах [1], в формализованные, измеримые и алгоритмически проверяемые показатели. Это позволяет автоматизировать процесс проверки, сделать его объективным, независимым от субъективного мнения отдельного эксперта и масштабируемым под любые объемы генерации. Формализация распространяется на семантические, дидактические и психометрические аспекты заданий.

**3. Принцип непрерывности верификации.** В условиях, когда ИИ-система генерирует новые варианты заданий постоянно и в реальном времени, процесс их проверки не может быть эпизодическим или выборочным. Каждое задание, независимо от времени и условий его создания, должно проходить полный цикл верификации перед использованием. Этот принцип обеспечивает стабильность и предсказуемость качества измерительных материалов на всем протяжении жизненного цикла образовательной системы.

**Дифференциация от традиционной тестологии.** Верификационная тестология не отрицает достижений классической тестологии, а надстраивается над ней, предлагая эволюционное развитие дисциплины. Ключевое отличие заключается в перераспределении акцентов и усилий между этапами жизненного цикла тестового задания. Если традиционный подход основное внимание уделяет итоговой валидации теста как целостного инструмента, то верификационная тестология фокусируется на сквозной верификации каждого элемента этого инструмента на этапе его создания. Это превращает искусственный интеллект из источника рисков, связанных с неконтролируемой генерацией, в надежный и предсказуемый инструмент педагогического проектирования, отвечающий строгим стандартам качества.

#### **4. Система критериев верификации качества тестовых заданий**

Ключевым элементом верификационной тестологии является комплексная система критериев, позволяющая осуществлять автоматизированную оценку качества ИИ-сгенерированных заданий. Данная система представляет собой формализованную операционализацию педагогических и психометрических требований, обеспечивающую их алгоритмическую проверку. Критерии сгруппированы в пять взаимодополняющих блоков, охватывающих все аспекты качества тестового задания.

##### **Критерии ясности и однозначности.**

Данная группа критериев направлена на оценку формальных характеристик формулировки задания, от которых напрямую зависит его валидность. Если задание допускает множественное толкование, его диагностическая ценность сводится к нулю.

- **Семантическая ясность** оценивает точность и однозначность передачи смысла. Для ее оценки применяется анализ векторных представлений текста (эмбеддингов). Рассчитывается индекс семантической ясности - косинусная близость между вектором задания и эталонными векторами, что позволяет выявить семантические аномалии и несоответствия.

- **Лексическая сложность** оценивает соответствие языковых конструкций возможностям целевой аудитории. Используются стандартизированные индексы удобочитаемости (например, Flesch-Kincaid, SMOG [8], [9], [10]) и коэффициент лексического разнообразия (Type-Token Ratio, TTR [11], [12]). Система настраивается на допустимые диапазоны этих показателей в зависимости от возраста и подготовки тестируемых.

- **Критерий отсутствия двусмысленности** выявляет скрытые многозначности. Задание признается соответствующим критерию, если доминирует единственная верная интерпретация, что подтверждается анализом вероятностного распределения возможных трактовок, выявляемых с помощью ИИ-моделей.

#### **Критерии дидактического соответствия**

Эта группа критериев обеспечивает содержательную валидность заданий, фокусируясь на их способности адекватно измерять запланированные образовательные результаты.

- **Таксономическая валидность** оценивает соответствие задания заявленному уровню познавательной деятельности согласно таксономии Блума (знание, понимание, применение, анализ, синтез, оценка). Для автоматической классификации используются специализированные ИИ-модели, обученные на размеченных корпусах педагогических текстов.

- **Содержательное соответствие** проверяет релевантность задания учебной программе. Формируется эталонный семантический профиль дисциплины на основе анализа учебных программ, учебников и методических материалов. Для каждого задания вычисляется **коэффициент соответствия программе** - косинусное сходство между его векторным представлением и векторами релевантных разделов курса. Для допуска к использованию этот коэффициент должен превышать установленный порог (обычно не менее 0,8).

#### **Критерии прогнозируемой эффективности**

Данные критерии решают проблему невозможности априорной психометрической оценки уникальных ИИ-генерируемых заданий, прогнозируя их ключевые характеристики до использования.

- **Прогнозируемая дискриминативность** оценивает способность задания дифференцировать испытуемых с высоким и низким уровнем подготовленности. Вместо

традиционных методов, требующих данных реальных ответов [13], [14], применяется метод «виртуального пилотирования». Создаются синтетические профили «сильных» и «слабых» виртуальных испытуемых, и алгоритм оценивает вероятность правильного ответа для каждой группы, вычисляя прогнозируемый коэффициент дискриминации.

- **Прогнозируемая сложность** определяет предполагаемую долю испытуемых, способных правильно ответить на вопрос. Для прогноза анализируется комплекс параметров: семантическая близость к эталонным простым/сложным заданиям, лингвистическая сложность, количество требуемых ментальных операций и структурная сложность. Использование нейросетевых моделей позволяет выявлять сложные нелинейные зависимости между этими признаками и фактической сложностью.

#### **Критерии тематической релевантности и чистоты.**

Данная группа представляет собой специализированное развитие семантического анализа, использующее методы вероятностного тематического моделирования [2], [3], [4] для оценки глубинного содержательного соответствия.

- Тематическая релевантность оценивает степень смыслового соответствия задания ключевым темам. Методология включает формирование эталонного тематического профиля на основе базы материалов, анализ тематического распределения задания и расчёт метрики близости (например, KL-дивергенции) между ними. Задание, чьё распределение существенно отклоняется от эталона, признаётся тематически нерелевантным.

- Тематическая чистота выявляет семантическую неоднозначность и «шум». Показателем служит энтропия Шеннона тематического распределения задания. Высокая энтропия указывает на «размытость» и принадлежность вопроса к нескольким несвязанным темам, что является признаком скрытой двусмысленности. Данные критерии также напрямую усиливают оценку качества дистракторов, позволяя анализировать семантическую дистанцию между тематическими профилями правильного и неправильных ответов.

#### **Критерии устойчивости и надёжности.**

Эти критерии обеспечивают стабильность результатов измерения независимо от незначительных вариаций в формулировках и условиях предъявления задания.

- **Устойчивость к перефразированию** проверяется методом семантического стресс-тестирования. Исходное задание автоматически преобразуется в несколько синонимичных вариантов, после чего анализируется сохранение семантической близости к оригиналу, идентичности эталонного ответа и таксономического уровня.

- **Качество дистракторов** оценивается по трем уровням: семантическая дистанция от дистракторов к правильному ответу (оптимальный диапазон схожести 0,4-0,6), внутренняя семантическая разнородность дистракторов и их соответствие типичным ошибкам учащихся.

### **Критерии внутренней согласованности ИИ**

Этот инновационный блок критериев реализует принцип «ИИ проверяет ИИ», создавая систему кросс-валидации внутри самой технологии.

- **Консенсус моделей** оценивает согласованность результатов анализа одного и того же задания несколькими независимыми языковыми моделями (например, GigaChat, YandexGPT, DeepSeek, GPT, Qwen, Claude). Низкий уровень консенсуса (менее 85%) сигнализирует о скрытых проблемах формулировки.

- **Обратимость генерации** проверяет логическую целостность связи «вопрос–ответ». Специализированная модель на основе эталонного ответа реконструирует исходный вопрос. Высокое семантическое сходство (коэффициент обратимости > 0,75) между оригинальным и реконструированным вопросом подтверждает устойчивость этой связи.

Представленная система критериев образует гибкий и модульный каркас для построения надежных верификационных контуров, адаптируемых к специфике различных образовательных контекстов и типов тестовых заданий.

### **5. Апробация методологии: платформа «ИИ-эксперт»**

Теоретическая и методологическая обоснованность верификационной тестологии была подтверждена в ходе практической апробации в рамках платформы «ИИ-эксперт», предназначенной для оценки управленческих компетенций руководителей образовательных организаций. Данная платформа служит референтной реализацией, демонстрирующей работоспособность предложенного подхода в условиях решения сложных диагностических задач. Платформа была ориентирована на оценку пяти ключевых блоков управленческих компетенций: «Управление кадрами», «Управление информацией», «Управление ресурсами», «Управление процессами» и «Управление результатами». Для обеспечения содержательной валидности и исключения «галлюцинаций» искусственного интеллекта была применена технология, аналогичная RAG (Retrieval-Augmented Generation). Ядро системы составляет база мини-документов (нормативные акты, письма, приказы, регламенты), на основе которой в реальном времени генерируются контекстуально обусловленные тестовые задания и кейсы. Это обеспечивает жесткую привязку генерируемого контента к актуальной нормативной и практической базе сферы образования. Процесс генерации и верификации заданий в платформе представляет собой многоэтапный конвейер, реализующий принципы внутренней согласованности ИИ и непрерывного превентивного контроля.

**1. Генерация:** Специализированный «агент-генератор» создает тестовые вопросы на основе извлеченного из базы документов контекста.

**2. Шлифовка:** «Агент-редактор» оптимизирует сгенерированные формулировки, улучшая их стилистику, проверяя соответствие дидактическим целям и оптимизируя структуру задания.

**3. Верификация:** «Агент-верификатор» выполняет ключевую операцию авто-верификации: ИИ самостоятельно отвечает на созданный вопрос и сравнивает полученный ответ с эталонным, заложенным в системе.

Для обеспечения баланса между креативностью и предсказуемостью outputa были тонко настроены параметры генерации: температура (0.3), top-k (50), top-p (0.9), Frequency Penalty (0.5) и Presence Penalty (0.4). Это позволило минимизировать повторы и поощрить лексическое разнообразие, сохраняя при этом семантическую стабильность. Внедрение описанного трехагентного конвейера позволило достичь выдающихся практических результатов. Изначальный уровень качества генерации, составлявший 0.93 (93% заданий соответствовали критериям качества после этапа шлифовки), был повышен до уровня 0.97 после полного цикла верификации. Это означает, что 97% заданий, поступающих к тестируемому, являются педагогически и содержательно корректными, что практически исключает риск использования дефектных измерительных материалов.

Помимо тестовых заданий, платформа продемонстрировала высокую эффективность в оценке кейсов, генерируемых на основе реальных ситуаций из практики образовательных организаций. Система анализирует предложенное тестируемым решение по множеству критериев, формируя развернутый аналитический отчет с выделением сильных и слабых сторон, потенциальных рисков и персональных рекомендаций по развитию компетенций. Апробация на выборке более 100 руководителей подтвердила способность системы не только точно оценивать уровень компетенций, но и выявлять специфические дефициты, характерные для разных типов образовательных организаций. Успешная реализация платформы «ИИ-эксперт» является эмпирическим доказательством состоятельности методологии верификационной тестологии. Достигнутый уровень качества, эффективность системы верификации и глубина аналитики создают прецедент для широкого внедрения подобных систем в различных областях образования.

## **6. Перспективы развития верификационной тестологии**

Успешная апробация принципов верификационной тестологии в платформе «ИИ-эксперт» открывает новые горизонты для трансформации образовательной диагностики. Дальнейшее развитие методологии видится в нескольких стратегически важных

направлениях, определяющих облик педагогических измерений в цифровую эпоху. Перспективным направлением является переход от эмпирически подобранных критериев и метрик к строгому доказательному подходу, обоснованному с помощью формальных математических моделей. В долгосрочной перспективе это может потребовать разработки специального логико-математического аппарата для описания и верификации сложных семантических правил и связей в тестовых заданиях. Подобные задачи находят отражение в фундаментальных исследованиях по теории алгоритмов и математической логике, направленных на анализ сложности формальных систем [5]. Интеграция таких подходов позволит перейти от эмпирических критериев к формально доказанным гарантиям качества и создать адаптивные системы верификации, способные автоматически калибровать пороговые значения в зависимости от специфики предметной области. Верификационная тестология обладает значительным потенциалом для интеграции в целостные образовательные экосистемы. На её основе могут создаваться интеллектуальные платформы, объединяющие контент-генерацию, адаптивное обучение, диагностику и аналитику в единый цикл. Такая система сможет автоматически создавать персонализированные образовательные траектории, где каждый учебный модуль сопровождается верифицированными диагностическими материалами, позволяющими объективно оценить прогресс в освоении компетенций. Интеграция с системами управления обучением (LMS) создаст замкнутый контур «обучение-диагностика-коррекция», существенно повышающий эффективность образовательного процесса.

Отработанные в рамках оценки управленческих компетенций методики и алгоритмы могут быть адаптированы для различных уровней образования - от школьного до дополнительного профессионального. Особый интерес представляет применение верификационной тестологии в областях, требующих оценки практических навыков и soft skills, где традиционное тестирование затруднено. Перспективным является развитие формата верификационных образовательных диалоговых систем, где искусственный интеллект не только генерирует контент, но и участвует в педагогическом диалоге, адаптируя его содержание и сложность в реальном времени на основе анализа ответов обучающегося.

Накопление данных о результатах верификации и диагностики открывает возможности для перехода от констатации знаний к предиктивной аналитике. Система сможет выявлять не только индивидуальные дефициты, но и закономерности в освоении учебного материала, типичные трудности, эффективные педагогические стратегии. Как показал опыт платформы «ИИ-эксперт», перспективным направлением является кластеризация обучающихся по результатам диагностики для создания групповых образовательных траекторий и

прогнозирования академической успеваемости. Широкое внедрение систем ИИ-генерации и верификации тестового контента требует разработки комплексной нормативной базы, регулирующей их использование в образовании. Необходимо установление стандартов качества, процедур верификации, требований к прозрачности алгоритмов и защите персональных данных. Важнейшей задачей является обеспечение интерпретируемости результатов для всех участников образовательного процесса - учащихся, педагогов, администрации.

Таким образом, верификационная тестология представляет собой не конечный продукт, а динамично развивающуюся методологическую платформу. Её эволюция в направлении большей доказательности, интегративности и персонализации создает основу для построения более справедливой, объективной и эффективной системы образования, отвечающей вызовам цифровой эпохи.

## 7. Заключение

Проведенное исследование позволяет констатировать, что верификационная тестология представляет собой закономерный и необходимый этап эволюции педагогических измерений в условиях цифровой трансформации образования. Преодолевая системный кризис доверия, вызванный массовой генерацией тестовых материалов искусственным интеллектом, предложенная методология предлагает переход от реактивного контроля качества к его системному доказательному обеспечению на уровне каждого отдельного задания.

Теоретическая значимость исследования заключается в разработке целостной методологической платформы, интегрирующей достижения педагогики, компьютерных наук и прикладной математики. Сформулированные принципы - приоритет превентивного контроля, формализация критериев качества и непрерывность верификации - создают новый стандарт для создания надежных диагностических инструментов. Разработанная комплексная система критериев верификации, охватывающая семантические, дидактические, психометрические и технологические аспекты тестовых заданий, позволяет операционализировать традиционные экспертные требования в формализованные, алгоритмически проверяемые показатели.

Практическая состоятельность верификационной тестологии получила убедительное подтверждение в ходе апробации на платформе «ИИ-эксперт». Достигнутый уровень качества генерации (0.97) демонстрирует не только техническую реализуемость подхода, но и его значительные преимущества перед традиционными методами разработки тестов. Многоагентная архитектура генерации и верификации, реализующая принцип «ИИ проверяет

ИИ», доказала свою эффективность для обеспечения содержательной валидности и семантической устойчивости создаваемых измерительных материалов. Перспективы развития методологии видятся в ее естественной эволюции в направлении верификационной диагностики и интеллектуальных диалоговых образовательных систем. Это открывает возможности для создания целостной образовательной экосистемы, где генерация контента, его верификация, адаптивная доставка и диагностика образуют единый автоматизированный цикл, обеспечивающий объективное отслеживание образовательного прогресса.

### Список литературы

1. Шмелев А.Г., Батурин Н.А., Костромина С.Н. и др. Российский стандарт тестирования персонала [Электронный ресурс]. Национальный совет по оценке и сертификации квалификаций персонала. 2014-2015. URL: <http://www.hrfederation.ru/projects/standards> (дата обращения: 20.10.2025).
2. Воронцов К.В., Потапенко А.А. Регуляризация, робастность и разреженность вероятностных тематических моделей. // Компьютерные исследования и моделирование. – 2012. Т. 4, № 1. С. 3-14.
3. Воронцов К.В. Аддитивная регуляризация тематических моделей // Машинное обучение и анализ данных. 2014. Т. 1, № 1. С. 39-62.
4. Воронцов К.В., Ирхин И.А. Сходимость алгоритма аддитивной регуляризации тематических моделей // Труды Института математики и механики УрО РАН. 2020. Т. 26, № 2. С. 140-157.
5. Bokov G.V. Undecidable Iterative Propositional Calculus // Algebra and Logic. 2016. Vol. 55. № 4. P. 274-282. DOI: 10.1007/s10469-016-9390-9.
6. Systematic Literature Review of Validation Methods for ИИ Systems [Электронный ресурс] / Lalli Myllyaho, Mikko Raatikinen, Tomi Männistö, Tommi Mikkonen, Jukka K. Nurminen // arXiv. 2021. URL: <https://doi.org/10.1016/j.jss.2021.111050> (дата обращения: 25.10.2025).
7. Sargent R.G., Balci O. Verification, Validation and Evaluation of Modeling Methods [Электронный ресурс] // Proceedings of the 2017 Winter Simulation Conference. 2017. P. 123-135.
8. Flesch–Kincaid readability tests [Электронный ресурс] // Wikipedia. 2023
9. Flesch Reading Ease and the Flesch Kincaid Grade Level [Электронный ресурс] // Readable.com. 2023.

10. Miller S. What Is the SMOG Index – A Complete Guide to Readability Scores [Электронный ресурс] // Medium. 2023. URL: <https://medium.com/@readabilitymatters/what-is-the-smog-index-a-complete-guide-to-readability-scores-efb6b5d0e7b5> (дата обращения: 21.10.2025).
11. Tempest E.L., The Validity of Type-Token Ratio as a Measure of Linguistic Ability: An Analysis Based on Clinical Data [Электронный ресурс]. 2022.
12. McCarthy P.M., Jarvis S. Metric of Textual Lexical Diversity (MTLD): An Advanced and Reliable Measure of Lexical Diversity [Электронный ресурс] // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2010». 2010.
13. Bonett D.G. Point-biserial correlation: Interval estimation, hypothesis testing, meta-analysis, and sample size determination [Электронный ресурс] // British Journal of Mathematical and Statistical Psychology. 2020. Vol. 73, Suppl. 1. P. 113-144.
14. Kadali B.R., Vedagiri P. Using logistic regression and point-biserial correlation, an investigation of pedestrian violations and their opportunities to cross at signalized intersections [Электронный ресурс] // IATSS Research. 2022. Vol. 46, Issue 3. P. 388-397.

## **VERIFICATION TESTOLOGY: A FORMALIZED SYSTEM FOR ASSESSING TEST ITEMS IN AI-POWERED EDUCATIONAL SYSTEMS**

**Sultanayev Yaudat Talgatovich**

Doctor of Sciences in Physics and Mathematics, Professor  
Department of Mathematics and Statistics,  
M. Akmulla Bashkir State Pedagogical University  
Ufa, Russian Federation

**Abzalimov Ramil Rafikovich**

Candidate of Sciences in Physics and Mathematics, Associate Professor,  
Director of the State Autonomous Institution of the Republic of Bashkortostan «Center for  
Assessment of Professional Skills and Teacher Qualifications»  
Ufa, Russian Federation

**Vildanova Venera Fidarisovna**

Candidate of Sciences in Physics and Mathematics,  
Associate Professor Department of Mathematics and Statistics,  
M. Akmulla Bashkir State Pedagogical University  
Ufa, Russian Federation

**Abstract.** The article addresses the systemic crisis of trust in the quality of test materials generated by artificial intelligence (AI) for pedagogical measurements. It introduces and substantiates the concept of verified testology – a methodology for evidence-based quality assurance through a multi-level automated verification system. A comprehensive system of formalized verification criteria is presented, including the assessment of semantic clarity, didactic relevance, predicted effectiveness, and robustness of test items. Special attention is paid to the «AI verifies AI» principle to ensure internal consistency. The successful practical implementation of the methodology within

the «AI-Expert» platform for assessing managerial competencies is described, demonstrating a generation quality level of 0.97. It is concluded that verified testology lays the foundation for a new paradigm of educational diagnostics, ensuring reliability, objectivity, and personalization in the context of mass AI use.

**Key words:** verification testology, artificial intelligence, pedagogical measurements, test quality, test item generation.

### References

1. Shmelev A.G., Baturin N.A., Kostromina S.N. and others. The Russian standard of personnel testing [Electronic resource]. The National Council for the Assessment and Certification of Personnel Qualifications. 2014-2015. URL: <http://www.hrfederation.ru/projects/standards> (accessed: 20.10.2025).
2. Vorontsov K.V., Potapenko A.A. Regularization, robustness and sparsity of probabilistic thematic models. // Computer research and modeling. 2012. Vol. 4, № 1. P. 3-14.
3. Vorontsov K.V. Additive regularization of thematic models // Machine learning and data analysis. 2014. Vol. 1, № 1. P. 39-62.
4. Vorontsov K.V., Irkhin I.A. Convergence of the algorithm of additive regularization of thematic models // Proceedings of the Institute of Mathematics and Mechanics of the Ural Branch of the Russian Academy of Sciences. 2020. Vol. 26, № 2. P. 140-157.
5. Bokov G.V. Undecidable Iterative Propositional Calculus // Algebra and Logic. 2016. Vol. 55. № 4. P. 274-282. DOI: 10.1007/s10469-016-9390-9.
6. Systematic Literature Review of Validation Methods for ИИ Systems [Electronic resource] / Lalli Myllyaho, Mikko Raatikainen, Tomi Männistö, Tommi Mikkonen, Jukka K. Nurminen // arXiv. 2021. URL: <https://doi.org/10.1016/j.jss.2021.111050> (date of access: 25.10.2025).
7. Sargent R.G., Balci O. Verification, Validation and Evaluation of Modeling Methods [Electronic resource] // Proceedings of the 2017 Winter Simulation Conference. 2017. P. 123-135.
8. Flesch–Kincaid readability tests [Electronic resource] // Wikipedia. 2023
9. Flesch Reading Ease and the Flesch Kincaid Grade Level [Electronic resource] // Readable.com. 2023.
10. Miller S. What Is the SMOG Index – A Complete Guide to Readability Scores [Electronic resource] // Medium. 2023. URL: <https://medium.com/@readabilitymatters/what-is-the-smog-index-a-complete-guide-to-readability-scores-efb6b5d0e7b5> (date of access: 21.10.2025).
11. Tempest E.L., The Validity of Type-Token Ratio as a Measure of Linguistic Ability: An Analysis Based on Clinical Data [Electronic resource]. 2022.
12. McCarthy P.M., Jarvis S. Metric of Textual Lexical Diversity (MTLD): An Advanced and Reliable Measure of Lexical Diversity [Electronic resource] // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2010». 2010.
13. Bonett D.G. Point-biserial correlation: Interval estimation, hypothesis testing, meta-analysis, and sample size determination [Electronic resource] // British Journal of Mathematical and Statistical Psychology. 2020. Vol. 73, Suppl. 1. P. 113-144.
14. Kadali B.R., Vedagiri P. Using logistic regression and point-biserial correlation, an investigation of pedestrian violations and their opportunities to cross at signalized intersections [Electronic resource] // IATSS Research. 2022. Vol. 46, Issue 3. P. 388-397.